УНИВЕРЗИТЕТ У НИШУ

ЕКОНОМСКИ ФАКУЛТЕТ

МАРИНА Б. МИЛАНОВИЋ

ИЗВОЂЕЊЕ ЗАКОНИТОСТИ ИЗ ЕКОНОМСКИХ ПОДАТАКА ПРИМЕНОМ *DATA MINING* ПРИСТУПА

докторска дисертација

Ниш, 2018. година

УНИВЕРЗИТЕТ У НИШУ

ЕКОНОМСКИ ФАКУЛТЕТ

МАРИНА Б. МИЛАНОВИЋ

ИЗВОЂЕЊЕ ЗАКОНИТОСТИ ИЗ ЕКОНОМСКИХ ПОДАТАКА ПРИМЕНОМ *DATA MINING* ПРИСТУПА

докторска дисертација

Ниш, 2018. година

UNIVERSITY OF NIŠ

FACULTY OF ECONOMICS

MARINA B. MILANOVIĆ

EXTRACTION OF REGULARITIES FROM ECONOMIC DATA USING DATA MINING APPROACH

Doctoral Dissertation

Niš, 2018

КОМИСИЈА ЗА ОЦЕНУ И ОДБРАНУ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

МЕНТОР:

Др Винко Лепојевић, ванредни професор,
Универзитет у Нишу,
Економски факултет

_____

ЧЛАНОВИ КОМИСИЈЕ:

_____

_____

Датум одбране:

_____

ИЗЈАВА МЕНТОРА О САГЛАСНОСТИ ЗА ПРЕДАЈУ УРАЂЕНЕ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Овим изјављујем да сам сагласан да кандидат Марина Милановић може да преда Реферату за последипломско образовање Факултета урађену докторску дисертацију под називом ИЗВОЂЕЊЕ ЗАКОНИТОСТИ ИЗ ЕКОНОМСКИХ ПОДАТАКА ПРИМЕНОМ *DATA MINING* ПРИСТУПА, ради организације њене оцене и одбране.

Ниш, 06.9.2018.

Ментор: Проф. др Винко Лепојевић

THE STATEMENT OF THE MENTOR'S CONSENT FOR THE SUBMISSION OF THE COMPLETED DOCTORAL DISSERTATION

Hereby, I declare that I agree that the candidate Marina Milanović, can submit the completed doctoral dissertation entitled EXTRACTION OF REGULARITIES FROM ECONOMIC DATA USING DATA MINING APPROACH to the officer for doctoral studies at the Faculty, for the purpose of its evaluation and defense.

Niš, 06.9.2018.

Mentor: Prof. Vinko Lepojević, PhD

Подаци о докторској дисертацији

Ментор:	Др Винко Лепојевић, ванредни професор, Универзитет у Нишу, Економски факултет
Наслов:	Извођење законитости из економских података применом *data mining* приступа
Резиме:	Развој информационе технологије и, последично, рапидно повећање расположиве количине података допринели су да *data mining,* као кључна компонента једног ширег интерактиваног, итеративаног и креативног процеса откривања знања из података, добије велики значај у економским истраживањима. Основна идеја *data mining*-а је ефикасно и ефективно идентификовање законитости, које су скривене у великим скуповима (мултидимензионалних) података, складиштеним у информационим репозиторијумима, помоћу софтверски подржаних метода и алгоритама. Узимајући у обзир наведено, у докторској дисертацији су разматрани најважнији теоријско-методолошки аспекти *data mining* приступа у анализи података и сагледане његове апликативне могућности у домену проучавање економских феномена. У том смислу, анализирани су трендови у савременој економији из перспективе растуће улоге података (као искористивог ресурса за генерисање вредности), фундаментални појмови повезани са концептом *data mining* и позитивни и негативни контексти његове примене. Задаци откривања знања из података, у функцији креирања *data mining* модела, посматрани су кроз призму широког дијапазона методолошких поступака за њихово спровођење. Посебна пажња је посвећена односу *data mining*-а и статистике, као науке која се традиционално бави откривањем законитости из података. Резултати истраживања указују на велики потенцијал интегрисане имплементације статистичког и *data mining* приступа у проналажењу иновативних методолошких решења конкретних проблема и, истовремено, сугеришу неопходност узајамног прилагођавања и модификовања базичних парадигми анализе података на обе стране. У емпиријском делу дисертације представљени су иновативни концептуално-методолошки оквири анализе података временских серија берзанских индекса и анкетних података о корисницима услуге. Емпиријски резултати, као егзактна сазнања издвојена из података, потврђују значај спровођења *data mining* анализе у проблемским контекстима економије, пословне економије и менаџмента. Спроведено истраживање представља погодну основу за профилисање будућих истраживачких усмерења у сфери примене *data mining*-а у изучавању економских појава.

Научна област:	Економске науке
Научна дисциплина:	Статистика
Кључне речи:	Економски подаци, знање, законитости, *data mining*, статистика
УДК:	330:519.21/25:004.6(043.3)
CERIF класификација:	S 180 Економија, економетрија, економска теорија, економски системи, економска политика
Тип лиценце Креативне заједнице:	**CC BY-NC-ND**

Data on Doctoral Dissertation

Doctoral Supervisor:	PhD Vinko Lepojević, Associate professor, University of Niš, Faculty of Economics
Title:	Extraction of regularities from economic data using data mining approach
Abstract:	The development of information technology and, consequently, rapid increase of the available amount of data have contributed to the fact that data mining, as a key component of a wider interactive, iterative and creative process of knowledge discovery from data, is of great importance in economic research. The basic idea of data mining is reflected in the efficient and effective identification of regularities, that are hidden in large sets of (multidimensional) data stored in information repositories, using software-supported methods and algorithms. Taking into account the foregoing, in this doctoral dissertation, the most important theoretical-methodological aspects of data mining approaches in data analysis are examined, as well as its applicative possibilities in the field of studying economic phenomena. In this sense, trends in the modern economy, from the perspective of the growing role of data (as a usable resource for generating values), fundamental terminology related to the concept of data mining as well as the positive and negative contexts of its application, have been analyzed. The tasks of discovering knowledge from data, in function of creating a data mining model, are viewed through the prism of a wide range of methodological procedures for their implementation. Special attention has been dedicated to the relationship between data mining and statistics, as a science that traditionally deals with the discovery of regularities from data. The results of the research indicate the great potential of integrated implementation of statistical and data mining approaches in finding innovative methodological solutions to specific problems and, at the same time, suggest the need for mutual adaptation and modification of basic paradigms of data analysis on both sides. Empirical part of the dissertation points out innovative conceptually-methodological frameworks for analyzing data from time series of stock exchange indices and survey data on service users. Empirical results, as the exact knowledge extracted from data, confirm the importance of implementing data mining analysis in the problem contexts of economics, business economics and management. The conducted research represents a suitable basis for profiling future research orientations in the field of data mining application concerning the study of economic phenomena.

Scientific Field:	Economic sciences
Scientific Discipline:	Statistics
Key Words:	economic data, knowledge, regularities, data mining, statistics
UDC:	330:519.21/25:004.6(043.3)
CERIF Classification:	S180 Economics, econometrics, economic theory, economic systems, economic policy

НАУЧНИ ДОПРИНОС ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Једна од највећих промена у сфери савремене економије, иницирана, али и вођена *ICT* развојем, односи се на растућу улогу података као извора знања, раста и значајног потенцијала за генерисање економске вредности. Да би се предности по основу поседовања и употребе података заиста искористиле, неопходна су одговарајућа усклађивања стварних потреба и аналитичких могућности за њихово процесирање и анализу. Стога су у докторској дисертацији елаборирана бројна питања, размотрене апликативне дилеме и осветљени кључни аспекти извођења законитости из економских података применом *data mining* приступа. У том смислу, теоријско-методолошки допринос докторске дисертације огледа се у: ▸ демистификацији *data mining*-а, као мултидисциплинараног приступа и методолошког оквира за разумевање, претраживање, обраду и анализу великих количина података, и ▸ образложеној и потврђеној оправданости примене *data mining*-а у анализи економских података. Практичне импликације истраживања у докторској дисертацији везују се за очекивање да ће презентовани резултати бити од користи са аспекта сагледавања могућности и формулисања смерница за ширу имплементацију *data mining* приступа од стране менаџмент структура српских предузећа. Са становишта унапређења научне мисли у области анализе података, посебан научни допринос реализованих теоријских и емпиријских истраживања односи се на сагледавање односа и утврђивање концептуалних сличности и разлике између статистичког и *data mining* приступа, уз апострофирање непоходности и значаја њихове интегрисане примене.

SCIENTIFIC CONTRIBUTION OF DOCTORAL DISSERTATION

One of the biggest changes in the modern economy, initiated, but also driven by ICT development, relates to the growing role of data as a source of knowledge, growth and significant potential for generating economic value. In order for the benefits, based on the possession and use of data, really to be used, an appropriate harmonization of real needs and analytical possibilities for their processing and analysis is necessary. Therefore, in the doctoral dissertation, numerous questions were elaborated, application dilemmas were discussed, and key aspects of the extraction of regularities from economic data using data mining approach were highlighted. In this sense, the theoretical and methodological contribution of the doctoral dissertation is reflected in: ▸ demystification of data mining, as a multidisciplinary approach and methodological framework for understanding, searching, processing and analyzing large amounts of data, and ▸ confirmed justification of the application of data mining approach in the analysis of economic data. Practical implications of the research in the doctoral dissertation are related to the expectation that the presented results will be useful in terms of considering possibilities and formulating guidelines for wider implementation of data mining approach by the management structures of Serbian enterprises. Regarding the advancement of scientific thought in the field of data analysis, a special scientific contribution of the conducted theoretical and empirical research refers to the understanding of relationships and determination of conceptual similarities and differences between statistical and data mining approaches, with emphasized necessity and importance of their integrated application.

С А Д Р Ж А Ј

УВОД

Најзначајнија промена у интелектуалној историји људског друштва односи се на улогу знања које је у свету савремене економије, као компонента интелектуалног капитала, постало критични фактор друштвено-економског развоја. У том смислу, успешно суочавање пословних система и државних институција, националних економија и, начелно, читаве светске економије са континуитетом у погледу изразите динамичности и непредвидивости привредног амбијента заснива се на различитим формама креирања, одржавања, унапређења, дистрибуције и примене знања.

Опште је позната чињеница да се све сфере људског деловања током последњих деценија налазе под снажним утицајем развоја и интензивне примене информационо-комуникационих технологија које су трансформисале економију како на макро, тако и на микро нивоу. Технолошки напредак је условио настанак структурних промена и развој нових привредних делатности које су на различите начине везане за рад са подацима и информацијама. Такође, дигитална технологија кроз аутоматизацију и информатизацију готово свих процеса и активности, узроковала је једноставно генерисање и складиштење огромних количина података и, консеквентно, донела нове изазове у погледу њиховог претварања у вредне информатичке садржаје и корисно знање.

Заправо, информатичка револуција је омогућила акумулирање великих количина мултидимензионалних сирових података који немају по аутоматизму употребну вредност за кориснике или власнике података и које је практично немогуће процесирати применом традиционалних приступа у анализи података. Стога се, у новонасталом окружењу, појавила потреба за иновативним научним приступима, технолошким решењима и методолошким оквирима који ће омогућити да се превазиђе јаз између расположиве количине података и степена њихове искоришћености за доношење квалитетних одлука у функцији креирања вредности.

У својству потенцијалног одговора за превазилажење наведених проблема, развијен је *data mining*, као мултидисциплиниран приступ за разумевање, анализу, интерпретацију и употребу података. Реч је о приступу који означава идентификовање значајних законитости дубоко скривених у великим скуповима података, заснован на примени софистицираних, компјутерски подржаних метода и алгоритама за анализу података. Суштински, фундаменталне промене у савременим условима привређивања, пословној филозофији и начину функционисања привредних субјеката условљене

обиљем и брзином протока података допринеле су да *data mining* постане питање од изузетног значаја са становишта ефективне и ефикасне трансформације података у вредне информације и корисно знање. Истовремено, с обзиром да се статистика традиционално бави анализом података, доступност великих количина података и *data mining* су статистику као научну дисциплину ставили пред нова искушења и изазове.

Имајући у виду наведено, предмет докторске дисертације под насловом „Извођење законитости из економских података применом *data mining* приступа" је истраживање и анализа најважнијих теоријско-методолошких аспеката *data mining* приступа и, сходно томе, сагледавање његових апликативних својстава, пре свега, у економским истраживањима и анализи макроекономских и микроекономских података.

Истина, истраживање и анализа података у функцији формулисања законитости о природи, тенденцијама развоја и међусобним релацијама између економских феномена, везаних како за функционисање и управљање економијом у целини, тако и за појединце и привредне субјекте, не представља новину. Међутим, с обзиром да расположива количина података рапидно расте, потреба за новим методолошким решењима која ће омогућити истраживање и проналажење смисла у великим (често веома хетерогеним) скуповима расположивих података кроз откривање (често веома малих делова) корисних информација за доношење квалитетних одлука представља кључни разлог који је усмерио како истраживачку, тако и менаџерску пажњу на *data mining*, као интегралну компоненту система управљања знањем заснованог на примени концепта (и технологије) пословне интелигенције.

Генерално, откривање законитости у великим количинама сирових података о појавама у области економије, пословне економије и менаџмента, чије је иманентно својство висока димензионалност, нужно захтева спровођење анализе подтака која укључује елементе *data mining* приступа. Сходно подели економије на макроекономију и микроекономију, посматрано са становишта макроекономских истраживања, *data mining* приступ омогућава откривање законитости о макроекономским кретањима и обезбеђује предвиђање процеса и догађаја у макроекономским системима. С друге стране, посматран са становишта микроекономских истраживања, овај приступ је усмерен на откривање законитости из података у циљу ефективног решавања пословних и управљачких проблема, при чему идентификоване законитости постају круцијални инпут у процесу пословног одлучивања.

У складу са дефинисаним и, са становишта значаја и актуелности, образложеним предметом истраживања у докторској дисертацији, формулисан је примарни циљ рада да се, сагледавањем и анализом (а) теоријских аспеката и методолошких поступака и процедура (увидом у релевантну и доступну научну и стручну грађу) и (б) резултата конкретних апликација, идентификују могућности и оцене ефекти примене *data mining* приступа за извођење законитости, односно обезбеђење корисних информација и стицање знања о економским феноменима.

На основу дефинисаног основног циља, с једне стране, и аналитичког карактера овог истраживачког усмерења, с друге стране, изведен је и прецизиран сет специфичних циљева, који обухватају следеће: ▸ указивање на кључне карактеристике савремене економије и анализирање улоге знања као кључног фактора њеног развоја; ▸ истраживање консеквенци револуције података и значаја откривања прикривених структура и законитости (глобалних и локалних модела) из огромних количина економских података; ▸ разматрање суштине и значаја управљања квалитетом података у функцији добијања корисних информација релевантних са становишта квалитета (пословног) одлучивања; ▸ идентификовање основних одређења *data mining* концепта и фаза процеса откривања знања из података; ▸ сагледавање мултидисциплинарне природе *data mining*-а као научне дисциплине, са посебним освртом на однос са статистиком; ▸ разматрање типичних питања која су повезана са применом *data mining*-а; ▸ истраживање и систематизовање кључних *data mining* задатака, метода и алгоритама и приказивање њихових суштинских карактеристика, уз сагледавање значаја истовременог коришћења низа метода при решавању конкретног проблема; ▸ разматрање аспектата примене *data mining* приступа у анализи временских серија; и ▸ реализацију *data mining* задатака применом одговарајућих метода и анализу откривених законитости из података за дефинисане конкретне проблемске ситуације.

Сходно опредељеном предмету и постављеним циљевима истраживања, формулисане су следеће хипотезе, које представљају темељ садржаја дисертације:

✓ Ефикасно и ефективно откривање законитости прикривених у великим скуповима података, карактеристичних за савремено економско окружење и пословање, нужно захтева анализу података која се базира на примени *data mining* приступа.

✓ Развијање и избор *data mining* модела, који ће послужити као основа за доношење информационо заснованих одлука и предузимање адекватних акција у решавању конкретних истраживачких проблема, заснива се на проналажењу

оптималног модела путем оцењивања карактеристика скупа модела креираних коришћењем (а) истог метода на различитим скуповима података и (б) различитих метода на истом скупу података, уз варирање параметара метода у оба случаја.

✓ Како у основи сви поступци за издвајање законитости из података садрже елементе статистике, искључиво методолошки примерена и валидна употреба статистичких постулата и метода у фазама *data mining* процеса омогућава побољшање квалитета *data mining* резултата, а самим тим и повећање поузданости формулисаних закључака о истраживачком феномену.

Такође, кореспондентно изложеном предмету и дефинисаном циљу (циљевима) његовог разматрања, изведена је структура дисертације, која, поред уводних и закључних разматрања, обухвата још четири комплементарна дела.

Разматрања у првом делу дисертације се односе на **концептуални оквир и апликативну релевантност** *data mining*-а. У том смислу, пажња је, најпре, посвећена карактеристикама савременог окружења, при чему је указано на значај информационо-технолошке револуције у стварању нове економије, односно економије знања. С обзиром да је развој информационе технологије омогућио брзо и лако складиштење великих количина података, истакнут је проблем проналажења смисла, односно потенцијално корисних и вредних информација скривених у подацима. При томе су елаборирана суштинска одређења појмова податак, информација и знање у циљу идентификовања концептуалних разлика и објашњења хијерархијске везе између њих. Такође, јасно је одређена позиција података као доминантног извора знања релевантног за стварање економске вредности и доношење стратегијских, тактичких, али и оперативних одлука. Фокус даљих истраживања у овом делу дисертације су теоријска сазнања о концепту *data mining*-а кроз обухватање различитих дефиниција, кључних елемената и еволутивног развоја (и као научне дисциплине чије је порекло везано за више дисциплина из различитих научних области и као савременог приступа у анализи података). Узимајући у обзир чињеницу да је стицање знања из података засновано на *data mining* приступу, суштински, потпроцес ширег процеса откривања знања, у наставку су сагледане фазе овог процеса. С тим у вези, посебно су разматрана питања повезана са компетенцијама активних учесника у реализацији идентификованих процесних фаза и наглашен значај статистичке едукације стручњака за *data mining* и, генерално, примене статистичког начина размишљања у свим фазама процеса извођења законитости из података. На крају првог дела одговарајућа пажња је посвећена аспектима примене *data mining*-а, обухватајући подручја примене (са

посебним истицањем потенцијалних могућности примене у домену економских истраживања), факторе од којих зависи реализација пројектних *data mining* задатака, потенцијалне користи и ограничења, као и најчешће заблуде, могуће пропусте и нове изазове.

Други део дисертације је посвећен међусобној повезаности **карактеристика података, задатака и метода у креирању *data mining* модела**. У том смислу, карактеристике појединих категорија података и *data mining* задатака су посматране у констелацији са проблемом избора метода за реализацију циљева откривања законитости из података. Узимајући у обзир чињеницу да су подаци срж *data mining* концепта, након осврта на кључне појмови повезане са подацима, затим, различите категоризације података и савремене облике њиховог организовања, указано је на значај квалитета података и истакнуто да је квалитет резултата било које анализе података у директној зависности од квалитета података. Даљим истраживањем у овом делу дисертације су обухваћена питања дефинисања и класификације *data mining* задатака и метода, као и креирања *data mining* модела. С обзиром да постоји велики број софтверских пакета за креирање *data mining* модела, поред већ наведених разматрања, представљен је кратак преглед карактеристика најчешће коришћених софтверских решења. Опште позната констатација је да статистика представља синоним за анализу података. Полазећи од тога и уважавајући чињенцу да се у свакој фази *data mining* процеса користе неки елементи и концепти статистичког начина размишљања, као и да су многи методи који се користе у креирању *data mining* модела у својој основи статистички, посебно је анализиран однос статистике и *data mining*-а. С тим у вези, након терминолошког прецизирања синтагме „*data mining* окружење", елаборирана је употреба статистике у новом окружењу, идентификоване сличности и разлике између статистичког и *data mining* приступа у анализи података и наглашена неоходност њиховог интегрисања у будућности.

У оквиру трећег дела дисертације разматрани су бројни м**етодолошки поступци за спровођење задатака претпроцесирања, процесирања и постпроцесирања података**. Након кратког осврта на значај задатака претпроцесирања са становишта квалитета резултата моделирања, представљени су најзначајнији методолошки поступци за припрему података, при чему се посебно обухватају елементи експлоративне анализе података. Будући да након претпроцесирања следи процесирање података, у даљем фокусу овог дела дисертације су дескриптивни и предиктивни методи за конструкцију *data mining* модела. С обзиром да је велико

интересовање узроковало огромно проширење расположивог методолошког спектра за реализацију *data mining* задатака, практично је немогуће све методе детаљно представити на једном месту. Сходно томе, сагледани су основни постулати и аспекти примене изабране групе метода. Избор метода за анализу извршен је са становишта сагледавања њиховог доприноса за остварење дефинисаних циљева дисертације. Како многи проблеми у домену економских истраживања укључују темпоралне аспекте, одговарајућа пажња је посвећена концептима који су релевантни за ефикасну примену *data mining*-а у анализи временских серија. Посебан акценат је стављен на представљање концепта сличности уз сагледавање значаја и прагматичне вредности истраживања сличности за (успешну) реализацију различитих задатака и циљева откривања скривених законитости у кретању појава представљених временским серијама података. Имајући у виду методолошку варијететност *data mining*-а, као и чињеницу да се различити методи могу користити за решавање истог задатака, у наставку је детаљно елаборирано питање оцењивања карактеристика и избора оптималног модела у односу на посматрани проблем, уз примену различитих критеријума и метода. При томе је наглашен значај издвајања оних законитости које су у посматраном контексту смислене и корисне.

Четврти део дисертације, који је насловљен **Емпиријско истраживање: анализа проблемских ситуација применом *data mining* приступа**, је посвећен оригиналним емпиријским истраживањима. Сходно изложеним теоријским одређењима и методолошким процедурама, примена *data mining* приступа је демонстрирана кроз следеће две ситуације у проблемском подручју економије: ▸ прва, на основу података временских серија о кретању вредности берзанских индекса истраживање је усмерено ка утврђивању сличности и класификацији берзи одабраних земаља у интерно хомогене и екстерно хетерогене групе, и ▸ друга, на основу анкетних података који се односе карактеристике испитаника и њихове ставове према обележјима квалитета услуге, истраживање је усмерено ка утврђивању сатисфакције корисника услуге у ресторатерском пословању и формирању тржишних сегмената на примеру једног одабраног ресторана. С тим у вези, након дефинисања проблемских ситуација и дизајнирања кореспондентних концептуално-методолошких оквира, пажња је посвећена операционализацији конкретних истраживања. За потребе реализације емпиријских истраживања коришћен је стандардни статистички софтверски пакет *IBM SPSS*, верзија 20.0, као и, у оквиру ове дисертације, посебно креиран специјализовани програм за потребе трансформације нумеричких у симболичке временске серије.

Логично, последњи део рада садржи детаљну анализу, интерпретацију и презентацију емпиријски утврђених резултата, који представљају не само комплемент изложених теоријских разматрања, већ и оригинални научни допринос у контексту дефинисаног предмета истраживања у дисертацији. Заправо, актуелност и значај добијених резултата у форми откривеног знања из великих количина података треба посматрати кроз призму како истраживачких изазова у научним круговима, тако и са аспекта практичне корисности у домену разматраних, али и потенцијално нових проблемских ситуација.

Током реализације истраживања коришћени су бројни извори у којима су представљена научна и стручна сазнања и тумачења о различитима аспектима интегралних елемената *data mining* концепта: од уџбеничке литературе, специјализованих научних часописа и радова, бројних тематских издања и зборника радова, техичких извештаја о расположивим софтверима, до Интернет извора. Такође, сходно опредељеном предмету и дефинисаним циљевима истраживања, за потребе аргументоване елаборације и верификације постављених кључних хипотеза, као и хипотеза које су дефинисане у емпиријском делу истраживања, у дисертацији је коришћена одговарајућа комбинација следећих општих научних метода: аналитичко-синтетичког метода, метода компарације и елемената индуктивног и дедуктивног закључивања. Поред наведених метода, у емпиријском делу истраживања коришћени су *data mining* методи, изабрани, пре свега, из групе метода примењене статистике.

Део I

КОНЦЕПТУАЛНИ ОКВИР И АПЛИКАТИВНА РЕЛЕВАНТНОСТ *DATA MINING*–а

1. ОТКРИВАЊЕ ЗНАЊА ИЗ ПОДАТАКА У КОНТЕКСТУ САВРЕМЕНЕ ЕКОНОМИЈЕ

Имајући у виду значај који знање као (глобални) економски ресурс има у савременој економији, као и промене које је донела револуција у домену информационих и комуникационих технологија, а односе се, пре свега, на афирмацију података као извора знања, у овом Поглављу су, најпре, сагледане кључне карактеристике савременог економског окружења, а затим и указано на генерисање података по изузетно високим стопама раста уз пратећи проблем све већег несклада између расположиве количине података и степена њихове искоришћености. Сходно томе, апострофирана је потреба за радикалним променама у начину процесирања великих количина података.

1.1. Кључне одреднице савременог економског окружења

Савремени свет се одликује изразито динамичним и непредвидивим променама, узрокованим, примарно, глобализацијским процесима и технолошком револуцијом. Неопходност суочавање актера привредног и друштвеног живота на свим нивоима организовања са изазовима који проистучу из ових промена економског, технолошког, политичког и социјалног карактера, а у правцу прилагођавања и опстанка у брзомењајућем амбијенту, афирмисала је улогу и значај знања, као основе укупног друштвеног и економског развоја. У том контексту, са економског становишта, успех пословних система и државних институција, националних економија и, начелно, читаве светске економије у прилагођавању и борби са растућом комплексношћу привредног амбијента заснива се на различитим формама креирања (и издвајања), одржавања, унапређења, дистрибуције и примене знања, које добија статус круцијалног стратегијског економског ресурса и кључног извора конкурентске предности привредних субјеката, региона и националних привреда.

Идеја о значају знања у економским активностима и процесима није нова. Знање, као економски ресурс, је одувек било један од основних фактора развоја друштвених заједница и националних привреда. Међутим, под утицајем треће научно-технолошке (информатичке) револуције, крајем XX века у теорији и моделима економског раста експлицитно је признато да знање представља фундаментални и неприкосновени извор економског раста и развоја. Наиме, у савременој економији базирање раста и развоја искључиво на материјалним (и финансијским) ресурсима није одрживо. Напредак

захтева улагање у нематеријалну имовину, чији су суштински елементи неопипљиви ресурси, односно информације и знање. *Đuričin & Janošević* (2009, стр. 5) истичу да су бројни докази који недвосмислено указују на чињеницу да на додату вредност и раст друштвеног производа највише утиче знање. Такође, исти аутори наводе да је секундарна улога земље, радне снаге и капитала (као традиционалних производних фактора) последица чињенице да се могу лако стећи уколико постоји знање.

Савремена економија, у којој је у потпуности афирмисана добро позната констатација, давно формулисана од стране *Francis*-а *Bacon*-а, „Знање је моћ", добија назив економија заснована на знању. Употреба наведеног термина је резултат суштинских структурних промена у развоју економије и преласка од индустријске ка информационо-технолошкој основици производње, као и потпунијег схватања и истицања улоге знања у економском расту. С обзиром да је концепт економије знања често разматран не само у стручној академској литератури, већ и на политичком нивоу од стране једног броја међународних организација и институција, заинтересованих за тенденције које се испољавају у оквиру новог типа економије, постоје бројне, мање или више прецизне, дефиниције самог појма. За потребе овог рукописа, издваја се дефиниција Организације за економску сарадњу и развој (енгл. *the Organization for Economic Cooperation and Development - OECD*), чија садржина довољно илустративно указује на суштинска одређења овог концепта. Заправо, као „економија базирана на знању", економија знања се дефинише као она „економија која је директно установљена на производњи, дистрибуцији и коришћењу знања и информација" (*OECD*, 1996, стр. 3). При том се знање препознаје као кључни покретач конкурентности и економског успеха и, истовремено, наглашава улога и значај истраживања, примене нових технологија и нових идеја (било у форми инвенција или нових примена постојећег знања), као и континуираног процеса учења у контексту економских резултата.

Да би се остварили користи по основу револуције знања, према приступу Светске банке, неопходно је јасно дефинисати стратегијски оквир који се односи на следећа четири стуба (области), критичних за транзицију једне земље ка економији знања (*http://web.worldbank.org/archive/website01503/WEB/0__CO-10.HTM*):

- економски и институционални режим који ће обезбедити подстицаје за ефикасно стицање, коришћење и примену знања и развој предузетништва,
- образовану и обучену популацију која ће креирати, дисеминовати и користити знање,

- динамичну информациону и комуникациону инфраструктуру која ће омогућити ефективно дисеминовање и процесирање информација, и
- ефективан систем иновација у оквиру организација и истраживачких центара спремних и способних да прихвате глобално знање, изврше његово прилагођавање локалним потребама и креирају нове технологије, односно знање.

Синтезом бројних покушаја који су усмерени на дефинисање концепта и елаборирање карактеристика економије засноване на знању, у контексту друштвене и економске реалности, могуће је идентификовати следеће кључне карактеристика ове економије (*Roberts* & *Armitage*, 2008, стр. 335-354):

- растући значај знања као инпута у економији,
- повећани значај информационих и комуникационих технологија,
- раст значаја знања као економског резултата / производа,
- раст комерцијализације знања кроз права интелектуалне својине,
- раст учешћа радника знања у структури радне снаге,
- повећани утицај знања у свим секторима привреде,
- успон праксе управљања знањем, и
- глобализација као покретачка снага за експанзију економије знања.

Као што се из наведеног може недвосмислено запазити, убрзани темпо трансформационих процеса који једну традиционалну / индустријску економију засновану на материјалним ресурсима воде ка економији знања представља последицу два међусобно узрочно-последично повезана фактора. Први се односи на феномен глобализације и настанак глобалне економије са високим нивоом интеракције и интеграције између људи, компанија и земаља кроз размену производа, информација, знања и културе. Други се односи на неслућене размере развоја информационих и комуникационих технологија (енгл. *Information and Communication Technologies – ICT*), које су омогућиле брзи пренос и генерисање огромних количина података у свим областима и на свим нивоима друштва и привреде. У складу са тим, економија знања није заснована само на развоју неколико сектора високе технологије. Све привредне гране морају интензивно користити знање и савремена достигнућа, као предуслов дугорочног економског развоја. Свакако, ова констатација не искључује подразумевану коњункцију и комбинацију материјалних и нематеријалних ресурса у креирању укупне вредности.

Улога и значај знања у контексту развоја економије знања може се сагледати како са макро, тако и са микро становишта. Разлике у знању и његовој технолошкој примени, вредноване кроз раст бруто домаћег производа и структуру спољнотрговинске размене, постају главни чиниоци националне конкурентности који деле развијене од неразвијених земаља. Наиме, разлика у нивоу развоја између појединих подручја и земаља данас се објашњава њиховом способношћу да креирају и примене нова знања, пре него стопом инвестирања и другим подстицајима (*Babić*, 2012, стр. 30). Сходно томе, у структури најразвијенијих привреда, заједно са услужним гранама, доминирају индустријске гране интензивне знањем. У том смислу, побољшање релативног положаја појединих земаља захтева стварање погодних услова за подстицање иновација и дифузију знања, односно креирање адекватног институционалног амбијента који ће омогућити успешно суочавање са насталим променама и преоријентацију производње на производе са високим садржајем знања.

Реализација концепта економије засноване на знању, уз афирмацију значаја организационог учења, мора бити праћена и кореспондентним прилагођавањима (променама) на нивоу свих економских субјеката, укључујући и ниво запослених појединаца, јер њихово индивидуално знање представља саставни део организационог знања. Ова прилагођавања обухватају промену пословног фокуса и, сходно томе, усвајање нових парадигми у вођењу предузећа, нових приступа у реализацији активности и организацији предузећа, нових начина коришћења људског капитала, нових технологија и нових програма за школовање, обучавање и подстицање спремности запослених да се суоче са императивом промена. Наравно, обим промена и прилагођавања треба ускладити са организационим потребама и могућностима, с једне, и захтевима окружења, с друге стране.

У оваквим околностима, креирање, примена и унапређење знања добијају епитет критичних активности предузећа. Да би се остварили дефинисани циљеви пословања и побољшале перформансе наведеним активностима је неопходно адекватно управљати. Стога, неизоставни део менаџемета сваког модерног и успешног предузећа представља управљање знањем. И поред опште сагласности о значају знања као извору конкурентности, ипак, између научника, истраживача и практичара не постоји консензус у погледу јединствене дефиниције концепта управљање знањем, што јасно указује на мултидисциплинарну природу (*Dalkir*, 2011, стр. 8) и различите аспекте посматрања ове проблематике. Због тога се, сходно контексту претходне дискусије, издваја дефиниција која представља сублимацију ставова гуруа управљања знањем

(*Evans et al*., 2014, стр. 85): „Управљање знањем обухвата систематске процесе за стицање, организовање, одржавање, примену, дисеминацију и обнављање свих облика знања како би се побољшале организационе перформансе и креирала вредност". У суштини, у фокусу управљања знањем је координација људи, процеса и технологија која се заснива на токовима знања са сврхом обезбеђења услова за континуирано учење, стратегијско планирање, доношење одлука и генерално коришћење знања за решавање проблема и његово брзо уграђивање у нове производе и технологије.

Примена савремених *ICT* решења снажно је подстакла лакшу и једноставнију реализацију свих процеса и циљева управљања знањем и допринела отварању нових могућности и генерисању нових идеја у том правцу. Наиме, континуирани *ICT* напредак и све масовнија примена нових технологија условили су убрзано стварање, складиштење и проток огромних количина комплексних података у свим сегментима живота и рада, на једној страни, и истовремено донели нове изазове у погледу њиховог претварања у вредне информатичке садржаје, односно корисно знање, на другој страни. Сходно наведеном, једна од највећих промена у сфери економије, подстакнута, али и вођена технолошким напретком односи се на улогу података као извора знања, детерминанате конкурентности и фактора без којег није могуће замислити савремено пословање. Подаци су постали неизоставни део сваке привреде, индустрије, организације, пословне функције и индивидуе, тако да њихов значај континуирано расте. Због знатног учешћу података у реализацији пословних процеса и разматрању алтернатива у процесу пословног одлучивања, савремена економија се често у литератури назива и економија вођена подацима (енгл. *data-driven economy*). У том смислу, подаци су центар будуће ере развоја економије и друштва знања (*EC*, 2014, стр. 4).

Количине података се свакодневно стварају по изузетно високим стопама раста. Последично, пословни системи су преплављени подацима који сами по себи немају вредност за кориснике и власнике података са становишта решавања конкретних проблема и које је практично немогуће применом традиционалних метода и алата за анализу података претворити у корисне информације и знања. Без радикалних промена у начину процесирања тих података веома је мала вероватноћа да ће они икада бити анализирани и употребљани од стране аналитичара и доносилаца одлука за сагледавање тренутне позиције и развојног потенцијала пословних система. Новонастале околности у погледу не само количине (обима), већ и других особина података (попут, разноврсности података) насталих као последица развоја рачунарске

технологије, довеле су до потребе за иновативним научним приступима, технолошким решењима и методолошким оквирима који ће омогућити да се премости јаз између растућег тренда генерисања података и опадајућег тренда учешћа оних података у структури генерисаних података који се заиста користе у својству ресурса за откривања знања и подршку одлучивању.

У контексту могућих одговора за решење проблема несклада између расположиве количине података и степена њихове искоришћености,[1] развијен је *data mining*[2], као мултидисциплинаран приступ за разумевање, претраживање, обраду и анализу мултидимензионалних података. Из његовог назива недвосмислено следи да је у питању приступ заснован на подацима. У суштини, *data mining* означава имплементацију широког спектра компјутерски подржаних метода за анализу података у циљу идентификовања корисних информација и стицања знања у форми значајних законитости скривених у великим скуповима података. Заправо, *data mining* је технологија и есенцијални сегменат процеса откривања знања из база података.

Сходно наведеном, произлази да *DM*, као једна од форми извођења и откривања знања из података уз примену савремених информационих решења, може представљати значајну подршку иницијативама и напорима за управљање знањем или, пак, њихов део. Управо, реализацијом *DM* процеса, а на основу утврђених трендова, законитости и релација у подацима, организације могу генерисати нове могућности за креирање и одржавање конкурентске предности која се може манифестовати у ефикаснијем доношењу одлука, превентивном деловању, дијагностификовању и решавању проблема, планирању, учењу и покретању иновација. Међутим, ова конкурентска предност захтева не само примену нових, високо софистицираних приступа, метода и алата за рад са подацима, већ и одговарајућу подршку компетентних експерата који поседују нова знања, аналитичке вештине и идеје за прикупљање, обраду, анализу и трансформацију сирових података у вредне информације и знање.

Генерално, у савременим условима нових изазова и извесне неизвесности, само „(са)знање о знању" је омогућило да се знање стиче и примењује на нове начине. У том смислу, упоредо са већ апострофираним процесима економске глобализације и

[1] Ова комплексна ситуација је описана познатом констатацијом *Naisbitt*-а из 1982. године да се свет дави у мору информација, а да ја жедан знања. Ипак, тренутно стање верније одражава модификована верзија ове констатације да се свет дави у мору података, а да је жедан информација (*Brown*, 2014), што јасно указује на значај издвајања (малих количина) знања из (велике количине) података.

[2] У наставку текста, поред пуног назива „*Data Mining*", који се задржава на местима где је то, према процени аутора, прикладније са становишта бољег разумевања текста, равноправно се користи и акроним ове синтагме – *DM*.

технолошког напретка, *DM* је постао важан приступ за анализу економских података (*Baicoianu* & *Dumitrescu*, 2010, стр. 185), а самим тим и за издвајања знања из (велике количине расположивих) података. У циљу проширења спектра решења проблемских ситуација у савременом економском и пословном амбијенту, сасвим је јасна потреба за интеграцијом управљања знањем и *DM*-а (као компонентом пословне интелигенције).

1.2. Пирамида знања

Схватање и појмовно одређење знања представља сложен проблем, јер знање је широка и апстрактна категорија која се може посматрати, а самим тим и интерпретирати, из различитих перспектива: као стање ума, објекат, процес, способност, сумарне информације, комбинација информација, људског искуства, интелигенције и експертизе, невидљива имовина, капитал, вредност. Мада изворно појам знања припада подручју епистемологије, овом приликом, не улазећи у далеку прошлост, биће изостављени филозофски аспекти тумачења и повезивања знања са сродним концептима, попут истинитости и веровања. За потребе овог рада (сходно дефинисаном предмету истраживања), концепт знања се посматра и интерпретира у контексту релације „подаци - информације - знање - мудрост".

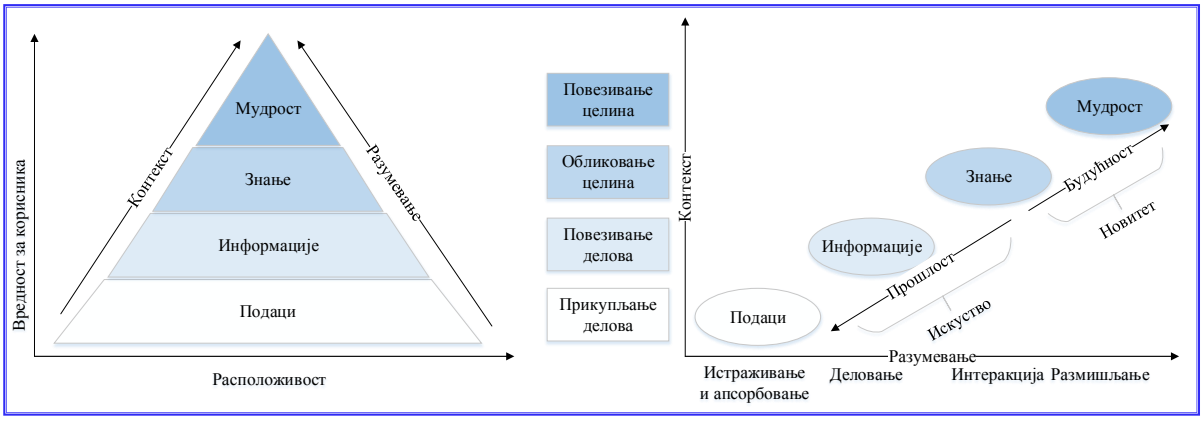
Заправо, разумевање појма знање представља неопходан услов за ефективно управљање знањем (*Allee*, 1997, стр. 71). У складу са тим, у наставку је дефинисање, анализа и јасно разграничење знања у односу на сродне категорије извршено према концепту хијерархијске надређености и подређености. Реч је о концепту који је врло популаран и уобичајено се користи у подручју рачунарства и управљања знањем као полазна тачка за објашњење релевантних односа и дефинисање знања преко података и информација.

У литератури су доступне различите верзије хијерархијских модела за дефинисање и истраживање природе знања. Наведено је сасвим разумљиво будући да у академским и професионалним круговима из различитих научних и апликативних подручја не постоји консензус у погледу конститутивних елемената хијерархијске структуре и њихових одређења. Већина истраживача у хијерархијску структуру укључује три основна елемента: податке, информације и знање. Констутутивни елемент са којим се најчешће проширује ова основа је мудрост, мада се, с обзиром на бројне дебате о концепту хијерархије, могу пронаћи и други елементи, углавном као компоненте између знања и мудрости. На пример, *Ackoff* (1989) укључује елемент разумевање (који није шире прихваћен као посебан концепт), док у свом раду новијег

датума, *Liew* (2013) проширује хијерархијску структуру увођењем елемента интелигенције.

Хијерархија „подаци - информације - знање - мудрост" (енгл. *Data-Information-Knowledge-Wisdom hierarchy - DIKW*) је позната и под називима хијерархија знања, информациона хијерархија, пирамида знања или пирамида мудрости. Генерално, *DIKW* концепт примарно служи за потребе контекстуализације елемената хијерархијске структуре узимајући у обзир њихове међусобне односе, као и за идентификовање и описивање трансформационих процеса елемената са нижег нивоа у хијерархији на виши ниво. При томе, „имплицитна претпоставка је да се подаци могу користити за креирање информација, информације за креирање знања, а знање за креирање мудрости" (*Rowley,* 2007, стр. 164).

DIKW логичка хијерархија често се приказују у форми пирамиде (или троугла) коју чине слојеви конститутивних елемената, указујући на њихову позицију у процесу стицања знања: у корену пирамиде се налазе подаци, следе информације, затим знање, а на врху мудрост. Једноставан приказ ових међусобних односа логичке надређености и подређености илустрован је на Слици 1 (лево).



Слика 1: Хијерархија „подаци - информације - знање - мудрост"

Извор: Приказ аутора прилагођен према *Clark* (2004)

Уколико се у разматрање хијерархијског међуодноса података, информација, знања и мудрости укључи и временска димензија, *DIKW* структура се може приказати у координатном сиситему у форми дијаграма који је познат под називом континуум разумевања. Дијаграм, који је преузет од *Clark*-а (а изворно базиран на претеча идеји *Cleveland*-а из 1982. године) представљен је на Слици 1 (десно) и приказује прелаз од податка преко информација и знања до мудрости, при чему се укључује и разумевање, али не као посебан концепт, већ као подршка на сваком од транзиционих нивоа.

Генерално, знање је детерминисано контекстом, искуством и разумевањем, тако да се стиче посредством давања смисла подацима кроз одговарајуће нивое транзиције. Такође, може се уочити да подаци и информације припадају прошлости, док се знање, повезано са садашњим временом, кроз дубље разумевање претвара у мудрост, која се, пак, односи на доношење одлука и будућност (*Clark,* 2004).

Одвојено посматрање сваког елемента *DIKW* хијерархије резултирало је постојањем бројних покушаја њиховог дефинисања, што је директна последица различитог схватања, значаја и позиције сваког конкретног елемента у различитим научним областима. У складу са тим, у наставку излагања пажња је усмерена на представљање дефиниција релевантних са становишта сврхе овог истраживања, а које довољно јасно указују на појмовно разграничење података, информација, знања и мудрости.

Подаци представљају опажања или чињенице изван контекста, тако да сами по себи, без обзира на извор из којег потичу, у сировом облику, немају конкретно значење. Према *Ackoff*-у (1989) подаци се дефинишу као симболи који репрезентују карактеристике објеката, догађаја и њиховог окружења. На пример, посматрани у организационо-пословном оквиру, подаци се односе на чињенице забележене о пословним активностима, трансакцијама и пословним феноменима. Имајући у виду чињеницу да је *ICT* развој знатно допринео проширењу концепта података, као и дисперзији њихових извора и појавних облика (укључујући не само низове нумеричких и алфанумеричких симбола), савремени (а истовремено довољно илустративан и свеобухватан) поглед на концепт података дао је *Liew* у форми следеће дефиниције: „Подаци су записани (снимљени, забележени - прикупљени и складиштени) симболи и сигнали који се могу прочитати" (*Liew*, 2013, стр. 49). При томе се такође наводи да се симболи могу појавити у различитим видовима, као што су речи, бројеви, дијаграми и слике, а сигнали се односе на сензорска очитавања светлости, звука, мириса, укуса и додира.

Како су подаци извор за креирање информација, информације се дефинишу као процесирани подаци, односно подаци којима је дато одређено значење кроз анализирање веза и односа између података. Другим речима, за разлику од података, информације имају одређено значење са аспекта конкретне ситуације или проблема, често у облику поруке. Заправо, информација је порука која садржи релевантно значење, импликацију или инпут за неку одлуку и / или акцију (*Liew*, 2013, стр. 50). Прецизније, уколико обрада организованих података резултира обликом који за

корисника има одређено (интерпретабилно, смислено и релевантно) значење у односу на конкретну ситуацију и вредност за решење проблема, доношење одлука и реализацију одређених акција, тада такав резултат добија својство информације. Метафорички, обрада података представља начин производње информација, при чему су подаци улазни, а информације излазни елементи овог процеса.

Знање представља резултат процесирања информација у умовима појединаца и непосредно је повезано са применом информација, односно доношењем одлука и покретањем и реализацијом кореспондентних акција, а у вези са конкретно разматраном проблемском ситуацијом. Поред информација, у дискусијама о појмовном одређењу знања често се додају, и са информацијама комбинују, друге компоненте, попут разумевања, способности, учења, искуства, акције и слично. У том смислу, према гледишту *Liew*-а, знање обухвата спознају или препознавање (знати - шта), способност за деловање (знати - како) и разумевање (знати - зашто) и садржано је у уму појединца. Сврха знања је да се, генерално, побољша живот људи, док у пословном контексту знање треба да омогући креирање или повећање вредности за предузеће и све његове стејкхолдере (*Liew*, 2013, стр. 50-51). Сходно наведеном, знање је информација која има контекст (то јест, која је обогаћена смислом и значењем) у конкретној ситуацији и на основу које се могу предузимати одређене активности. Међутим, веома је битно истаћи да знање не произлази директно из скупа прикупљених информација у одређеној области, већ настаје кроз процесе рада са информацијама који су појачани искуством и компентенцијама доносилаца одлука у анализирању и решавању сложених проблема (*Vercellis*, 2009, стр. 7), а који управо информације чине корисним.

Коначно, мудрост представља елемент у хијерархији знања који је рангиран на највишем нивоу, али, којем је посвећена најмања пажња у литератури. У покушајима да се дефинише и одреди садржај појма мудрост укључују се различити елементи, као што су маштовитост, креативност, интуиција, моралне норме, етички кодекси, способност прилагођавања потпуно новим ситуацијама кроз логичко мишљење, холистичко сагледавање проблема и слично. Најчешће се под појмом мудрост подразумева акумулирано знање (обогаћено ускуством), које омогућава да се постојећи концепти и знање из једног подручја примене при разматрању нових ситуација и решавању нових проблема, изван конкретног подручја и постојећих шаблона. Сходно томе, мудрост се може окарактерисати као знање са својством универзалности.

Дакле, према *DIKW* концепту, подаци су основни ниво у хијерархији, информација додаје контекст податку, знање дефинише како употребити информације, а мудрост када и зашто употебити знање (*Jifa*, 2013, стр. 715). Ипак, разграничење фундаменталних елемената *DIKW* хијерархије није једноставно, јер се оне међусобно прожимају, а неке од њих и често користе као синоними. Истина, информација се састоји од података, али подаци нису нужно информације, мудрост јесте знање, али знање није увек мудрост, као и што знање једне особе за другу особу може представљати само информацију, и обратно. У том смислу, полазећи од чињенице да је управо дати контекст оно што омогућава разликовање ових елемената, *Liew* (2007) илуструје контекстуалну зависност кроз питање: Шта је књига?, и наводи следећи одговор: књига је знање из перспективе аутора, информација из перспективе потенцијалног читаоца, а пошто је већ смештена на неку врсту медија за складиштење, такође, представља и податак[3].

Посебно значајно питање у дискусији о знању односи се на различите типове знања. И поред постојања бројних класификација, са становишта доприноса развоју области управљања знања, а самим тим и са аспекта функционисања организација и управљања разноврсним ресурсима (изворима) знања, најзначајнија и широко прихваћена типологија је она коју је формулисао *Nonaka*, а на основу које се прави разлика између имплицитног и експлицитног знања (*Славковић,* 2013, стр. 102). Имплицитно знање је персонално знање које се налази у уму појединца. Рефлектује се у облику искуства, вештина и комуникација, а непосредно је засновано на учењу, личним уверењима, ставовима и систему вредности. Ову форму знања, као иманентном делу сваке личности која га поседује, је веома тешко формализовати, а самим тим и делити са другима. Експлицитно знање је знање садржано у документима, базама података и свим другим формама складиштења изван људског мозга. Ова форма знања се може јасно и формално изразити (кодификовати), документовати, архивирати и, последично, једноставно размењивати. Логично, лакше је управљати експлицитним него имплицитним знањем. Међутим, експлицитно знање није у потпуности одвојено од имплицитног знања. Заправо, ове две форме су међусобно комплементарне, јер без имплицитног знања је тешко, чак и немогуће, разумети експлицитно знање (*Uriarte,* 2008, стр. 1-12).

Осим доприноса у смислу концептуалног разумевања знања и његовог диференцирања у односу на сродне категорије, *DIKW* хијерархија, у пословном

[3] На пример, из угла библиотекара и одређивања библиотечког фонда.

контексту, омогућава да се разуме и објасни процес стицања знања кроз процесе трансформације и кретања од података, као мање компактне форме, до компактнијих, апстракнијих и, са становишта благовременог доношења одлука и спровођења адекватних акција, кориснијих форми знања. С обзиром да је за савремену организацију од виталног значаја да учи како да стицање, дисеминацију и, генерално, управљања знањем остварује кроз повезивање са *ICT* решењима, у том смислу *DM* је још један приступ који стоји на располагању менаџерским структурама за стицање (пре свега експлицитног[4]) знања, обезбеђујући тиме подршку за доношење мудрих одлука у функцији успешног пословања и побољшања организационих перформанси.

1.3. Улога података у савременој економији и пословању

У ери знања и технолошких иновација велике количине података се стварају и складиште у сваком моменту и у свим сферама људског деловања. Генерално, континуирани продор и масовна примена Интернет и информационо-комуникационих технологија допринели су да количина расположивих података током последње декаде расте по експоненцијалној стопи, тако да се очекује да ће до 2020. године бити генерисано више од 16 зетабајта корисних података (*Cavanillas et al*., 2016, стр. 3). Заправо, дигитална револуција је омогућила стварање и складиштење огромне количине разноврсних мултидимензионалних података, које, са становишта могућности анализе, разумевања и управљања, знатно превазилазе људске спознајне способности и перцептивне механизме.

Многе организације (како пословни системи, тако и државне институције) су у праћењу технолошких трендова улагале изузетно велике ресурсе за прикупљање и складиштење података и превасходно биле усмерене на оперативне активности и ефикасност складиштења података (перспектива база података). При томе, не поклањајући довољно пажње анализи и употреби велике количине високо димензионалних, хетерогених и дистрибуираних података, последично, нису сагледане потенцијалне опасности у смислу да прикупљени подаци могу постати пасивне архиве које се ретко или никада не користе. Стога се за врло кратко време појавио проблем како из обиља расположивих података извући вредне информације и корисно знање

[4] Међутим, овде не треба занемарити претходно изнету констатацију о комплементарности експлицитне и имплицитне форме знања. Наиме, *DM* анализу, коју карактерише разноврстан и моћан методолошко-алгоритамски инструментаријум за генерисање експлицитног знања и достизање врха пирамиде знања, практично је немогуће спровести без појединаца који поседују адекватна персонална (имплицитна) знања из домена статистике, машинског учења и вештачке интелигенције, будући да су иста неопходна за разумевање комплексних алгоритамских решења и њихових апликативних могућности у конкретним проблемским ситуацијама.

које ће омогућити доношење квалитетних одлука и допринети остварењу серије позитивних импликација на организационе перформансе.

Међутим, важно је напоменути да су подаци увек имали одређени значај са аспекта функционисања организација, с тим што је у данашњем динамичном окружењу потенцијал њиховог утицаја знатно већи. Такође, треба истаћи да њихов значај није препознат само у пословним (производним и услужним) системима, већ и институцијама државне управе (чак су владе појединих земаља формулисале националне стратегије у овој области), али се у даљем тексту корисност података и њихових ефеката посматра, пре свега, кроз призму економских и пословних импликација. Заправо, многи институти, компаније из *IT* сектора, медији (на пример, *McKinsey* институт, *IBM* компанија, часопис *The Economist,* респективно), као и аутори у научној заједници у својим извештајима и чланцима дају легитимитет начинима размишљања и пословања заснованим на подацима, саопштавајући различите процене, резултате истраживања, коментаре и ставове који јасно указују на потенцијалне предности које се могу постићи правилном употребом података. При томе се подаци често упоређују и изједначавају са златом и нафтом, тако да у ресурсном смислу коришћење велике количине података може да генерише значајну економску вредност (*Kennedy,* 2014). *McKinsey* институт је 2011. године објавио извештај у којем се апострофира улога великих количина података као окоснице за побољшање иновативности, конкурентности и продуктивности (*McKinsey Global Institute,* 2011), док се у *OECD* извештају из 2013. године о подацима као новим изворима раста, наводе потенцијалне улоге података у домену развоја нових производа и услуга, унапређења производних процеса и ланаца снабдевања, унапређења истраживања и развоја, побољшања маркетинг активности и развоја нових менаџмент приступа (укључујући и значајна унапређења постојећих пракси организационог менаџмента) (*OECD,* 2013, стр. 4). *McAffe* & *Brynjolfsson* (2012, стр. 61-68) истичу да су, са становишта побољшања пословних перформанси, одлуке засноване на подацима (чињеницама) једноставно боље у односу на одлуке засноване на интуицији, тако да подаци поседују потенцијал за револуцију менаџемента у смислу ширења нове културе доношења одлука.

Преплављеност савремених предузећа подацима, чија се количина непрекидно повећава, ипак, у значајној мери усложњава њихово функционисање, тако да неопходна прилагођавања усмерена на коришћење потенцијала података не представљају једноставан задатак. У том контексту могу се разматрати следеће

констатације, које су резултат једног истраживања ставова менаџера о присуству и прихватању „поплаве података" (*Kantardžić*, 2011, стр. 11):

- 61% менаџера сматра да је преоптерећеност подацима присутна на њиховом радном месту;
- 80% менаџера сматра да ће се ситуација погоршати;
- више од 50% менаџера игнорише податке у процесу доношења одлука;
- 84% менаџера складишти податке и информације за будући период и не користи их у текућим анализама;
- 60% менаџера сматра да су трошкови прикупљања података већи од њихове вредности.

Међутим, са дате листе, оне констатције које указују на архивирање и игнорисање података свакако не могу бити оправдане.

Као што се на основу наведеног може запазити, неспорно је да постоје значајне предности које се могу постићи путем коришћења велике количине података. Другим речима, велике количине података креирају велике пословне могућности. Али, такође постоји све дубљи јаз између количине прикупљених података и њихове употребе у процесу стицања / издвајања знања и пословног одлучивања, о чему сведоче и процене Светског економског форума, према којима се данас се користи свега 0,5% свих расположивих података (*Солдић–Алексић & Chroneos Красавац,* 2016, стр. 240). Наиме, као што је већ речено, подаци сами по себи нису интересантни и уколико се не користе они за предузеће представљају трошак. Све док запослени не почну да употребљавају податке при обављању својих задатака, трошкови које узрокује одржавање података ће бити већи од вредности стварних или потенцијалних информација које се из њих могу извести (*Panian & Klepac*, 2003, стр. 45). Суштински, овај проблем се може изразити кроз питање: Како искористити предности поседовања података?. Односно, другим речима, да ли ће се организације заиста „удавити у мору" података или ће велика количина података постати вредан извор компаративне предности?

Решење проблема свакако није игнорисање и занемаривање података. Напротив, како су подаци, као стални производ технолошког развоја, постали фактор без којег је тешко замислити функционисање у савременом окружењу (чија је доминантна карактеристика неизвесност), за привредне субјекте је од посебне важности да се оспособе да користе податке у процесима креирања знања и доношења пословних одлука. Управо, тежња за разумевањем великих скупова података, као и за

превазилажењем дубоког јаза између расположивих података и њихове искоришћености у процесу одлучивања условили су потребу за прихватањем и развијањем потпуно нових начина прикупљања, организације, моделирања и анализе података. У том смислу, супституција традиционалних приступа у анализи података новим приступима, уз одговарајући, *ICT* решењима подржан, методолошки инструментаријум, који, сходно томе, располаже интелигентним и аутоматским могућностима за процесирање и трансформацију податка у корисне информације и знање, постаје императив. Наиме, да би се пословни сиситеми адаптирали на нове захтеве и промене које доноси расположивост велике количине података у контексту њихове употребе за унапређење различитих аспеката пословања, јасно је да су, поред препознавања и схватања значаја података као извора корисних информација и знања, неопходна и одговарајућа усклађивања стварних потреба и аналитичких могућности (људских способности и технолошких предуслова и капацитета) за процесирање и анализу података.

Заправо, у савременом окружењу многи проблеми су повезани са коришћењем традиционалних приступа у обради података. Сходно томе, стварање и развој нових научних приступа и методологија за претраживање, обраду и анализу података представља један од потенцијалних начина за прилагођавање новом окружењу и стицање ширег увида у његове карактеристике. У том контексту, као директна последица растућег значаја података, иницира је и вођен развој *DM*-а, као мултидисциплинарног приступа у анализи мултидимензионалних података из различитих перспектива. Осим наведеног, покретачка снага која је, такође, допринела растућем интересовању за *DM* приступ, како у оквиру академских истраживања, тако и у домену практичних апликација, је експанзија економских активности и, последично, потреба за применом *DM* метода у економским и пословним анализама. Будући да се у савременој економији и пословном свету данашњице подаци сматрају значајним ресурсом, који са развојем и проширењем *DM* методолошког оквира добија све већу вредност, способност издвајања корисног, али углавном скривеног знања из података, као и способност предузимања конкретних акција на основу таквог знања постаје круцијални фактор конкурентности, опстанка и пословног успеха.

Са становишта макроекономских истраживања, *DM* приступ омогућава откривање знања у форми законитости о структури, односима и тенденцијама у кретању макроекономских феномена и обезбеђује предвиђање процеса и догађаја у макроекономским системима. С друге стране, са становишта микроекономских

истраживања, овај приступ је усмерен на откривање законитости из података у циљу ефективног решавања пословних и управљачких проблема на свим организационим нивоима. Обезбеђујући квалитетне информације из сирових података и њихову трансформацију у корисно знање о различитим аспектима пословања, *DM* постаје неопходан сегмент система за подршку одлучивању. Међутим, не треба занемарити чињеницу да људски фактор и даље има кључну улогу у процесу одлучивања, а да напредни, софтверски подржани методи олакшавају тај процес.

Као што је случај и са сваком иновативном апликацјом, *DM* је „ прилично лако лоше спровести" (*Larose*, 2005, стр. xii), а недовољно знања и аналитичког искуства и вештина може бити посебно опасно. На пример, погрешни закључци могу настати као последица неадекватне припреме података за анализу. Стога, приликом примене *DM* методологије треба бити јако обазрив. Реч је о процесу који се не сме по аутоматизму применити, нити је *DM* примерено средство за решавање свих пословних проблема. За валидан избор и успешну примену одређеног *DM* метода (или комбинације метода) у конкретној ситуацији, неопходно је познавати како карактеристике метода, тако и проблема који се решава. Само на тај начин може се обезбедити компатибилност проблемске ситуације и примењеног метода, а самим тим и корисност крајњих резултата у форми уочене и моделиране законитости о разматраном феномену, јер различити проблеми траже различите начине њиховог решавања. При томе, важна претпоставка за оптимално коришћење података у функцији креирања знања применом *DM*-а, као моћног алата пословне интелигенције, је да сви актери (то јест, учесници) у реализацији *DM* процеса, поседују, у складу са својом улогом у појединим фазама процеса, одређени ниво знања како о разматраном феномену, тако и о примењеном методолошком приступу.

2. КОНЦЕПТ *DATA MINING*-а

Током последњих деценија, као последица опште компјутеризације и потребе за екстракцијом вредних информатичких садржаја из гигантских колекција (дигиталних) података, генерисаних кроз обављања свакодневних активности и трансакција у свим сферама људског деловања, феномен *data mining* је постао актуелна тема како у пословном свету, тако и у друштву у целини. Сходно томе, у овом Поглављу је, најпре, објашњено изворно терминолошко одређење *DM* синтагме, а затим су приказане различите дефиниције, кључни елементи, настанак и развој *DM* концепта.

2.1. Терминолошко одређење појма *data mining*

Услед огромног интересовања и истраживачког и медијског фокуса на различите аспекте *DM*-а, у релевантној литератури могу се пронаћи бројне дефинције овог појма. Стога, без претензија ка потпуном обухвату различитих приступа, у наставку следи кратак осврт на разматрани феномен у контексту значења сегмената ове синтагме, а затим и приказ репрезентативних, а довољно хетерогених дефиниција, укључујући и тумачења пратећих елемената, релевантних за правилно разумевање суштинских одређења *DM*-а.

Дакле, поставља се питање: шта јесте (а шта није) *data mining*?

Сам термин *data mining* је проистекао из сличности која постоји између трагања за корисним информацијама у великим количинама расположивих података и активности ископавања у циљу екстракције конкретних ресурса скривених у земљи. Оба процеса захтевају претрагу изузетно велике количине сировог материјала у циљу проналажења драгоценог грумења. Идеја наведене аналогије је врло једноставна. Прогрес у аквизицији дигиталних података и технологији њиховог складиштења резултирао је стварањем „планина података". Оне представљају метафору за изворе (руднике) драгоцених и квалитетних информација за пословне субјекте. Међутим, да би се екстраховале вредности из ових извора неопходно је трагати, односно спровести дубинску анализу података. Дакле, повезаност појмова подаци (енгл. *data*) и ископавање (енгл. *mining*) у контексту синтагме *data mining* указује на дубинску анализу велике количине података у циљу откривања корисних информација. Једноставно, *DM* се односи на екстрацију знања из огромних количина података.

Осим термина *data mining*, за откривање потенцијално корисних информација у подацима, често се користе и други називи, као што су: екстракција знања (енгл. *knowledge extrction*), откривање информација (енгл. *information discovery*), жетва информација (енгл. *information harvesting*), археологија података (енгл. *data archeology*) и анализа образаца у подацима (енгл. *data pattern processing*) (*Fayyad et al*., 1996, стр. 37; *Cios et al*., 2007, стр. 10). Међутим, без обзира на извесну терминолошку конфузију, *DM* је општеприхваћен термин који верно описује процес чији је циљ откривање мале количине драгоценог грумења из великих количина сировог материјала. Енглеска синтагма *data mining* код нас се најчешће преводи као: дубинска анализа података, интелигентна анализа података, истраживање података, трагање кроз податке, откривање законитости у подацима или, пак, рударење података.

У оквиру терминолошког расветљавања и објашњења *DM* феномена интересантно је осврнути се на још један аспект употребе самог термина. Наиме, за добијање информација из података одавно се користи статистика, тако да су и у статистичким круговима постојали напори за претраживање података на начин како се то чини у оквиру *DM* приступа. Као синоними за *DM* термин коришћени су називи попут, прекопавање података (енгл. *data dredging*), њушкање по подацима (енгл. *data snooping*) и пецање по подацима (енгл. *data fishing*) (*Hand et al*., 2001, стр. 22).

Заправо, статистичари су *DM* концепт прихватили са значајном резервом, тако да наведени термини носе извесну негативну конотацију која је проистекла из чињенице да ће довољно дуго и интензивно, компјутерски подржано, претраживање велике количине било каквог скупа података (чак и случајно генерисаног) омогућити не само идентификовање извесних небитних открића и проблематичних резултата, већ ће их учинити и статистички значајним (*Hand et al*., 2000, стр. 111; *Hand*, 2009, стр. 443). С обзиром да је, нарочито као последица развоја рачунарства, број потенцијалних испитивања великих скупова података из различитих перспектива у сушини неограничен, увек се, након исцрпног трагања, могу пронаћи одређени модели који су арбитрарни и довољно добро прилагођени подацима. Али, многи од генерисаних модела представљају резултат случајних флуктуација или су карактеристични за одређени временски тренутак, тако да се отвара проблем генерализације формулисаних правилности изван оквира расположивих података. Наведено схватање *DM*-а је у складу са статистичком досетком економисте *Coase*-а да „уколико довољно дуго мучите податке природно је да ће признати све или било шта" (*Scarpa*, 2011, стр. 337). Међутим, *DM* термин се не односи само на техничике, мање или више комплексне, аспекте анализе, независне од значаја и садржаја анализираних података. Напротив, експертиза у домену идентификовања и евалуације потенцијалних открића (резултирајућих правилности) из угла власника и корисника података је пресудан фактор у обезбеђивању екстракције корисних информација. Отуда, упркос присутној дози скепсе у прошлости, данас, са статистичког становишта, *DM* термин носи позитивно значење процеса усмереног на конструкцију (статистички) валидног модела екстрахованог из података, односно идентификовање релевантних законитости о разматраним феноменима.

DM се дефинише на различите начине. Из богате ризнице дефиниција, у контексту ових разматрања, наводе се оне за које се претпоставља да јасно указују на есенцијална, концептуална одређења *DM*-а.

Концизна и често цитирана дефиниција је: „*DM* представља откривање значајних, интересантних и неочекиваних структура у великим скуповима података" (*Hand,* 1999a, стр. 433). Дужа и целовитија верзија дефиниције *DM*-а, коју је такође дао *Hand et al.* (2001, стр. 1), гласи: „*DM* је анализа (често великих) опсервационих скупова података, путем иновативних, софистицираних начина, у циљу проналажења неочекиваних релација, веза и сумарних приказа података, који су и разумљиви и корисни за власника података".

Fayyad et al. (1996, стр. 39-40) посматрају *DM* као корак у процесу откривања знања, чија је срж примена специфичних алгоритама за анализу података, који, уз прихватљива ограничења у погледу ефикасности израчунавања, резултирају читавим низом образаца (или модела) у великим количинама података.

Према експертима водеће светске компаније која се бави истраживањима на пољу информационих технологија и пружањем консултатнтских услуга *Gartner Group, DM* је процес откривања значајних нових веза, правила и трендова кроз испитивање великих количина података складиштених у репозиторијумима података, користећи при томе технологије препознавања образаца (енгл. *pattern recognition*), као и статистичке и математичке технике (*http://www.gartner.com/it-glossary/data-mining*).

Tufféry (2011, стр. 4) истиче да је *DM* сет метода и техника за истраживање и анализирање скупова података (који су по правилу велики), на аутоматски или полуаутоматски начин, у циљу проналажења непознатих или скривених правила, веза или тенденција у подацима.

Berry & Linoff (2004, стр. 4) дефинишу *DM* као процес истраживања и анализе велике количине података уз помоћ аутоматских и полуаутоматских алата, у циљу откривања значајних образаца и правила.

У предговору другог издања уџбеника чији су аутори *Shmueli et al.* (2010, стр. 17), *Pregibon* истиче да је *DM* уметност екстракције корисних информација из великих количина података, указујући истовремено на његов растући значај у данашњем свету.

Довољно комплетна дефиниција, коју је предложио *Giudici* (2003, стр. 2), гласи: „*DM* је процес избора, истраживања и моделирања велике количине података како би се открили законитости или релације које су у почетку непознате, са циљем добијања јасних и корисних резултата за власника базе података".

Слично, *Azzalini & Scarpa* (2012, стр. 5) наводе да *DM* обухвата рад на процесирању (графичком и нумеричком) велике количине или непрекидних низова података са циљем да се екстрахују информације корисне за оне који поседују податке.

Према *Kantardžiću* (2011, стр. 6), „*DM* је процес откривања различитих модела, законитости и вредности изведених из скупа прикупљених података".

Имајући у виду варијететност презентованих дефиниција, условљену пре свега изабраним аспектима посматрања, *DM* концепт се може генерално определити и размотрити као:

- вишеетапни, интерактиван, итеративан и креативан процес проналажења интересантних вредних информација у великим репозиторијумима података;
- централна фаза у процесу откривања знања из података, у којој се путем софистицираних алгоритама идентификују потенцијално корисне и карактеристичне структуре у подацима; и
- скуп компјутерски подржаних метода дизајнираних за претраживање и дубинску анализу велике количине података у циљу проналажења законитости.

На основу наведеног, такође, недвосмислено произлази да *DM* није смањење обима података „по сваку цену", „слепа" апликација алгоритама и софтверских решења, покушај проналажења значајних веза тамо где оне не постоје, презентација и визуелизација података на различите начине, постављање упита и тумачење одговора из база података, или, тешко разумљива технологија која се базира искључиво на високом нивоу информатичког знања. Да би се разумела суштина *DM*-а, често се истиче да *DM* није тражење броја у телефонском именику или *Google* претрага на бази кључних речи.

Будућа истраживања ће сигурно довести до нових тумачења *DM* појма укључујући и, тренутно недостајућа, прецизнија дефинисања концепцијских фундамената *DM* као самосталне научне дисциплине и јаснију дистинкцију у односу на сродна подручја (попут, науке о подацима и пословне интелигенције). У суштини, дефинисање научне дисциплине је увек контроверзни задатак, што неизбежно прати и *DM,* као релативно младу дисциплина која се изузетно брзо развија и мења. У овом истраживању, сходно његовој сврси, полази се од дефиниције *DM*-а као науке која се бави извођењем законитости (корисних информација и знања) из великих скупова или база података (*Hand et al*., 2001. стр. xxvii).

2.2. Кључни елементи *data mining* концепта

Разматрање бројних дефиниција *DM*-а из различитих перспектива омогућава да се идентификују есенцијални елементи који представљају основу за интерпретацију овог концепта. Јасно се могу издвојити следећи елементи *DM* приступа у анализи

података: ▸ екстраховане смислене правилности (законитости), ▸ конкретно подручје на чији се домен издвојене правилности односе, ▸ велика количина података, и ▸ заснованост анализе података на процесном приступу у решавању разматраних проблема.

Опште је познато да се свакодневно, као последица нових *ICT* решења и непрекидног унапређења компјутерских механизама и капацитета за аутоматско складиштење и обраду података, генеришу велике количине података. Зато се и каже да је савремено доба, заправо, доба података. Међутим, из перспективе власника података, поседовање података само по себи нема вредност. Стога, разумевање података и процеса који их стварају, могућност издвајања информација (скривених у подацима) и способност стицање нових знања у конкретној области, у данашњем животном амбијенту и пословном свету, све више добија на значају. На том путовању од података ка знању, једно од превозних средстава јесте *data mining*. Сходно томе, како је проналажење смисла у подацима основни циљ *DM*-а (*Cios et al*., 2007, стр. 3), издвојене, вредне и смислене, правилности представљају један од круцијалних концептуалних *DM* елемената.

Правилности које настају као резултат *DM* анализе дефинишу се као структуре у подацима. *Hand et al*. (2001, стр. 9-10) праве разлику између два типа структура у подацима: глобалних модела и локалних образаца / правила (енгл. *patterns*). Модел, као глобална структура се односи на све тачке у простору података, а образац, као локална структура се односи на ограничене регионе (сегменте) простора података, што указује да се само неки делови (сегменти / слогови) података понашају на одређени начин. У *DM* контексту образац се једноставно интерпретира као локални модел, односно локално правило. Другим речима, глобални модел се односи на глобалну дескрипцију скупа података, док локални модел репрезентује локалну карактеристку података и односи се на део јединица посматрања и / или варијабли. Заправо, правилности представљају компактне приказе карактеристика сирових података који омогућавају откривање новог знања скривеног у великој количини података. Примери откривених правилности су: линеарни регресиони модел, класификациони модел, модел груписања, графички прикази, рекурентни обрасци у временским серијама и слично.

Основни проблем у процесу идентификовања законитости у подацима налази се у чињеници да је потенцијално њихов број огроман, односно, *DM* систем може генерисати на хиљаде правила, што знатно отежава валидацију сваког од њих. Сходно томе, поставља се питање да ли је откривена структура података, односно, правилност

довољно интересантна, необична или значајна да би била вредна пажње. Заправо, све идентификоване правилности не представљају знање. Само довољно интересантне, поуздане и значајне правилности репрезентују знање. У складу са наведеним, *Han et al.* (2012, стр. 21) постављају још три битна питања која се односе на резултирајућа *DM* правила:

- Шта откривено правило чини интересантним?
- Да ли *DM* систем може генерисати сва интересантна правила?
- Да ли *DM* систем може генерисати само интересантна правила?

Да би се правилност идентификована у нетривијалном процесу откривања знања оквалификовала као интересантна мора да поседује следећа есенцијална својстава (*Han et al.*, 2012, стр. 21; *Cios et al.*, 2007, стр. 3): разумљивост, валидност, иновативност и корисност. У датом контексту, термин нетривијалан се не односи на једноставно израчунавање и одређивање статистичких мера (попут, аритметичке средине, модуса и слично), већ указује на својеврстан вид истраживања, анализе и закључивања који је усмерен на идентификовање интересантних, вредних и смислених структура, правила у великим скуповима података. Наведена својства интересантности имају следећа значења (*Fayyad et al.*, 1996, стр. 41):

- валидност: односи се на захтев да су откривена правила генерално одржива и да са одређеним степеном поузданости важе и за нови скуп података, односно не представљају специфичност својствену само подацима који су коришћени за њихово извођење;
- иновативност: односи се на захтев да откривена правила буду нова са становишта постојећег система знања или, пак, нова са становишта самог корисника;
- корисност: односи се на захтев да откривена правила буду потенцијално корисна са становишта корисника (или власника података) и реализације конкретног задатка у процесу откривања знања, тако што омогућавају да корисник или (пословни) систем оствари постављене циљеве;
- разумљивост: односи се на захтев да откривена правила буду довољно разумљива за корисника како би се кроз правилне, коректне и логичне интерпретације разматраних феномена у форми знања омогућило остваривање одговарајућих користи.

Поред наведених својстава, у објашњењу интересантности откривених правила неопходно је указати на још једну димензију: уколико је конкретно *DM* истраживање базирано на провери хипотезе која је дефинисана од стране корисника у смислу да ли

познато, већ откривено правило важи и за нови скуп података, тада се утврђено правило може окарактерисати као интересантно (значајно) било да су потврђена очекивања (установљена и већ дефинисана правила) било да је утврђени резулатат на новом скупу података неочекивано контрадикторан уобичајеним законитостима (веровањима) (*Han et al*., 2012, стр. 21).

Дакле, у сваком скупу података могу бити издвојене бројне законитости, али само оне које се сматрају довољно интересантним представљају знање. Квантификовање интересантности, односно, степена валидности, новитета (иновативности), корисности и разумљивости спроводи се путем објективних мера заснованих на статистици и одређивању критичних вредности за дефинисане мере. Међутим, многа правила која су значајна према објективним критеријумима могу бити опште позната (на пример, уобичајено је да гондоле са слаткишима или играчкама посебно привлаче пажњу деце и утичу на њихово понашање) и, стога, неинтересантна у *DM* контексту. Услед наведеног, осим објективних мера, за вредновање квалитета откривених структура користе се и субјективне мере које одражавају потребе и интересе конкретног корисника. На пример, правила, која описују карактеристике купаца у *Zara* продавницама (једног од највећих модних малопродајних ланаца на свету) треба да интересују менаџере маркетинга, док та иста правила су од малог значаја са становишта кадровске политике и аналитичара који проучава податке трагајући за правилима која се односе на карактеристике запослених. Једноставно, субјективне мере базирају се на мишљењу и уверењу корисника о вредности и апликативности откривених правила. Сходно томе, при одређивању вредности правила и дефинисању иницијалних хипотеза за покретање процеса откривања знања неопходно је, у свако од претходно дефинисаних својстава интересантности, инкорпорирати елементе постојећег знања из конкретне области, као и *a priori* знање експерата о проблему који се анализира. Заправо, да ли је и колико откривено знање значајно зависи од конкретног корисника и домена апликације.

Одговори на друго и треће питање представљају, такође, велики изазов у домену *DM*-а, при чему питање генерисања свих интересантних правила одражава проблем комплетности *DM* алгоритма (мада је често нереално, али и неефикасно да *DM* систем генерише сва могућа правила), док је питање идентификовања само интересантних правила оптимизациони проблем (*Han et al*., 2012, стр. 22).

Како је знање кориснички оријентисан концепт, један од концептуалних *DM* елемената је област (домен) којој откривено правило припада (*Cios et. al*., 2007, стр. 5).

При томе, коначан суд о томе да ли откривена структура поседује вредност и смисао у оквиру разматраног проблема доноси експерт из конкретне области интересовања и посматрања, тако да се отвара питање релативне вредности резултата *DM*-а. Због тога је од изузетног значаја да *DM* истраживачи тесно сарађују са експертима из области којој предмет анализе припада, јер управо те особе својим знањем и искуством могу бити способне да нека на први поглед бесмислена правила интерпретирају на корекатан начин и иста претворе у вредне информације.

Велика количина података је, такође, есенцијални елемент *DM* концепта. По правилу, *DM* анализа се спроводи на великим скуповима опсервационих података који су претходно прикупљени за неку другу сврху, због чега се ова анализа често назива секундарном анализом података (*Han et al.*, 2012, стр. 1-2).

Као што је већ истакнуто, способност производње и прикупљања (складиштења) података је знатно повећана као директна последица *IT* развоја. Међутим, поставља се питање: колико је заиста велика „велика" количина података?. За разлику од података највећих светских компанија из педесетих година XX века који су заузимали неколико десетина мегабајта складишног простора (*Shmueli et al.*, 2010, стр. 32), размере савремених скупова и складишта података илуструју следећи примери (*Cios et al.*, 2007, стр. 4):

- *AT&T* компанија, један од светских лидера у пружању телекомуникационих услуга, реализује преко 300 милиона позива дневно за приближно 100 милиона корисника и складишти податке, између осталог, о времену и дужини трајања позива, у мулти-терабајтна складишта података;
- *Wal-Mart*, у својим продајним објектима, реализује око 21 милион трансакција дневно и складишти податке о обављеним трансакцијама у репозиторијуме капацитета око 10 терабајта;
- Систем за осматрање Земље америчке Националне ваздухопловне и свемирске администрације (*NASA*) пројектован је да генерише и до 50 гигабајта података по сату;
- *Mobil Oil*, нафтна компанија, складишти стотине терабајта података који се односе на истраживање и експлоатацију нафтних извора;
- Америчка агенција за унутрашњу безбедност прикупља петабајте података о грађанима својих и других земаља.

Интересантно је у овом контексту навести и резултате студије из 2000. године која је спроведена на Универзитету у Калифорнији под вођством истраживача *Lyman*-а

и *Varian*-а, а према којој се у свету годишње произведе између 1 и 2 ексабајта нових информационих садржаја, што је око 250 мегабајта по свакој особи на планети Земљи. У поновљеном истраживању из 2003. године, оцењено је да је 2002. године произведено 5 ексабајта података, што је двоструко више у односу на 1999. годину (*Lyman & Varian,* 2003).

Према неким другим изворима, током 2010. године предузећа су складиштила више од 7, а клијенти више од 6 ексабајта нових података, а процењени кумулативни обим података у истој години је више од 1000 ексабајта, док ће се до краја 2020. године количина података повећати 40 пута (*OECD*, 2013, стр. 8). Такође, пример који довољно илустративно указује на трендове у револуцији података односи се малопродајни гигант *Wal-Mart* где се сваког сата обраде и складиште подаци за више од милион трансакција купаца (око 2,5 петабајта), који су по обиму 167 пута већи од укупне количине података садржане у књигама Конгресне библиотеке у САД (*The Economist,* 2010).

Често се у дебатама о овако великим скуповима истиче да су то монструозне, чудовишне, бесмислене категорије људском перцептивном механизму тешко схватљиве. Међутим, рад са скуповима података који припадају категорији изразито великих скупова је постао реалност. Стога, изнете процене, представљене занимљиве релације и примери великих скупова (без критичког осврта на њихову веродостојност) јасно указују да се савремени свет сусреће са све већом количином података и да је од изузетне важности да се актери, нарочито у пословном свету, оспособе да искористе расположиве податке у процесу стицања знања и пословног одлучивања. У том смислу, са повећањем расположиве количине података не само да расте значај *DM*-а, већ се јавља потреба за развојем нових технологија у функцији оперативног коришћење велике количине података, које су засноване на умреженим рачунарима и паралелном процесирању, а чије разматрање и дубље елаборирање прелази оквире истраживања у овој дисертацији.

Генерално, квантификовање количине података, заиста, није једноставан задатак, тако да се при овим мерењима често спекулише, а саопштени резултати (и процене) бројних студија су не ретко контрадикторни и нереални, при чему не постоји могућност провере њихове тачности. Осим тога, величина количине података је сама по себи релативна категорија. Заправо, количина података која се данас сматра великом убрзо може постати уобичајена или мала, као што и величина која је за један пословни систем велика, за други може бити мала.

Имајући у виду да су предмет *DM* истраживања велики скупови података, јасно је да не постоји начин на који појединац може самостално и без аутоматизације одговарајућих поступака обрадити ове податке. Међутим, при употреби термина аутоматски треба бити јако обазрив. Без аутоматског процесирања немогуће је спровести *DM* анализу, али је погрешно закључити и веровати да је *DM* производ који се може купити. Заправо, то је приступ којим треба овладати и процес који треба разумети. Постоји потенцијална опасност да се интерпретација читавог концепта сведе на примену аутоматизованих алата за откривање значајних информација и знања из података. Због наведеног, треба нагласити чињеницу да пресудну улогу у спровођењу *DM* анализе има људски фактор, као и да су софтверска решења помоћни (али неизоставни) алати који аутоматски не решавају *DM* проблеме.

Сходно томе, још један веома важан аспект посматрања и кључни елемент у концептуалном одређењу *DM*-а јесте термин процес. *DM* се не може сматрати једнократном активношћу нити једноставним скупом изолованих алата чијим се избором и применом аутоматски добијају решења дефинисаних проблема, већ је реч о креативном, итеративном и интерактивном процесу откривања законитости из података, који се реализује кроз тимски рад *DM* аналитичара и експерата из области на коју се односи предмет истраживања.

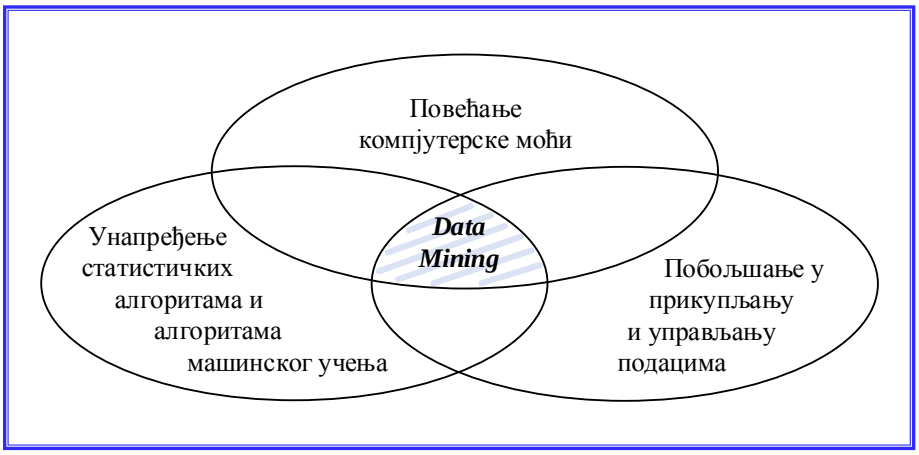
2.3. Развој *data mining*-а и еволуција у анализи података

Интензивно коришћење компјутерски заснованих система у свим аспектима пословног и друштвеног живота резултирало је акумулацијом огромних количина података. Истовремено, потреба за анализом, разумевањем и коришћењем ових података представља кључни разлог који је условио револуцију у анализи података и допринео развоју широке лепезе софистицираних алата за издвајање корисних информација (знања) из великих скупова (пре свега) опсервационих података.

У том контексту, *Han et al.* (2012, стр. 2), истичу да се *DM* може сагледати као резултат природне еволуције информационе технологије. Генерално, идеја о природној еволутивној путањи информационе технологије која је резултирала развојем *DM*-а прихваћена је од стране многих истраживача у овој области. Према *Thearling*-у ова еволуција је започела од тренутка када су пословни подаци први пут прикупљени у електронској форми и складиштени у меморији рачунара (*Thearling,* 2003). Еволутивни процес се континуирано настављао, најпре, у правцу побољшања приступа подацима, а затим кроз развој технологија које омогућавају корисницима да ефикасно управљају

подацима у реалном времену, до предвидиве и проактивне екстракције и испоруке информација. Заправо, упоредо са технолошким развојем, расле су и могућности складиштења и процесирања све веће количине података. У таквим околностима, на том еволутивном путу, ефективна и ефикасна анализа података нужно је захтевала супституцију традиционалних техника за анализу напредним *DM* аналитичким поступцима и алгоритамским решењима, с тим што су се од 2000. године на пољу рачунарске технологије одиграле иновативне промене које је било готово немогуће замислити и пројектовати. Импликације насталих промена у погледу не само обима података, већ, касније и њихове структуре, тока, извора и типова отвориле су читав низ, и данас актуелних, питања и проблема, али и могућности и идеја на пољу унапређења постојећих и развоја нових технологија за управљање и анализу података, у оквиру којег *DM* заузима посебно место.

Често се наводи да је, у еволутивном контексту, *DM* резултат развоја следећих технологија, које су омогућиле и подстакле спровођење *DM* процеса у пракси: ▸ снажне мултипроцесорске рачунарске технологије, ▸ технологије складиштења огромних количина података и управљања подацима, и ▸ алгоритамских техника и алата за претраживање података (*Thearling*, 2003). Њихова спона је илустрована на Слици 2.



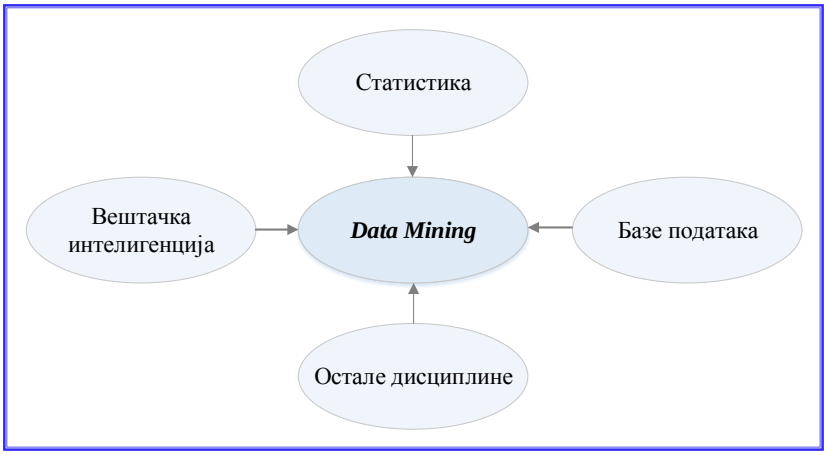
Слика 2: *Data mining* и конвергенција три технологије

Извор: Приказ аутора прилагођен према *Thearling* (2003)

Наиме, задатак откривања законитости у подацима уопште, а самим тим и у економским подацима није нов, тако да су процедуре многих *DM* техника одавно познате. Међутим, како сложеност алгоритама зависи, између осталог, од анализиране количине података, са порастом количине података бројни *DM* алгоритми су, практично постали неприменљиви у решавању реалних проблема. Све снажнији и доступнији рачунари су знатно изменили овакву ситуацију. Наиме, раст процесорске

снаге доступних рачунара је утицао на повећање не само обима активности у обради података, већ и на измену постојећих, као и појаву нових концепата и алата за претраживање података. Дакле, с једне стране, потреба за *DM* настала је као последица акумулирања изненађујуће брзо растуће количине података. Истовремено, с друге стране, напредак рачунарске технологије је омогућио примену *DM*-а. Сходно томе, одговарајућа технлошка основа и препознавање *DM* потенцијала довели су до тога да *DM* постане интегрални део пословне праксе и развојних и пословних стратегија.

DM је релативно млада мултидисциплинарна научно-истраживачка област, која је повезана са читавим низом различитих, добро утемељених области, као што су: статистика, вештачка интелигенција и машинско учење, базе података, теорија информација, математика, логика, управљање подацима, експертни системи, визуелизација података, као и читав низ придружених области и подобласти. Свака од ових области се карактерише бројним специфичностима, што је, последично, утицало и на креирање специфичних карактеристика које се искључиво односе на *DM*. Из читавог спектра различитих доприноса развоју *DM*-а ипак је могуће издвојити три генеричка корена на којима се заснива методолошка и терминолошка основа *DM*-а. Слика 3 илуструје корене *DM*-а који се везују за развој статистике и вештачке интелигенције, као академских дисциплина, али и за праксу обраде велике количине података уз примену технологија управљања базама података и система за подршку одлучивања.



Слика 3: *Data mining* корени

Извор: Приказ аутора прилагођен према *Gorunescu* (2011, стр. 2)

Статистика је најстарији *DM* корен који има виталну улогу у свим фазама *DM* процеса. Осим примене у моделирању конкретне проблемске ситуације, класични статистички методи и процедуре се интензивно користе и у другим фазама за бројне

сврхе, попут, уклањања ирелевантних и редундантних атрибута, детекције шумова у подацима, тестирања превелике прилагођености модела подацима за учење и оцењивање конструисаних модела.

Вештачка интелигенција је област из које потичу многе технике моделирања правилности у подацима засноване на хеуристици и симулацији људских способности и понашања (перцепција, реаговање, понашање, резоновање, закључивање и чињење). Подобласт вештачке интелигенције која има изузетно важну улогу у развоју *DM*-а позната је под називом машинско учење. Односи се на развој алгоритама који омогућавају рачунарима да унапређују сопствено понашање кроз учење на основу емпиријских података из база података или са сензора. Међу базичним алгоритмима за учење (прва генерација алгоритама), који су кроз широку употребу у решавању *DM* проблема дали добре резултате, издвајају се: неуронске мреже, правила повезивања, генетски алгоритми (енгл. *genetic algorithms*) и закључивање на бази случајева (енгл. *case-based reasoning*).

Системи за управљање базама података се сматрају трећим *DM* кореном, који се везује за практичне аспекте обраде велике количине података. Ови системи олакшавају спровођење *DM* анализе и обезбеђују основне изворе сирових података, повезујући их, сходно специфицираним захтевима корисника, на логичан начин. Упоредо са развојем нових генерација система база података отвара се простор и за унапређење перформанси *DM* техника. На пример, могућност паралелног извршења упита (као врло битна и пожељна карактеристика базе података о којој треба водити рачуна приликом креирања самих упита) може знатно допринети скраћењу времена потребног за претраживање великих скупова података при формулисању потенцијалних хипотеза и њихове верификације на подацима за учење. Чак и незнатна побољшања на пољу база података су врло корисна са становишта ефикасности *DM* процеса.

Претходна разматрања имплицитно указују да за потребе *DM* анализе није довољн само статистички приступ, нити било који парцијални приступ из перспективе осталих *DM* корена. Заправо, у *DM* су инкорпорирани доприноси свих наведених дисциплина. При томе, базе података, машинско учење (као подобласт вештачке интелигенције) и статистика се могу посматрати као *DM* перспективе које истичу ефикасност, ефективност и валидност примењених техника и добијених резултата, респективно (*Zhou*, 2003, стр. 139). Међутим, то сигурно не значи да машинско учење и статистика не воде рачуна о питањима ефикасности, односно, базе података и статистике о питањима ефективности, или пак, базе података и машинско учење о

питањима валидности. Само симултано респектовање сва три аспекта, укључујући и ефекте њихове интеграције, може довести до корисних *DM* резултата.

Дубља анализа сваке од наведених области превазилази обим и контекст ових истраживања. Међутим, због значаја који статистика има у обезбеђењу теоријске основе за анализу података и чињенице да статистичке процедуре и поступци имају главну улогу у скоро свим фазама *DM* процеса (укључујући и евалуацију алгоритама машинског учења), однос између *DM*-а и статистике је издвојен, детаљно разматран и дискутован у Поглављу 7.

3. ОТКРИВАЊЕ ЗНАЊА И *DATA MINING*: ПРОЦЕСНИ МОДЕЛИ

Сходно чињеници да откривање знања није једнократна активност, већ итеративан, интерактиван и креативан процес, у овом Поглављу су приказани различити процесни модели, као стандардизоване процедуре за откривање знања из података. Такође је указано на значај одговарајућих експертских знања у свим фазама и активностима овог процеса. Посебно је апострофиран значај интеракције свих учесника процеса у остваривању пословних и *DM* циљева.

3.1. Процесни приступ у откривању знања из података

Откривања знања из (база)[5] података (енгл. *Knowledge Discovery in Databases – KDD*) представља врло динамично истраживачко и развојно подручје. Раст истраживачког интересовања је узрокован насталом потребом и повећаном тражњом за алатима који ће помоћи у анализи и разумевању великих количина података. Међутим, једноставно познавање и пука примена софистицираних поступака за анализу података нису довољни за издвајање нових, потенцијално корисних и разумљивих резултата у форми законитости које су скривене у подацима. Овакве особине резултата могу се постићи само помоћу дефинисања јасног и једноставног процедуралног оквира за откривање знања из података. Дакле, пре него што се приступи издвајању знања из података, неопходно је формално структурирати општи приступ, односно стандардизовати процес откривања знања.

Кључни мотив за увођење процесног приступа и дизајнирање процесних модела за откривање знања из података је дефинисање општег оквира који ће, кроз интеграцију свих процесних активности, осигурати да крајњи резултат има

[5] Процес откривање знања из података, генерално, односи се на различите изворе података, мада се у оваквој формулацији истиче улога база података као доминантног извора података у том процесу.

одговарајућу употребну вредност за корисника. Процесни модел садржи низ процесних корака, који конституишу процес откривања знања из података, од дефинисања проблема до коришћења откривеног знања, а које практичари треба да следи при реализацији задатка откривања знања. Кроз детаљан опис процедура у сваком конститутивном кораку, процесни модел треба да омогући актерима да разумеју релевантне аспекте процеса и, истовремено, обезбеди смернице за планирање и реализацију пројектних активности по етапама. Наиме, процесни модели за откривање знања обезбеђују општи оквир и смернице за реализацију *DM* пројеката. Консеквентно, остварује се значајно смањење трошкова, скраћује време реализације пројекта и повећава учешће успешно реализованих пројеката.

Узимајући у обзир чињеницу да је добро структурирање процеса од суштинског значаја за екстракцију корисног знања и успешну примену *DM* метода, многи *DM* истраживачи и практичари су предложили бројне моделе, као стандардизоване процедуре за откривање знања. Они се крећу од врло једноставних, који садрже неколико процесних корака, до врло софистицираних и детаљних модела.

Иницијални напори на пољу развијања процесних модела потекли су из академских кругова. О концепту процесног модела у контексту структурирања процеса откривања знања из података као стандардизованог процеса, први пут је дискутовано током инаугуративне радионице, одржане 1989. године[6] у *Detroit*-у, када је и први пут употребљена фраза „Откривање знања из база података". Том приликом је истакнуто да је знање крајњи резултат подацима вођеног процеса откривања знања и уједно апострофирана улога *DM* метода у том процесу. Базичну структуру процесног модела, у форми *KDD* модела, развили су *Fayyad et al.* (1996, стр. 39). Убрзо након иницијалних академских напора уследили су и индустријски одговори. Наиме, на основу заједничких напора, искустава у практичној примени и знања стручњака из више светских компанија[7], уз подршку одговарајућих Европских комисија, предложен је и развијен општеприхваћени оквир и *de facto* стандард за откривање знања и реализацију *DM* пројеката под називом *CRISP-DM* модел (енгл. *CRoss Industry Standard Process for Data Mining*) (*Chapman et al.,* 2000).

[6] Након прве, одржане су још три *KDD* радионице, а од 1995. године редовно (сваке године) се одржавају међународне *KDD* конференције.

[7] Језгро *CRISP-DM* конзорцијума чиниле су следеће четири компаније: *NCR Systems Engineering Copenhagen* (провајдер база података), *ISL-Integral Solutions Limited,* сада део *SPSS Inc*., (провајдер комерцијалних *data mining* решења), *DaimlerChrysler*, тада *Daimler-Benz* (аутомобилска, авио, телеомуникациона и консултантска компанија) и *OHRA* (независна холандска осигуравајућа компанија). Последње две компаније послужиле су као извори података и студије случајева.

На темељу ова два модела предложени су бројни процесни модели, који се грубо могу класификовати на академске, индустријске и комбиноване моделе (*Cios et al.*, 2007, стр. 11-16). Сходно значају оригиналног *KDD* и *CRISP-DM* модела, *Mariscal et al.* (2010, стр. 141-142) разликују три групе приступа у развијању процесних *KDD* и *DM* модела: приступи засновани на оригиналном *KDD* моделу, приступи засновани на *CRISP-DM* моделу и остали, независни приступи.

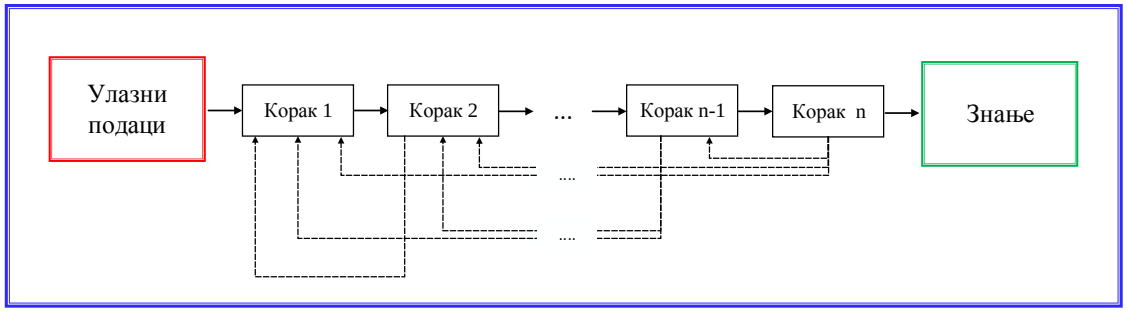
Независно од категорије, према истраживањима која су спровели *Kurgan & Musilek* (2006), свеобухватним поређењем водећих процесних модела за откривање знања идентификоване су извесне заједничке карактеристике, које се током времена не мењају и представљају инхерентно својство сваког процесног модела. Те карактеристике су:

- Сви процесни модели се састоје од низа корака, који се серијски извршавају са укљученим повратним спрегама и итерацијама. Сваки наредни корак у низу се покреће након звршетка претходног корака, тако да генерисани резултати претходног корака представљају улаз за наредни корак.
- Сви процесни модели обухватају серију активности које покривају читав животни циклус пројекта откривања знања. Заправо, дијапазон ових активности полази од задатка разумевања апликативног подручја и циљева пројекта, преко разумевања, припреме и анализе података, до евалуације, разумевања и примене, односно коришћења добијених резутата.
- Сви процесни модели су по својој природи итеративни. Високо итеративни карактер сваког корака (фазе) у серији, парцијално, као и целине процесног модела произлази из постојања вишеструких повратних спрега и понављања која се активирају поступком ревизије.

Општа структура, основни ток корака, заједничке карактеристике и улазни и излазни елементи процесних модела откривања знања су представљени на Слици 4. Узлазни елементи се односе на различите форме података, а излазни представљају генерисано ново знање у форми модела, релација, трендова и слично. Истовремено, визуелно је потврђена и снажна итеративна природа процеса са повратним спрегама између било која два корака, а која се суштински базира на бројним одлукама и континуираном експериментисању уз промене параметара у сваком кораку процеса.

Kurgan & Musilek (2006, стр. 5), такође, истичу да се главне разлике између бројних процесних модела односе на број предложених корака и обим активности у оквиру њих. У том контексту, важан аспект сваког процесног модела је време потребно

за реализацију процесних корака. Процена потребног времена треба да омогући прецизно планирање појединих активности и дефинисање битних рокова за њихову реализацију. Конкретне процене зависе од многих фактора, као што су комплексност проблема који се разматра, постојеће знање о проблему, расположивост људских ресурса, експертско знање, вештине и способности. Анализа процена многих *DM* истраживача и практичара указује на следеће закључке (уз напомену да постоје и одступања од ових процена): прво, више од половине напора, мереног утрошеним временом за реализацију процесних корака, односи се на припрему података и, друго, непосредна реализација *DM* корака захтева релативно мало напора, односно утрошеног времена наспрам његовог значаја у поређењу са другим процесним корацима (*Kurgan & Musilek*, 2006, стр. 17). Кључни разлог за овакав распоред потребног времена везује се за проблеме варијација квалитета података, које у кораку припрема података треба решити, а затим, на већ припремљене податке, применити одговарајуће *DM* методе.



Слика 4: Структура и ток процесних модела откривања знања

Извор: *Cios et al.* (2007, стр. 11)

Анализа тенденција у примени процесних модела, а према резултатима анкетних истраживања која су спроведена и презентована на страницама *web* сајта *KDnuggets.com*[8], *CRISP-DM* модел је најчешће коришћен модел за реализацију *DM* пројеката.[9] Такође, један од индикатора степена прихватања модела су и подаци о

[8] *KDnuggets* је водећи *web* извор у подручју пословне аналитике, *big data*, *data mining*-а и науке о подацима (енгл. *data science*). Као *on-line* платформа (чији је покретач *Gregory Piatetsky Shapiro*), омогућава не само повезивање истраживача, већ обезбеђује и покрива све релевантне аспекте повезане са *DM* анализом: вести, софтвере, курсеве, занимања, едукацију и *web* семинаре у овој области.

[9] Резултати анкета (доступни на: *www.kdnuggets.com/polls/*), спроведених 2002, 2004, 2007. и 2014. године, су показали да *CRISP-DM* модел користи највећи број испитаника. Установљено је, такође, значајно учешће испитаника који користе сопствене моделе креиране за спровођење процеса откривања знања, док је на трећем месту *SEMMA* процесни модел. Акроним *SEMMA* (енгл. *Sample, Explore, Modify, Model, Assess*) се односи на петоетапни модел који обухвата: узорковање (избор и подела података на део на којем ће бити изграђен модел, део за оптимизацију параметара и део за тестирање модела); истраживање (дескриптивна статистика и визуелизација), модификовање (припрема података и варијабли), моделирање (креирање модела уз одговарајућу софтверску подршку) и процену (компарација модела коришћењем одговарајућих критеријума). Овај модел је развијен у *SAS* институту (једној од водећој компанији у области пословне интелигенције, аналитичких софтвера и услуга) и уграђен у комерцијалну софтверску платформу *SAS Enterprise Miner*. Детаљан приказ сличности и разлика, кроз компаративну анализу *KDD*, *CRISP-DM* и *SEMMA* модела видети у: *Azevedo & Santos* (2008).

цитираности научних и стручних публикација и радова у којима су модели презентовани. Резултати претраге неколико, са аспекта броја индексираних радова, водећих индексних база, указују да је најчешће цитиран иницијални *KDD* модел (*Kurgan & Musilek*, 2006, стр. 14). Оправдано се може констатовати да је велика цитираност последица чињенице да готово сви радови који тангирају проблематику процесних модела садрже кратак осврт и дефиниције из првих радова *Fayyad*-а и његових сарадника у којима је овај модел представљен.

Имајући у виду информације о прихватању процесних модела (то јест, о примени у индустријским и академским пројектима и цитираности у научној, професионалној литератури), као и чињеницу да су послужили као основа за креирање многих модела, у наставку је пажња усредсређена на оригинални *KDD* и *CRISP-DM* модел.

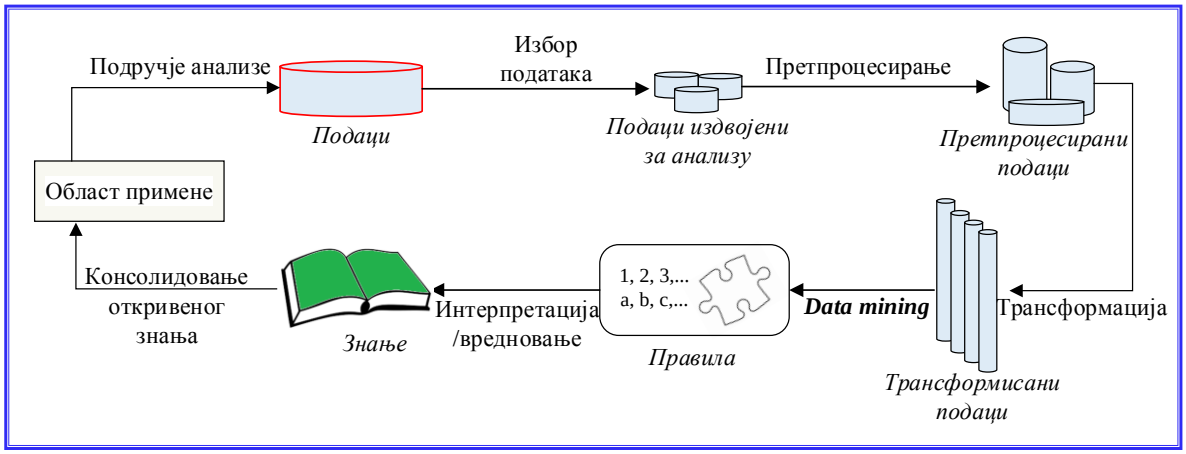
3.2. *CRISP-DM* модел

Процес откривања знања из података је комплексан, вишетапни, интерактиван и итеративан процес са серијом одлука које треба донети у сваком кораку, укључујући значајне итерације и повратне спреге између било која два корака. Изворно, *KDD* процес се дефинише као нетривијални процес идентификовања валидних, нових, потенцијално корисних и разумљивих законитости у подацима (*Fayyad et al.,* 1996, стр. 40-41). Реч је о процесу који се односи на целокупан процес откривања знања из података, а централни корак тог процеса је *DM* и односи се на примену специфичних алгоритама за екстракцију корисних резултата у форми законитости из података. Међутим, у пракси, *DM* и процес откривања знања из података се често користе као синоними, иако је *DM* само један од корака у *KDD* процесу.

Иницијални процесни модел откривања знања, који су предложили *Fayyad et al.* (1996, стр. 42) се састоји од следећих корака (представљених на Слици 5):

- први, односи се на развијање и разумевање подручја примене, стицање релевантног, већ постојећег, знања и идентификовање, из перспективе корисника, циља процеса откривања знања;
- други, односи се на креирање циљног скупа података и укључује избор подскупа променљивих и података (узорка), који ће се користити за спровођење задатака откривања знања;
- трећи, односи се на чишћење и претпроцесирање података, и обувата откривање и уклањање нестандардних опсервација и шумова у подацима, решавања питања недостајућих вредности у подацима итд;

- четврти, односи се на редукцију података и пројекцију, и обухвата, у зависности од циља конкретног *DM* задатка, проналажење промењивих које ће верно презентовати податке, уз примену метода за редукцију димензионалности и трансформацију података;
- пети, односи се на избор *DM* метода, сходно дефинисаним циљевима процеса откривања знања у првој етапи;
- шести, односи се на избор *DM* алгоритма и укључује експлоративну анализу и доношење одлуке о адекватним моделима и параметрима;
- седми, односи се на процесирање, односно *DM*, а подразумева генерисање законитости у одговарајућој форми (на пример, класификационо правило одређено применом метода стабло одлучивања, регресиони модел итд.);
- осми, односи се на интерпретацију и оцену сигнификантности, интересантности, необичности и сличности откривених законитости, укључујући и визуелизацију података и екстрахованих правила или модела, као и одлуку о враћању на било коју претходну етапу;
- девети, односи се на деловање (односно, предузимање конкретних акција) на основу откривеног знања и његово консолидовање, што подразумева директно коришћење знања, инкорпорирање знања у постојећи систем знања у циљу будућих деловања (укључујући проверу и решавање потенцијалних конфликата, неслагања новооткривеног са претходно екстрахованим знањем), документовање и презентовање откривеног знања заинтересованим странама.



Слика 5: Процес откривања знања из база података (*KDD* процес)

Извор: *Milanović & Stamenković* (2011a, стр. 7)

CRISP-DM процесни модел за откривање знања је дефинисан као хијерархијска структура, која се састоји од скупова задатака представљених на четири нивоа

апстракције (од главних фаза, преко генеричких и специјализованих задатака, до нивоа крајњих акција, одлука и резултата сваке фазе).[10] У наставку текста следи, најпре, кратак осврт на сваку фазу појединачно, а затим и дискусија о целини процеса, који су у великој мери засновани на изворном приказу представљеном од стране *Chapman et al*. (2000). У хијерархијској структури, први ниво модела састоји се од следећих шест фаза, које су представљене на Слици 6:

- Прва фаза, којом започиње *CRISP-DM* процес, је разумевање проблема који треба решити. Суштински, у овој фази спроводи се конверзија пословних проблема и циљева у *DM* проблеме и циљеве, укључујући, уз нужну процену свих релевантних фактора који могу утицати на процес и коначни резултат откривања знања, дизајнирање прелиминарног пројектног плана за реализацију дефинисаних циљева. Стога, ову иницијалну фазу сачињавају следећи генерички задаци: ▸ дефинисање пословних циљева, ▸ процена ситуације, ▸ дефинисање *DM* циљева, и ▸ генерисање пројектног плана.

- Друга фаза је разумевање података, односно иницијално прикупљање и стицање увида у природу података, као примарног ресурса целог процеса откривања знања. Уз обавезно састављање извештаја о оствареним резултатима на крају сваког обављеног задатка, ова фаза се разлаже на следеће генеричке задатке: ▸ иницијално прикупљање и организација података, ▸ дескрипција података, ▸ истраживање података, и ▸ оцена квалитета података.

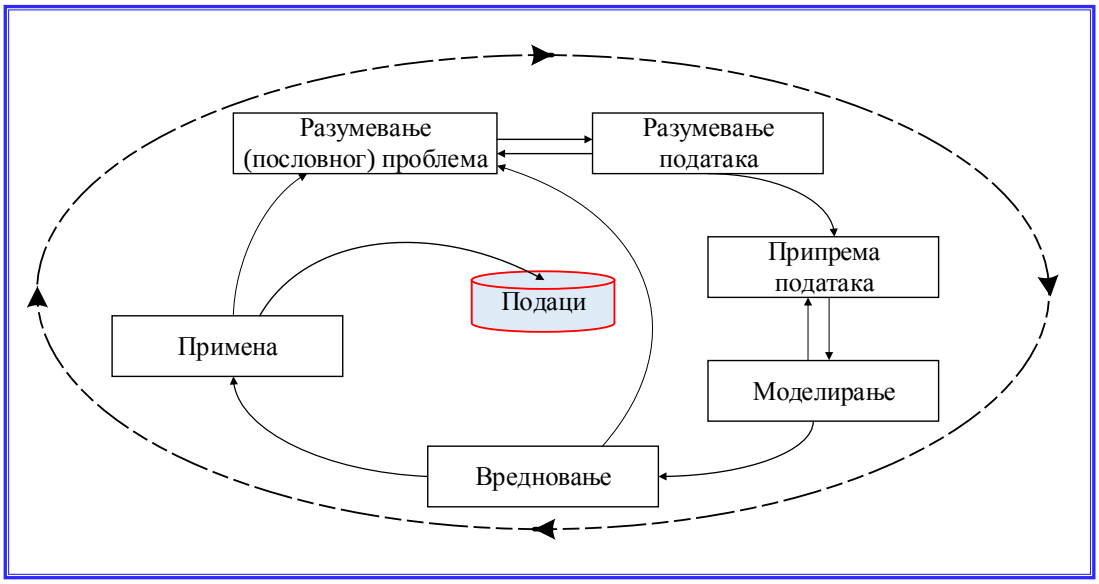
- Трећа фаза је припрема података, и обухвата све активности преуређења података неопходне за креирање коначног скупа података који ће бити коришћен за моделирање, а у складу са захтевима процеса моделирања и примене изабраног *DM* алгоритма у наредној фази. Ова фаза се може рашчланити на следеће генеричке задатке: ▸ избор података, ▸ чишћење података ▸ креирање нових варијабли (и података), ▸ интеграцију података, и ▸ форматирање података.

- Четврта фаза је моделирање, и подразумева избор различитих техника моделирања и њихову примену на улазни претпроцесирани скуп података. Ову фазу сачињавају следећи генерички задаци: ▸ избор технике (или техника) за моделирање, ▸ састављање плана за тестирање генерисаних модела, ▸ конструкција модела, и ▸ оцена модела.

[10] На пример, први ниво (фаза): припрема података; други ниво (генерички задатак): чишћење података; трећи ниво (специализовани задатак): решавање питања недостајућих података; четврти ниво (резултат процеса): аритметичка средина нумеричке варијабле.

- Пета фаза је евалуација, и односи се на оцењивање модела из перспективе разматраног пословног проблема и дефинисаног циља, за разлику од претходног корака у којем се оцењивање квалитета модела спроводи из перспективе анализе података (поузданост, тачност, логичка исправност итд). Генерички задаци у овој фази су: ▸ оцењивање резултата, ▸ провера и накнадно преиспитивање процеса, и ▸ утврђивање наредних активности.

- Шеста фаза је примена резултата, и односи се на организовање и представљање знања у форми која је разумљива кориснику и омогућава његово (сврсисходно) коришћење. Ова фаза може да буде врло једноставна и да се састоји само од састављања извештаје. С друге стране, може бити и врло комплексна, и сходно захтевима, подразумевати спровођење поновљеног процеса откривања знања. Генерички задаци у оквиру ове фазе су: ▸ састављање плана примене модела, ▸ састављање плана за праћење и одржавање модела, ▸ састављање финалног извештаја, и ▸ општи преглед пројекта са предлозима за побољшање.



Слика 6: *CRISP-DM* процесни модел

Извор: *Chapman et. al.* (2000, стр. 10); *Nisbet et. al.* (2009, стр. 35)

Секвенцијално кретање од једне ка наредној фази није стриктно одређено. Често је, у зависности од добијених резултата, потребно враћање на претходне фазе. Спољни круг на Слици 6 симболизује циклични природу самог процеса екстракције знања из података, а стрелицама је означен ток процеса и међузависност фаза. *CRISP-DM* модел има три кључна *feedback* механизма: од разумевања података до разумевања пословних проблема, од моделирања до припреме података и, од евалуације до разумевања

пословног проблема. Овај модел је апликативно неутралан у односу на алат и област у којој се примењује.

Као одговор на стално растуће потребе и промене захтева при анализи података упоредо са променама у окружењу долази до унапређења постојећих и предлога нових процесних модела. Сходно томе, очекује се и побољшана верзија *CRISP-DM* модела која ће обухватити не само промене, попут додавања нових и брисање постојећих фаза, укључивања додатних *feedback* механизма, већ и активности које се односе на проблеме рада са новим типовима података, организовање мултидисциплинарних тимова, управљање пројектима, управљање променама и управљање квалитетом. На томе ради *DM* конзорцијум, у који су, осим продаваца, провајдера и даваоца података, укључени и истраживачи и крајњи корисници (*Mariscal et al.*, 2010, стр. 150). У том смислу, веома је важно истаћи да не постоји универзални процесни модел, подједнако добар у свим апликативним подручјима и сагласан са свим дефинисаним циљевима, тако да у конкретној проблемској ситуацији сваки модел има своје позитивне и негативне стране и примењују се са различитим степеном успеха.

3.3. Улоге експерата у *data mining* процесу

За реализацију процеса откривања знања из података потребна су одговарајућа експертска знања. Сходно томе, валидна примена изабраних метода (и алата) и квалитет добијених резултата првенствено зависи од стручности, различитих вештина и знања особа које су активни учесници у реализацији овог процеса.

Заправо, фазни, *feedback* циклус откривања законитости из података једна особа не може самостално спровести. Успешна решења захтевају тим експерата са релевантним знањима и кореспондентним професионалним улогама у спровођењу *DM* процеса. Као чланови тима обавезно морају бити укључени експерти из домена апликативног подручја, информационих система и анализе података. Њихове улоге распоређене су на следећи начин (*Vercellis,* 2009, стр. 89-90; *Чубукова,* 2006):

- Експерти из домена апликативног подручја (енгл. *domain experts*) су стручњаци који поседују одговарајуће знање о конкретном предметном подручју и анализираном проблему који припада том подручју. Наиме, стручно знање ових експерата односи се на законитости карактеристичне за разматрани феномен, (пословно) окружење, процесе, клијенте, конкуренте, као и хипотезе о могућим везама између појава, процеса и догађаја, укључујући и процедуре за решавање типичних проблема. Од њих се, као добрих познавалаца проблема, очекује да иницирају процесе

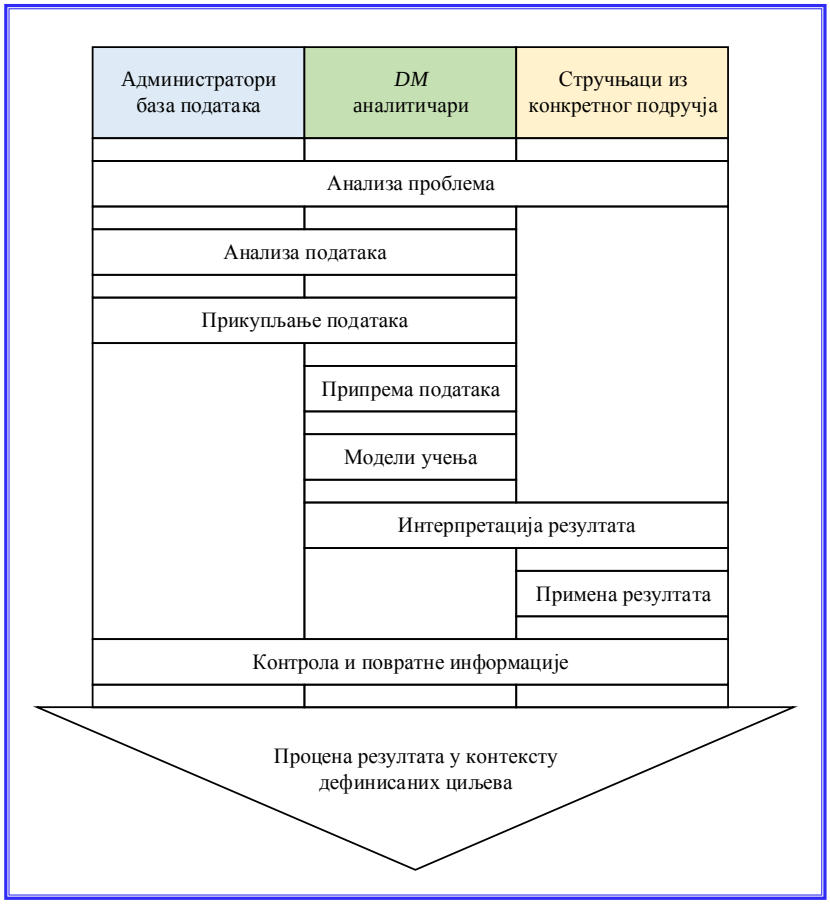
за откривање знања из података, дефинишу иницијалне циљеве анализе, учествују у интерпретацији и потенцијалном коришћењу резултата, предлажу активности на темељу добијених *DM* резултата, као и да током одвијања процеса, на захтев осталих чланова тима, обезбеде додатна тумачења проблема из угла података и примењених метода. Обзиром да иницирају *DM* процесе и користе добијене резултате, особе које обављају ове задатке (из *DM* перспективе) називају се и *DM* корисници. При том, углавном не поседују (техничке) вештине потребне за активно учешће у спровођењу методолошки захтевних фаза процеса, попут припреме података и моделирања.

- Експерти за информационе системе и технологије (*IT* аналитичари) су стручњаци који, пре свега, добро познају базе и складишта података. Ови експерти поседују одговарајуће знање о томе где и како чувати података, како приступити подацима и како податке повезати међу собом и интегрисати *DM* алате, процесе, моделе и резултате. Одговорни су за пројектовање и развој база података, њихово ефикасно коришћење и одржавање, прикупљање, модификовање и заштиту података. Осим ангажовања око обезбеђења приступа подацима, ови експерти располажу са техничким знањима релевантним са становишта одговарајућих софтверских решења и покретања *DM* алгоритама, тако да су, сходно селектованом методолошком решењу проблема, активни учесници нарочито у припреми података.

- Експерти за *DM*, односно, аналитичари за истраживање података (енгл. *data miners*), као стручњаци за анализу података су у могућности да на основу поседованог (примарно статистичког) знања спроведу експлоративну анализу, изаберу и примене *DM* методе, креирају (дескриптивне и предиктивне) моделе и (методолошки, пре свега) интерпретирају добијене резултате. Кроз непосредну сарадњу са *IT* експертима, *DM* аналитичари за своје акције обезбеђују приступ различитим изворима података. Такође, с друге стране, кроз комуникацију са експертима из апликативног подручја, долазе до важних информација о конкретном подручју, што им у великој мери омогућава да у потпуности схвате пословне циљеве корисника и, последично, знатно олакшава превођење тих циљева у *DM* проблеме, постепени развој *DM* модела и праћење добијених резултата.

Очигледно, без наведених експертских улога *DM* процес не може бити реализован. Посматрано у контексту учешћа у различитим активностима, поставља се питање интеракције ових актера у остваривању пословних и *DM* циљева. У начелу, неопходно је да сваки учесник детаљно познаје оне сегменте и аспекте процеса за које је директно одговоран. Истовремено, пожељно је да, примерено својој улози у процесу

откривања знања, сваки учесник поседују одређени ниво знања о целини процеса из перспективе проблема, методологије и података. Међутим, изузетно је тешко стриктно распоредити улоге учесника у процесу према *DM* активностима, јер се подручја њиховог деловања често преплићу. На Слици 7 приказане су експертске улоге, оквирно распоређене, сходно подразумеваним компентенцијама, према фазама *DM* процеса.



Слика 7: Актери у *KDD / DM* процесу и њихове улоге

Извор: *Vercellis* (2009, стр. 91)

На бази ове илустрације јасно се може уочити да је нужна међусобна сарадња и размена мишљења експерата током целог процеса реализације *DM* активности. Такође, јасно су видљиве кључне тачке пресека у којима се „укрштају" и интерактивно укључују поједини експерти у постизању пословних циљева. Заправо:

- интеракција *IT* експерта и експерта за *DM* остварује се у фази прелиминарне анализе података, укључујући и прикупљање података, при чему, као што је већ напоменуто, припрему података може спроводити *DM* аналитичар самостално или у интеракцији са *IT* експертом;
- интеракција експерта за *DM* и експерата из апликативног подручја остварује се у фази анализе проблема (у којој активно учествује и *IT* експерт) приликом

дефинисања циљева и процене конкретне ситуације, као и у фази (методолошке и проблемске) интерпретације резултата моделирања;

- интеракција *IT* експерта, експерта за *DM* и експерта из домена апликативног подручја остварује се, осим у првој фази анализе разматраног проблема, приликом контроле и накнадног преиспитивања целокупног процеса и идентификовања потреба за додатним итеративним понављањем одређених захвата и фаза процеса, а у правцу побољшања саме анализе. На крају, интеракција ових експерата остварује се приликом финалне процене новог знања, односно резултата и њихове усаглашености са дефинисаним циљевима.

Свака од ових улога, у зависности од кадровског потенцијала конкретне организације, може бити делегирана интерном или екстерном експерту. С обзиром да су пословни и *IT* експерти присутни у скоро свакој организацији, за прве две улоге најчешће се ангажују сопствени кадрови, док се за активности које се односе на трећу улогу углавном ангажује екстерни консултант. Након стицања релевантног знања из домена *DM*-а, активности треће улоге могу се поверити сопственим експертима или пак ангажовати и запослити нове особе, отварањем радних места за *DM* послове.

DM процес се реализује као пројектни задатак. Стога из менаџмент перспективе, не сме бити запостављена улога руководиоца пројекта, који сноси укупну одговорност за планирање, организовање, имплементацију и примену резултата реализованог пројекта. Обим пројектног задатка одређује број чланова пројектног тима за реализацију наведених улога, а област на коју се односи *DM* задатак често захтева укључивање у тим и других експерата са уско специјализованим знањима. Степен комуникације и координације активности и експерата из различитих области, утиче на време реализације пројекта и квалитет добијених резултата. Управо, домен одговорности руководиоца пројекта је повезивање и комбиновање сегмената из различитих области у један процес, као и усмеравање, координација и контрола експертских активности. При томе, руководиоцима стоји на располагању гантограм, као врло моћан алат за управљање пројектима који илуструје распоред фаза и активности пројекта и омогућава да се сагледа (и током реализације пројекта прати) потребно време и ресурси за финализацију пројекта.

4. РАЗЛИЧИТИ АСПЕКТИ ПРИМЕНЕ *DATA MINING*-а

DM је научно-истрживачка област са растућим интересовањем и значајем, али и област чијом се применом, кроз процес експлоатације потенцијала великих количина

расположивих података, добијају резултати који пословним системима могу обезбедити значајну конкурентску предност. Међутим, истраживање и примена *DM*-а се не ограничава само на пословно окружење. *DM* се може применити у свим оним подручјима где постоје велике количине података и у њима скривене потенцијално значајне законитости и релације. У том смислу, у овом Поглављу су сагледани следећи аспекти директно тангентни са реализацијом пројектних *DM* задатака: подручја, критични фактори, као и позитивне и негативне стране примене *DM* концепта.

4.1. Подручја примене *data mining*-а

DM технике и методи успешно се примењују у многим подручјима, од пословања и науке до спорта и забаве, тако да се разликују пословне, научне и остале *DM* примене. У оквиру сваког подручја постоји врло широка листа општих и специфичних задатака који представљају велики изазов са аспекта *DM*-а. У наставку се даје краћи преглед одабраних подручја и проблема који се могу решавати коришћењем *DM* методологије.

Маркетинг и продаја су прве пословне функције у оквиру којих су иницијално покренути развој и примена *DM*-а.[11] Са становишта ових функционалних подручја примена *DM*-а је значајна за многе проблеме и намене, попут: спровођење директног маркетинга, детерминисање маркетинг стратегија у односу на конкуренте, сегментацију тржишта, анализу и предвиђање понашања клијената, спречавање и откривање превара, побољшање унакрсне продаје, анализу и избор оптималних канала продаје итд. У савременој, дигиталној економији маркетинг је у све већој мери фокусиран на појединачног корисника кроз стратегије управљања односима са купцима (енгл. *Customer Relationship Management - CRM*). Комбинација *CRM* концепта са одговарајућим софтверским решењима и *DM* методима омогућава организацији да интегрише и анализира податке о својим клијентима, открије знања која су скривена у подацима и, сходно томе, да „скроји" понуду „по мери", односно у складу са посебним захтевима својих клијената. Стога, типичне примене *DM*-а у области *CRM* су изградња и анализа профила клијената и креирање специфичних понуда сходно појединачним захтевима клијената.

[11] Случај „пиво и пелене" је постао опште позната илустрација која говори о значају пословне аналитике и *DM*-а у проналажењу и издвајању оних законитости из пословних података које се на први поглед не могу идентификовати. Један трговачки ланац је на основу података са рачуна, користећи одговарајући софтвер за анализу података, дошао до сазнања да су мушкарци, обављајући куповину четвртком, најчешће уз пелене куповали и пиво. Захваљујући овом открићу, постављањем витрине са пивом у близини полице са пеленама, трговачки ланац је повећао продају. Наравно, реч је о једноставном примеру који илуструје сржа *DM*-а, али свакако распон његових примена и значај су знатно већи.

Идентификовање профила клијената је активно подручје примене *DM* не само у пословном свету и кориснички оријентисаним делатностима, већ у свим областима које укључују односе са клијентима. Аналитичари проучавају понашања клијената и идентификују, на пример, профил типичног клијента, профил незадовољног клијента, профил кључног клијента, профил профитабилног клијента, профил ризичног клијента, али и профил бирача на политичким изборима, радника, потенцијалних терориста, пореских преступника, студената, пацијената и слично.

На подручју телекомуникација, посебно последњих година услед жестоке конкуренције, интензивно се користе предности *DM* приступа у анализи података. Компаније које послују у овој области *DM* методе користе за идентификовање профила лојалног / профитабилног клијента, откривање превара у коришћењу телефонских картица, анализу постојећих и осмишљавању нових производа и услуга, анализу узрока одлазака клијената код конкурената, процену ризика који се односи на инвестиције у нове технологије (на пример, нано-технологије, оптичка влакна), идентификовање разлика у производима и услугама између конкурената итд.

Банкарски и финансијски сектор је подручја интензивне примене *DM*-а. У банкарству се *DM* методи успешно користе за оцену ризика финансијске институције (на пример, анализа кредитне способности клијената и предвиђање нивоа лоших пласмана), тренд анализу, анализу профитабилности, подршку у маркетиншкој кампањи итд. На подручју финансијских тржишта *DM* примена се односи на предвиђање цена акција и других хартија од вредности, предвиђање кретања девизног курса, оптимизацију трговања хартијама од вредности, предвиђање финанасијских криза, откривање илегалних финансијских трансакција итд.

У савременом пословању је постало популарно користити термине који имају префикс „*e*" и који се односе на различите видове пословања базираним на могућностима савремених технологија и *IT* апликација. Пословање које се обавља преко рачунарских мрежа, односно електронско пословање је изразито погодно и перспективно за примену *DM*-а. За организације које послују на Интернету, *DM* омогућава да се идентификује и анализира профил *веб* посетилаца, процени сличност садржаја прегледаних *веб* страница, анализира понашање купаца у *on-line* куповини, управља маркетиншким акцијама у *веб* окружењу, идентификују конкуренти, изврши избор најбољих пословних партнера и дефинише одговарајућа политика цена. Међутим, *DM* се не користи само у делу електронског пословања који се односи на куповину и продају роба и услуга, већ и у обављању других активности преко

Интернета: учење на даљину (енгл. *e-learning*), електронско банкарство (енгл. *e-banking*), електронска управа (енгл. *e-government*), научна сарадња која се остварује путем *Web*-а (енгл. *e-science*)[12], електронско библиотекарство (енгл. *digital libraries or bibliomining*) итд.

Спортске активности су изузетно погодне за примену *DM* метода, јер огромне количине података се прикупљају за сваког спортисту, екипу, догађај и спортску сезону. Употреба резултата ових метода у свакодневној пракси стручних штабова, спортиста и спортских тимова омогућава: побољшање процеса ангажовања играча, откривање талентованих појединаца и избор оптималног тима, праћење развоја спортиста, унапређење тренинга и спортских резултата, предвиђање перформанси појединаца и тима итд.[13]

Законодавство је једно од најкритичнијих подручја примене *DM*-а. У напорима привредних и непривредних организација, државних органа, обавештајних агенција и влада на спровођењу закона, *DM* методи се успешно користе за откривање необичних и сумњивих активности и трансакција и неовлашћених упада.[14] Необична примена *DM*-а, скоријег датума, односи се на процену правног ризика (*Tufféry*, 2011, стр.10).[15]

Као последица уоченог потенцијала који произлази из интеракције *DM*-а и биоинформатике, последњих година посебно активна област истраживања је примена и развој *DM*-а у решавању биолошких проблема, попут, анализе и истраживања еволуције и структуре генома, дешифровања генома, класификације гена, анализе протеина и протеинске структуре, дијагнозе и прогнозирање болести и слично.

У медицинској науци постоји велики простор за примену *DM*-а, од анализе профила пацијената, постављања компјутерски подржаних дијагноза, груписања

---

[12] Овај тип сарадње илуструје следећи пример: полазећи од тога да је статистика граматика науке, професор Ловрић је успео да реализује пројекат међународног карактера тако што је је електронским путем повезао 620 аутора из 105 земаља са 6 континената и који су узели активно учешће у припреми и писању радова обједињених у форми научног дела „Међународна енциклопедија статистичких наука".

[13] Пионирске кораке у имплементацији *DM*-а у области спорта направио је амерички национални кошаркашки савез (енгл. *National Basketball Association - NBA* лига). Већина тимова у *NBA* лиги користи софтверски пакет: *IBM Advanced Scout Data Mining software*, развијен у *IBM*-у као резултат тимског рада под руководством *Inderpal Bhandari*, ватреног навијача тима *New York Knicks* (*Larose*, 2005, стр. 3). Интересантно је навести, такође, да се сваке године најбољем бацачу у првој америчкој бејзбол лиги додељује „*Cy Young*" награда. Одлука о избору појединца коме ће награда бити уручена се у највећој мери базира на статистичким подацима прикупљеним током бејзбол сезоне, а добитник ове награде се предвиђа применом *Bayes*-овог класификатора.

[14] Често се наводи да је путем *DM*-а метода америчка војно-обавештајна служба успела да идентификује девет од једанаест терориста који су учествовали у нападу на *Twin Towers* 11. септембра 2001. године, и то чак годину дана пре него што се напад догодио.

[15] У Уједињеном Краљевству постоји систем (*Offenders Assessment System - OASys*) који има за циљ да процени ризик настанка поновног прекршаја у случају „условног пуштања на слободу" особе која је начинила прекршај. Процена ризика се спроводи на основу информација које се односе на: породично стање, место боравка, ниво образовања, изјаве сарадника и пријатеља, криминалног досијеа, извештаја социјалних радника и понашања особе током периода истраге.

пацијената према степену ризика за настанак одређених болести, дефинисања превентивних напора и сугерисања терапеутских алтернатива, до предвиђања трошкова здравствене заштите и решавања проблема складиштења података и дефинисања процедура у здравственим установама. Такође, фармација је позната по спровођењу квантитативних анализа, како за потребе проналажења нових медикамената, тестирања њиховог дејства и клиничких испитивања, тако и за потребе истраживање тржишта и доношења пословних одлука у циљу пласирања медикамената директним и индиректним корисницима, односно пацијентима и лекарима. Стога, она представља добар пример преплитања научних и пословних аспеката *DM* примене.

У области астрономије *DM* се може успешно користити за идентификовање небеских тела и одређивање њихове приближне старости. Такође, путем *DM* техника, на бази хиљада фотографских плоча направљених телескопима, анализом великог броја небеских објеката и разматрањем великог броја њихових карактеристика, могуће је сваки објекат класификовати као одређени тип звезде или галаксије.

Из презентованог прегледа непосредно произлази да се, независно од подручја примене, путем правилног коришћења *DM* методологије могу открити релевантне законитости. Управо снага примене *DM* и лежи у чињеници да се акценат ставља на податке, а не на подручје анализе (*Panian & Klepac*, 2003, стр. 250). Заправо, интерпретацијама добијених резулата се баве експерти из конкретног подручја, који уз помоћ аналитичара, тумаче и објашњавају откривене законитости, тако да специфичности подрчја примена долазе до изражаја у фази анализе и интерпретације добијених резултата. Стога се може очекивати даље повећање интересовања за примену *DM*-а у реалним проблемским ситуацијама, као и интензивирање академских истраживања и разматрања различитих аспеката ове проблематике на универзитетима и другим научно-истраживачким институцијама.

4.2. Критични фактори за реализацију *data mining* пројеката

У пословно оријентисаном приступу *DM*-у, један од захтева је да резултирајуће знање до којег се долази анализом података организације буде профитабилно, то јест, да допринесе остваривању једног или више пословних циљева, попут повећања обима продаје, прихода и профита, смањења трошкова, оптималног коришћење постојећих ресурса, смањења времена за повраћај улагања, повећања тржишног учешћа, побољшања задовољства клијената и унапређења квалитета. Другим речима, *DM* „напор” мора имати финансијско оправдање (*Kantardžić*, 2011, стр. 18). Резултати

анализе трошкова и користи јасно и недвосмислено треба да укажу на економску оправданост (или неоправданост) развоја *DM* апликација за решавање неког конкретног пословног проблема или стварања нових пословних могућности са перспективом интегрисања *DM* иницијатива у све организационе процесе.

Заправо, данас, у веома ризичном и неизвесном пословном окружењу, предузећа се све више фокусирају на стицање конкурентске предности кроз ефикасну употребу података. Сходно томе, *DM* је идентификован и препознат као значајан концепт и технологија за унапређење процеса доношења одлука анализом и претварањем великих количина података у вредне и корисне законитости, а самим тим и као средство за остварење пословних циљева. Међутим, упркос чињеници да су многа предузећа препознала потенцијал *DM*-а, јаз између способности анализе и способности складиштења податка доводи до тога да многи *DM* подухвати постану неуспешни, а подаци недоступни и недовољно искоришћени. *Bole et al*. (2014, стр. 253) указују на то да резултати многих истраживања, како у академској заједници, тако и у пракси, потврђују све веће интересовање за *DM* технологију, али се, истовремено, скреће пажња на несклад (заостајање) између тражње за *DM* технологијом и њене стварне имплементације. При томе се апострофира потреба за применом једног свеобухватног приступа који ће омогућити превазилажење ових неусаглашености.

Као и случају сваке друге *IT* иновације или компоненте технологије пословне интелигенције, имплементација *DM*-а је комплексан подухват који захтева одговарајућу инфраструктуру и ангажовање значајних ресурса. Услед наведеног, да би се повећала вероватноћа успешне реализације *DM* иницијатива у форми пројектног задатка, организације морају разумети и посебну пажњу посветити критичним факторима успеха (енгл. *critical success factors - CSF*). У супротном, имплементација *DM*-а је веома ризична.

Критични фактори успеха, према *Rockart*-у, представљају „ограничени број подручја у којима ће резултати, ако су задовољавајући, обезбедити успешне конкурентске перформансе организације" (*Rockart,* 1979; цитирано у: *Sim*, 2014, стр. 67). Једноставније, термин критични фактори успеха означава скуп фактора који утичу на реализацију одређеног пројекта и које треба узети у разматрање како би се извршила оптимална алокација ограничених ресурса са фокусом на она подручја, активности и задатке који ће, сходно проценама, имати највећи утицај на успешну реализацију циљева. Наиме, разумевање критичних фактора успеха (људских фактори и фактора организационе и технолошке природе) представља кључ за успешну

имплементацију сваког пројекта у области пословне интелигенције (*Yeoh & Koronios,* 2010, стр. 31), а сами тим и сваког *DM* пројекта.

Упркос њиховом очигледном значају, постоји веома мали број истраживања посвећених проблему фактора који су критични са становишта успешне *DM* имплементације. Према *Sim*-у (2014, стр. 66), главни разлози за то су тешкоће повезане са квантификовањем оних предности *DM* пројекта које су последица истраживања фактора успеха, као и чињеница да не постоји спремност учесника у реализацији активности да поделе детаље, сазнања и тајне о успешним пројектима. Стога се, у радовима новијег датум у циљу дефинисања концептуалног оквира и формирања листе критичних фактора успеха *DM* пројеката углавном примењује следећа методологија истраживања: најпре се, прегледом релевантне академске литературе и истраживачких радова из домена *DM*-а и повезаних области, саставља (дуга) прелиминарна листа потенцијлних критичних фактора, а затим у емпиријском истраживању, кроз анкетирање компетентних стејкхолдер партиципаната о степену значају сваког фактора за успешност *DM* пројеката, тестира релевантност сваког од њих, врши њихово рангирање и, коначно, категоризација.

Према општој, могло би се рећи грубој, категоризацији критичних фактора успеха *DM* пројекта, коју је предложио *Tufféry* (2011, стр. 617-621), могуће је издвојити пет група фактора повезаних са контекстом: предмета истраживања, података, људских ресурса, информационих система и пословне културе, а у оквиру којих је могуће идентификовати бројне подгрупе и димензије. У наставку текста укратко се представљају наведене категорије.

Предмет истраживања, односно разматрани проблем мора бити дефинисан на начин да његово решење заиста захтева примену *DM* приступа и употребу *DM* метода и алата. У том смислу потребно је разликовати, на пример, задатак откривања 20% купаца који генеришу 80% профита и примену дескриптивне статистичке анализе од стварних *DM* задатака усмерених на дефинисање профила садашњих и будућих купаца. Дакле, дефинисани предмет (идентификовани пословни проблем) *DM* пројекта мора бити компатибилан са суштинским карактеристикама процедура анализе великих количина података и потенцијалом *DM* анализе. У непосредној вези са дефинисањем предмета *DM* пројекта јесу и питања која се односе на детерминисање циљне популације и прецизирање значајних и реалних циљева саме апликације, крајњих корисника, као и пословних циљева.

Као што је већ истакнуто у Потпоглављу 3.3., вештине, стручност и знање појединаца на различитим организационим нивоима, који као чланови *DM* тима учествују у припреми, спровођењу и интерпретацији резултата *DM* пројекта, представљају један од кључних фактора за успешну реализацију *DM* задатака. При томе, сваки члан тима може имати више улога (задатака), као што и више појединаца може бити одговорно за једну улогу (*Myatt & Johnson*, 2014, стр. 7). Осим тога, учесници у пројекту своје задатке реализују у различитим временским периодима, односно фазама пројекта. Консеквентно, активна сарадња свих актера укључених у пројекат је од суштинског значаја са становишта откривања законитости које поседују вредност и смисао у контексту посматраног проблема. Посебно значајн аспект реализације *DM* пројекта, а који је индиректно повезан са људским ресурсима и тимским радом, односи се на чињеницу да сваки пројекат мора бити праћен и подржан одговарајућом пословном одлуком. Заправо, подршка топ менаџмента представља веома важну димензију реализације пројекта.

Карактеристике скупа података су уско повезане са реализацијом *DM* пројекта. Ове карактеристике се односе на обим, тачност, комплетност, доступност, временску одредницу, употребљивост и комплексност прикупљених података, као и велики број њихових извора. При томе је неопходно укључити и правни аспект коришћења и заштите података. Пошто за потребе *DM* анализе није једноставно обезбедити податке одговарајућег квалитета, а квалитет резултата примене *DM* метода и алата директно зависи од карактеристика података, у имплементацији *DM* пројеката подаци представљају критичан фактор успеха. Изузетно значајна димензија у разматрањима о подацима са становишта успешности у спровођењу *DM* активности се односи на избор метода и софтвера путем којих се обезбеђују *DM* модели (резултати). Наведеним проблемима посебна пажња је посвећена у тексту који следи.

За реализацију *DM* пројеката од пресудне важности је развој и имплементација адекватне *IT* платформе, јер, као што је раније већ речено, *DM* је и настао као резултат еволуције информационе технологије. Настојања организација да прикупе, складиште и анализирају податке и информације уз напредак и брзину усвајања нових технолошких иновација, резултирао је развојем постојећих *IT* система и појавом нових решења у форми софистицираних система за подршку процесу доношења одлука, попут система за подршку одлучивању, складиштења података, управљања знањем, управљања корпоративним перформансама, као и самог *DM*-а. Из перспективе *IT* система организације, *DM* процес је знатно олакшан уколико су подаци из различитих

извора интегрисани у оквиру складишта података. Мада се *DM* анализа може спроводити и уколико не постоји изграђено складиште података, у том смислу, једна од корисних опција са становишта имплементације *DM*-а је паралелна, пре него одвојена, реализација *DM* пројекта и пројекта складиштења података.

Способност организације да се прилагоди новонасталим околностима у пословном окружењу све више је директно зависна од развијања и инкорпорирања података и елемената *DM*-а у културни образац организације. Отуда, *DM* мора постати део вредности и норми при решавању проблема, а самим тим и део миљеа организационе културе. Сваким пословним подухватом који представља предмет интересовања и примене *DM*-а неопходно је управљати не само током фаза његовог спровођења, већ обезбедити и адекватну контролу, правилну интерпретацију и визуелизацију откривених резултата у форми законитости, као и њихово документовање и употребу од стране стејкхолдер корисника. Наиме, непосредни корисници морају схватити итеративну природу *DM*-а и уградити у своје системе вредности уверења о користима које се могу очекивати од примене *DM* приступа у решавању проблема и доношењу одлука, а искуство и сиситематизована претходно откривена знања користити као средство за реализацију актуелних и будућих *DM* пројката и унапређење компетентности запослених у том домену.

Полазећи од наведеног, може се констатовати да успешна реализација *DM* процеса у форми пројектног задатка зависи од бројних фактора, при чему сарадња између свих учесника са одговарајућом одговорностима, компетенцијама и вештинама заузима посебно место. Будући да пословање данашњице карактеришу огромне количине дигиталних података, савремени приступи анализи података, попут *DM*-а, постају круцијални извори за обезбеђење додате вредности и побољшање пословања. У том смислу, *DM* процеси се морају интегрисати са пословним процесима у организацији, а компетентни учесници, сходно својој улози у процесу откривања знања из података, како би могли допринети доношењу валидних пословних одлука, морају поседовати одговарајући (висок) ниво и (широк) распон знања о разматраном проблему и коришћеној методологији.

4.3. Ефекти, митови и реалности везани за *data mining*

Са становишта међународног и локалног пословања и, генерално, решавања економских проблема, *DM* апликације откривају и обезбеђују значајне развојне могућности у правцу унапређења ефикасности, продуктивности, профитабилности,

потрошачке сатисфакције, процеса одлучивања и конкурентске позиције, али истовремено са собом носе бројне изазове и потенцијалне негативне импликације на сигурност података, утрошак времена и новца, укључујући и губитак кредибилитета. Стога се, у овом делу текста, указује на позитивне и негативне стране примене *DM*-а.

Baicoianu & *Dimitrescu* (2010, стр. 186) истичу следеће главне предности *DM*-а:

- обезбеђује информације о пословним процесима, купцима и тржишном понашању,
- у стању је да искористи податке који су расположиви у репозиторијумима,[16] и
- открива правилности садржане у подацима које доприносе акумулацији пословног знања и предвиђању будућих догађаја, трендова и понашања на економским тржиштима.

Бројне користи од *DM*-а могуће је посматрати на нивоу организације (пословања), појединаца и друштва (*Wang*, 2003, стр. 406; *Bal et. al.,* 2011). Са аспекта организационог пословања, *DM* омогућава: ▸ идентификовање производа и услуга и њихових карактеристика који су важни за клијенте, ▸ идентификовање и креирање одговарајуће понуде која је у складу са специфичним потребама купаца, ▸ идентификовање купаца (сходно њиховим досадашњим преференцијама) који ће бити заинтересовани за нове производе и услуге (одноно, боље разумевање и предвиђање промена потреба купаца), ▸ идентификовање купаца које карактерише висока стопа куповине одређених производа, ▸ прилагођавање маркетинг кампање специфичностима конкретног тржишта, ▸ идентификовање, привлачење и задржавање изузетних (топ) купаца, ▸ идентификовање нових тржишних могућности, анализу канала дистрибуције и њихову оптимизацију, ▸ да се унапреде односи са клијентима, ▸ да се повећа продуктивност, ▸ да се редукује ризик, ▸ да се уштеди време и новац итд.

Доприносећи генерално побољшању услуга, задовољства и животног стила појединаца, *DM* користи са аспекта појединаца се односе на: ▸ брзи приступ интегрисаним системима и информацијама, ▸ брзину одговора на захтеве клијената, ▸ пружање бољих услуга корисницима, ▸ боље разумевање и прилагођавање захтевима клијената, ▸ креирање бољих односа са клијентима, ▸ обезбеђење резултата и закључака који не би били идентификовани применом једноставних анализа и класичних метода итд. Користи за друштво су повезане са: ▸ обезбеђењем информација за превентивно деловање у домену тероризма, и ▸ идентификовањем недозвољених активности.

[16] *DM* спречава да велике количине података остану неискоришћене архиве, јер се анализирањем подаци трансформишу у корисне информације и знање, додајући вредност самим подацима.

Међутим, у једном ширем смислу, не треба превидети друштвене користи од дигитализације и електронских информација са аспекта функционисања владе и грађана једне земље.

Експандирајући значај *DM*-а и настојање да се из података извуче корист за организацију, донео је потпуно нове изазове у погледу пословних улога запослених, као и вештина, знања и аналитичких способности које се од њих траже за рад са великим количинама података. Наиме, пословање и одлуке засноване на подацима захтевају ангажовање запослених који, поред одређених знања и вештина из домена статистике, машинског учења, програмских језика, примењене математике, али и управљања, тимског рада и комуникације, морају поседовати *X* фактор који у овом случају представља интелектуалну радозналост, односно визију о подацима (*Lukić,* 2013, стр. 329). Раст интересовања за рад са подацима, консеквентно, допринео је развоју широког спектра занимања и дефинисању нових радних места и позиција које су директно повезане са подацима. Данас у свету ова занимања припадају категорији најтраженијих, али и најплаћенијих занимања. Такође, у огласима српских организација су све чешће садржани захтеви за занимања из категорије рада са подацима. С обзиром да не постоји једнозначни српски термини за ова занимања, у огласима се могу пронаћи следећи називи позиција: пословни аналитичар, *DM* менаџер, специјалиста за интеграцију података. специјалиста за развој решења пословне интелигенције, стручњак за *DM* и анализу података, асистент статистичке обраде података и слично. Међутим, једно је сигурно: професионалци из ове области су дефицитаран кадар.

Иако *DM* обезбеђује низ предности, ипак постоје потенцијални недостаци и опасности које су последица ограничења дефинисаних у форми критичних фактора успеха. Најчешће се говори о недостацима који се односе на време, високе трошкове и сложеност реализације *DM* пројеката, као и недостатак стручности у спровођењу *DM* активности. О недостацима *DM*-а, као и у случају наведених користи, такође се може говорити са аспекта организације, појединаца и друштва (*Wang*, 2003, стр. 406). На сва три нивоа, недостаци се односе на нарушавање приватности и сигурност података, неадекватно коришћење прикупљених података, нетачност података и формулисање закључака (и предвиђања) засновано на таквим подацима. Наиме, појава великих база података и Интернета намеће нове (етичке) изазове по питању сигурности податка и информационих система и заштите права на приватност појединаца. У настојању да се ови проблеми ублаже непрекидно се морају развијати одговарајуће технике, као и

релевантни законски оквири, директиве, процедуре и политике о заштити података и приватности на различитим нивоима.

Нема сумње да резултати спроведених *DM* апликација могу бити добра основа за постизање низа користи. Међутим, многе организације не искористе могућности за успех због тога што чине круцијалне грешке у планирању и развоју *DM* апликација. Свест менаџмента, а посебно руководилаца пројеката о могућим грешкама и њиховим консеквенцама знатно смањује ризик од неуспеха *DM* пројеката.

Larose (2005, стр. 10-11) наводи следеће најчешће заблуде које узрокују грешке и неуспех у спровођењу *DM* пројектног задатка:

- Постоје *DM* алати који се могу (механички) применити на репозиторијум података и аутоматски идентификовати скривене законитости и обезбедити решења актуелних проблема.
- Процес *DM*-а је аутономан и не захтева знатан људски мониторинг.
- *DM* сам себе исплати веома брзо.[17]
- *DM* софтверски пакети су интуитивни, једноставни за коришћење и не захтевају фундаментално познавање метода обухваћених тим софтверским пакетима.
- *DM* идентификује узроке пословних и истраживачких проблема.
- *DM* аутоматски чисти и ажурира несређене базе података.

На основу личног искустава у решавању конкретних пословних *DM* проблема за клијенте компаније *Elder research*, *Deal* (2013), пак, идентификује следећих десет најчешћих *DM* грешака, које узрокују неуспех *DM* пројеката:

- Грешка 1: Нејасно и непрецизно дефинисање циљева и планова њиховог извршења. Ова грешка је последица настојања да се што пре примени нова технологија, без јасног увида у могућности које *DM* пружа у контексту решавања конкретног реалног проблема.
- Грешка 2: Суочавање са проблемом „превише и пребрзо". Реализација *DM* иницијативе представља велики подухват који, између осталог, захтева значајне ресурсе, огромну количину организационе енергије и промене у организационој култури, тако да је нереално очекивати велике резултате за кратко време и, у том смислу, неосновано убрзати (или чак и прескочити неке) фазе *DM* процеса.
- Грешка 3: Одсуство подршке власника података. Често власници кључних података у организацији нису спремни да у потпуности ставе податке на располагање

[17] Брзина и стопа повраћаја инвестиција су варијабилне категорије које зависе од низа фактора, тако да их је врло тешко (временски) генерализовати.

за потребе *DM* анализе. Заправо, понекад је дозвољен или обезбеђен само ограничени приступ подацима кроз достављање непотпуних скупова података, без пратећег речника о значењу појединих поља у табели и објашњења како су подаци прикупљени.

- Грешка 4: Чекање савршених података. Многе организације, и поред добро започетог *DM* пројекта (у смислу дефинисања циљева, одређивања приноса на инвестиције, састављања плана пројекта, обезбеђених и одобрених средстава, формирања *DM* тима), нерадо настављају да раде на пројекту док не обезбеде потпуне и добро организоване податке. Будући да је савршене податке готово немогуће обезбедити, чекање на такве податке непотребно продужава период рада и доводи у питање предвиђено време завршетка пројекта.

- Грешка 5: Веровање да су подаци организације савршени. Ниједна организација нема савршене податке. Међутим, веровање да су подаци организације савршени имаће за последицу недовољну посвећеност припреми („већ савршених") података што ће довести до нереалних очекивања у погледу времена и трошкова који су потребни за завршетак *DM* пројеката, али и до лоших резултата моделирања.

- Грешка 6: Превелико ослањање на софтверске пакете. Један од кључних проблема у *DM* апликацијама јесте избор адекватног софтверског решења. Међутим, то не значи да ће се избором и покретањем софтвера подаци сами моделирати. Спровођење *DM* анализе како би се решили одређени типови проблема у организацији, уз одговарајућа софтверска решења, захтева укључивање *DM* стручњака и стручњака из домена анализе података који поседују потребна знања о својствима *DM* метода, као и о начинима њихове креативне примене. У том смислу, експертска апликација софтвера је подједнако битна као и избор софтвера.

- Грешка 7: Неразумевање различитих нивоа аналитике. У данашњем окружењу, готово у свакој организацији постоји свест о важности коришћења аналитичких техника за остварење пословних циљева. Међутим, свака техника има своју примену, тако да у одређеним случајевима оне не поседују корисност са аспекта конкретне организације. Наиме, примењени алати морају бити усклађени са потребама организације, а експертска анализа, у функцији побољшања пословног одлучивања, усклађена са аналитичким способностима у организацији или могућностима ангажовања екстерних консултаната.

- Грешка 8: Искључивање експерата из конкретног пословног подручја којем припада анализирани проблем. Поред стручњака из домена анализе података,

стручњаци из пословног подручја представљају веома важну карику у решавању практичних проблема и успешној реализацији *DM* пројеката. Њихово укључивање доприноси бољем разумевању пословног проблема, потпунијем и прецизнијем процесу моделирања, логичкој провери и ефективној примени коначних резултата.

- Грешка 9: Избегавање и одлагање одговора на питања у домену имплементације модела и планирања развоја. Процес откривања знања из података применом *DM* приступа се не завршава креирањем *DM* модела. Заправо, смисао целог овог процеса је активирање креираног модела, што често захтева огроман напор. Стога некомплетно сагледавање процеса од креирања модела до његове примене може изазвати велике губитке са аспекта трошкова, времена, али и кредибилитета како конкретне организације, тако и консултантске фирме.
- Грешка 10: Убрзавање процеса. Изостављање било које фазе у *DM* процесу, под изговором да је непотребна, сувишна и да представља губљење времена и новца доводи до незадовољавајућих резултата. Дугорочно посматрано, позитивни ефекти се могу постићи само под условом да се члановима *DM* тима ставе на располагање неопходни ресурси (независно д тога колико они изгледали велики), укључујући и време, како би стручно обавили свој посао.

Насупрот овим заблудама и грешкама, *Khabaza*, један од коаутора *CRISP-DM* методологије и оснивач Друштва за *DM* аналитичаре, формулисао је девет закона (максима) који представљају истине (реалности) о *DM*-у, а у којима су садржане кључне претпоставке за успешну имплементацију *DM* концепта. У питању су следећи закони (*Khabaza*, 2010):

1) Закон пословних циљева: Пословни циљеви представљају темељ сваког *DM* решења. *DM* је примарно процес (а не технологија) који у својој основи има један или више пословних циљева.

2) Закон пословног знања: Пословно знање има централну улогу у сваком кораку *DM* процеса. Без пословног знања ниједан корак *DM* процеса не може бити ефективан, јер не постоје технички чисти кораци овог процеса.

3) Закон припреме података: Припрема података је више од половине сваког *DM* процеса. Према неформалним проценама, учешће припреме података у укупним *DM* напорима варира од 50% до 80%.

4) Закон правог модела: Прави модел за конкретну проблемску ситуацију може се идентификовати само путем експеримента, или (метафорички) „нема бесплатног доручка” за *DM* аналитичаре. (Сваки *DM* аналитичар је свестан чињенице да не постоји

универзални алгоритам који је подједнако добар у различитим неструктурираним проблемским ситуацијама, које представљају предмет *DM* анализе. Стога, тражење решења (модела) за сваки проблем представља озбиљан задатак кроз систем „покушаја и грешке", а софтверски алати су само помоћно средство у спровођењу *DM* процеса.)

5) Закон о законитостима: Законитости увек постоје као неизбежни производ сваког *DM* процеса. Поред корисних законитости, постоје и оне које су присутне, али не и релевантне за решење посматраног проблема.

6) Закон појачаног разумевања: Мада пословне проблеме решавају људи, а не алгоритми, *DM* методи доприносе бољем разумевању и решавању проблема, интегришући природан људски перцептивни механизам и законитости које би било тешко или немогуће идентификовати без адекватне софтверске подршке.

7) Закон предикције. Предикција, путем генерализације, повећава информације на локалном нивоу. *DM* омогућава замену неформалних очекивања са конзистентним, прецизнијим и на подацима заснованим оценама. Наиме, на основу сличности са претходним случајевима чији је исход познат, добијају се информације о вероватним исходима за нове случајеве. Важно је напоменути да ове нове информације нису подаци у смислу расположивих сирових података, већ информације у статистичком смислу, односно законитости откривене у процесу моделирања.

8) Закон вредности: Вредност *DM* резултата није (доминантно) детерминисана тачношћу и стабилношћу предиктивних модела. Тачност и стабилност су веома важни аспекти за вредновање *DM* модела, али у фокусу *DM* аналитичара примарно морају бити питања која су тангентна са смислом, разумевањем и коришћењем модела у контексту конкретног пословног проблема.

9) Закон промене: Све законитости су подложне променама. Откривене законитости одражавају не само промене у свету, већ и промене у самом разумевању откривених модела. Модел који данас даје добре прогнозе, већ сутра може бити ирелевантан.

Сходно претходно наведеном, очигледно, *DM* је присутан у свим аспектима и активностима на свим нивоима деловања, независно од тога да ли су људи тога свесни или не. Због тога се говори о свеприсутном и невидљивом *DM*-у (*Han et al.*, 2012, стр. 618-620). У том контексту, у овом делу излагања представљен је, истина, кратак осврт на веома битне аспекте *DM*-а што, свакако, не умањује њихов значај, већ доприноси целовитом сагледавању комплексности подручја *DM*-а. Фокус даљих излагања су методолошки аспекти *DM*-а.

Део II

МЕЂУЗАВИСНОСТ КАРАКТЕРИСТИКА ПОДАТАКА, ЗАДАТАКА И МЕТОДА У КРЕИРАЊУ *DATA MINING* МОДЕЛА

5. ПОДАЦИ КАО КЉУЧНИ ЕЛЕМЕНТ *DATA MINING* КОНЦЕПТА

Имајући у виду чињеницу да при спровођењу било које процедуре за анализу података карактеристике података детерминишу ток и квалитет резултата анализе, у овом Поглављу су сагледани кључни концепти повезани са концептом подаци, спроведене различите категоризације података, разматрани начини њиховог организовања и, коначно, анализирани одређени аспекти (високог и ниског) квалитета података, као фактора са значајним импликацијама на ток пословних процеса, ефикасност пословног одлучивања и организационе перформансе.

5.1. Фундаментални концепти повезани са појмом подаци

Подаци представљају базичну компоненту процеса откривања знања из података. Сам појам податак употребљава се у разним контекстима о којима је било речи у Потпоглављу 1.2. У ширем смислу, релевантном са становишта *DM*-а, као централне фазе процеса откривања знања, подаци нису само цифре; заправо, елементи података, као дескрипција феномена, могу бити и други записи у форми речи (текста), дијаграма (фигуре), слике, звука и видео записа.

Дакле, податак може бити било која наведена форма знакова или симбола у којој је физички забележен неки догађај, запажање или чињеница, али са одговарајућим контекстом тако да може пренети одређену информацију. Другим речима, подаци представљају сирови материјал који се користи за генерисање информација. У етимолошком смислу, реч податак (енгл. *data*) је множина латинске речи „*datum*”, која потиче од глагола „*dare*” (енгл. *to give*) и у српском језику преводи се глаголом давати, тако да се појам податак, оквирно, односи на оно што је дато, то јест, на дату вредност.

Мултидисциплинарно и хибридно порекло *DM*-а довело је до тога да се често од стране истраживача и практичара користе различити термини да означе исти појам или контекст, или пак, исти термини, а различити контекст. Поред тога, статистички термини и изрази могу имати значења која се разликују од њихових уобичајених и свакодневних употреба. Како би се услед наведеног отклониле евентуалне терминолошке непрецизности и нејасноће, у наставку се укратко разматрају и илуструју основни концепти који су непосредно повезани са појмом подаци.

Јединице посматрања су истоврсни појединачни случајеви, односно ентитети (субјекти или објекти) о којима се прикупљају, сређују, складиште и обрађују подаци. У статистичком смислу, скуп свих јединица посматрања конституише основни скуп, а

део јединица основног скупа за сврхе статистичке анализе конституише узорак (*Lovrić*, 2009).

Јединице посматрања[18] поседују бројне квалитативне и квантитативне карактеристике (особине) интересантне са становишта истраживања. Карактеристика према којој се јединице посматрања међу собом разликују и која представља предмет проучавања назива се обележје. Осим термина обележје, често се у литератури за карактеристике јединица посматрања као синоними користе и следећи термини: димензија, својство, атрибут, варијабла или променљива (односно, енгл. *dimension, feature, attribute, variable*, респективно). Термин димензија уобичајено се користи у области складиштења података, у литератури из области машинског учења доминира коришћење термина својство, *DM* професионалци преферирају термин атрибут, а статистичари термин варијабла (променљива) (*Han et al*., 2012, стр. 40). Обележја / променљиве[19] се представљају путем скупа кореспондирајућих вредности. Различити видови у којима се једна променљива може испољити називају се модалитети или вредности посматране променљиве, а вредност променљиве (или променљивих) која се односи на једну јединицу посматрања назива се податак или опсервација.

Постоји неколико начина за класификацију обележја / променљивих / варијабли. Сходно оквиру овог истраживања, примарни критеријум за њихово разликовање представљају потенцијалне вредности испољавања карактеристика јединица посматрања. Према овом критеријуму променљива може бити квантитативна (нумеричка) и квалитативна (категоријска, атрибутивна). Заправо, квантитативна карактеристика јединица посматрања назива се нумеричко обележје, а квалитативна карактеристика атрибутивно (категоријско) обележје. Атрибутивна обележја се изражавају описно (речима), а варијабилитет се испољава кроз припадност елемената (јединица посматрања) различитим категоријама (две или више). Квантитативна обележја су такве карактеристике елемената које се изражавају бројчано. Унутар ове групе разликују се прекидна (дискретна) и непрекидна (континуирана) нумеричка обележја. Модалитети прве групе могу бити само изоловане вредности, односно цели бројеви на мерној скали, док су модалитети друге групе било које нумеричке вредност унутар неког интервала, исказане целим бројевима или децималним записима.

[18] За јединице посматрања у статистичком смислу, у *DM* литератури се користе и следећи називи: објекти (енгл. *objects*), записи (енгл. *records*), примери (енгл. *examples*), јединице (енгл. *units*), случајеви (енгл. *cases*) итд.

[19] У складу са статистичком терминологијом, карактеристика појединачних елемената посматрања која представља предмет анализе је променљива величина и узима различите вредности „од случаја до случаја", односно од једне до друге јединице посматрања. Будући да узима вредности „на случај", које се не могу тачно унапред предвидети, она се назива случајна променљива или варијабла (енгл. *random variable*) (*Lovrić*, 2009, стр. 94-95).

Свако прикупљање података о карактеристикама јединица посматрања подразумева мерење и коришћење одређених мерних скала. Мерење представља, у скаладу са дефинисаним правилима, придруживање бројева или одређених ознака јединицама посматрања (*Lovrić*, 2009, стр. 26), при чему се придружени бројеви или ознаке односе не на елементе, већ на њихове карактеристике. Са становишта прецизности, разликују се четири нивоа мерења (од најнепрецизнијег до најпрецизнијег) и четири примарне мерне скале: ▸ номинална, ▸ ординална, ▸ интервална, и ▸ релациона.

Номинална скала се користи за прикупљање података о категоријским варијаблама, које се могу класификовати према броју и типу модалитета, при чему било какво рангирање нема значаја. Вредности на номиналној скали су листе категорија, симбола или назива по којима се јединице посматрања разликују. Овим вредностима могу се придружити и бројеви, али они немају метричка својства и нису намењени за квантификовање међусобних релација између елемената из различитих група. Дакле, релације које се могу успоставити на номиналној скали су „једнако" или „није једнако" са елементима у групи. Заправо, на номиналној скали, осим пребројавања, нема смисла примењивати математичке операције, чак и када су вредности представљене бројевима. Путем пребројавања идентификује се припадност елемената одређеним категоријама и формира одговарајућа расподела фреквенција. На тај начин могуће је добити важне информације о структури појаве и учешћу појединих категорија. Осим тога, у случају номиналне скале, једина мера централне тенденције коју има смисла одредити је модус.

Ординална скала се користи за категоријске варијабле чије је модалитете могуће рангирати према интензитету или значају у складу са утврђеним критеријумима, али магнитуда (величина растојања и степен разликовања) између суседних вредности је непозната, односно не може се прецизно утврдити. Јединицама посматрања се, према степену одређеног својства, придружују бројеви, словне ознаке или други симболи. Наиме, ординалној скали су својствене особине бројчаног система да бројеви следе релацију поретка (ранга), тако да је 5 веће од 4, а мање од 6. Међутим, бројеви само означавају ранг, а не и величину разлике између рангова. Ординална скала није обавезно линеарна; разлика у посматраном својству између елемената са рангом 5 и 6 не мора бити једнака разлици између елемената са рангом 10 и 11. Дакле, релације које се могу успоставити на ординалној скали су „веће од", „мање од" и „једнако са". Централну тенденцију ординалне варијабле репрезентују модус и медијана.

Интервална скала је нумеричка скала која се користи за прикупљање података о нумеричким варијаблама. Вредности на интервалној скали су рангиране и могу бити позитивне, нула или негативне. Поред рангирања, могуће је утврдити и квантификовати разлику између вредности варијабли које се мере на интервалној скали. За разлику од претходно наведених скала, ова скала омогућава одређивање величине одстојања између нумеричких вредности. Такође, омогућава да се упореде разлике између преференција као изразито психолошких категорија, тако да разлике између нумеричких величина на интервалној скали не показују и једнаке и апсолутно прецизне разлике измерених карактеристика јединица посматрања. Карактерише је и арбитрарни почетак, услед чега се не може користити за изражавање количинских (релативних) односа. Путем интервалне скале мере се и нумеричка обележја чији арбитрарни почетак може бити нула, али нулта вредност не значи одсуство појаве, већ представља само произвољно изабрану почетну тачку. Релације које се могу успоставити су „веће од", „мање од", „једнако са" и „различито", уз смисленост примене математичких операција сабирање и одузимање. За податке мерене на овој скали могуће је, осим модуса и медијане, одредити и аритметичку средину.

Релациона скала је, такође, нумеричка скала која показује и редослед модалитета и меру њиховог разликовања. Ова скала је са инхерентном нултом тачком, која има прави смисао, јер нулта вредност у случају конкретне јединице посматрања указује на одсуство карактеристике која је предмет мерења. За ову скалу, као највиши ниво мерења, важи правило да једнаке разлике између бројева на релационој скали означавају и једнаке разлике између утврђених вредности мерене карактеристике јединица посматрања, односно прецизно квантификује разлику између вредности. Поред тога, релациону скалу карактерише и употреба јединице мерења, при чему се може вршити превођење података са једне скале и одређеном јединицом мере на скалу са другом јединицом мере. Ова скала омогућава исказивање пропорционалних односа између модалитета посматраног обележја. Релације које се могу успоставити су „веће од", „мање од", „једнако са", и „различито", а математичке операције сабирање, одузимање, множење и дељење има смисла применити. За податке мерене на овој скали могуће је одредити све три претходно поменуте мере централне тенденције.

Генерално, пре примене *DM* метода потребно је добро познавати својства података, варијабли, као и мерних скала којима подаци припадају, јер сходно томе детерминисане су могућности примене смислених поступака над подацима.

5.2. Типологија података

У једном ужем контексту посматрано, различити типови, односно категорије података су детерминисани скупом могућих вредности карактеристика јединица посматрања (то јест, доменом података) и дозвољеним рачунским и логичким операцијама над тим вредностима. Сходно претходној класификацији карактеристика јединица посматрања, разликују се: ▸ квантитативни, и ▸ квалитативни подаци. Подаци прикупљени о квантитативној варијабли су квантитативни (нумерички) подаци. Насупрот томе, квалитативни подаци изражавају квалитативне концепте.

Специјални случај квалитативне променљиве са две вредности је бинарна (дихотомна) променљива. За ову променљиву постоје само две различите вредности које могу бити, на пример, „да и не", или у нумеричкој форми, најчешће 0 и 1. Као што се квалитативни подаци могу исказати нумерички, с друге стране, нумерички подаци се могу представити у квалитативној форми. Уколико се прикупљају подаци о радном стажу запослених, у извештај се може уносити учешће запослених са радним стажом испод и изнад 5 година. У овом случају реч је о нумеричким подацима, али је начин њиховог презентовања категоријски. Такође, питање границе између нумеричких и симболичких података илуструје и прикупљање података о степену задовољства потрошача квалитетом новог производа, на скали, на пример, од 1 (врло незадовољан) до 5 (врло задовољан). Заправо, без познавања природе посматране променљиве и суштине проблема који се решава, често је тешко направити јасну разлику између ове две категорије података.

Полазећи од четири примарне мерне скале, разликују се и четири кореспондентне категорије података: ▸ номинални, ▸ ординални, ▸интервални, и ▸релациони подаци. Такође, према броју променљивих, које имају функцију дескриптора јединица посматрања, подаци се могу класификовати на ▸ једнодимензионалне, ▸ дводимензионалне, и ▸ мултидимензионалне податке. Ови термини представљају ситуације у којима је, у зависности од циља истраживања, пажња усмерена на представљање и проучавање јединица посматрања путем једне, две или више променљивих, респективно.

Из перспективе *DM*-а и складиштења податка, ниво апстракције је веома важна карактеристика података. У разматраном контексту, ниво апстракције се односи на кретање од сирових (необрађених, трансакционих, оперативних) података ка вишим нивоима обраде, сумирања, агрегирања и консолидовања података, до коначног

дефинисања пословних законитости и правила (*Berry & Linoff*, 2004, стр. 475-484). Са повећањем нивоа апстракције смањује се количина података у смислу заузетости капацитета меморијског простора информационих сиситема.

Сходно нивоу апстракције разликују се следеће категорије података (Слика 8): ▸ трансакциони подаци (су сирови подаци са најнижим нивоом апстракције који се односе на карактеристике појединачних трансакција бележене у одређеном оперативном систему обезбеђујући информације о томе ко, шта, где, када и колико), ▸ сумарни подаци (означавају сумирање трансакционих података из различитих перспектива, укључујући и временску димензију), ▸ подаци складиштени у базама (кроз одговарајуће шеме логичког и физичког структурирање података), ▸ метаподаци (или, „подаци о подацима" имају функцију документације о структури складишта података, указујући на логички модел, физички распоред података и локацију локалних складишта података, хијерархију, дозвољене или могуће вредности података, информације о датумима уноса, изведене податке и слично, тако да се често називају и речником података), и ▸ пословна правила (као највиши ниво апстракције података, односе се на законитости, које су, кроз апликацију софистицираних алгоритама, идентификоване у поступку процесирања података на различитим нивоима детаљизирања и агрегирања, генерално, показујући шта је научено из података).

Слика 8: Хијерархијски приказ типова података према нивоу апстракције

Извор: *Berry & Linoff* (2004, стр. 475)

Посебно значајан критеријум за категоризацију података који се користе у *DM* процесу је структурираност, то јест, начин и тип организовања података узимајући у обзир њихове компјутерске формате (односно, шеме или приказе) и могућност коришћења за откривање законитости. Заправо, структура података је дефинисана шемом података и описује тип и редослед мањих јединица података унутар већих јединица. Према овом критеријуму, подаци се могу класификовати у три категорије (*Kantardžić*, 2011, стр. 11): ▸ структурирани, ▸ полуструктурирани, и ▸ неструктурирани (енгл. *structured data*, *semi-structured data* и *unstructured data*, респективно).

Структурирани подаци, који се често називају традиционалним подацима, организовани су у форми јасно уређених и, са становишта управљања, погодних приказа. Реч је о подацима за које је унапред дефинисан модел података или су организовани на унапред дефинисан начин. Наиме, постоји стриктно одређена шема којом се дефинише тип података, структура података и њихове релације. Типичан пример структурираних података су подаци у табеларним приказима (попут прорачунских *Microsofr Office Exsel* табела), у датотекама података многих софтвера за статистичку обраду података и традиционалним, релационим базама података (у којима „све *n*-торке података имају исти формат").

Полуструктурирани и неструктурирани подаци, као подаци исказани или сачувани у специфичним форматима (или пак, неформатирани подаци), заједно су означени као нетрадиционални подаци, а називају се и комплексним типовима података. Могу се појавити у разнородним, различитим и нестандардним (текстуалним и нетекстуалним) формама, попут: текстуалних докумената, *web* страница, цртежа, табела, графикона, дијаграма, мапа, *e-mail*-ова, *SMS* порука, телефонских разговора, презентација и других мултимедијалних записа, као и садржаја. *Han et al.* (2012, стр. 586) су ову широку групу комплексних података класификовали у следеће три категорије: секвенцијални подаци (временске серије, симболичке и генетичке секвенце), графови и подаци друштвених и информациних мрежа и група осталих типова података (попут просторних података, просторно и временски зависних података, мултимедијалних података, текстуалних података, *web* података, података у кретању (енгл. *data streams*)).

Полуструктурирани подаци не захтевају стриктно дефинисање одређене структуре. Међутим, то не значи да шема није могућа, већ је прилично флексибилна. Врло брзо се могу трансформисати у форму погодну за конвенционалне аритметичке операције. Неструктурирани подаци су подаци који не могу бити представљени у

форми претходно дефинисане структуре и на њима се не могу спровести класичне аритметичке операције. У контексту релационих база података, то су подаци који се не могу приказати у форми редова и колона релационих табела. Дакле, неструктурирани садржаји немају стандардизовану структуру мета података. Међутим, са становишта њихове организације и захтева система за складиштење могуће је дефинисати извесне екстерне формате који описује семантику података, али не припадају структури. На пример, електронска пошта може бити уређена по датуму, времену, пошиљаоцу, примаоцу или предмету, али садржај („тело") поруке остаје неструктуриран.

Према извештајима бројних консултантских компанија, услед развоја *ICT* система, раст, нарочито, неструктурираних података је знатно убрзан, тако да они чине преко 80% пословних (*Marr*, 2015, стр. 43), а и глобалних података. Међутим, екстракција вредних информација садржаних у овим подацима захтева обимна и компликована истраживања, која су често и веома скупа. За разлику од структурираних података који су спремни за интеграцију у базе података или добро структуриране датотеке, у циљу обраде и сажимање неструктурираних садржаја велике комплексности у садржаје мање комплексности, захтевају се специјална претпроцесирања (кодирање и форматирање) како би се могле применити посебне технологије које омогућавају оперативно коришћење ове врсте података.

Са појмом подаци непосредно су повезани и извори података. Сходно изворима и начинима прикупљања података, разликују се: ▸ примарни, и ▸ секундарни подаци. За разлику од примарних података који су резултат директног прикупљања података која се реализују у складу са истраживачким задатком за постизање одређеног, претходно дефинисаног циља, секундарни подаци се односе на постојеће податке који су раније прикупљени и објављени за потребе неких других истраживачких пројеката и циљева. У основи, секундарни подаци могу потицати из различитих извора, тако да је у склопу пословног система могуће идентификовати три категорије секундарних података: ▸ интерни, ▸ екстерни, и ▸персонални подаци (*Vercellis*, 2009, стр. 44).

Интерни подаци се односе се на оне податке који су директно проистекли из активности предузећа. У том контексту пословни подаци могу бити финансијски, производни подаци, маркетинг подаци, подаци о квалитету и слично. Екстерни подаци се односе на активности изван предузећа. Од изузетног значаја су за процес доношења стратегијских одлука, јер омогућавају да се идентификује фактори релевантни за потребе *SWOT* анализе. У ову групу података могуће је сврстати податке о конкурентности, макроекономске податке, податке о технолошким и маркетиншким

трендовима, финансијске податке, робне податке (на пример, цене сировина), демографске податке, психометријске податке итд. Персонални подаци се односе на чињеницу да доносиоци одлука при спровођењу анализа користе информације, резултате личних истраживања и субјективне процене складиштене унутар датотека и локалних база лоцираних на персоналним рачунарима.

Са становишта дефинисаних циљева ове дисертације, претходно наведени критеријуми поделе могу се применити и на категорију економских података. Такође, у истом контексту, а сходно подели економије на микроекономију и макроекономији, веома је важно направити разлику између микроекономских и макроекономских података (*Feelders*, 2002, стр. 169).

Начелно, микроекономски подаци се односе на јединице посматрања као што су појединци, домаћинства и предузећа, док макроекономски подаци представљају резултат агрегирања (или упросечавања карактеристика) јединица на националном нивоу (друштвени бруто производ, индекс цена, стопа незапослености, национални берзански индекс и друго). Међутим, макроекономски подаци се могу односити и на друге географске контексте (локалне, регионалне и интернационалне), али и на ниво привредних сектора. Такође, предмет посматрања може бити продаја једног производа, али и агрегатна продаја групе производа (на пример, нутриционистичких или текстилних производа) или подгрупа у систему категоризације који је прилагођен пословним потребама (на пример, формирање класа производа према нивоу финанасијског или транспортног ризика). При употреби термина микро и макро треба имати у виду да је реч о категоријама релативног значења и да одговарајуће значење имају у односу на виши или нижи ниво посматрања и агрегирања.

DM приступ примењен на економске податке омогућава откривање законитости о макроекономским кретањима и обезбеђује предвиђање процеса и догађаја у макроекономским системима. Посматран са становишта микроекономских истраживања, овај приступ је усмерен на откривање законитости из података у циљу ефективног решавања пословних и управљачких проблема, при чему идентификоване законитости постају круцијални инпут у процесу одлучивања. Укључивање временске димензије знатно проширује могућности анализе економских података.

5.3. Организовање података

Са аспекта квалитета анализе економских појава, једна од кључних претпоставки за обезбеђење корисних информација и знања из података у правом тренутку, на

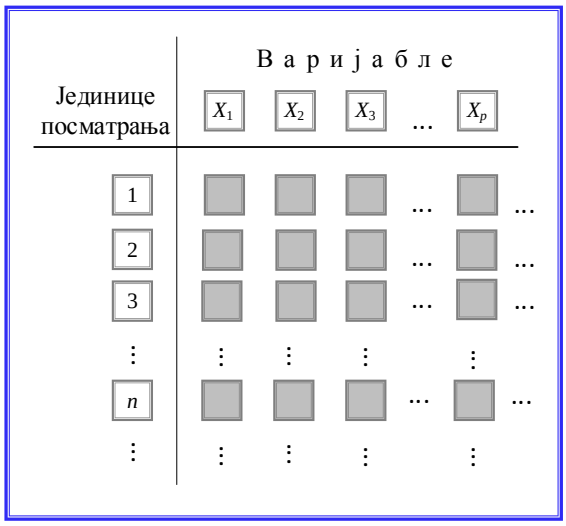
правом месту и у погодном облику односи се на ефикасно организовање података. У савременим условима валидно организовање великих количина података остварује се коришћењем компјутерске меморије, тако да се од информационог система очекује да осигура податке чији је садржај и облик прилагођен потребама корисника.

Под организацијом података, у техничком смислу, подразумева се физички и просторни распоред података у компјутерским меморијама пословног информационог система, укључујући елементе логичког представљања и повезивања података. Заправо, путем организације, подаци се доводе у одређени ред и на тај начин ствара основа за ефикасно спровођење целокупног поступка од прикупљања и одабира података за анализу, преко њихове обраде, до презентације резултата корисницима.

Фундаментална јединица за организовање (структурираних) података у компјутерским меморијама, на којој се базирају многи *DM* (и статистички) поступци, је табела, односно матрица података састављена од n редова и p колона, типа $n \times p$. Уобичајено је да сваки од n редова кореспондира са једном јединицом посматрања, а да свака од p колона приказује неку од заједничких карактеристика по којој се јединице посматрања међусобно разликују. Дакле, као резултат формалне припреме података за анализу, комбинацијом редова и колона настаје уређена и јединствена табела (енгл. *flat table*), а у сваком пресеку реда и колоне (односно ћелији или пољу табеле) добија се податак, као конкретизација (вредност, опсервација) одређене променљиве за јединицу посматрања. При томе, ове табеларне приказе карактерише стриктна структура организације података, јер су све јединице посматрања описане на истоветан начин и према истом редоследу карактеристика. Низ података који се односи на једну јединицу посматрања назива се слог или запис.

У *DM* контексту, скуп јединица посматрања представљених путем истих варијабли и организованих на одговарајући начин (најчешће у облику прецизно дефинисане табеларне структуре) назива се скуп података за анализу (енгл. *data set*). Поједностављени табеларни приказ скупа података и његових основних конститутивних елемената представљен је на Слици 9. Садржај генеричког елемента (поља) табеле (i, m), где је i=1, 2, . . ., n, а m=1, 2, . . ., p, означава i-ту статистичку јединицу посматрања која је класификована према нивоу m-те варијабле. Сваки ред кореспондира са једном јединицом посматрања, тако да се елементи у једном реду односе на различите карактеристике једне јединице посматрања и формирају њен профил. Заправо, ако се матрица података обележи симболом X, тада елемент матрице x_{im} представља вредност m-те променљиве мерене на i-том објекту.

Презентовани табеларни приказ је често тачка од које почиње *DM* (*Giudici & Figini*, 2009, стр. 9). Осим ове изворне форме, спровођењем одговарајућих трансформационих процеса, а сходно аналитичким потребама формирају се и изведене табеле. Такође, циљеви *DM* апликација често захтевају узимање у обзир и других димензија података, као што су простор и време, тако да дводимензионалне матрице података могу бити проширене укључивањем додатних димензија.



Слика 9: Матрица података

Скупови података организовани у табеларном облику се складиште и чувају као једноставне датотеке, с тим што се матрица података може генерисати и из комплекснијих облика организовања података. Мада су датотеке, пре свега због своје једноставности, популарни облик за чување података, при њиховом одржавању, са повећањем броја записа (количине података), проблеми редундантности, неконзистентности и изолације података постају све израженији. Стога су повећање количине података, као и појава сложенијих структура у форми полуструктурираних и неструктурираних података условили изналажење нових облика организовања података у рачунарским меморијама, попут: база података, складишта података, напредних база података и *WWW*[20] база података.

Са становишта оперативног вођења пословања погодан облик организовања података у информационим системима организација су оперативне базе података засноване на релационом моделу података (који је утемељен на концепту претходно приказане табеларне форме организовања података). Генерално, концепт база података је еволуирао из концепта датотека података, тако да представља његову надоградњу.

[20] Акроним енглеске синтагме *Word Wide Web*.

Као добро организован скуп података, базе података се од стране корисника могу претражити, испитати и искористити за решење многих оперативних *DM* проблема.

Док традиционалне базе садрже оперативне податке, који репрезентују резултате свакодневних активности обезбеђујући подршку за тактичко одлучивање кроз постављање и извршење (структурираних) упита, складиште података је подршка при решавању, пре свега, неструктурираних и мултидимензионалних проблема, то јест, при доношењу дугорочних, стратегијских одлука за које је, како би квалитет одлучивања био задовољавајући, неопходно размотрити велику количину података са инкорпорираном временском димензијом. При томе се не сме занемарити чињеница да су оперативне базе један од извора података за складиште података.

Суштина филозофије складиштења података је интеграција података из различитих извора и квалитетна трансформација мултидимензионалних података у корисне информације. У ширем контексту схваћен, процес развоја складишта података обухвата процесе изградње и коришћења складишта података (*Cios et al*., 2007, стр. 106; *Han et al*., 2012, стр. 127). Изградња складишта података односи се на анализу извора података, припрему података и питања физичке изградње и стварања складишта, као и непосредног ускладиштења података. Ипак, кључни разлог развоја складишта податка је могућност коришћења, јер систем који нема кориснике (аналитичаре, менаџере и друге пословне кориснике) не вреди ни креирати (*Berry & Linoff*, 2004, стр. 492). Коришћење података обухвата следеће активности: постављање упита над подацима који се налазе у складишту података, анализу тих података, израду извештаја, проналажење непознатих и скривених законитости међу подацима и презентацију (укључујући и визуелизацију) добијених резултата. Наиме, валидна структура система складишта података, утемељена и подржана одређеним хардверским, софтверским и мрежним сиситемима, обезбеђује пренос података од њихових извора до крајњих корисника. У том контексту, *Berry & Linoff* (2004, стр. 486) истичу да су подаци попут воде, а ток података упоређују са током воде.

У процесу развоја складишта података припрема података за њихово коришћење је једна од најбитнијих компоненти. Реализује се кроз низ припремних процесних активности, које се у стручним круговима поједностављено називају *ETL* процеси, а односе се на процесе екстракције, трансформације и уноса (енгл. *Extract*, *Transform* и *Load*) података (*Panian & Klepac*, 2003, стр. 86). Проблеми у вези са припремом података у процесима складиштења су веома слични са проблемима припреме података за *DM*. Са становишта повезаности *DM*-а и складишта података, треба истаћи

да складиште података није услов за *DM* анализу. Наиме, *DM* систем може да функционише и на подацима организованим у форми датотека и оперативних база података. Међутим, многе припремне активности које се спроводе приликом увоза података из различитих извора у складиште, такође, претходе *DM* процесирању. Због тога, интегрисање складишта података и *DM* алгоритама доприноси поједностављењу *DM* процеса, пре свега, у домену претпроцесних активности и, на тај начин, обезбеђујући припремљене податке, олакшава реализацију *DM* задатака (нарочито у великим компанијама) (*Kantardžić*, 2011, стр. 19).

Подаци напредних (специјализованих) база података, као репозиторијума комплексних структура, полуструктурираних и неструктурираних података, су исказани у специфичним форматима, тако да је неопходно извршити њихово форматирање ради примене или класичних *DM* метода или специфичних метода и технологија креираних за такве типове података. Овој категорији припада велика група информационих репозиторијума, попут: трансакционих база [које се складиште као *flat* датотеке и садрже листу записа о реализованим трансакцијама (на пример, о обављеној куповини), при чему свака трансакција укључује јединствени идентификатор, скуп елемената и временску одредницу]; просторних база података [које, поред уобичајених, складиште и просторне податке (географске карте и сателитске снимке), односно информације о географском размештају и позицији, као додатне димензије у анализи јединица посматрања]; темпоралних база података [као репозиторијума секвенцијалних (уређених и индексираних) података, при чему су временске серије најбројнија класа секвенцијалних података, због чега се темпоралне базе често називају базе временских серија]; текстуалних база података [као репозиторијума великих количина текстуалних података]; мултимедијалних база података [складиштење података који укључују видео снимке, фотографије и аудио податке]. Посебну категорију база чини *WWW* база, односно огроман дистрибуиран репотизоријум података организованих у форми међусобно повезаних интернетских докумената.

Имајући у виду претходно речено, јасно је да валидна организација података представља један од кључних фактора за спровођење ефикасне *DM* анализе. Осим тога, организација података је са становишта времена и ресурса најскупљи сегмент процеса откривања законитости из података.[21]

[21] Развој одговарајућих организационих облика за прикупљање и складиштења података није предмет разматрања ове дисертације, али због изузетног значаја овог питања, као и учесталих потреба да се комбинују подаци из различитих извора, за даљу дискусију видети у: *Berry & Linoff* (2004); *Cios et al*. (2007); *Han et al*. (2012).

5.4. Квалитет података

Концепти квалитет, систем квалитета и менаџмент укупног квалитета (енгл. *Total Quality Management - TQM)* су већ дуже време присутни у академској заједници са врло успешним и афирмативним, али неретко и оспораваним, имплементацијама у пословној пракси. С обзиром чињеницу да је савремени свет одређен и вођен подацима, ови концепти су прихваћени и у домену података.

Квалитет представља задовољство корисника, односно погодност за употребу (енгл. *fitness for use*) (*Juran & Gryna,* 1993, стр. 3). Дакле, корисник је тај који одређује квалитет. Полазећи од ове опште познате дефиниције и њеним једноставним прилагођавањем на податке, *Wang & Strong* (1996, стр. 7) концептуализују основне аспекте и дефинишу квалитет података као „погодност података за употребу од стране корисника података". Међутим, значење појма квалитет се не односи само на перформансе, већ укључује и димензије квалитета. У том контексту, димензије квалитета података представљају скуп свих карактеристика података које репрезентују одређене аспекте и могућности података да задовоље исказане и подразумеване потребе, очекивања или захтеве корисника података (прилагођено према: *Janošević i drugi*, 1999, стр. 19).

У литератури из области квалитета података предложене су бројне карактеристике, односно димензије квалитета и њихове класификација према различитим критеријумима. Разликују се, у суштини, три основна приступа при одређивању свеобухватне листе димензија квалитета података: теоријски, емпиријски и интуитивни (*Batini & Scannapieca*, 2006, стр. 19-42). У Табели 1 су представљене најчешће коришћене димензије квалитета података које су утврђене на бази анкетирања корисника и експерименталних истраживања (емпиријски приступ).

На скупу димензија базира се мерење и оцена квалитета података. Избор димензија за мерење нивоа квалитета је почетна тачка било које активности повезане са квалитетом података (*Batini & Scannapieca*, 2006, стр. 12). Међутим, не постоје стриктна правила и најбоља решења за избор димензија. Стога, свака организација, полазећи примарно од свог пословања и конкретних циљева истраживања, али узимајући у обзир и усклађеност са дефинисаним и познатим пословним правилима, међународним стандардима, организационим уредбама и законским регулативама које се односе на одређене димензија квалитета података, самостално одређује структуру и начин мерења димензија. Како су многе димензије по природи вишезначне, у оквиру

сваке од њих, такође, морају бити јасно идентификоване и дефинисане варијабле као субдимензије које ће бити предмет евалуације. Избор специфичних варијабли је често знатно комплекснији проблем него начин мерења и дефинисање кореспондентних показатеља (*Lee et al*., 2006, стр. 55).

Табела 1: Листа димензија квалитета података

Димензија	Дефиниција: У којој мери су подаци ...
Доступност	доступни
Одговарајућа количина	одговарајући са становишта количине
Уверљивост	истинити и веродостојни
Потпуност	потпуни
Концизност	сажети, језгровити, компактни
Конзистентност	конзистентно презентирани (у истом формату)
Једноставност коришћења	једноставни за коришћење
Тачност	тачни (коректни и поуздани)
Интерпретабилност	јасно дефинисани и смислено приказани (језик, симболи, јединице)
Објективност	објективни / непристрасни
Релевантност	релевантни, примењиви, „од помоћи”
Репутација	респектабилни (сходно извору и садржају)
Сигурност	заштићени од неовлашћеног приступа
Правовременост	правовремени, ажурни
Разумљивост	разумљиви
Додата вредност	корисни и додају нову вредност

Извор: *Kahn et al*. (2002, стр. 187); *Pipino et al*. (2002, стр. 212)

Квалитет се увек процењује у односу на неку референтну вредност, а ниво квалитета изражава колико подаци испуњавају очекивања корисника или колико одговарају захтевима / спецификацијама квалитета. Један од првих приступа у оцени квалитета података уводе *Pipino et al*. (2002) и указују да се свака димензија може оценити на субјективан и мерити на објективан начин. Показатељи објективног мерења могу бити: ▸ једноставни односи (рацио форме) жељених и укупних исхода, ▸ агрегатне форме које се користе у случају представљања димензије као функције више варијабли, и ▸ пондерисани просеци варијабли.

Субјективно оцењивање се спроводи путем анкетирања при чему групе стејкхолдера, то јест, корисника исказују став, перцепције и своје мишљење о посматраној димензији квалитета, користећи, најчешће, *Likert*-ову скалу. Међутим, за потребе међусобног поређења различитих димензија, неопходно је показатеље субјективног оцењивања прилагодити (трансформисати) и исказати на истој скали као и показатељи објективног мерења (на пример, од 0 до 1).

Такође, исти аутори истичу да организације настоје да синтетизују и представе оцене свих димензија у форми једног агрегатног показатеља квалитета - индекса квалитета података[22], али уз напомену да при његовом коришћењу треба бити јако обазрив, нарочито услед потенцијалних проблема узрокованих субјективним одређивањем варијабли и система пондера, различитих типова мерних скала и слично.

Мерење квалитета служи да се оцени актуелни квалитет података и да се, након тога, дефинишу и предузму мере за његово побољшање. Заправо, квалитетом података је неопходно управљати. У том контексту, по аналогији са *TQM* концептом, чији је основни циљ базиран на непрекидном побољшању, развијен је и концепт менаџмента укупног квалитета података (енгл. *Total Data Quality Management - TDQM*).

TDQM је резултат примене *TQM* концепта на податке, имплементације идеје „погодност за употребу" и коришћења прилагођене варијанте *Deming*-овог циклуса побољшања квалитета. У основи, *TDQM*, као процес управљања квалитетом података, не сме бити сведен на пуко коришћење механизама за отклањање идентификованих недостатака, већ се побољшане верзије система и процеса (са пратећим процедурама) које спречавају настанак података лошег квалитета морају институционализовати. Стога, *TDQM* процес је и процес стицања знања за обезбеђење новог и вишег стандарда квалитета података.

ICT развој омогућио је организацијама да прикупе и складиште огромне количине података. Међутим, са повећањем количине података упоредо расте и ризик настанка проблема који су изазвани неквалитетним подацима, јер велика количина података не значи, по аутоматизму, и висок квалитет података. Заправо, присуство неквалитетних података је иманентно својство скоро сваког информационог система.[23] То је реална опција коју увек треба узети у обзир и сагледати последице њиховог присуства, које могу варирати од незнатних поремећаја у функционисању до великих финансијских губитака. Другим речима, пословање организације у великој мери зависи од квалитета података прикупљених и складиштених у постојећим информационим системима.

Негативне консеквенце лошег квалитета података су вишеструке. *Loshin* (2011) наводи четири групе пословних проблема и негативних утицаја који могу настати као последица квалитета података који је испод очекивања корисника:

- финанасијски утицаји: повећање оперативних трошкова, броја пропуштених прилика, казни и других накнада, као и смањење прихода и / или новчаног тока;

[22] Попут, индекса цена на мало или *Dow Jones Industrial Average* индекса у домену берзанског пословања.
[23] Неквалитетни подаци се често називају „прљави" или погрешни подаци (енгл. *„dirty" or erroneous data*).

- утицаји повезани са поверењем и сатисфакцијом: смањење сатисфакције интерних и екстерних стејкхолдера, смањење поверења у организацију, низак степен поверења у предвиђање будућих тенденција, неконзистентно оперативно управљање и извештавање, закаснела и неправилна реализација одлука;
- утицаји који се односе на продуктивност: повећање обима посла, повећање времена за обраду и смањење квалитета финалног производа;
- утицаји који су повезани са ризицима: лоше процене кредитне способности, инвестициони ризици (на пример, нетачан финансијски извештај може утицати да одлуку одличног инвеститора да одустане од пројекта), конкурентски ризици, преваре и губици, као и утицаји који се везују за усклађеност са законском регулативом.

Такође, лош квалитет података има негативан утицај и на ефикасност процеса одлучивања. Квалитет пословних одлука је директно пропорционалан квалитету информација које су изведене из података. Одлуке базиране на лошим подацима изазивају финансијске губитке, подривају лојалност клијената и угрожавају репутацију организација на тржишту. Истовремено, и сами доносиоци одлука, уколико су подаци сумњивог квалитета, губе поверење у њихову веродостојност, тако да у процесу одлучивања више користе своју интуицију (*Redman*, 2013). Осим тога, у ситуацијама када губе поверење у расположиве податке, доносиоци одлука и други корисници су спремни да одбаце добре иницијативе, пројекте и предлоге само зато што су засновани на подацима. Лош квалитет података узрокује повећање оперативних трошкове и кроз повећање времена и додатног ангажовања осталих ресурса за откривање и корекцију грешака у подацима (*Haug et al*., 2011, стр. 173). Менаџери и доносиоци одлука утроше скоро 50% свог времена за „трагање и лов" по подацима, идентификовање грешака, њихово исправљање и претрагу различитих извора података како би проверили достављене податке (*Redman*, 2013).

Сходно наведеном, јасно је да подаци представљају важан стратегијски ресурс у пословању савремених организација, који је, попут осталих ресурса, повезан са одређеним трошковима и користима. У релевантној литератури могу се наћи бројне класификације трошкова квалитета података. Уобичајено је да општа класификација трошкова обухвата три категорије: корективне, превентивне и трошкове побољшања квалитета података, које се даље могу рашчланити на различите начине. С друге стране, квалитетни подаци доприносе повећању сатисфакције корисника, прихода и профита, представљају вредну имовину и значајан извор за стицање конкурентске предности.

Међутим, тачну величину утицаја квалитета података на перформансе организационих јединица, ефикасност пословног одлучивања, свакодневне активности, трошкове пословања и остварене користи је изузетно тешко одредити. У процедуралном смислу, још увек не постоје јасно дефинисани методи и стандарди за мерење величине утицаја било које од наведених категорија утицаја квалитета података. Интересантно је истаћи да су истраживања квалитета података углавном усмерена на процену негативних импликација лошег квалитета, док су врло ретке студије које се баве проценом величине позитивних утицаја квалитетних података. На пример, *Larry English*, један од највећих ауторитета у области квалитета података, истиче да „...трошкови пословања настали услед рада са неквалитетним подацима, (укључујући неповратне-ненадокнадиве трошкове, трошкове дораде и преправке производа и услуга и изгубљени и пропуштени приход) могу износити чак 10% до 25% укупног прихода или укупног буџета организације" (*English,* 1999, стр. 12). Према *Redman*-у процењени износ трошкова лошег квалитета података је најмање 2% прихода, при чему нису укључени невидљиви губици организационе репутације и сатисфакције клијената (*Redman*, 2001, стр. 17).

Неспорно, већина организација препознаје огромне предности доброг квалитета података и признаје да економски ефекти малих недостатака могу бити веома значајни. Упркос тој чињеници, резултати студија, које су спроведене од стране индустријских експерата у оквиру истраживачких кућа *Gartner Group, Price Waterhouse Coopers* и *The Data Warehousing Institute*, јасно указују на кризу у управљању квалитетом података, као и на противљења доносилаца одлука на вишим нивоима да довољно пажње и рада посвете овом питању. *Marsh* (2005, стр. 106-107) је сумирао и у свом раду презентовао листу резултате оваквих истраживања. Ради илустрације озбиљности проблема у вези са квалитетом података, без дубље дискусије и критичког осврта на њихову валидност (пре свега, са аспекта потпуности исказа и међусобне компатибилности), овом приликом издвајају се неки резултати са поменуте листе:

- 88 % свих пројеката везаних за интеграцију података или пропадне или значајно надмаши предвиђене трошкове;
- 33% организација је отказало или одложило увођење нових *IT* система због лоших података;
- 611 милијарди долара је годишњи губитак америчких привредних субјеката због лошег слања поште и додатног ангажовања особља на решавању тих проблема;

- мање од 50% компанија је веома сигурно у квалитет расположивих података;
- свега 15% компанија је сигурно у квалитет достављених екстерних података;
- организације углавном прецењују квалитет својих података и потцењују трошкове грешака;
- огромна количина времена и новца се троши на „гашење пожара" и ублажавање постојеће кризе која постоји у погледу квалитета података, пре него на суочавање са проблемом и његово дугорочно решавање.

Упркос чињеници да су проблеми квалитета података прескупи да би били игнорисани, у пракси се често занемарују, тако да је реални квалитет података изузетно скроман (*Panian & Klepac*, 2003, стр. 30). Стога се као кључно питање поставља како обезбедити задовољавајући ниво квалитета података. Решење проблема сигурно није у: толерисању присуства погрешних података, прихватању чињенице да је квалитет података скроман, избегавању података при доношењу одлука, предузимању *ad hoc* (по идентификовању грешака) корективних мера од стране самих корисника података, или, пак, делегирању задатака запосленима у информатичком одељењу да открију узроке грешака и предузму мера да се грешке отклоне. Решење овог проблема захтева јасно дефинисање и стандардизовање процедура за мерење квалитета података и његових утицаја на реализацију пословних циљева, покретање једнократних пројеката за побољшање и, коначно (за трајно праћење и побољшање квалитета података), успостављање система за управљање квалитетом података, као саставне компоненте система за подршку одлучивању. Заправо, холистички и интегрисани приступ који омогућава дугорочно решење проблема квалитета података јесте проактивно управљање укупним квалитетом података. Његова оријентација базира се на процесном приступу, односно претпоставкама да је управљање квалитетом и његово побољшање континуирани и превентивни процес, а не само ретроспективна, корективна активност (*Marsh*, 2005, стр. 110).

Будући да се у модерном пословању подаци складиште у компјутерским базама података, постоји тенденција да се квалитет података третира као информатичко-технолошки проблем и да се, последично, одговорност за квалитет података пренесе искључиво на *IT* запослене (који и нису власници пословних процеса који производе податке). Међутим, квалитет података није питање које се односи само на складиштење, компјутерску обраду и софтверска решења. Из перспективе квалитета, у животом циклусу података постоје два критична тренутка: тренутак настајања и

тренутак ангажовања, односно употребе података (*Redman*, 2013). При томе, квалитет података је детерминисан у тренутку стварања података, док се питање процене квалитета података поставља у тренутку употребе података преко екстрахованих информација приликом реализације пословних циљева. Управо због чињенице да је квалитет података одређен у тренутку креирања података (на пример, у одељењу за истраживање тржишта), исти аутор, такође, наводи да успех у управљању и решавању проблема квалитета од стране *IT* сектора често изостане. Стога, један од најједноставнијих начина за подршку проактивном управљање укупним квалитетом је реформа међусобних односа и успостављање боље комуникације између креатора и корисника података (независно од њихове организационе повезаности), преношење одговорности за квалитет података са *IT* особља на линију менаџера и тимски рад *IT* особља, креатора и корисника података по свим питањима квалитета података.

Према томе, концепт квалитета податка се односи на широку лепезу питања, која указују на техничке и нетехничке аспекте квалитета података. Одговори на питања захтевају истовремено сагледавање проблема из различитих углова и интегрисање резултата који су из таквог приступа проистекли. Уосталом, и из истраживачке перспективе, квалитет података се, као мултидисциплинарни проблем, разматра у различитим научним подручјима, укључујући, пре свега, статистику, менаџмент (менаџмент квалитета) и информатику.

6. DATA MINING ЗАДАЦИ И МЕТОДИ

Основни циљ *DM* методологије јесте откривање скривених законитости из података, које корисницима могу помоћи у разумевању конкретних појава, процеса и проблема. Остварење овог циља се заснива на примени одговарајућих метода и конструкцији (дескриптивних и / или предиктивних) модела реалних система базираних на релевантним и расположивим подацима који потичу из различитих, интерних и екстерних, извора. Сходно томе, у овом Поглављу су разматрани општи аспекти и класификације *DM* задатака, метода и модела и апострофирана њихова међузависност у постизању *DM* циљева.

6. 1. Дефинисање и класификација *data mining* задатака

Примарна сврха *DM*-а је откривање законитости, односно тражење карактеристичних (глобалних и локалних) структура у постојећим подацима које су у форми знања усмерене на решавање реалних проблема. У основи, циљеви откривања

знања и *DM*-а могу се класификовати у следеће две категорије (*Fayyad et all*., 1996, стр. 43): ▸ верификација, од стране аналитичара, дефинисаних хипотеза о значају релација међу одређеним подацима и конфирмација или не њихове утемељености и оправданости, и ▸ откривање нових законитости (сазнања), које човек, као актер у процесу откривања знања, не би могао, нити својим искуством нити интелектуалним способностима, да открије и докучи без адекватне компјутерске подршке.

DM циљеви се остварују путем *DM* задатака. Суштински, *DM* задаци су задаци анализе података усмерени на откривање законитости из великих скупова података за решавање реалних проблема. Другим речима, под *DM* задацима подразумевају се реалне (пословне) проблемске ситуације преточене у *DM* контекст до чијих се решења долази применом бројних метода на различитим типовима података. Сходно диверсификованости реалних проблемских ситуација, у релевантним изворима могу се пронаћи различите класификације *DM* задатака. На пример, *Hand et all*. (2001, стр. 12) издвајају следеће типове *DM* задатака:

- Експлоративна анализа података: У питању је једноставна претрага података без претходно јасно одређене идеја шта се конкретно тражи. Ова анализа, статистичким речником, може се окарактерисати као подацима вођено генерисање хипотеза, насупрот процедури тестирања статистичких хипотеза која се базира на адекватним теоријским основама. Технике експлоративне анализе података су интерактивне и визуелне.
- Дескриптивно моделирање: Има за циљ да опише све податке или процесе који су повезани са генерисањем података. Дескриптивно моделирање резултира моделима за оцену функције дистрибуције података, поделу мултидимензионалног простора на групе и опис веза између варијабли.
- Предиктивно моделирање: Омогућава да се одреде / предвиде вредности једне категоријске или нумеричке променљиве на основу познатих вредности осталих променљивих.
- Откривање локалних структура (правила): У овој групи налазе се задаци који имају за циљ откривање необичних и нестандардних образаца понашања (укључујући и идентификовање тачака података које се значајно разликују од осталих, а које се, статистички речено, називају екстремне вредности) или правила која указују на честе комбинације ентитета у скупу података (најчешће је реч о трансакционим базама података и идентификовању правила повезивања).

- Проналажење по садржају (енгл. *retrieval by content*): Код ове групе задатака истраживач / корисник претражује одређени садржај, који је најчешће у форми текста или слике. На пример, путем кључних речи, корисник претражује огромне количине информација које су садржане у електронским часописима, каталозима или се налазе на приватним и комерцијалним сајтовима.

У складу са типом законитости коју генерише *DM* систем, општеприхваћена подела *DM* задатака моделирања је на дескриптивне и предиктивне. Путем дескриптивних задатака генеришу се законитости о својствима података и релацијама између њих у дефинисаном скупу, док се путем предиктивних задатака на бази постојећих података (то јест, слогова са познатим вредностима) генеришу законитости за потребе предвиђања вредности зависних варијабли или будућег понашања одређених ентитета. Дубљом класификацијом ове две групе задатака утврђена је листа примарних *DM* задатака (*Fayyad et al.*, 1996; *Berry & Linoff*, 2004; *Larose*, 2005; *Kantardžić*, 2011), који се у наставку текста укратко представљају.

Сумаризација, као дескриптивни задатак, односи се, пре свега, на иницијалну експлоративну анализу података. Углавном се спроводи у раним фазама реализације *DM* пројекта применом метода дескриптивне статистичке анализе и метода визуелизације распореда података. Путем сумарних информација, овај задатак обезбеђује сажет преглед и повећава степен разумевања карактеристика и структуре скупа података релевантних за конкретни проблем. Кроз различите комбинације нивоа сажимања и димензија, сумаризација резултира једноставним компактним описима скупа (или делова скупа) података из различитих перспектива, који се користе као њихови репрезенти. Како обезбеђује законитости које омогућавају стицање општег увида у податке, сумаризација се назива и апстракција или генерализација.

Класификација је задатак који подразумева креирање класификационог модела за позиционирање сваког новог елемента, сходно његовим карактеристикама, у једну од две или више претходно дефинисаних класа зависне варијабле. Кључне карактеристике класификационог задатка су: ▸ припада категорији предиктивног моделирања, ▸ зависна варијабла је категоријска, и ▸ генерализацијом познате структуре у форми класификационог модела, нови податак се класификује, као функција улазних варијабли, у једну од претходно дефинисаних класа.

Регресија је *DM* задатак који се, насупрот класификацији, односи на креирање предиктивног модела за зависну нумеричку варијаблу, при чему је код логистичке регресије реч о категоријској зависној варијабли. Циљ регресионе анализе је да се, кроз

(надгледани) процес учења, одреди регресиони модел који најбоље описује везу између унапред идентификоване зависне променљиве и једне или више објашњавајућих променљивих (регресора), а затим да се на основу тога модела оцене и предвиде вредности зависне варијабле за одабране вредности објашњавајућих варијабли.

Предвиђање је још један важан предиктивни *DM* задатак. Заснива се на коришћењу модела, креираном на бази познатих података, за предвиђање будућег понашања и будућих вредности посматраних карактеристика са што мањом грешком. При томе је могуће предвидети ознаку класе новог елемента, предвидети интервалне вредности нумеричких карактеристика, одредити будуће вредности у функцији времена итд. Сходно наведеном, читав низ модела може се прилагодити за употребу у функцији реализације овог задатка, као што су класификациони модели, регресиони модели, модели временских серија, модели базирани на мишљењу експерата и слично.

Анализа временских серија је задатак повезан са откривањем корисних законитости у структури временских серија и, сходно томе, креирањем модела за прогнозирање будућих вредности посматране појаве.[24] Обзиром да је велика количина расположивих података данас представљена у форми временских серија и да, последично, многи *DM* проблеми укључују временску димензију, анализа временских серија, као *DM* задатак, добија све више на значају.

Груписање је *DM* задатак који се односи на класификацију елементарних јединица анализе (објеката, ентитета или случајева) у две или више група, с обзиром на њихову сличност, односно различитост према низу посматраних карактеристика (обележја). За разлику од класификације, код анализе груписања не постоји претходно знање о структури елемената и припадности било којег елемента некој групи. Такође, ниједна посматрана карактеристика нема својство зависне варијабле, већ се све улазне карактеристике равноправно третирају. Самим тим у питању је индуктивни процес учења који не захтева иницијалне податке за вођење процеса учења. Кључне карактеристике задатка груписања су: ▸ припада категорији дескриптивног моделирања, ▸ може се односити на различите типове података (нумеричке, текстуалне), и ▸ групе нису унапред дефинисане, већ се формирају из података. Интересантно је истаћи да се специјална примена анализе груписања односи на идентификовање елемената чије вредности посматраних карактеристика знатно одступају од уобичајених вредности, што чини суштину претпроцесног задатка откривања нестандардних вредности.

[24] У контексту временских серија уместо термина предвиђање користи се термин прогнозирање.

Идентификовање фреквентних образаца (енгл. *frequent patterns*) је још један веома значајан и популаран задатак у *DM* истраживањима, који се односи на откривање комбинација и веза између података које се често појављују заједно (*Han et al*., 2012, стр. 243). Најчешћа форма овог задатка је откривање правила повезивања (или, асоцијативних правила), која резултирају „ако-онда" тврдњама са одговарајућом вероватноћом о симултаном појављивању одређених комбинација елемената.

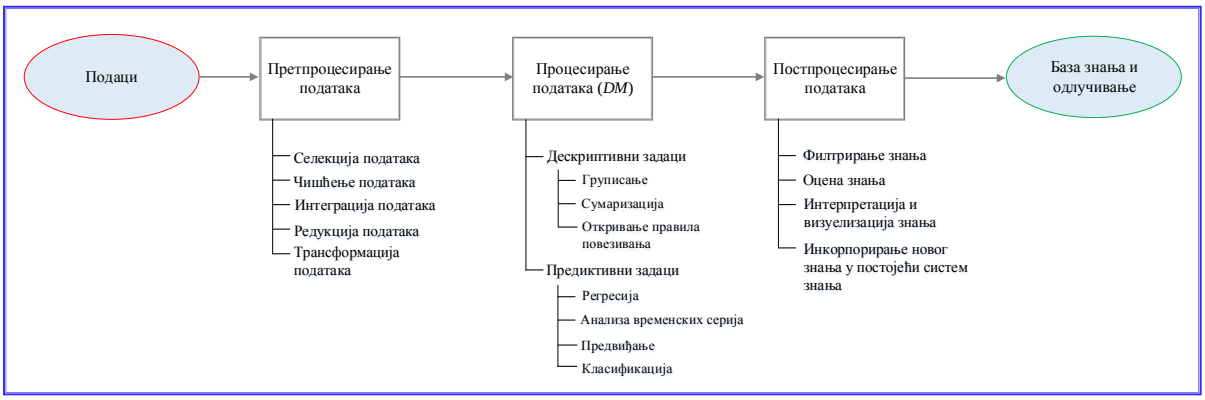
Анализа редоследа (секвенцијална анализа или откривање секвенци) се користи за идентификовање секвенцијалних правила у подацима који зависе од редоследа појављивања елемената. Секвенце се дефинишу као уређени низ елемената (или догађаја), који могу бити различитог типа и могу се срести готово у свим областима.[25] Анализа редоследа припада групи дескриптивних *DM* задатака, с тим што се идентификована секвенцијална правила могу успешно користити и за предвиђање наредних (будућих) елемената у низу.

Наведени задаци представљају уобичајене спецификације општих *DM* проблема који се могу односити на различита подручја примене. Свакако, наведена листа није коначна. Са интензивнијом употребом *DM* технологије појављују се нови проблеми и шири лепеза задатака. Осим тога, укључивање *DM*-а у нова апликативна подручја довело је до појављивања нових *DM* задатака који се односе на специфичне проблеме у домену конкретног подручја.

Међутим, *DM* представља једну фазу у процесу откривања знања из података, тако да су и друге фазе од великог значаја за успешно екстраховање знања. Стога је могуће говорити не само о *DM* задацима, већ о задацима укупног *KDD* процеса. Наиме, будући да се у свакој фази *KDD* процеса акценат ставља на различите аспекте анализираног проблема, а имајући у виду значај квалитета података и правилне интерпретације добијених резултата моделирања, сходно потребама овог истраживања, извршена је класификација задатака откривања знања у следеће три групе: ▸ задаци претпроцесирања података, ▸ задаци процесирања података (или задаци моделирања или *DM* задаци), и ▸ задаци постпроцесирања података, односно изведених законитости. Хијерархијски, свака наведена група садржи читав низ специфичних

[25] Разликују се два облика секвенци: категоријске (дискретне) и темпоралне. Категоријске секвенце се односе на симболе (и стање) и не укључују временске аспекте посматрања релација између елемената. За разлику од њих, темпоралне секвенце су засноване на временском редоследу различитих категорија (симбола и стања), активности и догађаја. Неки примери секвенцијалних података су: подаци о (хронолошком) редоследу куповине артикала, подаци о ефектима медицинских третмана, подаци повезани са природним катастрофама (на пример, подаци о јачини серије земљотреса), подаци о дневним температурама и нивоу угљен-моноксида у атмосфери, подаци о телефонским позивима, подаци о ценама акција и токовима на финансијском тржишту (берзанске секвенце), подаци о саобраћају и посетама на *веб* страницама (*web log* анализа секвенци - редоследа кликова током корисничких сесија), подаци о структури гена и слично.

задатака на наредном нивоу (Слика 10), при чему је њихова реализација заснована на примени различитих методолошких поступака.



Слика 10: Задаци откривања знања из података

Претпроцесирање података је од круцијалног значаја за корисност и валидност закључака у контексту дефнисаних циљева истраживања. Подаци апсорбовани из различитих интерних и екстерних извора су обично непотпуни, несистематични и неконзистентни (са аспекта математичке и логичке исправности). У циљу отклањања или редукције недостатака изворних података, а полазећи од чињенице да је квалитет података пресудан фактор за успешну анализу, пре процесирања података, неопходно је спровести бројне задатке претпроцесирања. Ови задаци се односе на анализу сирових података и обухватају активности попут прикупљања, чишћења, редукције и трансформације података. Задаци процесирања се односе на примену алгоритмизираних метода у циљу екстракције законитости. Резултати, који су добијени путем индуктивног *DM* процеса и изабраних, сходно циљевима истраживања и типовима анализираних података, алгоритама, често нису погодни са становишта комерцијалних апликација и њихове употребе од стране крајњих корисника. Услед тога, добијени резултати, по правилу, морају бити накнадно обрађени, односно постпроцесирани. Поступци постпроцесирања представљају неку врсту „филтрирања" издвојених законитости у циљу елиминације шумова, непрецизности и корисничког неразумевања (*Bruha & Famili*, 2000). Стога, постпроцесирање укључује задатке као што су филтрирање, евалуација, интерпретација, инкорпорирање новог знања у постојећи систем знања и слично. На тај начин заокружује се целокупан процес анализе података у функцији извођења законитости из података, то јест, откривања корисног знања. При томе, постизање софистицираних резултата у решавању реалних проблема заснива се на успешном комбиновању и спровођењу појединачних задатака.

6.2. Класификација *data mining* метода и проблем њиховог избора

Циљеви откривања знања из података преточени у *KDD* задатке реализују се коришћењем бројних метода, који своје порекло вуку из различитих дисциплина. Отуда, у методолошком смислу, основна карактеристика анализе података у *DM* контексту садржана је у комбинацији широке апаратуре, од математичких метода, класичне статистичке анализе и кибернетских метода до метода машинског учења и вештачке интелигенције, укључујући и последња достигнућа у области информационих технологија.

Упркос мултидисциплинарном пореклу, ипак се може говорити о општеприхваћеном скупу метода који су у стручној литератури означени као *DM* методи. При томе, велики број тих метода је развијен пре него што је и употребљена синтагма *data mining*, схваћена као скуп метода и поступака који имају за циљ откривање законитости у маси података (*Panian & Klepac*, 2003, стр. 247-248). Осим тога, као последица интензивне истраживачке активности, популарности и проширења домена примене *DM*-а долази до побољшања постојећих, али и развоја нових метода за откривање законитости у подацима.

Међутим, треба истаћи да је готово немогуће повући јасну границу између *DM* метода и „осталих" метода. На пример, многи методи које се користе у скоро свим фазама процеса откривања знања су у основи статистички методи. Свакако да коришћење статистичких метода у *DM* окружењу не умањује њихова статистичка својства нити их декларише као искључиво *DM* методе. Наведно не представља недостатак, већ указује на адаптибилну природу овог подучја и могућност избора и коришћења бројних метода корисних са становишта остварења циљева анализе.

Постоји више начина и критеријума за класификацију *DM* метода, и то: *DM* циљеви, *DM* задаци, област којој метод изворно припада (на пример, из домена статистике и машинског учења), извори и начин организовања података (методи за анализу података релационих, темпоралних и других форми база података).

Једна од најчешће презентованих подела је са становишта циљева *DM* анализе, при чему се разликују следеће две групе метода (*Maimon & Rokach*, 2010, стр. 5-6):

- верификационо оријентисани методи: усмерени су на проверу, од стране корисника дефинисаних, хипотеза о значајности односа међу одређеним подацима и укључују најчешће методе традиционалне статистике, као што су тестирање хипотеза о параметрима и моделима расподела вероватноћа и анализа варијансе; и

- методи откривања иновативног знања, то јест, дескриптивно и предиктивно оријентисани методи. (Будући да је примарни циљ *DM*-а откривање новог знања, пре него тестирање хипотеза, у фокусу даљих разматрања је углавном ова група метода.)

Као и *DM* задаци, *DM* методи оријентисани ка откривању нових знања, деле се на предиктивне и дескриптивне. Циљ дескриптивних метода је разумевање, (логичко) објашњење и интерпретација (група) података, док су предиктивни методи оријентисани на креирање модела путем којег се на основу комбинација улазних варијабли доносе закључци о излазној варијабли. Често примена дескриптивних метода претходи предиктивном моделирању. У литературе из машинског учења уобичајено се за дескриптивне методе користи термин ненадгледано, а за предиктивне надгледано учење. У ненадгледаном моделирању (енгл. *unsupervised learning or modeling*) без издвајања зависне варијабле откривају се извесне опште релације између посматраних варијабли. За разлику од овог приступа, у случају надгледаног моделирања (енгл. *supervised learning or modeling*) увек постоји зависна варијабла која се класификује, оцењује или предвиђа, односно објашњава путем других варијабли или прати њена еволуција током времена.

Предиктивни методи, генерално, односе се на предиктивне и класификационе задатке. И поред јасне границе између класификације и предикције (предвиђања[26]), ипак, при употреби настаје извесна терминолошка конфузија и неразумевање суштинске разлике. Стога се поставља питање у чему је разлика између класификационог и предиктивног типа предиктивног *DM* проблема. Кључна разлика је у природи излазне (зависне) варијабле: ако је иста квалитативна, у питању је класификација, док је у случају нумеричке варијабле реч о (нумеричкој) предикцији. Ако је резултат предиктивног моделирања модел којим се детерминише припадност одређеној класи, тада се предиктивни проблем назива класификација. С друге стране, ако је резултат предиктивног моделирања нумеричка вредност зависне варијабле, тада се предиктивни проблем назива нумеричка предикција (или регресија). Другим речима, предвиђање класе је класификација, а предвиђање нумеричке вредности је предикција (*Han et al.*, 2012, стр. 328). Иако постоји читав низ метода за сврхе нумеричке предикције, будући да се најчешће користи статистичка регресиона анализа, постоји тенденција да се термини регресија и нумеричка предикција користе као синоними.

[26] Термин предвиђање (предикција) вредности зависне варијабле у функцији једне или више улазних варијабли треба разликовати од термина предвиђања (прогнозе) у контексту временских серија и кретања појава у времену (енгл. *prediction vs forecasting*). Обе категорије припадају предиктивном *DM* моделирању.

Терминолошку и семантичку испреплетаност класификације и предикције потврђује и одговор на постављено питање који је дао *DM* конуслтант *Piatetsky-Shapiro*: ако се класификују постојећи подаци, односно јединице посматрања (објекти) за које су већ познате вредности излазне варијабле, тада је реч о класификацији, а уколико се класификациони модел користи за нове податке, то јест одређивање категорија нових јединица посматрања, тада се говори о предикцији (*http://www.kdnuggets.com/faq/classification-vs-prediction.html*).

Практично, када је у питању класификациони тип предиктивног проблема, задатак је да се, на основу карактеристика објеката чија је припадност категоријама зависне варијабле унапред позната (на пример, подносиоци захтева за кредит, класификују се у категорије клијената са ниским или високим кредитним ризиком), применом одговарајућег метода формира модел и дефинише скуп правила на основу којих ће се вршити класификација нових објеката или одређивати вероватноћа будућих догађаја (на пример, вероватноћа преласка клијената у конкурентске компаније). У случају предиктивног типа проблема, на основу одређене констелације улазних варијабли формира се модел и одређује не класа, већ нумеричка вредност излазне варијабле (на пример, цена апртмана за летовање као функција више улазних варијабли). Модели који се формирају при решавању првог типа проблема називају се класификатори, а могу имати функцију разврставања објеката или предвиђања категорије нових података. Резултирајући модели при решавању предиктивног типа проблема називају се предиктори.

Групи класификационих метода припадају бројни методи, као што су: методи засновани на стаблу одлучивања, методи засновани на неуронским мрежама, методи засновани на *Bayes*-овој логици, методи засновани на подржавајућим векторима (енгл. *support vector machine*), методи засновани на одстојању, попут *K*-најближих суседа (енгл. *K-nearest neighbors*) итд. За решавање предиктивног типа проблема предвиђања на располагању је богата палета статистичких метода који се могу применити, али примарно место припада методима регресионе анализе. У зависности од типа варијабле која се предвиђа (у статистичкој терминологији та варијабла се назива зависна), као и типова варијабли на основу којих се врши предвиђање (називају се независне варијабле, објашњавајуће варијабле, предиктори или регресори) разликују се линеарна, нелинеарна и логистичка регресија. У случајевима када је појава која се предвиђа дискретна, а предиктор варијабле непрекидне, дискретне или њихова комбинација, примењује се логистичка регресија. Такође, логистичка регресија, се

често користи и за решавање *DM* проблема класификације. Коначно, групи предиктивних метода, суштински припада и специфична група *DM* метода за анализу појава са укљученом временском димензијом (анализа временских серија).

Група дескриптивних метода се односе на редукцију, сумаризацију, груписање и визуализацију података. Њихово основно својство је одсуство зависне варијабле. За разлику од предиктивних модела, где није примарно познавање начина одређивања параметара и функционисања модела, већ, пре свега, добијени резултати у погледу тачности модела на тестним подацима, код дескриптивних модела веома је важно познавати начин на који систем идентификује законитости како би се обезбедила њихова логичка интерпретација. У најпознатије дескриптивне методе убрајају се анализа груписања и асоцијативна анализа.

У складу са извршеном поделом *KDD* задатака према фазама процесног приступа у откривању знања, разликују се: ▸ методи за реализацију задатака претпроцесирања, ▸ методи за реализацију задатака (предиктивног и дескриптивног) моделирања (процесирања), и ▸ методи за оцену и селекцију модела генерисаних из података (односно за постпроцесирање), при чему у оквиру сваке групе постоји мноштво методолошких поступака за спровођење серије различитих задатака. Наведена класификација, иако оквирна, јасно указује на разноврсност метода откривања знања, комплексност анализе података у *DM* окружењу и потребу за комбинованим коришћењем више метода.

Избор адекватног метода је један од најтежих проблема у целом процесу откривања знања из података, јер се често одређени задатак може решити помоћу неколико различитих метода, као што и један метод може бити употребљен за реализацију више задатака. Наиме, не постоји усаглашени приступ у погледу спровођења конкретног *DM* задатка по одређеном поступку. Осим тога, развијени софтверски алати отварају велике могућности комбиновања бројних метода што додатно захтева широк распон стручног знања о својствима метода у циљу сагледавања могућности за њихову валидну примену при решавању конкретног проблема. Заправо, сваки метод наспрам конкретне проблемске ситуације има своје предности и недостатке. Стога, основни критеријум избора треба да буде однос основних карактеристика посматраног проблема (на пример, дескриптивни или предиктивни тип проблема, ниво неизвесности који се везује за проблем и слично) према специфичним карактеристикама различитих метода. Услед свега наведеног, а имајући у виду утицај примењеног метода на квалитет резултата моделирања, питању

избора и евентуалне ревизије прелиминарно изабраног метода треба студиозно приступити.

У начелу, коначан избор одређеног метода (или комбинације метода) за конкретни реални проблем детерминишу следећи параметри (*Gibert et al*., 2010; *Klepac & Mršić,* 2006, стр. 31): ▸ главни циљ проблема који треба решити, ▸ структура расположивих података, и ▸ преференције, вештине и знања *DM* истраживача.

Претходна разматрања експлицитно показују да је мултидисциплинарна природа *DM*-а, подржана *IT* прогресом, резултирала развојем богате методолошке апаратуре расположиве доносиоцима одлука за извођење знања из великих скупова података. Истовремено, бројност метода изнедрила је питање правог избора и правилне употребе одређених метода у решавању конкретног проблема током свих фаза итеративног процеса откривања знања. Претпоставка квалитетне анализе податка засноване на *DM* приступу је познавање карактеристика самих метода (услова коришћења, предности и недостатака) и разматрање њихове апликативности у констелацији са карактеристикама конкретног проблема, што захтева непрекидну сарадњу стручњака из области у којој се метод примењује и аналитичара података који користи *DM* методе. У супротом, знање екстраховано из података доводи до погрешних закључака.

6.3. Алгоритми и софтверски пакети за креирање *data mining* модела

Сваки метод, као начин решавања проблема, карактеришу јасна и прецизна одређења у истраживању феномена теоријског, практичног и управљачког карактера. Сходно томе, примена сваког *DM* метода се базира на специфичним и добро дефинисаним процедурама које се називају алгоритми. У том смислу, *DM* је алгоритамски заснован процес решавања проблема, састављен од низа уређених активности (корака) путем којих се подаци трансформишу у одговарајуће резултате, у форми локалних структура и глобалних модела.

Концепт алгоритма се појавио знатно пре рачунара, али, компјутерски подржани алгоритми су данас основа за решавање многих практичних и теоријских проблема у различитим сферама људског деловања. Такође, већина алгоритама који се користе у оквиру *DM*-а су већ познате (математичке, статистичке и друге) процедуре чија је могућност употребе у решавању конкретних задатака путем анализе велике количине података знатно повећана са појавом рачунара. Својство аутоматског откривања законитости, које је споменуто при појмовном дефинисању *DM*-а, управо се односи на софтверску алгоритмизацију метода.

Спецификација *DM* алгоритма, као механизма за креирање *DM* модела, подразумева дефинисање његове структуре. *Hand et al.* (2001) идентификују следеће структурне компоненте *DM* алгоритма:

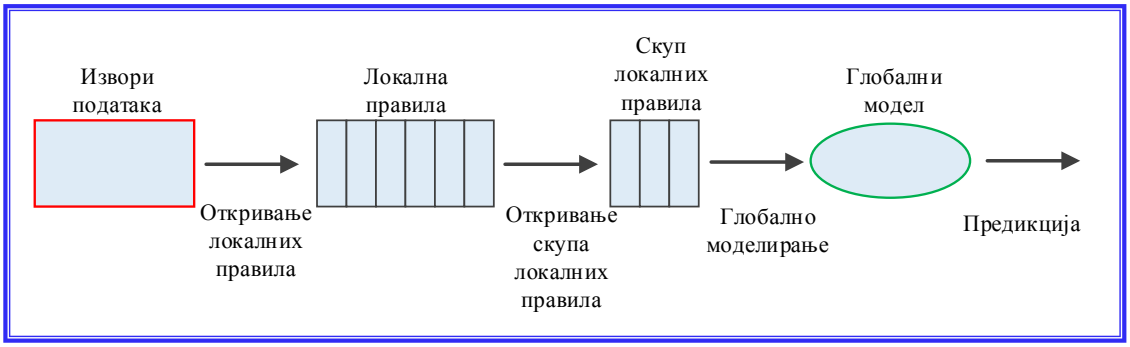
1) *DM* задатак: Потпуно је природно да су за решавање различитих типова *DM* задатака потребни различити типови алгоритама.

2) Модели и локалне структуре података: Ова компонента се односи на утврђивање основне структуре или функционалних форми података, при чему се примењени методи моделирања битно разликују по коначном облику резултата, односно облику глобалних модела и локалних образаца. Нема сумње да је корисно направити разлику између ове две категорије. Често је у пракси граница недовољно јасна и прилично произвољна, тако да није упутно, а ни потребно увек исту истицати. О интерпретацији концепата „модел и образац" већ је дискутовано у Потпоглављу 2.1.

Скуп локалних образаца обезбеђује глобалне информације о подацима. Услед наведеног, обрасци се често користе за изградњу корисног глобалног модела, као очекиваног резултата *DM* процеса. *Knobbe et al.* (2008) су посветили посебну пажњу начину конверзије скупа образаца у глобални модел. Они су дефинисали оквир процеса за изградњу глобалних модела на бази локалних образаца (Слика 11), који је назван „Од локалних образаца до глобалних модела", идентификујући три фазе у том процесу: ▸ откривање локалних образаца, ▸ откривање скупа локалних образаца, и ▸ креирање глобалног модела. У суштини, полазећи од извора података и њихове припреме за *DM,* прва фаза се везује за примену експлоративне анализе података на ограничене регионе података и, у односу на дефинисани циљ, откривање скупа потенцијално корисних образаца кандидата. У другој фази се оцењују претходно откривени скупови образаца и бира компактан скуп релевантних образаца. У трећој фази се изабрани скуп конвертује у добро избалансирани модел за решење конкретног проблема и остварење дефинисаног циља.

Наглашавање термина „глобални и локални" указује на чињеницу да се глобални модел односи на свеобухватну структуру или већину случајева посматрања, а путем локалних образаца снимају се само одређени аспекти података и идентификују подскупови података са поседованим карактеристикама које их јасно издвајају од осталих комплемената. Заправо, модел је генерална карактеристика (законитост) већег дела (или свих) података, а образац се дефинише као локална карактеристика (законитост) која важе за одређени, мали део података. Сходно томе, обрасци су локалне законитости, односно представљају локалне моделе. Из наведеног произлази

круцијална релација: глобални модел је скуп локалних модела, односно, локални модели су компоненте глобалног модела.



Слика 11: Процес развоја глобалних модела на бази локалних образаца

Извор: *Knobbe et al.* (2008)

Полазећи од претходних разматрања, у наставку овог истраживања користиће се термин модел (у ширем и ужем смислу), као репрезент глобалних или локалних карактеристика података, то јест, законитости већег или мањег дела података, респективно.

3) Циљна функција (енгл. *score function*): Реч је о интерној функцији вредновања апроксимације (модела) која има за циљ да оцени и квантификује квалитет предложеног модела са становишта прилагођавања датим подацима. Циљна функција треба да одражава корисност модела. Међутим, како је то у пракси тешко остварити, циљна функција има одређено значење које се односи на грешке / ризике / трошкове. Циљне функције се разликује од метода до метода и могу укључивати вероватноћу, суму квадрата грешака, стопу погрешне класификације и слично.

4) Методи претраживања, односно оптимизације: Методи претраживања односе се на претраживање простора решења и проналажење најбољих модела тако да се максимизира или минимизира циљна функција. Процедура претраживања је кључна компонента *DM* алгоритма, јер за дати облик модела функционише као оптимизацијски алгоритам који се одликује хеуристичким приступом претраживању, комплексношћу претраживања и контролом процеса претраживања путем критеријума заустављања.

5) Стратегија управљања подацима: Ова компонента *DM* алгоритма се односи на бирање стратегије управљања подацима којом се одређују начини за чување, индексирање, организацију и приступ подацима.

На први поглед наведене структурне компоненете *DM* алгоритма изгледају једноставно. Међутим, са становишта међусобног односа и начина на који су поједине

компоненте уобличене, њихове импликације су дубоке. Наиме, комбиновањем различитих приказа модела са различитим циљним функцијама, различитим методима претраживања и различитим техникама за управљање подацима могуће је генерисати потенцијално бесконачан број алгоритама, од модификације постојећих до креирања потпуно нових идејних решења. Имајући у виду разлике у својствима алгоритмизираних метода која настају услед различитих комбинација структурнх компоненти, последично, један од основних проблема на који се наилази у пракси је избор одговарајућег алгоритма у конкретној ситуацији.

Развој и примена *DM* алгоритама подразумева коришћење моћних софтверских алата, који морају бити дизајнирани тако да подрже итеративну природу *DM* процеса. Постоји велики број компанија које нуде софтверска решења за *DM* и, сходно томе, читав спектар *DM* алата. *Kantardžič* (2011, стр. 480-495) и *Mikut & Reischal* (2011) су приказали исцрпне, али не и кончне, листе комерцијалних и јавно доступних (отворених) алата, уз представљање битних информација о њима, попут основних својстава, *web link*-а и *web site* адреса издавача, продаваца и консултантских компанија. Истовремено, у оба наведена извора је указано да су промене на тржишту ових алата врло честе и да се непрекидно појављују нови софтверски производи. Такође, за праћење (ажурираних) информација о употреби постојећих и појави нових алата, сугерисани су следећи Интернет извори: *www.kdnuggets.com*, *www.knowledgestorm.com* и *www.the-data-mine.com*.

Компарацију и категоризацију *DM* софтверских решења могуће је извршити према различитим критеријумима. Неки од тих критеријума су (*Mikut & Reischal*, 2011, стр. 433-436):

- корисничке групе (алати могу бити развијени сходно потребама различитих корисничких група, тако да постоје алати за пословне апликације, примењена истраживања, едукацију, развој нових и модификацију постојећих алгоритама);
- структура података (димензионалност различитих типова података захтева различите алате, од структурираних података у форми дводимензионалних табела које су подржане у готово свим постојећим алатима, преко временских серија, до примене *DM* метода и развоја софтвера за рад са екстремно великим и високо димензионалним просторним и мултимедијалним подацима);
- *DM* задаци и методи (за реализацију *DM* задатака и, генерално, задатака откривања знања из података постоје бројни методи, али, логично, фреквенција са којом су методи инкорпорирани у различите софтверске алате варира, тако да се

разликују: методи садржани у скоро свим алатима (углавном методи класичне статистичке анализе), методи који се појављују у већини алата (стабло одлучивања, анализа груписања, неуронске мреже, факторска анализа и методолошки поступци за трансформацију података и унакрсну валидацију), методи који су расположиви у неким алатима (попут, асоцијативне анализе, метода *K* најближих суседа и *Bayes*-ових мрежа) и ретко расположиви методи (метод случајних шума и генетски алгоритам);

- могућност извоза и увоза података и модела између различитих софтверских алата (имајући у виду различите формате података и бројност извора за генерисање података, важно питање са аспекта функционалности *DM* алата је развој компоненти и стандарда за повезивање софтверских решења);
- визуелизација и стил интеракције између корисника и алата (при чему се разликују алати који имају чисто текстуални *interface* уз коришћење програмског језика, графички *interface* са одређеном структуром менија и графички кориснички *interface* где корисници бирају алгоритме из палете за избор, дефинишу параметре и креирају цео *DM* ток);
- тип платформе (самостална или клијент / сервер решења);
- модели лиценцирања (сходно различитим верзијама доступности алата говори се о комерцијалним софтверима и софтверима отвореног кода).

Не разматрајући дубље ову систематизацију алата, у наставку следи кратак осврт на популарне софтверске пакете, као колекције *DM* алгоритама.

Три најчешће коришћена *DM* алата на данашњем тржишту која припадају групи статистичких софтверских пакета и који поред стандардних укључују и низ посебних програмских модула за *DM* су: *IBM SPSS Modeler*[27], *SAS-Enterprise Miner* (*SAS-EM*) и *Statistica Data Miner* (*StatSoft*) (*Nisbet et al*., 2009, стр. 197).

Поред наведених комерцијалних пакета, на основу анкетног истраживања корисника од стране реномираних институција и путем *site*-ова из домена аналитике, као најчешће коришћени софтверски алати отвореног кода (бесплатно доступни)

[27] *SPSS* (акроним синтагме *Statistical Package for the Social Sciences*) је софтверски пакет који се користи за статистичку анализу и пружа изузетне могућност за решавање *DM* проблема и задатака. Реч је о пакету за стастистичку анализу компаније *SPSS Inc*., који је након припајања ове компаније *IBM* компанији, званично називан *IBM SPSS Statistics*. Прва верзија пакета за *DM* анализу је оригинално (1994. године) развијена у британској компанији *Integral Solutions Limited (ISL*), која је (1998. године) купљена од стране компаније *SPSS Inc*., а која је, пак, касније (2009. године) припојена *IBM* компанији. Овај софтвер је оригинално назван (*SPSS*) *Clementine*, затим преименован као *PASW* (*Predictive Analytics SoftWare*) *Modeler* и последично, његов актуелан назив је *IBM SPSS Modeler*. Многе консултанске фирме су анализом тенденција на *DM* тржишту идентификовале *IBM SPSS Modeler* као водећи *DM* алат. Као један од главних разлога оваквог позиционирања аналитичари, генерално, наводе помоћ *SPSS* алата у побољшању оперативних процеса и решавању кључних пословних проблема, до помоћи у доношењу одлука стратегијског карактера. Детаљан приказ концепта и примене *IBM SPSS Modeler*-а видети у: *Wendler & Gröttrup*, 2016.

идентификовани су: *RapidMiner* (претходно познат под називом *Yale*), *KXEN* (*Knowledge Extraction Engine*) и *Weka* (*Waikato Environment for Knowledge Analysis*). У питању су пакети који се одликују релативно широком палетом инкорпорираних алгоритама за реализацију *DM* задатака. Такође, изузетно популарне апликације су засноване на програмима као што су *MatLab* (комерцијални програм) и *R* (бесплатни програм), који су уједно и програмски језици. Мада изворно нису намењени за *DM* анализу, ови програми садржи уграђене математичке функције, статистичке функције (методе), као и функције за визуелизацију које подржавају имплементацију *DM* алгоритама. Компанија *Microsoft* је значајно допринела развоју *DM* подручја, између осталог, развојем *SQL Server Analysis Services* алата као дела (модула) програмског система за управљање базама података. Иако *Microsoft Excel* не представља погодно окружења за рад са изузетно великим количинама података (хиљадама колона и милионима редова), може се користити као радна платформа за подршку другим алатима. Стога, листи *DM* алата свакако треба додати и специјалне додатке за *Microsoft Excel*, као што је *XLMiner* пакет, који поред стандардних *Excel* функција користи и додатне функције, а који се може врло успешно користити за сврхе едукације и реализације *DM* пројеката мањих размера (*Shmueli et al.*, 2010). Поред наведених, врло често су у употреби и пакети за опште и пословне примене у којима доминирају алгоритми машинског учења: на пример, *DBMiner* и *IBM DB2 Intelligent Miner.*

Евидентно је да се софтверски алати појављују у различитим облицима: самостални програми који подржавају један метод, компонента програмских система или статистичких програмских пакета, самостални алати са инкорпорираном серијом алгоритама или алати за конкретна проблемска подручја и специфичне апликације. Како број расположивих алата континуирано расте, за кориснике постаје све тежи избор „најкориснијег“ алата са аспекта конкретног пословног сценарија. Комплексност *DM* проблема и комплетност анализе захтевају примену и повезивање различитих метода. У том смислу, подаци се често обрађују у ланчаном процесу који, применом низа метода садржаних у различитим софтверским пакетима из домена *DM*-а, треба да резултира постизањем дефинисаних циљева. У овом процесу, софтвери служе као помоћни алати за конструкцију модела, а пресудну улогу имају *DM* експерти.

6.4. Креирање *data mining* модела

Примена *DM* метода у конкретној ситуацији, при решавању конкретних задатака, резултира одговарајућим моделом (или, моделима) из података, тако да се *DM* може

сагледати као процес креирања модела. У начелу, сходно класификацији *DM* задатака и метода, разликују се дескриптивни и предиктивни *DM* модели.

Под моделом се подразумева експлицитно поједностављени приказ реалних објеката, процеса и појава. Модел, као селективна апстракција реалности, се често пореди са ауто картом, која није савршен приказ пута, али је користан водич. Управо, путем сумирања главних карактеристика реалних феномена, предност употребе модела потиче из његове једноставности у поређењу са истраживаним делом реалности. Модел може бити исказан у виду математичких формула или једначина, симбола, речи, слика, скупа правила и слично, али у сваком случају мора бити користан са становишта употребе за разумевање и управљање променама истраживаних феномена.

Поред питања избора адекватног метода / алгоритма (или портфолија метода / алгоритама), пре самог моделирања (односно, примене метода), важно питање у процесу развоја модела (законитости) из података односи се одлуку над којим подацима ће бити спроведено креирање и тестирање модела. За те сврхе, уобичајено је да се релевантни расположиви подаци поделе (партиционирају) на два или више одвојених, међусобно искључивих делова, који немају заједничке елементе. Свакако, постоји могућност да се *DM* анализа спроведе и без раздвајања података по посебним групама. Међутим, са појавом великих скупова података, независно од коришћеног метода, креирање ефикасних модела имплицира тестирање њихове поузданости и, консеквентно, поделу података за моделирање. Дакле, основна идеја поделе података у *DM* анализи је да се из анализе привремено искључи део података који ће касније бити коришћен за валидацију и тестирање модела.

При томе, подаци на којима се спроводи моделирање ретко када обухватају читаву популацију. Много чешће је реч о узорку података. Нарочито, уколико је скуп доступних података велики, да би се омогућила њихова ефективна обрада потребно је издвојити један део који ће бити анализиран. Осим тога, често физички није могуће обухватити цео скуп. Управо, због чињенице да је у бројним ситуацијама анализа на основу узорка погоднији или, пак, једини могући вид анализе, изузетно је важно да узорак буде репрезентативан, односно да на најбољи могући начин представља фундаментална својства основне популације и заступљеност појединих модалитета обележја. У супротном, резултирајући модел неће бити квалитетна симплификација извесних аспеката реалних проблема и система. Стога је за сврхе избора дела података за анализу и поделу података пожељно користити статистичке методе узорковања, као и методе процене релевантности и репрезентативности узорака.

У пракси се обично подаци за креирање модела[28] деле на три дела (партиције) (*Shmueli et al.*, 2005, стр. 19):

- део података за учење (тренажни део) - користи се за изградњу иницијалног модела или више модела;
- део података за валидацију - користи се за процену, оптимизацију и побољшање параметара (перформанси) сваког модела како би се обезбедило адекватно поређење и избор најбољег модела; и
- део података за тестирање - користи се за оцену перформанси и мерење ефикасности изабраног модела.

Дакле, свака формирана група података се користи за тачно одређену сврху у процесу моделирања. Међутим, у погледу процентуалне заступљености појединих делова у структури расположивих података не постоји универзално правило. Начелно, стандардне поделе података су засноване на случајном узорковању са различитим учешћем појединих партиција: од аутоматски подешених процената (на пример, однос између три наведене партиције може бити 40% : 40% : 20%, респективно), преко процената дефинисаних од стране корисника, до подједнаке процентуалне заступљености појединих партиција (33,3%). Осим поделе на три дела, у пракси се често користи и подела података на два дела: део података за учење (за генерисање модела) и део података за тестирање (за евалуацију резултирајућег модела).

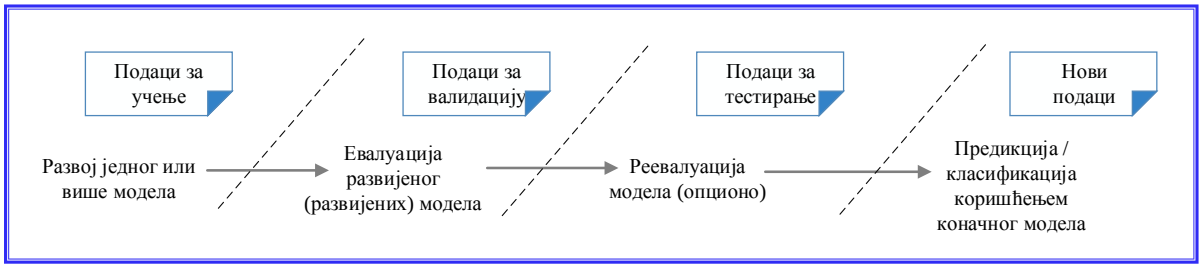
По правилу, већи је проблем поделити мање него веће скупове / узорке података. У случају малих скупова података користе се посебни методи поделе података, које су компјутерски захтевнији и сложенији, али омогућавају максимално искоришћавање свих података, попут метода изостављања једног елемента, метода поновљених узорака и различитих форми унакрсне валидације са *k* преклапања (*Cios et al.*, 2007, стр. 472). Такође, узорковање и подела небалансираних скупова који садрже ретке случајева (догађаје) захтевају посебан истраживачки приступ.

[28] Сегменти скупа података за креирање модела, према актуелној и општеприхваћеној терминологији усвојеној у *DM* заједници, називају се (под)скуп података за учење, (под)скуп података за валидацију и (под)скуп података за тестирање (енгл. *train set, validate set, test set,* респективно). За наведене три групе података за моделирање раније су коришћени изрази тренажни, тестни и евалуациони скуп. У случају предиктивног моделирања, креирани модел, примењује се на новом скупу података који се назива циљни скуп (енгл. *score set / production set*). Посматрано са статистичког становишта, веома је важно осврнути се на употребу термина „скуп / узорак". Наиме, у текстовима о *DM*-у уочена је терминолошка непрецизност и неусаглашеност приликом употребе ових термина. Не улазећи у анализу разлога за такво стање (вероватно један од њих је и мултидисциплинарна природа *DM*-а), недоследна, комбинована и паралелна употреба термина скуп (за учење, валидацију и тестирање) и узорак (за учење, валидацију и тестирање) је пре свега са статистичког, а и суштинског становишта недопустива. У контексту истраживања у овом раду, узимајући у обзир важност суштинске терминолошке дистинкције, овом приликом се инсистира на значењу и употреби термина скуп и узорка у њиховом изворном статистичком смислу: термин скуп односи се на све елементе на којима се извесна варијабилна појава испољава и посматра, а термин узорак на део скупа (чак и када узорак чини 99% од величине скупа).

Са становишта провере квалитета модела изграђеног на бази дела података за учење, од изузетне важности је обезбедити да се процеси учења, валидације и тестирања модела одвијају на независним деловима података, са случајним избором јединица посматрања у појединим деловима. У том контексту, често се поставља питање зашто је потребно спроводити и валидацију и тестирање. Оправданост и објашњење произлазе из потребе решења проблема пренаучености, односно превелике прилагођености модела подацима (енгл. *overtraining, overfitting*) (*Shmueli et al.*, 2005, стр. 19; *Berry & Linoff*, 2004, стр. 71). Заправо, модел је могуће, кроз процес учења, а касније и кроз подешавање параметара у валидационом процесу, толико специјализовати и прилагодити подацима да постаје дискутабилна валидност његове примене на новим подацима. Међутим, издвајањем дела података за тестирање обезбеђује се непристрасна оцена перформанси модела на расположивим, а до тада неискоришћеним подацима. Уколико се приликом тестирања идентификује присуство проблема превелике прилагођености као последица обухватања и моделирања случајних варијација присутних у валидационим подацима, а које вероватно неће бити присутне у новим подацима на којима ће модел бити примењен, приступа се новој корекцији модела. При томе, могуће је да су, након „подешавања" на делу података за тестирање, перформансе модела слабије, али се на тај начин постиже генерализација перформанси и испуњава захтев да модел буде довољно општег карактера.

Коначно, изабрани модел примењује се на потпуно новим подацима, који у тренутку креирања модела нису расположиви. Дакле, ови подаци не представљају део података за креирање модела. На Слици 12 је приказан однос између свих наведених делова података, као и њихове улоге у *DM* процесу. Полазећи од једног ширег оквира посматрања, процес тражења и креирања (скупа) модела може бити схваћен као процес проучавања карактеристика и дескрипције релација својствених расположивим подацима, који, као форма прелиминарног моделирања, примарно треба да обезбеди предуслове за квалитетено предиктивно моделирање, као и за бројне друге могућности примене *DM* модела. При томе, још једном треба нагласити важност везе између креирања и примене *DM* модела, обзиром да су тип и начин конструкције модела у великој мери одређени сврхом за коју ће модел бити употребљен. За разлику од предиктивних модела, дескриптивни модели се не примењују на новим подацима, већ описују и откривају законитости које важе за постојеће анализиране податке у форми визуелних приказа и хомогених група. Такође и тестирање статистичких хипотеза често резултира дескриптивним моделима (*Berry & Linoff*, 2004, стр. 52). Генерално,

резултате дескриптивног моделирања је изузетно тешко оценити, тако да се ефекти могу анализирати тек са одређене временске дистанце.



Слика 12: Партиције података за моделирање и њихове функције у *DM* процесу

Извор: Приказ аутора прилагођен према *Shmueli et al.* (2005, стр. 20)

Важно је напоменути да са протоком времена од креирања модела до његове употребе, независно да ли је реч о дескриптивним или предиктивним моделима, долази до деградирања перформанси модела. Заправо, сваки *DM* модел има своје трајање, односно, свој животни циклус. Временска зависност и лимитираност модела значи да креирани модел који је у конкретној ситуацији изабран на основу високог степена поузданости као адекватан, временом губи на поузданости упркос поновљеним процесима учења на новопристиглим подацима. У неким областима примене откривене законитости су релативно стабилне и не захтевају често креирање нових верзија модела. Међутим, знатно су чешће ситуације у којима се услови примене модела изразито променљиве природе. Због тога је, као значајан део примене модела, путем аутоматизованих процеса, неопходно стално пратити функционисања модела и спроводити његово побољшање и прилагођавање конкретним условима и свежим подацима. Ново моделирање се може вршити или променом структуре модела у оквиру коришћеног метода или, креирањем модела уз помоћ неког другог метода (или скупа повезаних метода, чијим се уланчавањем постижу синергијски ефекти, нарочито при реализацији изразито комлексних циљева анализе) (*Klepac & Mršić,* 2006, стр. 26).

Креирање *DM* модела је динамичан и итеративан процес који захтева понављање појединих корака све док се не добије модел задовољавајућег квалитета. Наиме, уколико су приликом тестирања модела дијагностификовани незадовољавајући резултати, поново започиње итеративни процес обраде података, што одговара специфичном спиралном приступу развоја *DM* модела.[29] Овај приступ омогућава непрекидну изградњу, праћење и унапређење модела путем минимизирања одступања и грешака модела, „минимизирајући ентропију уз помоћ повратне везе са уграђеним

[29] О спиралном приступу развоја система пословне интелигенције видети у: *Panian & Klepac*, 2003, стр. 104-109.

корективним чиниоцима" (*Klepac & Mršić,* 2006, стр. 235). Једноставно, када будућност постане прошлост, узимајући у обзир стечена искуства, отпочиње нови циклус креирања, прилагођавања и адаптације модела новонасталим условима како би се извршиле нове процене са што мањим степеном грешке. Међутим, корекција модела није, и не сме бити, само питање техничког коришћења неког готовог софтверског производа и пуког механичког „пуштања података у обраду". Нити се модели могу самоадаптирати и самокориговати, нити се ови процеси могу самопокренути. Заправо, још једном се мора истаћи да у овим процесима доминанатни улогу има људски фактор, односно *DM* аналитичар који дубоко разуме проблематику и основна својства примењених метода (и алгоритама) за креирање модела.

Из свега наведеног произлази да је *DM* подручје које пружа изузетан потенцијал за креативно, а самим тим и квалитетно решавање проблема. *Nisbet et al*. (2009, стр. 46) наводе да је *DM* својеврсна уметност, а процес креирања модела и описивање релација присутних у подацима упоређују са вајањем скулптуре. У контексту ове метафоре битно је доћи до следећег циља: исклесати скулптуру, односно креирати модел. Вајар свој рад на мермерној скулптури отпочиње клесањем комада камена сходно својој јасној визији и идејним решењма о коначном облику скулптуре. Након неколико „удараца" у блок мермера, уметник се повлачи у страну и посматра „својих руку дело". Овај итеративни процес „клесања и посматрања" се наставља све док скулптура не добије коначан облик. Попут вајара, *DM* истраживач, при креирању модела, полазећи од сирових података, најпре, спроводи иницијално чишћење података, решава проблем недостајућих вредности и кроз трансформационе процесе формира изведене варијабле. Након тога, применом одговарајућих алгоритама, спроводи иницијално моделирање, а затим наставља модификовање параметара и варијабли стварајући нове „оплемењене" верзије модела. Овај итеративни процес се одвија док се не исцрпе могућности за побољшање перформанси модела, односно док се не креира коначан модел одговарајућег квалитета.

DM, као фаза и срж моделирања према *CRISP-DM* процесу, је снажно и директно повезан са припремом података и оцењивањем квалитета модела, то јест, са фазом која непосредно претходи и фазом која следи након *DM* фазе. Стога је у наредном Делу овог истраживања анализа усредсређена на методолошке аспекте и поступке моделирања схваћеног у ширем смислу, а које управо обухвата припрему података, непосредно креирања модела и оцену његовог квалитета.

7. СТАТИСТИКА *versus DATA MINING*

Будући да су велике количине података условиле дубоке и озбиљне промене у домену анализе података, а самим тим и у примени статистичких метода у смислу њиховог прилагођавања новим приступима који произлазе из развоја информационих технологија, у овом Поглављу пажња је посвећена дискусијама и различитим тумачењима релација између статистике и *DM*-а.

7.1. Статистика у *data mining* окружењу

Анализа података и стицање знања путем издвајања законитости из података није нова истраживачка идеја. Наиме, позната је констатација да статистика, као научна дисциплина са веома дугом историјом, представља синоним за анализу података. Током развоја људске цивилазције, пре нове ере у Вавилону, Египту, Римској републици и Кини забележене су многе акције које су се по свом обиму и начину спровођења могле назвати статистичким, док се зачеци статистике као научне дисциплине везују за енглеску и немачку школу статистике из XVII века. У том контексту, статистички корени анализе података потичу из далеке прошлости.

Традиционално статистика се бави подацима. На почетку развоја задатак статистичких акција сводио се на прикупљање и евиденцију података о имовини и бројном стању становника, војника и пореских обвезника у циљу сагледавања финансијске и војне моћи тадашњих владара и држава. Заправо, статистика се првобитно односила на скуп нумеричких података о стању посматраних појава. Међутим, кроз дугу историју, не само да се мењало схватање и дефинисање самог појма статистике, већ су настајали и нови методи, као и нова размишљања о задацима, улози и значају статистике. У савременом свету, статистичко размишљање и примена статистичких метода су постали окосница у решавању истраживачких и пословних проблема, доношењу одлука од стране влада и менаџера, остварењу дефинисаних циљева, уз растући значај у свакодневном животу грађанства. Грубо, развој статистике се може пратити кроз следеће три етапе (*Lovrić i drugi*, 2017, стр. 20): ▸ једноставно прикупљање података о стању посматраних појава и сачињавање државописа у функцији сагледавања моћи владара и формулисања политика на државном нивоу (на пример, пореске политике), ▸ развој теорије вероватноће који је омогућио развој теорије статистичког закључивања, и ▸ револуцију у развоју и доступности електронских рачунара, која је допринела знатном помаку у развоју статистике и

отворила изузетне могућности са аспекта њене универзалне корисности у скоро свим научним областима, као и примени нових, компјутерски интензивних метода.

ICT иновације и развој донели су многе користи и ризике у савременом свету. У том контексту, *Straf* (2003), члан Америчке Статистичке Асоцијације, истиче да не постоји погоднија научна дисциплина од статистике која може преузети водећу улогу у суочавању са изазовима технолошких промена. Неспорно је да све научне дисциплине имају одређену улогу у технолошком развоју, али се поставља питање зашто је улога статистике посебна. Дискутујући о томе, овај аутор истиче, између осталог, да се статистика не бави само применом одређених технолошких решења, већ и праћењем и мерењем различитих аспеката технолошког развоја у свим областима науке, привреде и друштва на свим нивоима организовања (локалном, државном и међународном). Сходно томе, статистика унапређује открића у другим наукама, тако да као технологија она представља „фундаментални и непроцењиви део инфраструктуре других наука". Управо због њене релевантности са аспекта развоја других наука, статистика може да постане интегративна сила између њих.

Међутим, нагли развој и широка доступност компјутера (и статистичких софтвера) омогућили су не само да се афирмише улога и омогући релативно једноставан начин употребе статистике у домену свих научних дисциплина и осталих области живота и рада, већ су, истовремено, допринели и развоју саме статистике. Наиме, да би одговорила претходно наведеним изазовима, упоредо са технолошким развојем развијала се и статистика у следећим правцима: ▸ проширење садржаја и повећање дисперзије статистичких активности, ▸ изналажење нових и прилагођавање постојећих метода за рад са великим количинама података, и ▸школовање статистички образованих кадрова уз стицање нових знања и вештина које се захтевају од нове генерације статистичара, укључујући њихово активно учешће у истраживањима и пројектима мултидисциплинарног карактера.

Сходно претходно разматраном контексту, потпуно је јасно да је револуција у примени рачунара од друге половине XX века до данашњих дана знатно утицала на промене у схватању статистике. Из једног ширег угла посматрања, статистика се дефинише са аспекта њене сврхе за повећање степена разумевања окружења и појава, повећање благостања и побољшање квалитета живота кроз ефикасно откривање и ефективну употребу знања из података. У том смислу, статистика није скуп метода или скуп података, већ пре свега активност путем које се генерише и ефективно користи знање из података (*Straf*, 2003, стр. 3).

У расправама о шансама и изазовима пред којима се налази статистика у годинама интензивних технолошких промена посебну пажњу завређују размишљања о односу статистике и математике, као и статистике и рачунарске науке. *Lovrić* (2009, стр. 4) указује на различите ставове многих статистичара о овој теми и, између осталог, наводи саопштење еминентних статистичара из целог света са почетка XXI века да „статистика више није подручје математике (ако је икада била), већ велики корисник математике, као и рачунарских метода". Наиме, у овом и ставовима сличног типа апострофира се мултидисциплинарна природа статистике и њена самосталност као научне дисциплине која се бави екстракцијом информација из података. Међутим, претходно изнете констатације свакако не значе да је математика нестала из статистичке теорије (*Efron &Tibshirani*, 1991, стр. 390). Напротив, математика и даље омогућује теоријску анализу статистичких процедура, али је традиционалне методе математичке анализе могуће заменити специјално конструисаним компјутерским алгоритмима. Такође, веома је важно напоменути и чињеницу да је моћ многих нових статистичких метода (али и процедура, попут постепене регресије) заснована на електронским израчунавањима. Стога, *Friedman* (1997) наводи изјаву *Efron*-а према којој „статистика представља најуспешнију информатичку науку", а „они који игноришу и запостављају статистику су осуђени су да је поново измисле (креирају)". Заправо, уз адекватну методолошку подршку компјутерски заснованих израчунавања и алата, статистика треба да се фокусира на решавање скупа проблема и анализу података са циљем извлачења закључака и законитости из података.

Интересантно је истаћи да је растући значај статистике у савременом свету потврђен и од стране Статистичке комисије при Уједињеним нацијама, која је, у циљу промовисања достигнућа статистичких система како на националном, тако и на глобалном нивоу, као и рада статистичара који обављају своју делатност у различитим околностима, културама и подручјима, прогласила 20. октобар Светским даном статистике (енгл. *World Statistical Day - WSD*), први пут обележен 2010. године широм света. Тим поводом, тадашњи генерални секретар Уједињених нација *Ban Ki-moon,* у писму упућеном шефовима држава / влада позива их да пруже потпуну и благовремену подршку напорима усмереним на спровођење одлуке Генералне скупштине о проглашењу Светског дана статистике и износи јасан став у прилог афирмације статистичког деловања: „Учинимо овај историјски Светски дан статистике успешним - признавањем и прослављањем улоге статистике у друштвеном и економском развоју

наших нација и усмеравањем даљих напора и ресурса ка унапређењу националних статистичких капацитета" (*https://unstats.un.org/unsd/wsd/News3.aspx*).

На значај статистике, која помаже истраживачима и корисницима да дођу (од података) до релевантних открића и знања о варијабилним појавама у функцији разумевања савременог окружења (света) чија су иманентна својства динамичност и неизвесност, упућује и често цитирана визионарска изјава енглеског писца *Wells*-а: „Статистички начин размишљања ће једног дана бити неопходан обичном грађанину исто онолико колико читање и писање". Управо, наведена изјава илуструје реалност. Заправо, статистички начин размишљања је постао окосница за доношење валидних одлука и побољшање процеса у условима неизвесности у свим подручјима и на свим нивоима људске активности.

Као што је у досадашњим разматрањима више пута истакнуто, поред неизвесности, битан аспект савременог окружења је доступност велике количине података. Наиме, развој компјутера је повезан са генерисањем и складиштењем велике количине података. Последично је условио и транзицију свих подручја (од науке, државне управе, привреде и живота грађана) из фазе недовољне расположивости до фазе обиља података. Потреба и изазови анализирања ових података резултирали су не само револуцијом у погледу позиције статистичке анализе, већ фомулисањем и нових методолошких приступа и оквира анализе. Један од њих јесте *DM* приступ.

У том контексту, компјутерски подржана анализа и рад са великом количином података (великим бројем јединица посматрања и великим бројем варијабли) путем комбинације метода из различитих области може се означити синтагмом *DM* окружење. Да би статистика допринела решавању и превазилажењу изазова који су повезани са радом у *DM* окружењу неопходно је спровести адекватна прилагођавања традиционалне статистичке методологије, обезбедити развој нових статистичких метода и инкорпорирати статистички начин размишљања, концепте и идеје у *DM* методолошке оквире. С тим у вези, потпуно је јасна констатација *Hand*-а (2001, стр. xxviii) да *DM* захтева разумевање и статистичких и рачунарских питања и концепата.

Став да је *DM* окружење комплексно представља уобичајену констатацију, али чињеница да пружа изузетне могућности за развој нових методолошких приступа и решења, укључујући и развој статистике, је неспорна. Међутим, при примени статистике у *DM* окружењу и, генерално, уз примену рачунарске технологије треба бити јако обазрив. Пратећа појава (а може се слободно констатовати све чешће примећена) је погрешна употреба статистике и, сходно томе, погрешно тумачење

резултата. Усвајање и разумевање статистичког начина размишљања, као и услова и начина примене статистичких метода независно од софтверске имплементације је претпоставка за њихово успешно интегрисање у нове методолошке оквире. У супротном, примена статистичке методологије уз помоћ рачунара постаје црна кутија (*Lovrić i drugi*, 2017, стр. 25). Стога је веома је важно истаћи да искључиво методолошки примерена и валидна употреба статистичких постулата и метода у фазама *DM* процеса омогућава побољшање квалитета *DM* резултата, а самим тим и повећање поузданости обезбеђивање научне заснованости издвојених законитости из података и формулисаних закључака о истраживачком феномену.

7.2. Сличности и разлике између статистике и *data mining*-а

Статистика и *DM* представљају подручја анализе података која су повезана са трансформацијом података у корисне информације и знање. Суштински, оба подручја анализе имају исте циљеве чије је остварење засновано на учењу из података. Стога се отвара проблем утврђивања сличности и разлика између ова два методололошка приступа. При томе, поређење карактеристика статистичког и *DM* приступа у анализи података може се базирати на различитим критеријумима.

Посматрано из угла циљева истраживања, јасно је да су статистика и *DM* оријентисани ка идентификовању одређених структура у подацима. Упркос овом преклапању, за разлику од статистике која је углавном фокусирана на дефинисање глобалних модела (користећи теорију узорака и статистичког закључивања), фокус *DM*-а је на идентификовању како глобалних модела тако и локалних образаца понашања (локалних модела). Заправо, откривање локалних модела није централни предмет интересовања статистичара, али уколико је модел откривен, од интереса је, пре свега, оценити његову „реалност" и поузданост (*Hand*, 1999a). За разлику од статистичара, за *DM* аналитичаре, примарно је идентификовање локалних модела, док се оцена њихове „реалности" или вредности углавном препушта власницима база података или експертима из конкретног подручја.[30] У *DM* ситуацијама, сходно расположивости података популације и дефинисаним циљевима истраживања, откривање и дефинисање тражених модела може се базирати на узорку података или претраживању комплетног простора података. Често су од стране статистичара упућене примедбе (које се могу прихватити или одбацити) да при идентификовању

[30] На пример, *DM* аналитичар може идентификовати групу предузећа са сличном тенденцијом кретања цена њихових акција у одређеном временском периоду, а економисти и добри познаваоци кретања на финансијском тржишту треба да пруже образложење добијених резултата.

глобалних структура *DM* аналитичари не користе довољно теорију узорка и процедуре статистичког закључивања (*Soldić-Aleksić*, 2004, стр. 46). Међутим, при идентификовању локалних модела (као образаца понашања ограничених сегмената простора података) анализа не може бити заснована на узорку, односно, захтева претраживање целог простора података, за које не постоји алтернативно решење. Стратегије проналажења локалних структура су посебно корисне за откривање екстремних вредности (*Hand*, 1998, стр. 117).[31]

При откривању структура важно питање се односи на оцену интересантности, необичности и значаја откривених структура. И док се у статистици овај проблем решава кроз концепте вероватноће, тестирања хипотеза и статистичке значајности, *DM* истраживачи су дефинисали одговарајуће мере ваљаности (односно, интересантности, необичности и сличности) откривених структура, што су операционализовали у форми циљне функције, као компоненте *DM* алгоритма.

У компарацији сличности и разлика између статистике и *DM*-а, често се истиче да је статистика примарна, а *DM* секундарна анализа података (*Hand*, 1998, стр. 112). Статистика добија атрибут примарне анализе података, зато што се подаци углавном прикупљају путем посебних стратегија за генерисање и прикупљање података и са већ унапред дефинисаним питањима, а затим се спроводи анализа да би се добили одговори на постављена питања и провериле дефинисане истраживачке хипотезе. Начелно, код *DM*-а, анализа података се спроводи на подацима складиштеним у различитим репозиторијумима, независно од сврхе за коју су изворно прикупљени. Међутим, наведена разлика се може окарактерисати као условна, јер као што се статистичка анализа може спроводити на подацима који потичу из секунадарних извора, тако се и *DM* приступ може применити на подацима из примарних извора.

Независно од тога да ли се посматрају као примарна или секундарна анализа података, статистичко и *DM* моделирање обухватају низ активности и покривају цео процес анализе података, од јасног дефинисања проблема и сврхе анализе, преко обезбеђења квалитетних података и креирања модела, до визуелне презентације и примене добијених резултата. Заправо, попут статистичке анализе, *DM* није једнократна активност, већ итеративан процес решења проблема који представља круцијални потпроцес *KDD* процеса. С обзиром да је начелно *DM*

[31] На пример, ако је циљ истраживања оцена броја или пропорције корисника банкарских кредита који не испуњавају редовно уговором преузете обавезе (плаћање рате кредита), тада се модел може дефинисати на бази узорка, а затим, формулисати закључак о траженом сумарном показатељу популације. Међутим, ако је циљ истраживања идентификовање корисника који не плаћају редовно рату кредита, јасно је да се мора претражити сваки слог базе података корисника банкарских кредита.

усмерен на откривање структура из постојећих база података, питања која су повезана са планирањем експеримената, прикупљањем података и дизајнирањем упитника имају већу важност у процесу статистичког моделирања. Међутим, оно што је од подједнаке важности у оба процеса јесте претпроцесирање и обезбеђење квалитета података, јер ако подаци нису на адекватан начин припремљени за фазу моделирања не постоји методологија која ће дати валидне резултате. Ипак, треба нагласити да је знатно теже обезбедити адекватан квалитет података и евентуално спровести корекцију података у случају *DM* моделирања и рада са великим базама података. Такође, сличност између ових процеса моделирања повезана је са и значајем иницијалне и експлоративне анализе података. Као прве фазе анализе, оне обезбеђују прелиминарне информације о подацима у форми сумарних статистичких показатеља и погодних графичких приказа, а самим тим опредељују даље правце анализе и детерминишу избор метода моделирања.

Често се, као једна од битних карактеристика *DM*-а, истиче да је реч о приступу који се не базира на унапред дефинисаним хипотезама о истраживачком феномену. Међутим, наведено не значи одсуство хипотеза у *DM* анализи. Прецизније речено, док многи традиционални статистички методи захтевају унапред дефинисање хипотеза о разматраном проблему, *DM* се бави самом конструкцијом хипотеза. Заправо, за разлику од класичног статистичког поступка тестирања хипотеза и верификације *a priori* добро дефинисаних, теоријски утемељених, очекивања, суштински, *DM* је усмерен на генерисање хипотеза директно проистеклих из самих података. У случају великих база података многе важне релације нису унапред познате, тако да не могу бити узете у обзир при дефинисању хипотеза. Због тога, за потребе *DM* анализе су неопходни како одговарајући алгоритми за хеуристичко претраживање простора података у функцији дефинисања потенцијалних истраживачких хипотеза, тако и одговарајуће процедуре за њихову валидацију.

Сходно претходно наведеном, сасвим је оправдано направити разлику између конфирмативног и експлоративног аналитичког приступа у *DM*-у. Конфирмативна анализа има за циљ, да коришћењем традиционалне статистичке процедуре, потврди или одбаци дефинисане хипотезе о истраживачком феномену. Експлоративна анализа, која је типична за *DM*, подразумева претраживање података и откривање претходно непознатих и неопажених корисних информација укључујући и њихово повезивање ради дефинисања извесних хипотеза (*Giudici*, 2003, стр. 6). Генерално, у *DM* контексту конфирмативна и експлоративна анализа су комплементарне, при чему је првом

случају реч је о *DM*-у базираном на верификацији знања (организационог или персоналног у одређеном подручју интересовања), а у другом случају о *DM*-у базираном на откривању знања о истраживачком феномену.

На основу изнетих разматрања, произлази да су разлике (а не супротности) између класичног статистичког и *DM* приступа у анализи података последица фундаменталне разлике која се односи на количину анализом обухваћених података. При томе треба напоменути да је појам велика количина података релативна категорија, јер оно што се данас сматра великом количином података знатно је „веће" у односу на велику количину на крају XX века. Осим тога, скуп од неколико хиљада опсервација из статистичке перспективе је велики скуп, док се из *DM* перспективе може третирати као скуп недовољне величине с обзиром да савремене базе садрже милионе и билионе података.

У непосредној вези са количином података су и питања која се односе на тип података и тип и број варијабли које се посматрају. За разлику од класичних статистичких метода путем којих се процесирају подаци исказани у форми бројева уз испуњеност строго дефинисаних претпоставки о њиховој дистрибуцији, *DM* алгоритми функционишу са знатно ширим распоном типова података (слике, аудио подаци, текстуални подаци, просторни подаци и слично) уз знатно мање или потпуно одсуство захтева у погледу претпоставки о њиховој дистрибуцији. Дакле, у *DM* контексту подаци се схватају много шире и не односе се само на вредности квалитативних и квантитативних варијабли у статистичком смислу. Такође, број варијабли са којима раде статистичари је релативно мали у односу на њихов број у раду *DM* аналитичара, тако да је избор важних и интересантних варијабли класичан *DM* проблем.

Коначно, врло често се при разматрању разлика између статистике и *DM*-а наводи и следеће: у статиситчкој литератури у фокусу је модел и његово извођење, док се у *DM* литератури акценат ставља на алгоритме и одговарајуће софтверске алате који стоје на располагању (*Hand et al*., 2000, стр. 112). Ова разлика представља директну последицу кључног одређења *DM*-а по којем се *DM* карактерише софтверски подржаним откривањем законитости из великих количина података. Међутим, то свакако не значи да се статистика не бави истраживањем велике количине података и применом софтверских алата. Стога је више реч о формалној разлици.

Претходним излагањем није формирана свеобухватна листа критеријума за поређење статистичке и *DM* анализе, али је указано на неке аспекте који су од посебне важности у сагледавању сличности и разлика између ова два приступа.

7.3. Критички осврт на однос статистике и *data mining*-а

Увидом у литературу која се бави одређеним аспектима *DM*-а, може се запазити да однос између статистике и *DM*-а није довољно разматран, што се делимично може објаснити чињеницом да истраживања у домену *DM*-а углавном спроводе информатичари, који се, по природи ствари, пре баве алгоритамским контекстима и питањима ефикасности израчунавања, него статистичким идејама, концептима и питањима (*Smyth*, 2001, стр. 36).

За статистичаре (али и економисте, такође) синтагма *DM* је дуго имала пежоративно значење (снимање, пецање, њушкање и слично) са јасном негативном конотацијом. Овакав став је последица следећа два кључна разлога: први, путем *DM* процедуре подаци се испитију из различитих углова што резултира великим бројем модела, тако да се увек може (лако) пронаћи одређени модел који ће се, независно од његове комплексности, добро прилагодити подацима и, други, анализа велике количине расположивих података може довести до тога да несигнификантне законитости добију статус неочекиваних, интересантних и значајних (*Giudici*, 2003, стр. 5). Заправо, ове ране, почетне перспективе и негативне критике усмерене ка *DM* приступу у анализи података су сублимиране у форми већ поменуте максиме „ако довољно мучите податке, природно је да ће увек признати". Ипак, протеклих година став према *DM*-у (не само од стране статистичара) се променио. *DM* је постао значајно мултидисциплинарно истраживачко и апликативно подручје са знатним потенцијалом за примену у економији, пословној екномији и менаџменту. У том смислу, кроз компентентну примену *DM*-а, обухватајући различите рачунарске и статистичке методе, економисти могу остварити знатне користи креирањем валидних модела, који су не само добро прилагођени подацима, већ обезбеђују и добро предвиђање (*Sapra*, 2014), а самим тим и квалитетно одлучивање.

С обзиром да су сваком *DM* процесу својствена софтверски подржана комплексна и обимна израчунавања, информатичари се често декларишу као „власници" *DM*-а (*Ganesh*, 2002). При томе се, заиста, не може оспорити чињеница да су многе идеје (потпуно нове и различите у односу на претходне) у подручју анализе података потекле из информатичке науке. Неки од тих изузетно значајних доприноса информатичке науке односе се на: развој флексибилних метода предиктивног моделирања, употребу модела скривених варијабли за проблеме

груписања великог обима података и предиктивне проблеме, развој алгоритама за проналажење локалних образаца у понашању података (у већем степену него за идентификовање глобалних модела), прилагођавање традиционалних алгоритама за рад са великом количином података (обухватајући и компјутерске и статистичке аспекте) и коришћење алгоритама за учење у циљу анализирања и разумевања структуре хетерогених типова података (попут, мултимедијалних, *web* и текстуалних података) (*Smyth*, 2001, стр. 37-38). Веома је важно истаћи да у сваком од ових подручја статистика има велику улогу, с тим што за то нису заслужни сами статистичари, већ информатичари и инжењери који су настојали да статистичке методе прилагоде новим проблемима и околностима. Међутим, како су многи *DM* методи у својој основи статистички, статистичари су почели да показују озбиљно интересовање за ово подручје, што свакако може допринети развоју статистике као науке. Заправо, статистика је, суштински, један од корена *DM*-а која има виталну улогу у свим фазама *DM* процеса: од припреме података (укључујући и откривање екстремних вредности и избор релевантних варијабли), преко управљања процесом учења током примене алгоритама, до валидације резултата надгледаног учења и оцене стабилности резултата ненадгледаног учења.

Ипак, сходно специфичним својствима *DM*-а (која се примарно односе на обиље података, моћ компјутерских израчунавања и функционисање *DM* алгоритама искључиво у условима велике количине података), примена статистичких метода у *DM* окружењу, а посебно метода статистичког закључивања, праћена је одређеним тешкоћама. Реч је о проблемима који су повезани са превеликим прилагођавањем модела подацима, узорковањем, тестирањем статистичких хипотеза у условима велике количине података, укључујући и проблем вишеструке компарације (*Lallich et al.*, 2006, стр. 325-326). У основи, све ове проблеме могуће је сублимирати у форми проблема „лажних, погрешних и несигнификантних" открића (веза, релација и слично). У класичном статистичком окружењу примену статистичких метода карактеришу ригорозне, добро дефинисане претпоставке, захтевна израчунавања и (неретко, нарочито у прошлости) недовољна количина података. Услед недовољне количине података, статистика је пронашла методолошко решење да се исти (а врло често и релативно мали) узорак користи за одређивање оцена параметара модела, али и одређивање степена поузданости добијених оцена. Међутим, акценат који класична статистика ставља на статистичко закључивање недостаје у *DM*-у. Наиме, у *DM* окружењу, где

доминирају велика количина података и моћ компјутерских израчунавања, креирање модела се спроводи на једној партицији података, а оцењивање његових перформанси на другој. У том контексту *DM* се представља као „статистика обима, брзине и једноставности”, при чему се једноставност не односи на једноставност функционисања алгоритама, већ једноставност и разумљивост логике закључивања базиране на коришћењу различитих партиција података за учење и тестирање (*Shmueli et al.*, 2010, стр. 32). Али, с друге стране, не треба превидети чињеницу да је *DM* приступ анализи података изложен великом ризику од сувише доброг прилагођавања модела подацима за учење, односно обухватања не само структурних карактеристика, већ и случајних варијација података. У том случају долази до нарушавања перформанси модела приликом његове примене на новим подацима, тако да унакрсна валидација или коришћење различитих партиција података за учење и тестирање представљају начине за решење овог проблема.

Теорија узорака, као добро развијена област статистике, у *DM* анализи се користи на базичном нивоу (*Benjamini & Leshno*, 2010, стр. 529), упркос бројним користима које обезбеђују различити модели узорковања. Као последица тога, статус јединица посматрања (у смислу да ли је реч о подацима скупа или узорка, ако су подаци узорка који начин узорковања је коришћен, који је однос величине узорка и величине скупа и слично) није увек јасан, што знатно отежава валидацију изведених резултата и закључака. Истина, количина података која се сматра релевантном за *DM* је знатно већа у односу на ону која се користи у традиционалним статистичким процедурама. Међутим, често се дефинисани циљеви анализе података у решавању конкретних проблема, са већом прецизношћу и мањим захтевима у погледу израчунавања, могу постићи употребом и знатно мање количине података у односу на цео скуп података. Чак и софистицирана и рачунски интензивна процедура примењена на узорку података може обезбедити супериорније резултате него мање софистицирана процедура примењена на комплетној бази података (*Friedman*, 1997, стр. 9). Такође, у многим ситуацијама није могуће анализирати или приступити комплетној бази података. Услед наведеног, примена методологије узорковања у *DM* апликацијама има потпуни смисао и оправдање и то у контексту: редукције димензионалности, добијања прелиминарних информација о подацима које детерминишу даље правце њихове анализе и, коначно, стицања релевантних сазнања о разматраним феноменима. Другачије речено, узорковање представља рационалан и ефикасан начин решавања

проблема процесирања велике количине података и давања смисла скуповима података невероватних размера.

Статистички концепт тестирања хипотеза представља предмет дискусије и озбиљних расправа у академским круговима (између статистичара) већ више деценија. Централни предмет и ток ових дискусија се временом мењао: од дефинисања кључних теоријских концепата (нулте и алтернативне хипотезе, критичне области теста, *p*-вредности), преко процедура на којима треба да се заснива тестирање хипотеза (углавном у ранијим радовима су разматране су *Fisher*-ове и *Neyman-Pearson*-ове идеје, као и интеграција ова два правца у тестирању хипотеза), до истицања ограничења и проблема у спровођењу процедура тестирања (нарочито у радовима који су се појавили у последњој деценији XX века).

Као најчешћи проблеми код статистичких тестова наводе се њихова погрешна примена и некоректно тумачење добијених резултата (*Lovrić i drugi*, 2017, стр. 249-250), вероватно услед неразумевања концепата, услова, ограничења и логике тестирања хипотеза и широке употребе ове процедуре (инициране развојем компјутерске технологије) нарочито од стране корисника који не поседују потребна знања из статистике. Ипак, при томе, најважнији проблем се односи на одлуку о нултој и алтернативној хипотези, јер са повећањем величине узорка свака нулта хипотеза ће бити одбачена (*Lovrić i drugi*, 2017, стр. 250). Како је исход тестирања унапред познат, сама филозофија тестирања хипотеза је доведена у питање. Сходно наведеном, са *DM* становишта постоји опште уверење да тестирање хипотеза није релевантан концепт. Заправо, у *DM* окружењу знатно се смањује ефективност, односно корисност класичне статистичке процедуре тестирања хипотеза, јер са значајним повећањем броја опсервација (што је иманенетно својство *DM* окружења) *p*-вредност тежи нули, тако да ће при довољно великом броју опсервација свака нулта хипотеза бити одбачена, односно сваки резултат тестирања, при довољно великој количини података, постаје статистички значајан.

Генерално, у *DM*-у се ретко тестира једна хипотеза. Углавном је реч о фамилији сличних хипотеза. Стога, као и у статистици, при анализи случајних и систематских варијација и тестирању хипотеза, у непосредној повезаности са претходним проблемом јесте проблем вишеструког тестирања и компарације. Овај проблем се јавља приликом избора интересантних варијабли у надгледаном учењу и интересантних правила у ненадгледаном учењу или, пак, поређењу ефикасности више алгоритама. У сваком од наведених случајева, избор варијабле, правила или алгоритма је резултат *m* поређења,

односно понављања одговарајућег теста. Сукцесивна примена истог теста на истим подацима систематски повећава ризик грешке прве врсте, односно број лажних открића. Многе варијабле или правила постају (статистички) значајни и ако у суштини то нису. У ствари, ако се спроведе m тестова са нивоом значајности теста α (то јест, вероватноћом одбацивања истините нулте хипотезе), чак и када не постоје разлике нити у једном поређењу, а варијабле или правила нису интересантни у *DM* контексту, процедура аутоматски креира $m \times \alpha$ лажних открића (*Benjamini & Leshno*, 2010; *Lallich et al.*, 2006). Због тога је потребно формулисати другачији начин поређења, код кога би стварни ниво значајности након свих поређења био једнак унапред постављеном. Најједноставнији начин за решавање проблема вишеструке компарације је коришћење *Bonferroni* процедуре, која се заснива на m тестова уз ниво значајности α / m. Заправо, ризик грешке се смањује, тако да постаје теже одбацити нулту хипотезу.

Имајући у виду претходно дискутоване проблеме у поступку тестирања хипотеза, логично је да се поставља питање употребе овог статистичког метода у *DM*-у. У проналажењу потенцијалних одговора, став је да тестирање хипотеза не треба искључити из *DM*-а. Стога се, овом приликом, апострофира валидна и коректна употреба тестирања, што подразумева да *DM* аналитичари морају бити упознати са ограничењима тестирања и, генерално, свим фундаменталним аспектима статистичког закључивања, јер *DM* алгоритми не могу бити супституит за филозофију статистичког размишљања. Такође, полазећи од тога да статистичка значајност показује да ли има довољно аргумената за одбацивање нулте хипотезе, што не значи да добијени резултат мора бити и практично значајан, да би се контролисала стопа лажних открића, у одлучивање о томе да ли откривена структура поседује вредност у контексту разматраног феномена (то јест, да није лажна) мора бити укључен и експерт из области којој предмет анализе припада. Наведена констатација још једном потврђује, већ више пута истакнуту, чињеницу да *DM* није једнократна активност, већ континуирани процес откривања законитости из података, њихове валоризације и интерпретације.

Да би се у потпуности стекао увид у везу између статистике и *DM*-а, као и варијететност аспеката посматрања тог односа, интересантно је указати на серију радова у којима аутори настоје да одговоре на следећа истраживачка питања: да ли *data mining* може бити део статистике; да ли је *data mining* статистика или више од статистике; да ли је *data mining* примена статистике у форми експлоративне анализе података; која је следећа „генерација" у примени статистиких метода; да ли је *data*

mining статистички *déjà vu*; да ли је *data mining* комплемент или супституту за статистичку анализу; да ли *data mining* треба укључити у „биографију" статистике.

У правцу тражења одговора на формулисна питања посебну пажњу завређују радови чувеног статистичара *Hand*-а (1998; 1999а; 1999б) који истиче да *DM* преузима многе идеје и методе из статистике (нарочито из домена експлоративне анализе података), али *DM* је и „нешто више" и „нешто другачије" од саме статистике. Такође, исти аутор наводи да *DM* није „лек за све" и да рад са великом количином података носи са собом низ опасности и проблема (*Hand et al.*, 2000; *Hand*, 2009), за чије решавање статистичари могу дати велики допринос. Међутим, овде треба напоменути и констатацију да се већина статистичара слаже са ставом да статистика постаје релативно мање утицајна у ери информатичке револуције, наводећи да је један од кључних разлога за такво стање маркетинг проблем статистике (*Friedman*, 1997, стр. 8). Стога, статистици предстоји озбиљна борба и конкурентско надметање са другим методологијама и подручјима у домену *DM*-а. Сходно томе, *Ganesh* (2002) наглашава да „биографију" статистике треба проширити елементима компјутерски засноване анализе података, односно *DM*-а. Услед свега наведеног, као и чињенице да су многи *DM* методи у својој основи статистички, не изненађује то што су главни статистички софтверски пакети (као што су *SPSS*, *SAS* и *Statistica*) промовисани као *DM* алати. Будући да се *DM* и статистика баве извођењем законитости из податка, *Kuonen* (2005) истиче да се само по себи намеће питање да ли је *DM* статистички *déjà vu* и констатује да би одговор „да" био апсурдан, јер и поред великог значаја статистике у *DM*-у, постоје многи *DM* аспекти који нису статистички. Статистичар *Scarpa* (2011, стр. 337) наводи да се *DM* знатно преклапа у одређеним сегментима са стандардним процедурама и техникама експлоративне статистичке анализе података, али се истовремено суочава са новим проблемима од којих су многи не само последица количине података, већ и нових типова података, што иницира развој нових метода и технологија за њихово процесирање (изван оквира класичног *DM*-а).

Уважавајући претходно изнете ставове о односу *DM* и статистике, оно што је важно за обе дисциплине јесте да је време међусобног игнорисања и негативних критика између статистичара и *DM* аналитичара прошло. У том смислу, *Smyth* (2001) истиче да су статистички и алгоритамски аспекти подједнако важни у *DM*-у и констатује да је статистика есенцијална и вредна компонента у било којој *DM* апликацији. Успешност у имплементацији *DM*-а у будућности ће критично зависити од способности интеграције статистичких метода у матрицу *DM* праксе. Из другог угла

посматрано, *DM* иницира и обезбеђује изузетне могућности за развој нових методолошких решења у домену статистике (попут метода поновљених узорака, који омогућава да се из оригиналног узорка извуче велики број узорака са понављањем - уз неизоставну рачунарску подршку - и формулишу закључци који су ослобођени претпоставки на којима се заснива традиционална статистика). Да би се потенцијал њихове интеграције заиста искористио неопходна су обострана прилагођавања уз извесне модификације базичних парадигми и оперативних принципа оба приступа у анализи података.

Укратко, *DM* је релативно ново истраживачко и апликативно подручје, али и млада научна дисциплина, која се, као производ компјутерске ере, континуирано мења и развија. С друге стране, статистика као утемељена научна дисциплина која има знатно шири оквир у контексту приступа учењу из података, такође се мора котинуирано прилагођавати новим околностима. Ове тенденције ће неизбежно усмерити обе дисциплине једну према другој (укључујући и одговарајућа терминолошка усаглашавања), јер као што *DM* неће бити ефективан у откривању знања из података без статистичког размишљања, тако и статистика без елемената *DM*-а неће бити успешна у раду са великим скуповима података.

Део III

МЕТОДОЛОШКИ ПОСТУПЦИ ЗА СПРОВОЂЕЊЕ ЗАДАТАКА ПРЕТПРОЦЕСИРАЊА, ПРОЦЕСИРАЊА И ПОСТПРОЦЕСИРАЊА

8. ЗАДАЦИ И МЕТОДОЛОШКИ АСПЕКТИ ПРЕТПРОЦЕСИРАЊА ПОДАТАКА ЗА *DATA MINING*

Резултати било које анализе података директно зависе од квалитета података. Сходно томе, значајну улогу у обезбеђењу и побољшању квалитета података и, консеквентно, *DM* резултата имају методолошки поступци претпроцесирања података. Управо, у овом Поглављу разматрају се питања, која, у оквиру стандардизованог *CRISP-DM* процеса, припадају фазама разумевања и припреме података.

8.1. Значај и задаци претпроцесирања података

Претпроцесирање података је фаза у процесу откривања знања која омогућава боље разумевање саме природе података и обезбеђује неопходне претпоставке за ефикасну анализу у наредним фазама овог процеса. Заправо, *DM* моделирању претходи мукотрпан и дуготрајан процес претпроцесирања података. Упркос подразумеваној примени рачунара, ова фаза захтева знатно ангажовање стручних кадрова који су у стању да, сходно дефинисаном пословном проблему, припреме квалитетне податке и тиме омогуће спровођење одговарајућих поступака анализе. Практична искуства показују да је претпроцесирање временски најзахтевнија фаза и да „конзумира" знатно више времена него сама *DM* фаза. Процене утрошеног времена варирају од 60% – 80% и више укупног времена потребног за целокупан *KDD* процес (*Kamel*, 2009, стр. 538).

Већина *DM* експерата нагласак ставља управо на претпроцесирање пре него на преостале фазе *KDD*-а. *Zhang et al.* (2003, стр. 377) потврђују значај претпроцесирања следећим констатацијама: ▸ реални подаци су нечисти, ▸ за постизање високих перформанси, то јест, побољшање ефикасности *DM* система, неопходни су квалитетни подаци, и ▸ квалитетни подаци доприносе креирању високо квалитетних модела. Због чињенице да поседују лоша својства, реални подаци могу прикрити корисне законитости. Претпроцесирањем, односно решавањем следећих проблема, генерише се нови скуп (мањи од оригиналног) релевантних и квалитетних података, који води до квалитетних модела (*Cios et al.*, 2007, стр. 37; *Han et al.,* 2012, стр. 84-85):

- волуминозност података – претраживање гломазних репозиторијума података са својством високе димензионалности може представљати озбиљан проблем у смислу времена егзекуције и других перформанси *DM* алгоритама и ефикасности *DM* система;
- динамичка природа података – будући да се подаци константно мењају, потребно је водити рачуна о степену ажурираности података у изворима из којих

потичу и, сходно томе, путем већих пондерационих фактора вредновати актуелне податке, односно процес претраживања усмерити ка новијим подацима;

- непотпуност података – односи се на недостатак одређених значајних варијабли, недостатак вредности варијабли и слично, а може се манифестовати и кроз чињеницу да подаци садрже само агрегатне вредности или, пак, уместо тачних вредности, недовољно прецизне изразе (попут висока, просечна или ниска вредност);
- неконзистентност података – односи се на несклад у дефинисању, означавању или називима појединих варијабли или категорија, као и дуплирање записа приликом интеграције података из више извора (подаци који се понављају добијају на значају, а њихов утицај на коначан резултат моделирања се мултипликује);
- нетачне вредности и неправилности у подацима – реч је о неисправностима које нису систематског карактера и настају услед неисправности и непрецизности коришћених мерних инструмената, отказа хардвера, софтверских грешака, грешака приликом уноса и преноса података, правописних грешака, коришћења скраћеница и слично, а манифестују се у форми случајних грешака (шум у подацима) или значајних и неочекиваних одступања (то јест, екстремних вредности, с тим што, наравно, нису све екстремне вредности резултат грешке);
- редундантност података – односи се на случајеве када између посматраних варијабли постоји јака корелација, тако да за анализу није потребно задржати све варијабле, затим на означавање исте варијабле различитим називима, као и на посебну категорију редундантних варијабли и података (постојећих или изведених) који су ирелевантни, сувишни или недовољно информативни за истраживани проблем.

Примена *DM* алгоритама на сирове или лоше претпроцесиране податке по правилу резултира лошим резултатима. Значај претпроцесирања из угла квалитета коначних резултата откривања знања потврђује да се често између ове фазе и фазе моделирања спроводи и међуфаза додатног прегледа података (која се уосталом може сматрати и компонентом претпроцесирања) (*Pyle*, 1999, стр. 89). При томе, не треба заборавити чињеницу да је за исправно претпроцесирање потребно познавати начин функционисања *DM* алгоритама, јер сви *DM* методи треба да се посматрају у строгој повезаности са методолошким поступцима претпроцесирања.

Активности у фази претпроцесирања се поклапају са активностима у оквиру раније поменутог *ETL* процеса. Уколико не постоји организовано складиште података, претпроцесирање се спроводи над доступним, у контексту циља анализе релевантним,

базама података или датотекама. Основни задаци који се спроводе и проблеми који се решавају (у форми питања) у фази претпроцесирања су (*Nisbet et al*., 2009, стр. 51):

- Фаза разумевања података
  - Прикупљање података: Како пронаћи податке потребне за моделирање?
  - Интеграција података: Како интегрисати податке из различитих извора?
  - Дескрипција података: Како „изгледају" подаци?
  - Оцена података: Колико је чист скуп података?
- Фаза припреме података
  - Чишћење података: Како очистити податке?
  - Трансформација података: Како исказати варијабле?
  - Унос (уметање) недостајућих података: Како третирати недостајуће податке?
  - Пондерисање и балансирање података: Да ли се сви случајеви третирају исто?
  - Филтрирање података: Шта урадити са неконзистентим подацима, екстремним вредностима и подацима који садрже шум?
  - Апстракција података: Како третирати темпоралне податке?
  - Редукција података: Да ли се може смањити количина података која ће се користити за моделирање?
  - Извођење нових података: Да ли се могу извести и формирати нове варијабле?

Узимајући у обзир неспорни значај наведених питања и задатака, у даљем излагању се разматрају кључни проблеми и задаци претпроцесирања уз приказ изабраних методолошких поступака за њихово решавање.

8.2. Интеграција, чишћење и трансформација података

Сходно дефинисаном (пословном) проблему, претпроцесирање података заправо почиње са прикупљањем података, који најчешће потичу из више хетерогених извора. Стога је неопходно, најпре, прикупљене податке прочистити и интегрисати у један извор, а затим према потребама даље анализе спровести њихову трансформацију.

- Интеграција података

Најчешћи проблеми који се јављају и кореспондентни методолошки поступци који се користе при спровођењу задатка интеграције (енгл. *data integration*) су (*Han et al.*, 2012, str. 94-99):

- шематска интеграција – односи се на интеграцију шема различитих извора података и решавање „конфликата" у одређивању назива ентитета и / или њихових

варијабли (на пример, установити да ли се варијабла са ознаком *customer_ID* у једној бази односи на исту варијаблу са ознаком *cust_number* у другој бази). Метаподаци (који садрже кључне информације за сваки ентитет / варијаблу, попут назива, значења, типа података, ранга дозвољених вредности и слично) знатно доприносе избегавању грешака у шематској интеграцији.

- уочавање редундантности у подацима – присуство редундантности може се открити одређивањем коефицијента корелације за нумеричке податке, док за категоријске податке међусобна зависност варијабли може бити откривена путем χ^2 теста. Откривање редундантности користи се и у функцији редукције података.

- откривање дупликата – као додатак откривању сувишних (редундантних) варијабли, врши се детектовање истих записа (*n*-торки) који се појављују у различитим изворима, како би се у интегрисани систем одређени запис унео само једанпут.

- откривање и решавање проблема вредносних конфликата – односи се на уочавање *n*-торки које описују исти ентитет у различитим системима, али се вредности варијабли тог ентитета разликују по појединим изворима. Ове разлике могу настати због разлика у приказима, мерним скалама или кодирању (на пример, тежина изражена у килограмима и британским фунтама, цене изражене у различитим валутама, вредности варијабле пол могу бити означене у једној апликацији са М и Ж, а у другој 1 и 2 итд.). Такође, приликом усклађивања варијабли из различитих извора посебну пажњу треба посветити структури података (на пример, цена са или без пореза, друштвени производ исказан у реалним или номиналним вредностима). Осим тога, варијабле могу имати исти назив у различитим изворима, а да се при томе односе на различите нивое апстракције (на пример, варијабла укупна продаја у једној бази података може се односити на један производ у свим продајним центрима, а у другој бази варијабла под истим називом на цео асортиман у једном региону). Метаподаци се често користе као начин за решавање вредносних конфликата.

Већина наведених проблема су врло суптилни и тешко се откривају било којим аутоматским поступком. Услед тога, пажљива интеграција је изузетно значајна, јер доприноси смањењу и елиминисању редундантности, неконзистентности и семантичке хетерогености у резултирајућем скупу података.

- Чишћење података

Полазећи од релације: Подаци + Анализа = Резултат (*Dasu & Johnson*, 2003, стр. 12), јасно је да било каква, чак и најсофистицираније анализа лоших података нема

смисла и своди се на познату максиму из рачунарске литературе „шкарт улази-шкарт излази”.[32] Стога се и у *DM*-у истиче значај чишћења података (енгл. *data cleansing*).

Чишћења података обухвата следеће три фазе (*Maletic & Marcus*, 2010, стр. 23): ▸ дефинисање и утврђивање врста грешака (односно, категорија прљавих података), ▸ претраживање података и идентификовање грешака, и ▸ исправљање откривених грешака. Свака од ових фаза односи се на комплексне проблеме, као и широк спектар специјализованих метода за њихово решавање. У основи, као претпроцесни задатак, чишћење података обухвата уочавање: недостајућих података, шумова у подацима, нестандардних опсервација, неконзистентности у подацима и слично, при чему свако од ових питања представља посебну и обимну истраживачку област.

Веома чест проблем са којим се суочавају истраживачи у процесу пречишћавања података приликом реализације *DM* подухвата су недостајући подаци (енгл. *missing data*). Реч је о подацима који нису познати за неке варијабле ентитета у *n*-торци или узорку податка. Важно је направити разлику између недостајућих и празних података (енгл. *empty data*) (*Pyle*, 1999, стр. 61). Недостајуће вредности неке варијабле су вредности које реално постоје, али нису унете у базу података, док су празни подаци они за које вредност у реалном свету не постоји или се не може претпоставити.[33]

Недостајуће и празне вредности се често називају нултим вредностима (енгл. *null values*). При њиховој интерпретацији треба бити обазрив и направити разлику између ових категорија, али и у односу на „праву” нулу као модалитет обележја и резултат мерења. С обзиром да често није могуће поновити мерење, нити из података закључити које вредности су недостајуће, а које празне, у разграничењу и валоризацији њиховог значаја, неопходна је сарадња са експертима из области којој проблем припада. Наиме, независно од тога што су у питању незабележене вредности, оне могу носити скривено значење које је важно приликом откривања нових законитости. При томе, посебно место припада откривању узрока настанка неодређених вредности, јер они опредељују начин њиховог третирања. Такође, неодређене вредности не указују аутоматски на

[32] Реч је о ефекту који је познат под називом „*GIGO*” ефекат. Сам назив је акроним израза„*Garbage-In-Garbage-Out*”.

[33] Следећи пример илуструје разлику између недостајућих и празних података (*Pyle*, 1999, стр. 61-62): ресторан брзе хране продаје сендвиче са ћуретином са додатком сира. Да би се установиле преференције потрошача, води се евиденција о куповини и при томе прате две варијабле: пол (вредности: мушки / женски) и тип сира (вредности: швајцарски / амерички). При праћењу захтева, један купац је тражио сендвич са ћуретином, али без сира. Том приликом, продавац није регистровао пол купца. Дакле, вредности посматраних варијабли при овој куповини нису евидентиране тако да су оба поља у бази података остала без садржаја. У разматрању датог случаја, постоји квалитативна разлика између ове две врсте неодређених вредности: аналитичар може да претпостави и „подеси” вредност варијабле „пол купца”, јер иста није записана, а реално постоји (недостајући податак), док за другу варијаблу вредност није измерена и реално не постоји (празан податак), тако да се мора другачије поступити, односно размотрити структура базе података и приступити, на пример, редизајнирању модела података.

грешке у подацима. Узроци који доводе до појаве неодређених вредности су бројни, попут: отказа мерних система, отказа рачунарских система, грешака приликом чувања, обраде и преноса података (услед техничких пропуста и људских утицаја), пружања одговора од стране респондената на неочекивани начин или њихово одбијање да открију персоналне информације итд.

Због значаја недостајућих података са становишта квалитета података, поставља се питање како поступити са недостајућим подацима? У литератури преовлађује следећи став: ако је проценат недостајућих података мањи од 1%, сматра се да то није велики проблем за *DM* процес; ако је учешће недостајућих података од 1% до 5%, проблем се може решити применом уобичајених традиционалних техника; ако је њихово учешће од 5% до 15% потребно је применити софистициране технике, док учешће изнад 15% може озбиљно да утиче на интерпретацију резултата (*Soldić-Aleksić,* 2013, стр. 474).

За решавање проблема недостајућих података, постоји велика група статистички заснованих метода, као и велика група метода која се заснива на алгоритмима машинског учења. У оквиру обе групе постоје бројне опције од једноставних (попут импутације средње вредности) до сложенијих метода базираних на релацијама између варијабли. Општеприхваћена класификација метода за третирање недостајућих података (предложена од стране аутора *Little*-а и *Rubin*-а, а која се наводи у готово свим академским радовима, обухвата следеће групе: ▸ игнорисање и брисање (одбацивање) недостајућих података, ▸ параметарско оцењивање, и ▸ методи замене недостајућих података (*Batista &Monard*, 2003, стр. 520-521).

Генерално, за решавање проблема недостајућих података не може се издвојити један метод као апсолутно најбољи. Наиме, различите ситуације захтевају различита решења (*Magnani,* 2004). Стога је прави избор метода у конкретном случају комплексан задатак и зависи од више фактора: конкретног *DM* задатка, релевантности атрибута који садржи недостајуће податке за анализу, узрока настанка недостајућих података, учешћа недостајућих података у посматраном скупу података, механизма генерисања недостајућих података, искуства истраживача, расположивости статистичких рачунарских пакета итд. Међутим, независно од степена комплексности избора, најмање добро решење је игнорисати присуство недостајућих података.

Концепт шум у подацима (енгл. *noise*) има важну улогу у статистичкој анализи података и, у терминолошком смислу, односи се на случајну грешку или варијације у

мерењу варијабле. Случајна грешка се математички концептуализује кроз следећу релацију: Измерена вредност = Тачна вредност ± Случајна грешка.

Реч је о нерегуларностима (или нетачностима) које представљају инхерентно својство података (нарочито великих база података) и резултат су утицаја случајних фактора. Заправо, њихово присуство у подацима је пре правило него изузетак. Генерално, разликују се два типа шума у подацима (*Zhu et al*., 2006, стр. 276):

- шум у варијаблама – репрезентује тип грешке и случајна одступања која се односе на вредности варијабли, а њихови извори могу бити и недостајући и редундантни подаци, и
- шум у класама – репрезентује тип грешке која се односи на: 1) погрешно означене (неконзистентне и контрадикторне) податке, и 2) некоректно класификоване случајеве (најчешће као последица приближних вредности варијабли по групама).

У методолошком смислу веома важна напомена подразумева да у поступку анализе података, шум треба елиминисати пре откривања екстремних вредности и неочекиваних законитости. За изоловање шума и разликовање тачних („добрих") од нетачних („лоших") података користе се различити статистички методи, методи визуелизације и алгоритми машинског учења. Уобичајени приступи за елиминацију шума су засновани на усклађивању података и обухватају (*Han et al.,* 2012, стр. 89-90):

- локално усклађивање (уједначавање) – заснива се на подели уређеног скупа података на *k* интервала једнаке ширине или једнаких фреквенција и замени сваке вредности у интервалу са аритметичком средином, медијаном или модусом тог интервала или граничном вредношћу интервала која је најближа тој вредности;[34]
- груписање – подразумева организовање сличних вредности у групе, тако да вредности које остану ван формираних група могу указивати на одступања која представљају шум или екстремну вредност;
- регресију – подаци се усклађују прилагођавањем и избором најбоље регресионе функције, тако да подаци који су ван функције могу указивати на присуство шума у подацима.

➢ Трансформација података

Трансформација података је претпроцесни задатак путем којег се изабрани подаци трансформишу у погодну форму за спровођење *DM* моделирања. Постоје бројни методи за трансформацију података засновани на аритметичким операцијама,

[34] За елиминацију случајних варијација у подацима (нарочито подацима временских серија) често коришћени методи су метод покретних просека и експоненцијалног усклађивања (енгл. *smoothing*).

усклађивању података, концепту хијерархије, превођењу једног типа варијабле у други, комбинацији и креирању композитних и нових варијабли итд. У основи, трансформисане варијабле су резултат математичких функција, које се крећу од једноставних (попут, $\sqrt{x_i}$, $log(x_i)$, $1/x_i$) до изразито сложених трансформационих форми (попут, *Box-Cox*, *Fourier*, *Wavelets* и *Kernel* трансформација). Њихов избор и коришћење у конкретним апликацијама зависи од врсте података, количине података, проблема који се решава и *DM* задатка који се извршава.

Основни типови трансформације података (која се често назива и консолидација података) су (*Han et al.*, 2012, стр. 112):

- усклађивање: овим поступком трансформације обезбеђује се елиминисање шума из података, али и смањење утицаја екстремних вредности;
- агрегација: подразумева примену операције сумирања над подацима у складишту података (на пример, анализа дневне, месечне и годишње продаја);
- генерализација номиналних података заснована на концепту хијерархије: подаци на нижем нивоу замењују се подацима који су на вишем концептуалном нивоу (на пример, ознаке категоријске варијабле град, могу бити генерализоване и замењене ознакама округа или државе);
- дискретизација: полазећи од концепта хијерархије, оригинални подаци нумеричке варијабле се замењују са малим бројем група (на пример, нумеричка варијабла године старости може бити замењена интервалима [до 10, 10–20, итд.] или старосним категоријама: млади, средовечни и стари);
- креирање нових варијабли: на основу познатог скупа података стварају се нове варијабле, које, између осталог, могу допринети и откривању недостајућих података (на пример, у анализу се укључује варијабла површина на основу трансформације варијабли дужина и ширина, а уколико за неки елемент није могуће одредити нову варијалу то може бити знак недостајућег података за дужину и / или ширину);
- нормализација (или стандардизација): односи се на скалирање свих варијабли тако да њихове вредности падају унутар специфичног интервала са малим размаком варијације (на пример, [−1, +1], [0, 1] или [−3, 3], независно од тога што оригиналне податке можда одликује велика разлика између максималне и минималне вредности), омогућавајући, при томе, да се избегне и утицај мерних јединица на резултате анализе.

Стандардизација је врло популаран вид трансформације података користан за неколико *DM* метода, попут, неуронских мрежа, метода најближег суседа и анализе

груписања. Најчешће се користе следећа три једноставна и ефикасна метода стандардизације (*Kantardžić*, 2011, стр. 26): ▸ метод децималног скалирања, ▸ *min-max* метод, и ▸ метод *z*-стандардизације. Међутим, при спровођењу било које форме трансформације треба бити обазрив, јер се може променити природа оригиналних података која ће довести до значајног губитка релевантних информација.

Посебно значајан вид трансформације је издвајање нових варијабли трансформацијом или комбинацијом постојећих варијабли. Трансформација ове врсте се назива екстракција варијабли (енгл. *feature extraction*). У појмовном смислу, важно је јасно разграничити екстракцију варијабли, као форму трансформације података усмерене на редукцију димензионалности, од избора подскупа варијабли из оригиналног скупа свих варијабли (енгл. *feature selection*), као „чисте" форме редукције података. Често се за сврхе стварање редукованих приказа података врши њихово комбиновање. Генерално, не само селекција и екстракција варијабли, већ сви кључни задаци претпроцесирања међусобно се преклапају и комбиновано користе: на пример, усклађивање је форма чишћења, али и трансформације и редукције (дискретизације) података, а дискретизација форма трансформације, али и редукције података.

8.3. Редукција података

Непрекидно повећање количине дигиталних података у смислу броја јединица посматрања и варијабли, а консеквентно и вредности, поставља нове и веће захтеве пред рачунарске ресурсе са становишта складиштења, преноса и обраде података. Истовремено, анализа таквих података постаје веома сложен задатак, разумевање скривене структуре података теже, а когнитивно оптерећење аналитичара све веће. Стога је у припреми података за потребе *DM* анализе неопходно смањити обим података, али уз минимални губитак релевантних информација.

Другим речима, са повећањем расположиве количине података,[35] редукција података (енгл. *data reduction*), као претпроцесни задатак, добија све више на значају. Она омогућава добијање скупа података знатно мањег обима од оригиналног, уз очување интегритета и задржавања његове структуре, тако да процесирање (примена *DM* метода) редукованог скупа доводи до истог (или скоро истог), а у неким апликацијама и побољшаног, квалитета аналитичких резултата. Заправо, редукција података не редукује квалитет резултата (*Kantardžić*, 2011, стр. 37).

[35] Када су у питању мали скупови података, често се сматра да су процеси трансформације описани у Потпоглављу 8.2. довољни за адекватну припрему података за анализу.

Редукција података укључује следећа три процеса (*Nisbet et al*., 2009, стр. 69): ▸ смањење броја варијабли (редукција димензионалности), ▸ смањење броја јединица посматрања (узорковање), и ▸ смањење броја вредности варијабли (дискретизација). Спровођења процеса редукције у било којем од наведених праваца захтева и сагледавање добрих и лоших страна сваког од њих. У суштини, за утврђивање ефеката редукције података на перформансе процеса генерисања модела и самог модела користе се следећи типични критеријуми: ▸ ефикасност, ▸ тачност, и ▸ једноставност генерисаних модела (*Vercellis*, 2009, стр. 101). Пошто је тешко пронаћи решење које је истовремено најбоље по сваком критеријуму, често се трага за компромисним решењем у функцији остварења примарне сврхе редукције података: да се кроз унапређење квалитета података обезбеди побољшање перформанси процеса генерисања модела и самог модела, а тиме и побољша квалитет пословних одлука.

- Редукција димензионалности

Скупови података за *DM* анализу садрже високо димензионалне податке, односно садрже стотине, чак и хиљаде варијабли, које са становишта конкретног *DM* задатка имају различити значај. Наиме, димензионалност података исказује број варијабли које поседују објекти анализираног скупа података. Висока димензионалност угрожава перформансе *DM* алгоритама (кроз повећање времена и простора који су потребни за процесирање) и знатно отежава откривање законитости из података.

Проблеми у домену анализе података који су повезани са коришћењем великог броја варијабли представљени су синтагмом „проклетство димензионалности" (енгл. *curse of dimensionality*)[36]. Очигледно да је реч о синтези две компоненте: једна је димензионалност и, као што је већ истакнуто, односи се на број варијабли, а друга је проклетство и односи се на тешкоће у анализи узроковане повећањем броја варијабли. *Hand et al.*(2001, стр. 193) овај проблем дефинишу као експоненцијалну стопу раста броја података у (математичком) простору са порастом броја варијабли. Са повећањем броја варијабли изузетно брзо расте и сложеност проблема у анализи података. У ширем смислу, проклетство димензионалности је израз свих феномена који се појављују у вези са високо димензионалним подацима и најчешће имају негативне последице на перформансе алгоритама учења.

Сходно наведеном, сасвим је оправдано настојање да се истраживачка пажња фокусира на варијабле које су значајне из перспективе разматраних пословних

[36] Творац израза „проклетство димензионалности" је амерички математичар *Richard Bellman* (1920-1984), познат по теорији динамичког програмирања, коју је формулисао половином XX века.

проблемских ситуација преточених у конкретне *DM* задатке. За потребе смањења димензионалности (енгл. *dimensionality reduction*) предложени су бројни методи, који су засновани на већ поменутим приступима селекције и екстракције варијабли.

Селекција варијабли је процес смањивања њиховог броја и избора оптималног подскупа варијабли уклањањем ирелевантних и / или редундантних варијабли из оригиналног скупа, а на основу одређеног критеријума. Ако се претпостави да је A оригинални скуп варијабли (који дефинише скуп података D), m број варијабли у скупу A, одабрани подскуп варијабли A', k број варијабли у подскупу A', а $S(A')$ критеријум за избор варијабли у подскуп, процес селекције се дефинише као: процес избора подскупа A' са k варијабли из оригиналног скупа A са m варијабли (то јест, $A' \subset A$; $k < m$, а у случају изразито димензионалних података $k << m$)[37]. Добијени подскуп је оптимално решење A'_{opt} ако је простор варијабли оптимално редукован према критеријуму $S(A')$.

Процес селекције варијабли обухвата следеће кораке (*Dash & Liu*, 1997, стр. 132):

- Први корак: Генерисање подскупа варијабли из оригиналног скупа применом алгоритма за претраживање простора варијабли. Почетна тачка овог процеса је избор шеме претраживања иницијалног скупа варијабли (односно, начина генерисања подскупа), која може бити (*Vercellis*, 2009, стр. 104): одабир варијабли унапред, елиминација варијабли уназад и претраживање варијабли у оба смера.

- Други корак: Евалуација подскупа варијабли и тражење одговора на питање да ли је дефинисани подскуп оптималан, тако што се генерисани подскуп варијабли оцењен према дефинисаном критеријуму пореди са претходно најбољим подскупом. Уколико је нови подскуп бољи од претходно најбољег, тада се претходно најбољи подскуп замењује са новим подскупом варијабли. Овај процес се понавља док одређени критеријум заустављања (енгл. *stopping criterion*) не буде задовољен.[38] Оптимални подскуп је релативна категорија: подскуп изабран коришћењем једне мере за евалуацију не мора бити оптималан и према другом критеријуму.

- Трећи корак: Провера испуњености критеријума заустављања, како процес претраживања не би трајао бесконачно. Критеријуми заустављања се разликују у

[37] Процес селекције варијабли је корисно применити чак и у скуповима података са мањим бројем варијабли, јер избор важнијих варијабли директно утиче на креирање квалитетнијег модела. Наиме, и мала побољшања (на пример, од неколико процената у предиктивној тачности) могу бити изузетно значајна.

[38] У литератури се најчешће говори о следећим мерама за оцењивање значаја варијабли и подскупа варијабли (*Das*h & *Liu*, 1997, стр. 135-136): ▸ мере удаљености или дивергенције; ▸ мере информација или неизвесности (путем којих се вреднује садржај и одређује информацијска добит варијабле, попут *Shannon*-ове мере ентропије); ▸ мере зависности (попут коефицијента корелације); ▸ мере доследности (путем којих се тражи минималан скуп варијабли који доследно, као и цели скуп варијабли, раздваја класе у подскупу, с тим што недоследност постоји ако два објекта са истим вредностима варијабли припадају различитим класама); ▸ мере грешке класификације (односе се на тачност у разврставању нових елемената према релевантним варијаблама).

зависности од тога да ли су дефинисани на основу начина генерисања подскупа варијабли или начина евалуације генерисаног подскупа. У оквиру прве групе критеријума проверава се да ли је: ▸ унапред одабран одређен број варијабли, и ▸ да ли је спроведен унапред одређен број итерација. Провером испуњености друге групе критеријума испитује се: ▸ да ли додавање или елиминисање варијабли из генерисаног подскупа даје боље резултате него сам подскуп, и ▸ да ли је према коришћеном критеријуму за вредновање добијен оптимални подскуп варијабли (на пример, испуњен критеријум минимална стопа класификационе грешке).

- Четврти корак: Валидација резултата која је усмерена на проверу ваљаности подскупа варијабли с обзиром на дефинисане потребе и проблем који се разматра. При томе се користе различите форме провере, попут поређења добијених резултата са резултатима других метода или са унапред дефинисаним и познатим оптималним подскупом (одређеним на бази *a priori* знања о подацима).

У основи, генерисање релевантног подскупа варијабли може бити засновано на појединачној евалуацији сваке варијабле, као и на евалуацији подскупа варијабли. У првом случају, на основу појединачних вредности дефинисане мере за евалуацију релевантности, спроводи се, најпре, рангирање свих варијабли, а затим елиминисање оних варијабли које не остваре одговарајући ниво релевантности исказан конкретном мером. С друге стране, избор подскупа варијабли подразумева претраживање скупа варијабли и, према дефинисаном критеријуму, избор оптималног подскупа. С обзиром да у типичним *DM* апликацијама број варијабли није мали, јасно је да проналажење оптималног подскупа представља тежак задатак (*Kumar & Minz,* 2014, стр. 214). Логично, смањењем простора претраживања, смањује се и време претраживања.

Осим одабира подскупа релевантних варијабли, екстракција варијабли је један од начина за суочавање са проблемом димензионалности скупа података за анализу. Значења појма екстракције варијабли може се разматрати у ужем и ширем контексту.

У ужем смислу значења, појам екстракције варијабли се дефинише као процес издвајања нових варијабли путем линеарних и нелинеарних функционалних форми трансформације оригиналних варијабли формирајући нови редуковани скуп трансформисаних варијабли. У ширем контексту посматрано, осим наведеног значења у ужем смислу, издвајање нових варијабли обухвата и такве комбинације оригиналних варјабли које резултирају конструкцијом нових композитних варијабли са одговарајућим реалним значењем (попут, природног прираштаја, индекса телесне масе,

берзанских индекса и слично), при чему се скуп варијабли проширује.[39] Међутим, новоформиране композитне варијабле повећавају експресивну снагу оригиналних варијабли, тако да се у анализи разматраног проблема могу користити уместо иницијалних варијабли из којих су изведене. Треба напоменути да се често у литератури при интерпретацији појма екстракције варијабли обухвата само значење у ужем смислу, док се стварање композитних варијабли третира као посебан вид трансформације усмерен на конструкцију нових вариајбли кроз комбиновање старих варијабли у функцији редукције података (*Motoda & Liu*, 2002*).*

Ако се претпостави да је оригинални скуп A чини m варијабли, то јест, $A = \{A_1, A_2,\ldots, A_m\}$, после екстракције / трансформације варијабли добија се нови скуп B са k варијабли, то јест, $B = \{B_1, B_2,\ldots, B_k\}$. При томе је $k < m$, $B_i = f_i\{A_1, A_2,\ldots, A_m\}$, (за $i = 1, 2,..., k$), а f_i трансформациона функција којом се подаци велике димензионалности из скупа A пресликавају у нови простор мање димензионалности, то јест, скуп B. Циљ ове трансформације је да се, сходно дефинисаном критеријуму за евалуацију екстрахованих варијабли, формира нови скуп са минималним бројем варијабли. Предиктивна тачност за класификационе задатке, а мере сличности или мере одстојања у анализи груписања могу пружити одговор на питање колико је добра спроведена трансформација (*Motoda & Liu*, 2002*).*

Трансформација у форми функционалног пресликавања може се реализовати на више начина. Како се пресликавање спроводи путем линеарних и нелинеарних трансформација, методи за екстракцију варијабли се деле на линеарне и нелинеарне. У групу линеарних метода спадају анализа главних компонената (и факторска анализа), анализа независних компонената и *Fisher*-ова линеарна дискриминациона анализа, а у групу нелинеарних се убрајају *Kernel* методи, вишеслојни перцептрони, методи засновани на подржавајућим векторима. И док се наведени методи углавном користе за временски независне податке, посебна, широка група метода екстракције користи се за претпроцесирање временске серије.

На бази презентираних разматрања, јасно је да се смањење димензионалности у случају екстракције заснива на трансформацији и комбиновању варијабли, а резултира креирањем скупа нових варијабли, док се у случају одабира подскупа релевантних

[39] На пример, композитне варијабле (као резултат једноставних процеса сумирања, одређивања диференција и рацио релација) у неким случајевима могу боље описати разматрани проблем од иницијалних, оригиналних варијабли. У демографији, композитна варијабла природни прираштај боље репрезентује проблем „беле куге" него иницијалне варијабле наталитет и морталитет, док у скуповима мединцинских података, нова варијабла индекс телесне масе, екстрахована као пондерисни однос између варијабли висина и тежина је бољи инпут у дијагностификовању проблема прекомерне тежине од оригиналних варијабли на основу којих је нова варијабла изведена.

варијабли заснива на елиминисању мање важних варијабли. Међутим, као што је већ истакнуто, екстракција и селекција варијабли нису два одвојена и независна питања.

- Редукција вредности

У домену редукције података, један од претпроцесних задатака јесте редукција вредности, односно дискретизација. Концепцијски, дискретизација је непосредно повезана са добијањем мањег броја различитих вредности варијабли, које се заснива на конверзији варијабли, с једне, и концепту хијерархије варијабли, с друге стране. При томе се један тип података трансформише у други.

Заправо, укључујући хијерархијске релације и различите нивое апстракција варијабли, дискретизација обухвата:

- прво, смањење броја вредности квантитативних варијабли поделом рангираних оригиналних вредности (попут нумеричких вредности између 1 и 20 за варијаблу дужина периода отплате кредита у годинама) у одређени број нумеричких интервала ([1, 6); [6, 11); [11, 16); [16, 20]) или њиховом конверзијом у категоријске варијабле и лимитирани број различитих дискретних категорија (Рок = кратак, средњи, дуг);
- друго, смањење броја различитих вредности категоријских варијабли путем замене сваке вредности на нижем нивоу са кореспондирајућом дискретном вредношћу која се налази на вишем нивоу хијерархијског концепта, при чему корисник или експерт може једноставно дефинисати концепт хијерархије потпуним или делимичним специфицирањем редоследа варијабли кроз одређени шематски приказ (који, на пример за варијаблу географска локација купаца, укључује следеће варијабле[40] и релације између њих: улица < град < регион < држава).

У начелу, замена бројних вредности варијабли са малим бројем дискретних категорија смањује и поједностављује оригиналне податке (*Han et al.,* 2012, стр. 113), а резултирајуће *DM* законитости постају компактније, и не ретко разумљивије и прецизније наспрам резултата добијених процесирањем оригиналних података. Успешна дискретизација значајно проширује апликативне границе многих алгоритама учења. Заправо, неки алгоритми нису прилагођени за рад са нумеричким варијаблама. Међутим, конверзијом квантитативних у квалитативне вредности проширује се коришћење таквих алгоритама у бројним ситуацијама и значајно унапређују њихове перформансе, као и перформансе креираних модела. Такође, чак и алгоритми који су прилагођени за рад са нумеричким варијаблама постижу боље перформансе (убрзава се

[40] Напомена: концепт хијерархије код категоријских варијабли укључује групу варијабли, при чему у случају, на пример, географске локације одреднице улица, град, регион и држава представљају посебне варијабле.

њихов рад) уколико се нумеричке варијабле на одговарајући начин дискретизују и представе у форми мањег броја дисјунктних нумеричких интервала.[41] Међутим, не сме се превидети чињеница да било који процес дискретизације доводи до губитка информација. Стога, циљ доброг дискретизационог алгоритма је да се неминовни губитак информација минимизира (*Jin et al*., 2009, стр. 2).

Најбоља дискретизација се често базира на логичком расуђивању, *a priori* знању о разматраним варијаблама, искуству и процени стручњака у одређеном подручју примене, или је, пак, резултат консезуса и општеприхваћених категорија поделе појединих варијабли. Уколико изостану ови субјективни предлози и праксом диктирана (дефинисана) решења, користе се бројни методи дискретизације. У постојећој литератури постоји више аспеката разликовања и димензија за класификацију дискретизационих метода. Најчешће су у питању следеће дихотомне категорије метода: методи раздвајања (поделе) *vs* методи спајања, глобални *vs* локални, надгледани *vs* ненадгледани, статички *vs* динамички, директни *vs* инкрементални (детаљније о наведеним категоријама видети у: *Cios et al*., 2007, стр. 234-235; *Liu et al*., 2002, стр. 394-395; *Yang et al*., 2010, стр. 104-105). Осим наведених подела, често се говори о следећим методима за дискретизацију нумеричких података: методи груписања, хистограм анализа, ентропијски заснована дискретизација, методи спајања засновани на χ^2 статистици и методи засновани на интуитивној подели.

Проблем редукције вредности дистрибуцијом опсервираних вредности у интервалне групе је оптимизациони проблем. До оптималног решења се долази хеуристичким процедурама кроз више итерација, уз дефинисање евалуационих мера и критеријума за прекид процеса. У том смислу, типичан процес дискретизације (за нумеричке податке) обухвата следећа четири корака (*Liu et al*., 2002, стр. 397):

- рангирање нумеричких вредности варијабле која се дискретизује;
- проналажење најбоље преломне тачке за поделу рангираних нумеричких вредности на интервале или најбољег пара суседних интервала за спајање (на основу бројних статистичких мера или мера ентропије, као мера нехомогености интервала);
- подела или спајање (под)интервала према дефинисаном критеријуму;
- заустављање процеса у одређеној тачки (сходно изабраном критеријуму, попут броја интервала, статистичких мера, мера ентропије и граничних вредности).

[41] Начелно, дискретизациони интервали могу бити приказани и путем вредности номиналних варијабли са бројем модалитета који одговара броју интервала дискретизације. Међутим, у том случају нису расположиве информације о уређењу између интервала у нумеричком смислу и домену.

Без дубљег разматрања, а у циљу потпунијег увида у сложеност питања дискретизације, апострофира се још један проблем који је комплементаран дискретизацији квантитативних у квалитативне податке. Реч је о дискретизација квалитативних података и њихово представљање нумеричким величинама.

При решавању проблема смањења категорија квалитативних варијабли и њиховог превођења у нумерички облик разликују се два типа трансформационих релација: ▸ трансформација номиналних у квантитативне варијабле, и ▸ трансформација ординалних у квантитативне варијабле. Ове трансформације се најчешће спроводе за потребе прилагођавања података посебном типу анализе. С обзиром на различита својства номиналних и ординалних варијабли са становишта поређења, то јест, рангирања опсервација, разликују се и начини њихове дискретизације. Начелно, процес комбиновања, спајања и означавања бројевима категорија квалитативних варијабли захтева инкорпорирање знања, логичког расуђивања и искуства експерата у цео процес трансформације. Трансформација се често спроводи заменом квалитативих варијабли са више вештачких варијабли (енгл. *dummy variables*), које узимају вредности 0 и 1, тако да се називају дихотомне или бинарне променљиве.

- Узорковање

Метод узорковања, традиционално повезан са статистичким процедурама, своју примену налази и у делокругу претпроцесирања података за откривање знања и *DM* анализу. Будући да се тренд повећања количине података у савременим базама података наставља, за *DM* истраживаче и практичаре, питање редукције података засноване на узорковању, односно смањењу броја јединица посматрања уз очување суштинских карактеристика целог скупа податка, добија све више на значају.

Главне користи које су резултат узорковања великих количина података односе се на брзину и ефикасност рада са мањом количином података. У спровођењу *DM* задатака посебна корист од узорковања је у домену поделе узорка на тренажни, валидациони и тестни део за сврхе креирања модела из података, при чему, као што је већ наведено, постоје различити методи за поделу података за анализу.

Генерално, случајност је статистички концепт који има фундаменталну улогу у креирању репрезентативних узорака. За *DM* сврхе уобичајено се користе следеће врсте случајних узорака: прост случајан, систематски, кластер и стратификовани узорак.

Са становишта примене, наведени методи представљају методе узорковања опште намене. Насупрот томе су методи за специфичне сврхе које захтевају

специјализована знања о: специфичностима апликативних подручја, типовима разматраних проблема (попут временски зависних података), узорковању небалансираних проблема и ретких догађаја[42], узорковању малих скупова података[43] и слично. Истина, савршено репрезентативан узорак не постоји, али пословни и *DM* аналитичари морају познавати природу података и разматраних скупова како би у конкретној ситуацији извршили довољно добар избор метода узорковања.

8.4. Експлоративни *data mining*

Посебан део статистике, који је изузетно важан у припреми података за сваки вид њихове анализе, укључујући и *DM* анализу, јесте експлоративана анализа података (енгл. *Exploratory Data Analysis - EDA*). Статистичар *Tukey* је седамдесетих година XX века указао на основне карактеристике ове анализе, извршио њено поређење са конфирматорном анализом података (енгл. *Confirmatory Data Analysis - CDA*) и истакао да је *EDA* (нумерички и графички) детективски посао. Наиме, улога истраживача (аналитичара) је да истражи (анализира) податке на што више начина и то све док се не појави довољно уверљива (веродостојна) „прича" о подацима (*Yu*, 2010, стр. 10).

Да би се боље разумеле фундаменталне карактеристике експлоративне анализе података, неопходно је указати на кључну разлику између *EDA* и *CDA* концепта. Насупрот класичној процедури тестирања хипотеза (и генерално статистичког закључивања), која се заснива на верификацији унапред дефинисаних хипотеза о разматраном проблему и везама између варијабли, *EDA* (по аналогији са детективским послом) ставља акценат на истраживање података у циљу креирања идеја за генерисање хипотеза о разматраном проблему и идентификовање систематских веза између варијабли када не постоје претходне информације и очекивања о природи тих веза, при чему се не захтева испуњеност одређених, стриктних и унапред дефинисаних претпоставки за спровођење ове анализе.

---

[42] У статистичком смислу, реч је о неравномерној расподели елемената према вредностима посматране променљиве. Неравнотежа оваквог типа може узроковати значајне проблеме у спровођењу класификационог *DM* задатка, јер у формираним узорцима елементи појединих ретких категорија могу бити изостављени што ће знатно смањити тачност креираног класификационог модела. Разликују се два ефикасна метода узорковања неравномерно распоређених података (који се називају и методи узорковања ретких догађаја):

- први, подразумева случајно елиминисање, то јест, смањење елемената бројније (већинске) класе (енгл. *majority class*) у узорку (енгл. *undersampling*), и
- други, базира се на поновљеном случајном узорковању кроз повећање елемената ретке категорије, то јест, мањинске класе (енгл. *minority class*) у узорку (енгл. *oversampling*).

Очигледно, равнотежа у узорку података која одражава структуру скупа у погледу релативне заступљености појединих вредности (категорија) циљне варијабле постиже се чешћим узорковањем ретких категорија и ређим узорковањем чешћих категорија.

[43] У случају мањих скупова података, користе се, између осталог, методи поновљених узорака, од којих најважније место заузимају самогенеришћи методи (енгл. *bootstrap methods*).

Најзначајније карактеристике експлоративне анализе података (које је јасно одређују у односу на друге видове анализе података) су (*Tukey*, 1977, цитирано у: *Liu*, 2014, стр. 10): ▸ реч је о приступу за анализу података, пре него о скупу формалних и ригорозних процедура, ▸ усредсређеност на свеобухватно разумевање података у циљу откривања њихових суштинских карактеристика, ▸ употреба једноставних дескриптивних мера у циљу сумирања и сажетог представљања података, као и трансформације података у циљу побољшања њихове интерпретабилности, ▸ доминантна улога графичких приказа података, ▸ флексибилност како са становишта спровођења анализе „по мери" структуре података, тако и са становишта реаговања на откривене обрасце у подацима, ▸ усредсређеност на изградњу привремених („пилот") модела и генерисање хипотеза засновано на интерактивном истраживању података.

Не улазећи у дубљу елаборацију наведених карактеристика, суштина овог приступа може се исказати путем шире, делимично модификоване дефиниције, која је оригинално представљена у *The Cambridge Dictionary of Statistics* (*Everitt*, 2006, стр. 145): *EDA* је приступ у анализи података у којем се примарно наглашава употреба графичких метода. Овај приступ се не заснива на *a priori* претпоставкама о структури података или на (у процедуралном смислу) строго формалним моделима. При томе се претпоставља да подаци поседују структуру састављену од две компоненте: регуларности (систематска веза) и нерегуларности. Путем експлоративне анализе посматрани подаци се разлажу на саставне компоненте како би се, уз минимално коришћење формалних статистичких и математичких метода, открила основна законитост, то јест, компонента регуларности. Дакле, *EDA* омогућава откривање и „извођење" образаца (правила) који долазе из података, а не из очекивања или претпоставки истраживача о подацима.

У начелу, као форма прелиминарних истраживања, *EDA* омогућава да се: ▸ максимира количина откривеног знања скривеног у подацима, ▸ открије основна структура података, ▸ издвоје важне варијабле, ▸ идентификују нестандардне вредности, ▸ утврде основне претпоставке, које ће у наредним фазама истраживања бити тестиране, ▸ развију једноставни модели, ▸ детерминишу оптимални параметри, ▸ предложе хипотезе које се односе на узроке посматраних појава, ▸ предложе одговарајући статистички методи за анализу расположивих података, и ▸ обезбеди знање неопходно за даља прикупљања података у функцији реализације започетих истраживања или експеримента (*http://www.itl.nist.gov/div898/handbook/eda/eda.htm*; *Gorunescu*, 2011, стр. 58).

Упоредо са развојем компјутерских ресурса који је праћен драматичним повећањем количине расположивих података, развијао се и *EDA* приступ у анализи података. Услед наведеног, многи *EDA* методи су усвојени у *DM*-у и прилагођени за прелиминарна истраживања великих база података, као што се и многи *DM* методи користе за потребе прелиминарних истраживања података и реализацију *EDA* задатака. Заправо, модерна анализа података и примена напредних, високо софистицираних метода довела је до терминолошке модификације *EDA* концепта и увођења синтагме експлоративни *data mining* (енгл. *Exploratory Data Mining - EDM*).

EDM, као савремени следбеник идеје која стоји у основи *EDM* концепта, дефинише се као прелиминарни процес истраживања податка у циљу откривања њихове структуре коришћењем метода дескриптивне статистике, метода визуелизације и напредних метода за анализу података (*Dasu & Johnson*, 2003, стр. 19). Дакле, *EDM* се заснива на употреби не само базичних квантитативних и класичних графичких метода, већ и „обогаћених", софистицираних и компјутерски интензивних метода за откривање дубоко скривених законитости и генерисање хипотеза из сирових података. Самим тим, оквири претраге података су знатно проширени. Међутим, са становишта примарне сврхе истраживања података, суштински, разлика између традиционалног и савременог концепта експлоративне анализе података не постоји (*Kamel*, 2009, стр. 539). Генерално, примарна сврха и традиционалне и модерне експлоративне анализе података је да се испитају релевантне карактеристике посматраних варијабли, као и да се идентификују основне релације између анализираних варијабли или јединица посматрања. Сходно броју варијабли обухваћених анализом разликују се три правца експлоративне анализе: ▸ једнодимензионала (енгл. *univariate*), ▸ дводимензионална (енгл. *bivariate*), и ▸ вишедимензионална (енгл. *multivariate*) експлоративна анализа.

Као што је већ истакнуто, један од фундаменталних стубова на којем се заснива *EDM* анализа јесте визуелизација података, која обухвата графичку презентацију анализираних података и добијених резултата анализе путем не само базичних приказа, већ читавог арсенала напредних, софистицираних графичких приказа, развијених на темељу најразличитијих *ICT* достигнућа.

Методи визуелизације повећавају степен разумљивости података и омогућавају да извесне „скривене" и незапажене карактеристике постану очигледне, јер људи имају невероватне способности да уоче скривене правилности и везе, а људско око представља ненадмашан алат у (*DM*) анализи података. Суштину примене метода визуелизације открива следећа фраза: (Добра) слика говори више од хиљаду бројева.

Посебно треба истаћи чињеницу да методи визуелизације имају значајну улогу не само у домену „упознавања са подацима" и реализације задатака експлоративне анализе и претпроцесирања, већ у свим фазама процеса откривања знања. Сходно томе, методи визуелизације се интегришу у процес откривања знања кроз: ▸ визуелизацију сирових, „очишћених" и претпроцесираних података, ▸ визуелизацију парцијалних резултата итеративних фаза, и ▸ визуелизацију коначних резултата (*DM*) анализе.

Приликом дискусије о визуелизацији у оквиру *EDM* задатака и, уопште, претпроцесних задатака, незаобилазни предмет разматрања се односи на нестандардне вредности. Заправо, визуелно истраживање је ефикасан начин за уочавање њиховог присуства. Будући да нестандардне опсервација могу представљати озбиљан проблем из перспективе добијања смерница за даљи ток анализе, а самим тим и из перспективе квалитета коначних резултата анализе, оправдано је констатовати да уочавање нестандардних вредности представља суштински задатак у *EDM* анализи и припреми података (у оквиру задатка чишћења података) за *DM* анализу. Осим тога, откривање нестандардних вредности се често реализује и као самостални *DM* задатак и апликација. У наставку текста истраживачка пажња се усмерава на проблематику нестандардних вредности.

8.5. Анализа екстремних вредности

Један од примарних корака у правцу обезбеђења предуслова за добијање валидних резултата у анализи података је откривање оних података који значајно одступају од вредности свих осталих података. Такви подаци који на неки начин нису конзистентни са преосталим делом података називају се нестандардним или нетипичним подацима (енгл. *outliers*), мада се у српском језику углавном користи изворни енглески термин или превод у форми синтагме екстремне вредности. Поред наведеног, у различитим апликативним подручјима као синоними за нетипичне вредности користе се и термини, попут, аномалије, девијације, занимљивости, специфичности, новитети, изненађења, изузеци, дефекти и слично.

Довољно општу и често цитирану дефиницију концепта нестандардног податка дали су *Barnett* & *Lewis* (1994, стр. 7) који под овим појмом подразумевају „опсервацију (или подскуп опсервација) која није конзистентна са преосталим делом посматраног скупа података". За *Pyle*-а (1999, стр. 71) „екстремна вредност је појединачна вредност или вредност променљиве са врло ниском фреквенцијом, која је значајно удаљена од већине преосталих вредности посматране променљиве". Другим

речима, екстремни податак је нетипична вредност која се знатно разликује (значајно већа или мања) од осталих вредности у скупу података, или, пак, од осталих вредности груписаних, то јест, нагомиланих по различитим зонама (групама) приближних вредности у скупу података. При томе, контекст „значајно већа или мања одступања" не односи се само на изразито различите вредности у посматраном скупу података, већ и на умереније (блаже) форме одступања од типичних вредности, које, такође, могу имати значајне негативне консеквенце на резултате анализе података. На Слици 13, која илуструје спектар типичних и атипичних (или, стандардних и нестандардних) података, јасно се уочава да је много већа област типичних, а мања област атипичних опсервација, што је једна од основних претпоставки на којој се заснива истраживање екстремних података. Заправо, екстремна вредност је догађај који се ретко појављује.

Слика 13: Област типичних *vs* атипичних података

Као истраживачки проблем, откривање екстремних вредности није једноставан задатак. Користећи термин аномалије за описивање екстремних вредности, *Chandola et al.* (2009, стр. 15:3) наводе следеће факторе који усложњавају њихово идентификовање:

- дефинисање области нормалног понашања која обухвата све могуће стандардне опсервације је веома тешко;
- посматране појаве током времена еволуирају, тако да постојећи спектар вредности не мора бити довољно репрезентативан и у будућности;
- граница између типичних и атипичних података је често нејасна и непрецизна;
- егзактно тумачење појма екстремни податак је различито за различита подручја примене (на пример, у области медицине, мало одступање од нормалне телесне температуре сматра се екстремним податком, док слична одступања у кретању цена акција на финансијском тржишту представљају уобичајена кретања);
- у процесу примене одговарајућих метода за откривање екстремних вредности и креирању валидних модела, расположивост означених, нестандардних података за учење и тестирање модела је често критично питање;

- подаци који су резултат недозвољених активности могу се толико прилагодити стандардним опсервацијама, тако да њихово разликовање постаје озбиљан проблем, што додатно отежава дефинисање зоне која се односи на типичне вредности;
- подаци могу садржати такве шумове који имају сличне тенденције са стварним екстремним вредностима тако да их је веома тешко разграничити.

Проблем истраживања екстремних података може се сагледати кроз следећа два субпроблема: (1) јасно дефинисање услова под којима се подаци могу сматрати неконзистентним у посматраном скупу података, и (2) избор ефикасних метода за (а) откривање екстремних података, и (б) третирање откривених екстремних вредности.

Одлука о укључивању или елиминисању екстремних података из даљег разматрања примарно зависи од узрока њиховог појављивања. У принципу, екстремни подаци могу настати услед механичких грешака, промена у понашању система, девијација и превара у понашању, људских грешака, грешака приликом мерења или, пак, због природних одступања у популацији (*Kantardžić*, 2011, стр. 42). Заправо, веома је важно направити разлику између следећих типова узрока појављивања екстремних вредности (*Batini & Scannapieca*, 2006, стр. 86): ▸ податак је резултат грешке настале приликом мерења или уноса података, ▸ податак потиче из друге популације, и ▸ податак репрезентује редак догађај и указује да се нешто неуобичајено догађа са посматраном појавом.

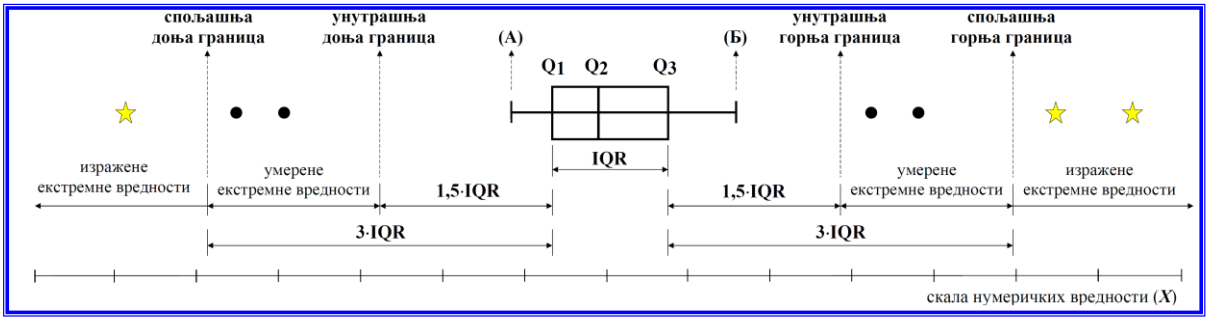
У првом и другом случају екстремни подаци представљају грешке које могу довести до погрешне спецификације модела, пристрасних оцена параметара и нетачних резултата анализе. Алтернативно, трећи тип узрока одражава оригиналне карактеристике разматраног феномена, као резултат њихове инхерентне варијабилности. Стога, приликом решавања проблема екстремних вредности треба бити обазрив, јер нису сви екстремни подаци грешке, пропусти или, једноставно, „лоши" подаци. Са становишта корисности информација које садрже, они могу бити најбогатији део скупа података, тако да се ни под којим околностима екстремни подаци не смеју аутоматски означити грешком и занемарити (*Stamenković & Milanović*, 2014, стр. 175). Из наведеног јасно произлази да ефикасно откривање неуобичајених вредности и идентификовање узрока њиховог настанка обезбеђује корисне информације и сазнања о посматраним појавама и смањује ризик доношења неквалитетних одлука заснованих на погрешним подацима.

У методолошкм смислу, откривање нестандардних вредности може се представити као следећи процес: од n разматраних тачака пронаћи k тачака које

највише одступају од осталих података, при чему је ($k << n$). Иако је у већини случајева реч о једној или неколико тачака (као потенцијалним екстремним вредностима), могу се појавити и групе тачака које се налазе изван дефинисаних зона уобичајених вредности. Сходно значају који има идентификовање екстремних вредности, у многим подручјима примене[44] развијени су бројни методи за њихово истраживање, укључујући визуелне и квантитативне оцене (индикаторе) дивергенције сваког податка.

Фундаментална класификација методолошких приступа који се користе у процесу откривања нестандардних вредности обухвата (*Kantardžić*, 2011, стр. 42): ▸ приступ заснован на визуелизацији, ▸ статистички приступ[45], ▸ приступ заснован на мерењу удаљености, и ▸ приступ заснован на моделима или мерењу одступања елемената од утврђених законитости. У прилог чињеници да је проблем екстремних вредности широка истраживачка област, посебно са методолошког аспекта, прави се разлика и између једнодимензионалних *vs* вишедимензионалних и параметарских *vs* непараметарских метода за њихово идентификовање (*Ben-Gal*, 2010).

У случају једнодимензионалних података, најчешће коришћени поступци за идентификовање (једнодимензионалних) нестандардних опсервација су: ▸ поступак заснован на емпиријском правилу „68-95-99,7", ▸ поступак заснован на *z*-трансформацији, ▸ *Grubbs*-ов тест (H_o: Посматрани скуп података не садржи екстремне вредности), и ▸ графички приказ сумарних статистичких мера посматране варијабле познат под називом *box plot* (илустрован на Слици 14).



Напомена: Коришћени симболи имају следећа значења: Q_1 = први квартил; Q_2 = други квартил; Q_3 = трећи квартил; *IQR* = интерквартилна разлика.

Слика 14: *Box plot* и екстремне вредности

[44] Анализа екстремних вредности добија на значају у бројним апликацијама, попут: идентификовање ризичних клијената, неуобичајених тржишних трансакција, неочекиваних промена у пословним процесима, нерегуларности приликом политичких гласања итд. Важно је истаћи да екстремне вредности не одражавају само нежељене промене, појаве и догађаје, већ то могу бити подаци са позитивном конотацијом. Листа подручја у којима се примењује анализа екстремних вредности дата је у: *Hodge & Austin* (2004).

[45] У начелу, статистику као науку, не интересују појединачни подаци, већ глобално понашање свих података. Међутим, имајући у виду значај појединачних података који одступају од преосталих података и њихов утицај на сумарне статистичке мере и показатеље, сасвим је разумљиво зашто су они предмет статистичких истраживања.

Методолошки поступци који се односе на случајеве када се посматра једна променљива могу послужити и за идентификовање мултидимензионалних нестандардних опсервација, с тим што се као једнодимензионална променљива користи главна компонента, која представља линеарну комбинацију оригиналних променљивих (*Kovačić*, 1994, стр. 211). Стога, анализа главних компоненти заузима посебно место у статистичким поступцима за откривање мултидимензионаланих екстремних вредности. Примена статистичких метода подразумева откривање оних опсервација које су лоциране релативно далеко од центра вишедимензионалне дистрибуције свих опсервација, при чему је важно истаћи да вишедимензионалне екстремне вредности не могу бити откривене посматрањем сваке варијабле одвојене. У идентификовању вишедимензионалних нестандардних опсервација користе се одговарајући индикатори одстојања, попут, *Mahalonobis*-овог одстојања, као и погодни графички прикази базирани на овом одстојању, попут, хи-квадрат *Q-Q* и *Beta Q-Q* дијаграма (*Ramzan et al*., 2014, стр. 256-257). У начелу, висока вредност *Mahalonobis*-овог одстојања сугерише да конкретна опсервација у мултидимензионалном простору (мултидимензионална опсервација) има екстремну вредност за једну или више променљивих (*Soldić-Aleksić*, 2015, стр. 224). Међутим, поступак заснован на овом одстојању не обезбеђује идентификовање варијабле (варијабли) која је условила већу вредност *Mahalonobis*-овог одстојања конкретне јединице посматрања (*Hair et al*., 2010, стр. 66). У статистичком смислу, *Mahalonobis*-ово одстојање следи приближно хи-квадрат распоред, са бројем степени слободе који је једнак броју варијабли. Поређењем израчунатих вредности *Mahalonobis*-овог одстојања за сваку јединицу посматрања и вредности 97,5% перцентила хи-квадрат распореда, као критичне вредности, идентификује се присуство мултиваријационих нестандардних опсервација. Уколико је вредност *Mahalonobis*-овог одстојања већа од одређене критичне вредности (у ознаци, $\chi^2_{(p;\,0{,}975)}$, где је p број варијабли), посматрани мултидимензионални податак је екстремна вредност (*Filzmoser et al.,* 2014).

Генерално, независно од примењеног метода, након идентификовања присуства екстремних вредности, следи даља анализа и откривање узрока њиховог појављивања, која, између осталог, аналитичарима треба да послужи као темељ за избор адекватног методолошког приступа у третирања екстремних вредности. Заправо, наредни (суб)проблем је како поступити са таквим подацима. Такође, и за његово решење постоје бројни методи, који, са методолошког аспекта, додатно доприносе повећању разуђености анализе екстремних вредности. Суштински ови методи су базирани на

једној од следећих опција: задржавању оригиналног податка, модификацији конкретне вредности која је оцењена као екстремна у циљу њеног приближавања осталим подацима, искључењу исте из даље анализе, као и трансформацији података у правцу „нормализације" расподеле и смањења утицаја екстремних вредности. Заправо, универзални одговор на постављено питање не постоји, јер који су начини погодни за формулисање оптималног решења у конкретној ситуацији зависи од читавг низа фактора, попут, типа података, типа екстремне вредности и узрока њеног јављања, експертског мишљења аналитичара, као и значаја идентификоване вредности, резултата емпиријских анализа, симулација и слично.

9. МЕТОДИ ЗА РАЗВОЈ *DATA MINING* МОДЕЛА

Бројност метода за сврхе предиктивног и дескриптивног *DM* моделирања је повезана са чињеницом да не постоји један универзални и супериорни метод који даје најбоље резултате у свакој проблемској ситуацији и у свим доменима примене. Наиме, сваки метод поседује карактеристична својства која исти декларишу као боље прилагођен од других у решавању конкретног проблема. Заправо, одговор на питања када и како применити одређени метод базира се на познавању његових карактеристика. Полазећи од наведеног, у овом Поглављу су разматрана суштинска својства, као и кључни имплементациони аспекти метода који су, из велике групе *DM* метода, изабрани узимајући у обзир, пре свега, њихов допринос и улогу у контексту остварења дефинисаних циљева докторске дисертације.

9.1. Анализа груписања

Анализа груписања, или кластер анализа (енгл. *cluster analysis*), као метод мултиваријационе статистичке анализе за алокацију јединица посматрања у смислене групе којима је могуће управљати, успешно се користи и у истраживању великих база података за реализацију *DM* задатака. С обзиром да групе нису унапред познате, а зависна варијабла није дефинисана, анализа груписања припада групи дескриптивних метода, то јест, метода заснованих на парадигми ненадгледаног учења. Са становишта *DM* анализе, анализа груписања се може интерпретирати на следећа два начина: ▸ као метод за откривање скривених (или слабо уочљивих) структура и екстракцију знања из (мултидимензионалних) података, и ▸ као претпроцесни корак у примени других *DM* алгоритама (у својству сегмента експлоративне *DM* анализе, метода за редукцију димензионалности и метода за откривање екстремних вредности).

9.1.1. Кључни концепти у анализи груписања

Анализа груписања је метод који се користи за груписање јединица посматрања у две или више непознатих, смислених и корисних група на основу њихове сличности према посматраним карактеристикама. Заправо, основна сврха груписања је идентификовање хомогених група јединица посматрања, тако да су јединице посматрања које припадају једној групи, са становишта вредности анализираних варијабли сличне међу собом, али се према вредностима истих варијабли знатно разликују од јединица посматрања укључених у друге групе. У суштини, овај метод омогућава не само стицање увида у дистрибуцију података по групама, већ обезбеђује погодну основу за дубља истраживања разматраних проблемских ситуација путем примене, на формираним групама, и других методолошких поступака. [46]

Фундаментални концепт у анализи груписања и формирању интерно хомогених и екстерно хетерогених група је концепт сличности. При томе, степен сличности, или, алтернативно, разлике између објекта одређује се путем одговарајућих мера блискости, на основу којих су развијени и бројни методи груписања података. Блискост–удаљеност (енгл. *proximity–distance*) и сличност–различитост (енгл. *similarity–dissimilarity*) су повезани концепти у смислу да су два објекта међусобно ближа уколико је удаљеност или разлика између њих мала, а сличност велика. Дакле, појам удаљеност, то јест, одстојање је супротан од појма сличности, тако да ако је веће одстојање између објекта они су мање слични, и обратно.

Полазна основа у анализи груписања је матрица података дата у Потпоглављу 5.3. Примена одређених алгоритама за груписање, захтева трансформисање ове матрице у матрицу блискости (енгл. *proximity matrix*). Наиме, неки методи груписања захтевају матрицу блискости и матрицу оригиналних података, док други захтевају само матрицу блискости. Квадратна и симетрична матрица блискости димензија $n \times n$ формира се на основу матрице података, а њену структуру чине индикатори сличности или разлика одређени за све парове n објеката.

[46] У складу са називом метода, развијена је и специфична терминологија из ове области. Групе које се формирају називају се кластери, а поступак разврставања података у групе сличних јединица посматрања кластерисање. Груписање се често назива и сегментација, а формиране групе сегменти (нарочито у домену маркетинга). Уобичајено је да се за јединице посматрања користи термин објекти, подразумевајући под тим појмом све видове јединица посматрања у изворном, статистичком смислу. У основи, термин кластер потиче од енглеске речи *cluster* која означава групу – скупину истоврсних ствари, грозд. Поред коришћења термина група–кластер и груписање–кластерисање у квантитативном смислу, постоје и други контексти коришћења ових концепата. У питању је облик тржишног повезивања пословних субјеката исте, сродних или различитих делатности, обухватајући сегменте од производње до пласмана, уз адекватну сарадњу са државним, образовним и научно-истраживачким институцијама ради остварења заједничких интереса. Неки од српских кластера су: Аутомобилски кластер Србија, Кластер сирева Југ, Кластер Шумадијски цвет (за подршку развоју цвећарства у Шумадији и Поморављу) итд.

Нека су профили *i*-тог и *j*-ог објекта представљени у форми *p*-димензионалних вектора: $\boldsymbol{x}_i = [x_{i1}, x_{i2}, ..., x_{ip}]$ и $\boldsymbol{x}_j = [x_{j1}, x_{j2}, ..., x_{jp}]$.

Уколико се мером блискости мери разлика између објеката, тада матрица блискости постаје матрица различитости, *D*, а њени елементи су мере разлике, d_{ij}, које изражавају одстојање између *i*-тог и *j*-ог објекта. Из математичког угла, свака функција различитости, d_{ij}, која испуњава следеће услове (метричке аксиоме) за све тачке *i* и *j*:

- $d_{ij} \geq 0$: услов не-негативности – одстојање је не-негативан број и то, већи од 0 ако се објекти разликују (али, ближи 0 када су објекти више слични), а једнак 0 само ако су објекти идентични,
- $d_{ii} = 0$: одстојање објекта од себе самог је 0,
- $d_{ij} = d_{ji}$: услов симетричности - одстојање је симетрично, и
- $d_{ij} \leq d_{ih} + d_{hj}$: услов триангуларности – директно одстојање између објекта *i* и објекта *j* не сме бити веће од одстојања од објекта *i* до објекта *j* преко објекта *h*,

назива се функција одстојања, мера одстојања или метрика (*Dalbelo Bašić*, 2011, стр. 397; *Kovačić*, 1994, стр. 259). Еуклидско одстојање (енгл. *Euclidean distance*) је једна од најпознатијих метрика. Важно је истаћи да је концепт различитости шири од концепта одстојања, тако да нису све мере различитости метрике, односно постоје мере које не поседују сва метричка својства.

Алтернативно, уколико се мером блискости мери степен сличности, тада матрица блискости постаје матрица сличности, *S*. Њени елементи су мере сличности (засноване, на пример, на коефицијенту корелације), s_{ij}, које изражавају сличност између променљивих, а које се могу користити приликом груписања објеката тако што се мера сличности прерачуна / конвертује (под одређеним условима) у меру одстојања између *i*-тог и *j*-ог објекта. Мера блискости је мера сличности између објеката *i* и *j* ако испуњава следеће услове (*Kovačić*, 1994, стр. 259):

- $0 \leq s_{ij} \leq 1$: услов нормираности и не-негативности[47], односно, мера сличности се креће у интервалу [0, 1], указујући на различите степене сличности између објеката,
- $s_{ij} = 1$: мера сличности је једнака јединици само ако оба објекта имају идентичне вредности за све променљиве, док вредност 0 имплицира да се објекти максимално разликују у погледу свих променљивих,
- $s_{ii} = 1$: из услова нормираности произлази да сличност објекта са самим собом је једнака јединици, и

[47] Да би се испунио услов нормираности, при коришћењу коефицијента корелације за сврхе утврђивања сличности узима се његова апсолутна вредност или се од његове вредности одузме јединица, а добијена разлика подели са 2.

- $s_{ij} = s_{ji}$: услов симетричности.

Симетричне матрице **D** и **S**, са нулама, односно јединицама на главној дијагонали, респективно, представљене су на Слици 15. При томе, важе следеће релације: $d_{ij} = d_{ji}$ и $d_{ii} = 0$, односно $s_{ij} = s_{ji}$ и $s_{ii} = 1$.

$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1p} \\ d_{21} & 0 & \cdots & d_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 1 & s_{12} & \cdots & s_{1p} \\ s_{21} & 1 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & 1 \end{bmatrix}$$

Слика 15: Матрица различитости (**D**) и матрица сличности (**S**)

Генерално, постоје бројне мере за изражавање различитости и сличности између парова објеката (често се говори о стотинама мера), а које се, оквирно, могу класификовати у две групе: мере различитости и мере сличности (*Romesburg*, 2011, стр. 264). Сличност / различитост која је утврђена применом једне мере не мора се подударити са резултатима и редоследом који произлазе по основу коришћења друге мере. Објективно најбоља мера блискости између објеката не постоји, као ни опште правило за њихову примену. У сваком случају, тип променљиве је кључни фактор који детерминише групу мера која се може применити у одређеној ситуацији. Наиме, уколико су све променљиве којима је дефинисан профил објеката истоврсне, потребно је применити неку меру карактеристичну за тај тип променљивих, а уколико је у питању комбинација различитих променљивих потребно је применити решења предложена за такве случајеве. Детаљну дискусију о различитим мерама блискости и формулама за њихово израчунавање видети у: *Jain & Dubes* (1988, стр. 11-19); *Shmueli et al*. (2005, стр. 208-214); *Han et al*. (2012, стр. 67-79).

У наставку текста укратко је представљено Еуклидско одстојање, као најчешће коришћена мера одстојања између објеката. Еуклидско одстојање између објеката *i* и *j* за *p* нумеричких променљивих се дефинише путем следећег израза:

$$d_{ij} = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2} \,. \qquad (1)$$

Ово одстојање се може одредити на основу оригинално измерених опсервација и стандардизованих података, али и као просечно и пондерисано одстојање. У практичним истраживањима често се користи квадрат Еуклидског одстојања. Обзиром да је Еуклидско одстојање широко коришћена мера одстојања, важно је апострофирати релевантност следећих њених особина (ограничења) (*Shmueli et al*. (2005, стр. 210-213):

▸ осетљивост на различитост мерних јединица у којима се исказују променљиве (стога је, пре израчунавања Еуклидског одстојања, конвертовање свих мерења на исту скалу путем стандардизације уобичајено решење), ▸ у потпуности игнорише везу између променљивих (тако да је у случајевима високе корелације између опсервација променљивих, бољи избор за мерење одстојања нека друга мера, на пример *Mahalanobis*-ово одстојање), и ▸ осетљивост (одсуство робустности) на утицај екстремних вредности и присуство шумова у подацима (заправо, уколико су присутне нестандардне опсервације препоручује се употреба робустније мере, попут *Manhattan* одстојања). Поред одређивања одстојања између два објекта, Еуклидско одстојање је могуће одредити и за сваки објекат у односу на његов центроид.

Поред мера блискости између објеката, за потребе анализе груписања дефинишу се и мере базиране на одстојању између група објеката. Ради презентације идеја које се налазе у основи различитих начина одређивања одстојања између група објеката, нека се посматрају две групе: група C_1 која садржи n_1 објеката (i=1, 2,..., n_1) и група C_2 која садржи n_2 објеката (j=1, 2,..., n_2). Нека се опсервације променљивих за објекте у групи C_1 означе са x_{im}, а објекте у групи C_2 означе са x_{jm}. Такође, нека су центроиди група C_1 и C_2, а њихови елементи: $\bar{x}_1 = \left[\bar{x}_{11}, \bar{x}_{12}, ..., \bar{x}_{1p}\right]$ и $\bar{x}_2 = \left[\bar{x}_{21}, \bar{x}_{22}, ..., \bar{x}_{2p}\right]$. Концепцијски, под центроидом кластера се подразумевају средње вредности (на пример, аритметичка средина или медијана) свих посматраних променљивих за све објекте у одређеној групи. Заправо, центроид је вектор средина чије су координате просечне вредности свих променљивих за објекте једне групе.

Широко коришћене мере одстојања између две групе (где је: $i \in C_1$, а $j \in C_2$) су:

- минимално одстојање – одстојање између две групе одређује се као најмање одстојање од свих одстојања између парова објеката који припадају овим групама:

$$d(C_1, C_2) = min(d_{ij}); \quad (2)$$

- максимално одстојање – одстојање између две групе одређује се као највеће одстојање од свих одстојања између парова објеката који припадају овим групама:

$$d(C_1, C_2) = max(d_{ij}); \quad (3)$$

- просечно одстојање – одстојање између две групе одређује се на основу просечног одстојања свих парова објеката из двеју посматраних група:

$$d(C_1, C_2) = \bar{d}_{ij} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{ij}; \quad (4)$$

- одстојање између центроида две групе:

$$d\,(C_1, C_2) = d\,(\bar{x}_1, \bar{x}_2), \tag{5}$$

при чему се међусобно одстојање група одређује као квадрат Еуклидског одстојања између њихових центроида путем следећег израза:

$$d_{12}^2 = \sum_{m=1}^{p} (\bar{x}_{1m} - \bar{x}_{2m})^2 . \tag{6}$$

За израчунавање прве три мере користи се матрица различитости, а одређивање одстојања између центроида група, поред ове, захтева и оригиналну матрица података.

Разумевање основног принципа груписања објеката, који гласи: минимизирање суме квадрата одступања унутар и максимизирање суме квадрата одступања између група (или, максимизирање сличности унутар и минимизирање сличности између група), захтева дефинисање суме квадрата одступања унутар група, суме квадрата одступања између група и укупне сума квадрата одступања (енгл. *within-cluster sum of squares* [*W*], *between-cluster sum of squares* [*B*], *total sum of squares* [*T*], респективно). Ради објашњења наведених концепата, нека је *n* објеката подељено у *k* група (*h*=1, 2,..., *k*) и нека је симболом n_h представљен број објеката у *h*-тој групи ($\Sigma n_h = n$), а центроид *h*-те групе са елементима за *p* променљивих у форми израза: $\bar{x}_h = \left[\bar{x}_{h1}, \bar{x}_{h2}, ..., \bar{x}_{hp}\right]$. Суштину ових концепата и релације између њих илуструје Слика 16.

Сума квадрата одступања унутар сваке групе *h* представља суму квадрата одступања опсервација свих n_h објеката групе, $\boldsymbol{x}_{im}$, од центроида групе, односно:

$$W_h = \sum_{i=1}^{n_h} \sum_{m=1}^{p} (x_{im} - \bar{x}_{hm})^2 . \tag{7}$$

По аналогији са анализом варијансе, ова сума је израз варијабилитета унутар група и назива се сума квадрата грешака (енгл. *error sum of squares*). Заправо, квадратна грешка *h*-те групе је сума квадрираних Еуклидских одстојања између сваког објекта у односној групи и њеног центроида.

Укупна сума квадрата одступања унутар група, то јест, укупна квадратна грешка целог простора који садржи *k* група је сума индивидуалних одступања сваке групе и представља се путем следећег израза:

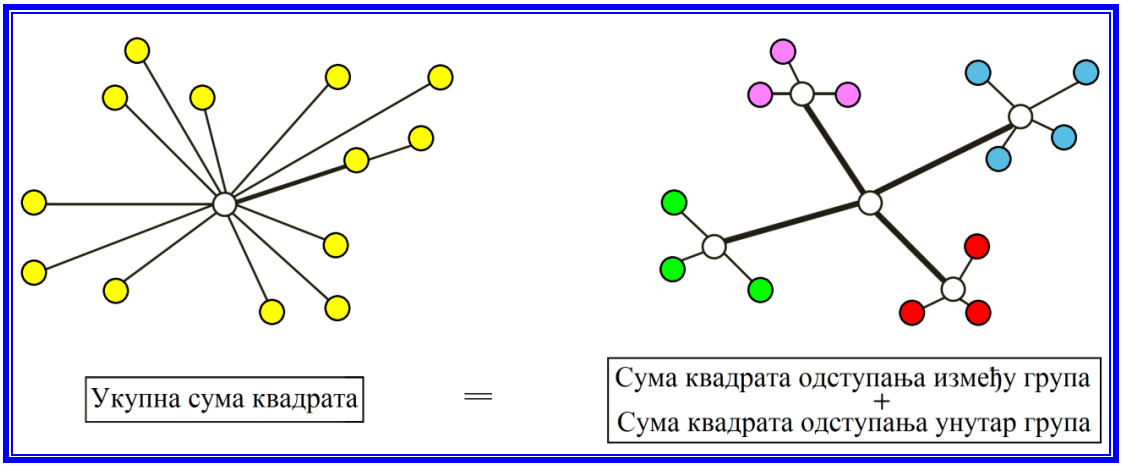
$$W = \sum_{h=1}^{k} W_h . \tag{8}$$

Сума квадрата одступања између група представља суму пондерисаног квадратног одступања средина група, $\bar{x}_{hm}$, од кореспондирајућег општег просека, $\bar{x}_m$, и представља се путем следећег израза:

$$B = \sum_{h=1}^{k} \sum_{m=1}^{p} n_h (\bar{x}_{hm} - \bar{x}_m)^2 . \qquad (9)$$

Укупна сума квадрата одступања је сума квадрата одстојања индивидуалних опсервација свих *n* објеката од кореспондирајуће опште средине, односно, симболички:

$$T = \sum_{i=1}^{n} \sum_{m=1}^{p} (x_{im} - \bar{x}_m)^2 , \text{ или } \quad T = W + B. \qquad (10)$$



Слика 16: Укупна сума квадрата и сума квадрата одстојања унутар и између група
Извор: *Tufféry* (2011, стр. 243)

Управо, на основама различитих начина мерења одстојања између група и претходно дефинисаних концепата издиференцирали су се бројни алгоритми за спровођење анализе груписања.

9.1.2. Методолошки оквир за креирање модела груписања

Спровођење анализе груписања, као метода за класификацију објеката у релативно хомогене групе, обухвата следеће кораке (*Hair et al.*, 2010): ▸ формулисање истраживачког проблема, ▸ одређивање оквира истраживачког пројекта, ▸ провера претпоставки, ▸ избор процедуре груписања и процена броја група, ▸ интерпретација формираних група, и ▸ оцењивање поузданости и валидности резултата груписања и профилисање група.

У оквиру првог корака спровођења анализе груписања неопходно је дефинисати истраживачки проблем, прецизирати циљеве анализе груписања у контексту дефинисаног истраживачког проблема и одредити променљиве које ће бити коришћене за груписање објеката. Заправо, при избору варијабли истраживач мора имати у виду значај укључивања оних променљивих које репрезентују својства објеката, а односе се на циљеве анализе груписања.

Дефинисање оквира истраживачког пројекта се односи на избор узорка за анализу, избор начина мерења блискости између објеката (имајући у виду природу података, изабрани метод за повезивање објеката у групе и циљеве истраживања), одређивање дескриптивних статистичких мера и идентификовање присуства екстремних вредности, доношење одлуке о трансформацији променљивих и третирању ненумеричких променљивих.

Важно својство анализе груписања се односи на чињеницу да она није метод утемељен на теорији статистичког закључивања. Међутим, уколико је у конкретном случају констатована теоријска и логична оправданост структурирања објеката у групе, да би резултати анализе груписања заиста били смислени и поуздани, неопходно је испунити одређене претпоставке у погледу репрезентативности узорка и одговарајућег степена корелисаности варијабли.

Након решења претходних питања, следи избор и примена одговарајућих метода за формирање група. Значај овог питања још више долази до изражаја с обзиром да различити методи груписања примењени на истим подацима углавном дају различите резултате. Сходно томе, правилан избор методе захтева познавање карактеристика и разумевање предности и недостатака различитих начина за спровођење анализе груписања. Са применом метода груписања непосредно је повезано питање избора коначног броја група који је оптималан из перспективе конкретног предмета истраживања. При томе, неопходно је постићи решење које представља компромис између захтева за формирањем хомогених група (који по правилу повећава број група) и захтева за лакшим управљањем великим количинама података путем формирања њихових репрезената у форми смислених група (који по правилу смањује број група).

Интерпретација резултирајућих група означава анализу сваке групе у смислу утврђивања њених основних карактеристика. Процес интерпретације најчешће започиње тумачењем, а затим и упоређивањем центроида група, како би се формулисали прелиминарни закључци и сугестије за даљу анализу. Од интерпретације резултата практично је неодвојиво вредновање поузданости и валидности добијених резултата, укључујући и профилисање група. Оцењивање поузданости и валидности резултата груписања подразумева анализу могућности примене резултата груписања на целу популацију и генерализовања закључака, као и оцену стабилности добијеног решења током времена. Профилисање група, као надградња и комплемент интерпретације резултата, означава наставак тумачења карактеристика сваке групе да би се потпуније објасниле разлике између група према релевантним варијаблама. Због

тога, за дефинисање профила група истраживачи, поред променљивих које су коришћене за изградњу модела груписања, користе и варијабле које претходно нису биле укључене у процедуру груписања. Такође, уобичајено је да се при интерпретацији и валидацији резултата истраживачи не ослањају искључиво на добијене резултата, већ укључују резултати других експеримената, као и постојећа сазнања експерата из конкретног подручја.

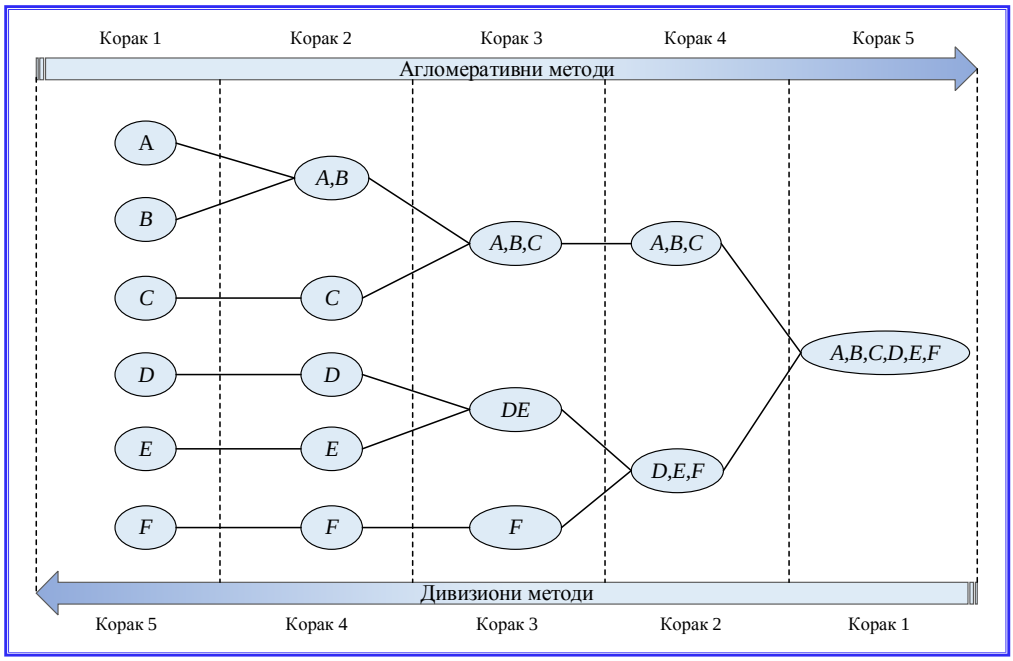
Нестручно спровођење представљеног поступка груписања и одсуство јасно дефинисаних истраживачких хипотеза може довести до безначајних резултата и погрешних закључака, јер ће, неспорно, свака процедура груписања увек, независно од смислености и логичности постојања било какве структуре међу варијаблама, резултирати одређеним групама. Стога, само правилна примена анализе груписања подржана стручним знањем истраживача омогућава откривање валидних законитости, то јест, скривених структура у улазним подацима у форми група којима подаци припадају. Због изузетног значаја добијања смислених и интерпретабилних група, у наставку овог Поглавља укратко су представљене класичне процедуре груписања (које карактерише висока фреквентност употребе и доступност у бројним софтверским алатима отвореног кода), док је у Потпоглављу 11.2. посвећена пажња критеријумима за оцену поузданости и валидности добијених резултата груписања.

У основи, разликују се две процедуре груписања: хијерархијска и нехијерархијска (енгл. *hierarchical and non-hierarchical*) (*Hair et al.*, 2010). Хијерархијско груписање подразумева формирање низа угњеждених група организованих у облику хијерархијског стабла. При томе, зависно од смера развијања хијерархијске структуре објеката, разликују се агломеративно и дивизионо груписање.

Методи агломеративног груписања (енгл. *agglomerative clustering*) су засновани на итеративном удруживању објеката и група. При томе се полази од претпоставке да је сваки објекат савршено хомогена група, тако да на почетку анализе има онолико група колико и објеката, то јест, n једночланих група. У следећем кораку, на основу израчунате матрице блискости, према утврђеном критеријуму, одређује се и спаја најсличнији пар објеката, а затим се, у наредним корацима формирају нове групе повезивањем већ формираних група или објеката. Поступак се понавља n-1 пута, све док сви објекти не буду елементи једне групе. Методи дивизионог груписања (енгл. *divisive clustering*) су засновани на итеративном дељењу (раздвајању) група и објеката, при чему се полази од једне групе који садржи све објекте, а завршава се са

формирањем онолико група колико има анализираних објеката, тако да на крају овог поступка груписања сваки објекат представља посебну групу.

Слика 17 илуструје оба приступа у хијерархијском поступку груписања на хипотетичком примеру од шест објеката означених симболима {A, *B*, *C*, *D*, *E*, *F*}: посматрањем приказа са леве у десну страну представљен је ток агломеративних метода, док се кретање у супротном смеру односи на методе раздвајања. Агломеративни хијерархијски методи се много чешће користе од метода раздвајања.



Слика 17: Агломеративни и дивизиони приступ у анализи груписања

Извор: Приказ аутора прилагођен према *Han et al.* (2012, стр. 460)

У складу са различитим начинима мерења блискости, развијено је више агломеративних метода, од којих се издвајају (*Vercellis*, 2009, стр. 307-308):

- метод једноструког повезивања (енгл. *single linkage method*) – базиран на минималном одстојању између парова објеката;
- метод потпуног повезивања (енгл. *complete linkage method*) – базиран на максималном одстојању између парова објеката;
- метод просечног повезивања (енгл. *average linkage method*) – базиран на просечном одстојању између две групе;
- метод центроида (енгл. *centroid method*) – базира се на критеријуму удруживања две групе које карактерише најмања међусобна удаљеност центроида;
- метод варијансе (енгл. *variance method*) – базира се на суми квадрата одступања унутар група и суми квадрата одступања између група. Код овог метода

најпре се, за сваку групу одређују аритметичке средине по свакој варијабли, затим, за сваки објекат одређује квадрат Еуклидског одстојања од аритметичке средине групе, а након тога сумирају резултирајућа одстојања за све елементе, то јест, објекте у оквиру групе. Спајају се оне групе за које је укупна сума квадрата свих одстојања најмања. Наиме, у свакој итерацији нова група настаје од оних двеју група чијим спајањем долази до најмањег повећања суме квадрата унутар група у односу на повећање суме квадрата до којег долази спајањем ма које две групе на конкретном нивоу груписања. Најпознатији представник ове групе је *Ward*-ов метод.

Процес хијерархијског груписања објеката и група објеката успешно се представља путем графичког приказа у облику стабла дијаграма, који се назива дендрограм, а који је приказан у емпиријском делу овог истраживања.

Када је реч о методима нехијерархијског груписања, објекти се разврставају у унапред одређен број група, мада постоје варијанте код којих се током поступка груписања број група мења. У начелу, број група се дефинише на основу искуства истраживача, резултата ранијих анализа, примене поступка хијерархијског груписања или одговарајућих софтверских решења, при чему се објекти (сходно циљним функцијама) могу кретати из једне у другу групу у различитим фазама анализе.

Типичан поступак нехијерархијског груписања започиње одређивањем центара група (коришћењем матрице оригиналних података), који представљају почетне тачке око којих се формирају групе. Иницијални центри група могу бити изабрани на различите начине, уз захтев да се изабране тачке међусобно довољно разликују. Након избора центара груписања, одређује се одстојање између сваког објекта и сваког центра групе, тако да се сви објекти чије је одстојање од центра једне групе мање од унапред постављеног критеријума укључују управо у ту групу. У току поступка груписања, свака реалокација објеката (прераспоређивање објеката из једне у другу групу како би се остварило побољшање у погледу интерне хомогености групе) је праћена итеративним поступком одређивања нових центара група и израчунавањем одстојања објеката наспрам нових тачака груписања. Поступак се наставља све док нова побољшања више нису могућа. Најчешће коришћен метод нехијерархијског груписања је метод *k*-средина.

Поред класичних метода груписања, развијени су бројни методи базирани на неуронским мрежама, *fuzzy* логици, генетском алгоритму и другим напредним концептима, у које се убрајају методи засновани на густини, формирању мрежне структуре, моделима, графовима, ограничењима детерминисаним од стране апликације

или корисника, затим, методи за груписање категоријских података и високо димензионалних података итд. Генерално, категоризацију метода груписања је тешко спровести, јер се многе категорије преклапају, тако да метод може поседовати својства неколико категорија.

У свакој од наведених категорија постоји мноштво подкатегорија и алгоритама, дизајнираних за побољшање одређених аспеката и решење специфичних проблема груписања (Детаљан приказ алгоритама груписања и њихових својстава представљен је у: *Andritsos*, 2002). Будући да је њихова анализа изван оквира овог истраживања, у наставку се, у информативном смислу, наводе неке од побољшаних верзије фундаменталних алгоритама груписања: ▸ *PAM* (*Partitioning Around Medoids*) алгоритам, познат под називом *k-medoids* алгоритам, представља проширење алгоритма *k*-средина и намењен је за ефикасно третирање екстремних вредности; ▸ *CLARA* (*Clustering LARge Applications*) алгоритам је усмерен на решавање проблема (временске) комплексности израчунавања *PAM* алгоритма; ▸ *CURE* (*Clustering Using REpresentatives*) алгоритам садржи примарно компоненте хијерархијског агломеративног груписања, али укључује и елементе нехијерархијске процедуре груписања, а његово главно својство је коришћење више од једне тачке у свакој групи као њених репрезената; ▸ *k-modes, CACTUS* (*Clustering Categorical Data Using Summaries*) и *ROCK* (*RObust Clustering using linKs*) су алгоритми оријентисани на груписање категоријских података; ▸ *Two-step clustering* је алгоритам специјално дизајниран и имплементиран у статистичком софтверу *IBM SPSS*, са примарном сврхом да се спроведе анализа великих база података, омогућавајући груписање објеката не само одвојено на основу квантитативних или категоријских променљивих, већ и кроз симултану обраду променљивих на различитм нивоима мерења.

Генерално, бројност алгоритама груписања је последица методолошког прилагођавања анализе груписања новим изазовима и захтевима процесирања велике количине података, односно условима *DM* окружења. При томе, један метод може резултирати добрим решењем ако се примени на једном скупу података, а дати лоше резултате применом на другом скупу са потпуном другачијим својствима. Независно од тога, ипак постоји сагласност истраживача да добар метод груписања у *DM* апликацијама треба да поседује следећа својства (*Han et al*., 2012, стр. 446):

- високу скалабилност – прилагођеност алгоритма за рад са великим скуповима;
- могућност рада са различитим типовима варијабли;
- способност откривање група које су специфичног облика;

- могућност да прихвати инкорпорирање знања експерата из домена анализираног проблема у форми улазних параметара пре почетка рада алгоритма;
- могућност рада са подацима који садрже екстремне, недостајуће, непознате и / или погрешне вредности;
- независност од редоследа уноса података у бази и могућност инкременталног груписања (сходно ажурирању података);
- способност груписања високодимензионалних података;
- инкорпорирање специфицираних и кориснички оријентисаних ограничења при формирању група;
- интерпретабилност, разумљивост и корисност финалних резултата груписања.

9.1.3. Примена анализе груписања

Примена анализе груписања може се посматрати кроз призму: ▸ улоге анализе груписања у спровођењу *DM* задатака и њеног односа са другим методима, ▸ домена примене у решавању конкретних проблема, и ▸ предности и недостатака саме примене. Као што је већ истакнуто, анализа груписња је један од првих корака у *DM* анализи, која се може користити као самостални алат експлоративне анализе за утврђивање дистрибуције и основних својстава података и претпроцесни корак за даља истраживање група хомогених података.

У односу на друге методе за редукцију података и класификацију јединица посматрања по групама, а за потребе прецизног методолошког разграничења, анализа груписања се најчешће пореди са факторском и дискриминационом анализом. За разлику од факторске анализе, где се врши редукција броја променљивих, код анализе груписања се врши редукција броја објеката. Сходно томе, уколико је број променљивих сувише велики, један од начина за њихово редуковање је да се, најпре, примени метод факторске анализе, а затим на добијене факторе, који представљају главне димензије - променљиве разматраног проблема, примени анализа груписања. Када је реч о проблему класификације, за разлику од дискриминационе анализе где је унапред позната припадност сваког објекта одређеној групи, код анализе груписања, разврставање објеката и формирање група се спроводи непосредно из података, без коришћења претходног знања о постојећој структури података.

У настојању да се правилно интерпретира решење проблема груписања и идентификују променљиве које су допринеле управо таквом решењу, могуће је користити више приступа и над формираним групама применити различите методе.

Осим тога, идентификовање доминантних карактеристика на којима се заснива профилисање сваке групе понекад је тешко или готово немогуће без спровођења додатних и дубљих анализа (*Klepac & Mršić*, 2006, стр. 48). Широко коришћени методи у интерпретацији група и дефинисању њиховог профила су стандардни методи дескриптивне и експлоративне статистике, дискриминациона анализа, анализа варијансе, стабло одлучивања и визуелизација.

Груписање објеката налази примену у многим областима, укључујући астрономију, археологију, медицину, едукацију, психологију, лингвистику, социологију, хемију, биологију и економију. Идеја груписања је успешно реализована у форми Мендељејевог периодног система у којем су по одређеној законитости елементи систематизовани и груписани у групе са сличним хемијским својствима, независно од појављивања и развоја анализе груписања у методолошком смислу.

Интересантну и необичну примену анализе груписања илуструју следећи примери: један из области астрономије, а други из области конфекцијског дизајна (*Berry & Linoff*, 2004, стр. 353-354). Почетком XX века астрономи су истраживали однос између светлости коју емитују звезде и њихових температура и установили да се звезде могу класификовати у три изразито хомогене групе у погледу овог односа, али са знатно великим разликама између група, јер, у основи, читав низ различитих процеса генерише светлост и топлоту. Током деведесетих година XX века, америчка војска је наручила истраживање са циљем редизајнирања војничке униформе за жене, како би се смањио број конфекцијских величина и тиме допринело смањењу залиха. Ово истраживање је резултирало у дизајнирању новог система величина униформи са 20 различитих величина (групе) на основу комбинација 6 различитих димензија тела (варијабле), попут, ширине рамена, обима груди, дужине руку итд.

У домену економских истраживања анализа груписања (као метод за откривање нових знања), такође, поседује велики потенцијал. Најпопуларнија употреба анализе груписања везује се за маркетинг истраживања и сегментацију тржишта. Актуелна примена анализе груписања се односи на Интернет претраживања и изучавање друштвених мрежа, што је резултирало и великим бројем *web* базираних алата груписања (*Neha & Vidyavathi*, 2015, стр. 8). Сваки покушај који има за циљ да се дочарају потенцијали и шароликост коришћења анализе груписања не може бити окарактерисан као успешан, уколико се не укаже на њену уобичајену примену која се односи на груписање земаља (и других територијалних јединица различитог административног нивоа) према бројним економским и демографским индикаторима

(укључујући и компоненте културе – обичаје, традицију, стил живота), као и компанија према оствареним перформансама.

Као и приликом спровођења сваког другог метода, метод груписања има своје позитивне и негативне стране. Будући да је велике количине података тешко, а често и немогуће, правилно интерпретирати без претходне смислене класификације у групе, кључне предности се односе на: ▸ редукцију података и, сходно томе, свођење карактеристика великог броја података на карактеристике формираних (репрезентативних) група уз минималан губитака информација, и ▸ формулисање нових и проверу већ дефинисаних хипотеза о структури података. С друге стране, као ограничења метода груписања наводе се: ▸ дескриптивни карактер и слаба теоријска утемељеност, која се пре свега односи на теорију статистичког закључивања, ▸ примена поступка груписања ће увек резултирати одређеним групама независно од смислености постојања било какве структуре у подацима, и ▸ решење анализе груписања је врло тешко генерализовати, јер у потпуности зависи од варијабли које су коришћене као основа за мерење сличности објеката (*Hair et al*., 2010).

Са наведеним ограничењима мора бити упознат сваки корисник методе груписања како би исту успешно применио. Неспорно, корисност резултата груписања примарно зависи од знања и способности истраживача да сагледа смисленост и могућност примене анализе груписања у оквиру разматраног проблема. Такође, одлуке о примени различитих метода и различитих алгоритама истог метода, о коначном броју група и избору варијабли за груписање, у великој мери зависе од субјективних процена истраживача. Међутим, став аутора је да коментари о квалитету добијених резултата са негативном конотацијом због доминације субјективних процена и одсуства заснованости методе на теорији статистичког закључивања не могу бити прихваћени, сем у смислу упозорења о недостацима које овај метод (као и сваки други метод) поседује и о којима треба водити рачуна приликом примене. Нелогично је истицати да овај метод није „довољно статистички”, јер већина питања, етапа и одлука везује се за статистичке концепте и примену прикладног статистичког метода и алгоритма. Штавише, при самом дефинисању метода истиче се да је реч о методу мултиваријационе статистичке анализе. Упркос огромним проблемима, у вођењу поступка груписања и избору коначног решења, знање и креативност истраживача долазе до изражаја и представљају управо ону линију која раздваја добру од лоше спроведене анализе и механичког коришћења готових софтверских производа.

9. 2. Стабло одлучивања

Стабло одлучивања (енгл. *decision tree*) је један од најчешће коришћених метода за креирање класификационих и предиктивних модела у решавању предиктивних *DM* задатака класификације, предвиђања, регресије и оцењивања. Такође, овај метод се успешно користи и при реализацији задатака дескрипције, визуелизације и редукције димензионалности података. У питању је метод индуктивног закључивања која припада категорији непараметарских метода надгледаног учења.

9.2.1. Концепт и структура стабла одлучивања

Док се у домену операционих истраживања и теорије одлучивања, концепт стабла одлучивања интерпретира као графички приказ проблема избора одлуке у условима неизвесности и представља хијерархијски модел скупа одлука и природних стања (догађаја) са могућим исходима (то јест, ефектима), с друге стране, у домену *DM*-а примарна сврха стабла одлучивања је класификација одређеног скупа улазних података. При томе, на основу спроведене класификације се не доносе појединачне одлуке, већ формулише читав низ правила која ће се применити при решавању посматраног проблема. Суштински, стабло одлучивања је хијерархијски модел који се састоји од низа правила за поделу хетерогеног улазног скупа података на групе које су хомогене у погледу категорија зависне променљиве.

Заправо, као форма анализе више варијабли, стабло одлучивања је метод моделирања података путем којег се, на основу изабраног критеријума и вредности улазних варијабли, врши подела (класификација) јединица посматрања хетерогене популације на одређени број мањих, претходно дефинисаних хомогених класа (категорија, група, сегмената) излазне варијабле. Улазне променљиве су независне (предиктор) променљиве, које могу бити категоријске или нумеричке, а излазна (циљна) променљива је зависна, која, такође може бити категоријска или нумеричка. Резултирајући модел међузависности улазних и излазне варијабле графички се представља у форми стабла, што се непосредно везује и за назив самог метода.

Структуру стабла одлучивања чини хијерархијски уређен скуп чворова (енгл. *nodes*)[48] (то јест, подгрупа / сегмената података), међусобно повезаних гранама стабла (енгл. *branches*). На почетку хијерархијске структуре налази се корен стабла. Он се односи на зависну променљиву и обухвата посматране податке узорка за учење које у

[48] Термин чвор је генерички назив за елементе стабла одлучивања.

процесу креирања модела треба класификовати у одређене класе. Корен стабла нема улазне, а може имати две или више излазних грана, осим уколико сви елементи припадају истој класи када чвор нема излаз. За разлику од кореног чвора (енгл. *root node*), сви остали чворови имају једну улазну грану (*Rokach & Maimon*, 2008, стр. 8).

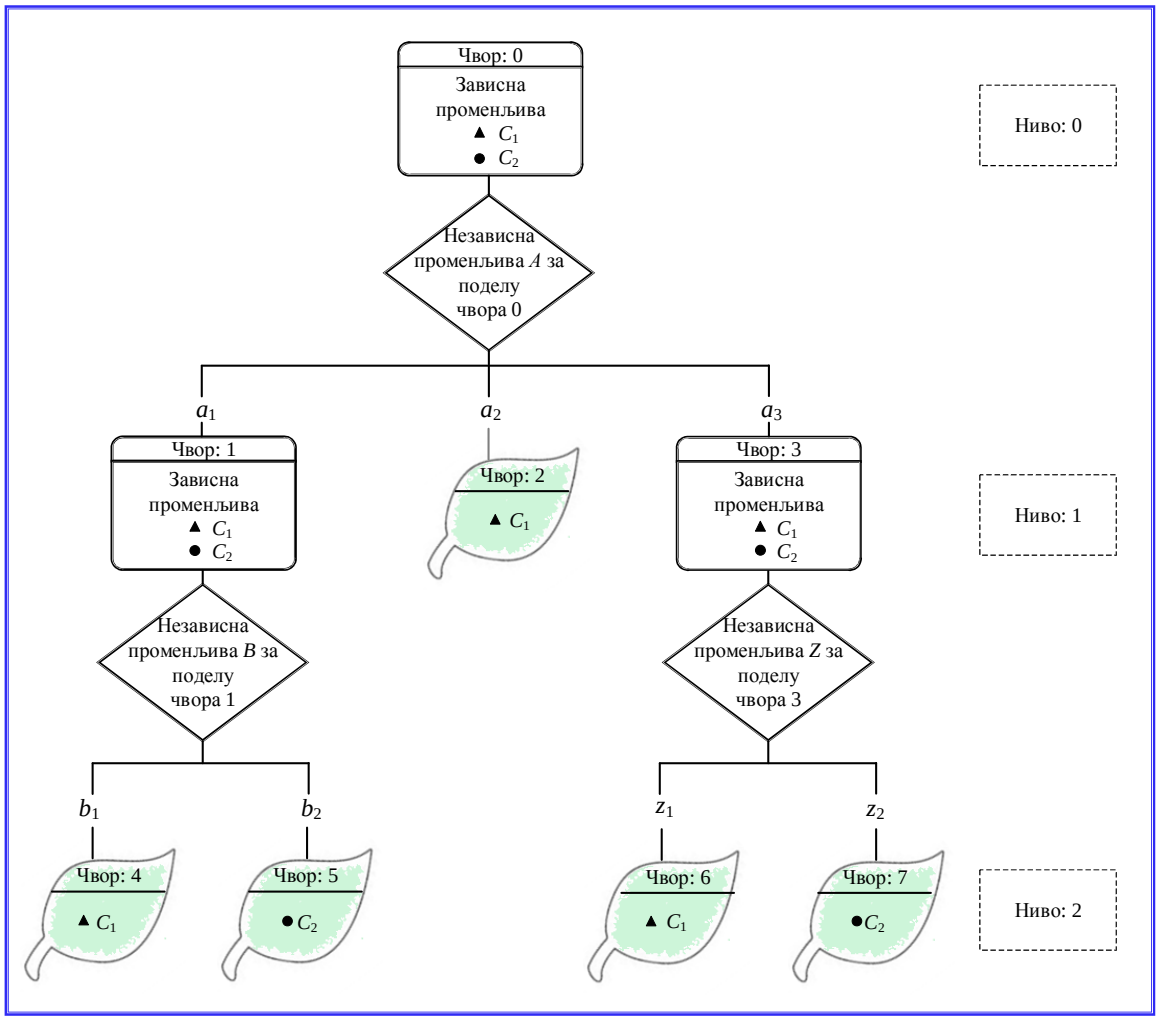
Чвор са излазним гранама назива се унутрашњи чвор (енгл. *internal node*). Сваки унутрашњи чвор кореспондира са једном улазном варијаблом и има једну улазну и, у зависности од могућих вредности (исхода) улазне варијабле, две или више излазних грана. Исходи, односно одговори (представљени гранама стабла) могу бити бинарног типа, као и избори између неколико могућих вредности или интервала вредности. Унутрашњи чворови се називају и тест чворови или чворови одлуке, јер се у сваком од њих испитује неки услов над одређеном варијаблом.

Чворови које карактерише поседовање једне улазне гране и одсуство излазних грана називају се завршни или терминални чворови (енгл. *terminal node*), а често се. називају и „листови" (engl. *leaves*) стабла. Завршни чворови представљају могућа решења проблема (то јест, вредности зависне варијабле).

Дакле, свака грана, као елемент структуре стабла се завршава чвором одлучивања или завршним чвором. Чворови и гране, као елементи структуре стабла, су истовремено и елементи знања о односу између зависне варијабле и улазних варијабли које се формулише у облику (класификационих) правила ако-онда. У основи, стабло одлучивања обухвата све посматране податке, а њихова подела на хомогене групе спроводи се на путањи од корена до неког од завршних чворова стабла. При томе, свака путања од корена до листа представља једно правило. Управо, сврха формирања стабла одлучивања јесте генерализација структуре података кроз утврђивање скупа логичких ако–онда правила, који ће омогућити прецизну класификацију постојећих, као и нових података и / или предвиђање вредности зависне променљиве.

Графички приказ стабла одлучивања, који се може посматрати као „структура секвенцијалних питања и одговора на та питања од врха до краја стабла" (*Panian i drugi*, 2007, стр. 201), представљен је на Слици 18. При конципирању ове илустрације претпостављено је да вредности зависне променљиве Y припадају двема класама, које су означене симболима C_1 и C_2. Из скупа независних варијабли, за формирање дисјунктних подскупова и поделу јединица посматрања изабране су варијабле A, B и Z. За сваку од њих су, из скупа вредности које оне појединачно узимају, одређене тачке поделе опсервација и то: $A=\{a_1,a_2,a_3\}$, $B=\{b_1,b_2\}$ и $Z=\{z_1,z_2\}$. Визуелном анализом лако се могу генеристи следећа правила: ▸ Ако је $(A=a_2)$, тада је $(Y=C_1)$; ▸ Ако је $(A=a_1)$ и

($B=b_1$), тада је ($Y=C_1$); ▸ Ако је ($A=a_2$) и ($B=b_2$), тада је ($Y=C_2$); ▸ Ако је ($A=a_3$) и ($Z=z_1$), тада је ($Y=C_1$); ▸ Ако је ($A=a_3$) и ($Z=z_2$), тада је ($Y=C_2$).



Слика 18: Елементи хијерархијске структуре стабла одлучивања

Стабла одлучивања се могу поделити према различитим критеријумима. Један од њих се односи на тип зависне променљиве, тако да се са тог становишта разликују класификациона и регресиона стабла одлучивања. Класификациона стабла се користе када је зависна променљива категоријска, док се за класификационе и предиктивне проблеме нумеричке зависне променљиве користе регресиона стабла, при чему је циљ пронаћи најбоље прилагођени регресиони модел зависне променљиве као функцију (значајних) предиктор варијабли. За разлику од класификационог стабла, где су вредности завршних чворова категорије зависне променљиве, у регресионом стаблу вредности завршних чворова се одређују као просечне вредности зависне променљиве оног дела узорка за тренирање који припада том чвору. Дакле, предвиђање се добија као просечна вредност зависне променљиве за елементе тог чвора.

У наставку текста пажња је усмерена на одређене аспекте процедуре дизајнирања класификационог стабла. Реч је о општој процедури која се примењује и при формирању регресионог стабла. Такође, поред наведене разлике између ова два типа стабла, апострофирају се и оне разлике које се односе на критеријуме поделе и мере (не)чистоће чворова. У непосредној вези са овим методолошким питањима је оцењивање перформанси креираних модела, о чему се дискутује у Поглављу 11.

9.2.2. Методолошки оквир и кључна питања у формирању стабла одлучивања

У методолошком смислу, формирање стабла одлучивања се спроводи путем процедуре рекурзивног дељења (енгл. *recursive partitioning*) података у групе, али тако да се максимизира хомогеност (минимизира ентропија) у погледу зависне променљиве у свакој од добијених група. Процес учења се одвија поређењем могућих подела, најпре, целог скупа посматраних података на основу сваке независне променљиве и избором најбоље поделе према одговарајућем критеријуму. Након тога, стабло се развија узастопним понављањем наведених поступака у сваком чвору на сваком нивоу хијерархијске структуре. Процес се завршава када се добију хомогене (чисте) групе или испуни неки од дефинисаних критеријума за заустављање рекурзије. Сагласно са наведеним, први слој грана и унутрашњих чворова приказује независну променљиву која је у најјачој вези са зависном променљивом. Дакле, са аспекта конкретног проблема, независне променљиве које су у најјачој вези са зависном променљивом налазе се обично при врху хијерархијског стабла (*Shmueli et al*., 2005, стр. 125).

Развој класификационог модела кореспондира са фазом учења модела и заснива се на рекурзивном поступку хеуристичке природе. За примену метода стабло одлучивања, потребно је да буду испуњени следећи услови (*Kantardžić*, 2011, стр. 173):

- Подаци који се анализирају морају бити приказани кроз формат *flat* табеле, тако да се свакој јединици посматрања придружује одређени број парова у облику „променљива - њена вредност”. У питању је коначан број променљивих са различитим дискретним или континуираним вредностима. При томе, све јединице посматрања се морају представити путем истих променљивих, тако да композиција (структура) променљивих остаје непромењена.
- Класе за деобу података морају бити *a priori* дефинисане, а њихов број коначан.
- Класе морају бити дискретне и међусобно јасно раздвојене, тако да одређена опсервација може припадати само једној од постојећих класа. Подразумева се да је број класа много мањи од броја података.

- Да би се идентификовале валидне правилности скривене у подацима мора се обезбедити расположивост довољне количина података. Количина података потребна за формирање поузданог модела зависи од бројних фактора, попут, броја променљивих, броја класа и комплексности самог класификационог модела.

- Метод стабло одлучивања заснован на индуктивном закључивању и резултирајући класификациони модел морају обезбедити логичан, сажет (редукован) и поуздан приказ класа у форми стабла и логичних правила (услова) одлучивања, чију основу чине констатације у погледу вредности појединих променљивих.

Представљени општи оквир рекурзивног дељења података за генерисање стабла одлучивања подразумева да буду решени одређени методолошки проблеми и специфицирани параметри модела пре или током имплементације конкретног алгоритма. Реч је питањима која се односе на (*Vercellis*, 2009, стр. 239): ▸ правила за поделу хетерогене популације на мање хомогене групе, ▸ критеријуме за заустављање рекурзивног процеса, и ▸ скраћивање раста стабла.

При формирању стабла одлучивања једно од кључних питања се односи на прецизирање критеријума за избор како променљиве која обезбеђује најбољу поделу посматраних опсервација на одређени број класа, тако и за дефинисање могућих вредности променљиве као тачака поделе опсервација које припадају конкретном чвору. Процес избора нове променљиве и поделе опсервација понавља се за сваки чвор одлучивања, при чему се узимају у обзир само опсервације које припадају том чвору. Сходно броју могућих вредности променљиве која има највећу моћ у подели одређеног чвора, одређује се број излазних грана чвора (две или више). У том смислу, подела променљиве на две класе назива се бинарна подела, а стабло одлучивања чији сваки чвор садржи не више од две излазне гране назива се бинарно стабло.

Начин поделе променљиве је детерминисан типом променљиве. За категоријску променљиву могуће су следеће варијанте:

- уколико променљива по природи ствари узима две могуће вредности (бинарна категоријска променљива), на пример, {0, 1} или {не, да}, тада се све вредности садржане у једном чвору деле у две класе, које одговарају наведеним модалитетима;

- уколико променљива узима више од две вредности, тада је могуће спровести поделу у: ▸ две класе кроз различите комбинације могућих вредности, ▸ више класа чији је број еквивалентан броју могућих вредности, или ▸ више класа формирањем група могућих вредности, нарочито ако је број вредности променљиве велики.

Уколико се ради о квантитативној променљивој, формирање класа се спроводи дискретизацијом, то јест, поделом рангираних опсервација на следеће начине:

- у две класе, тако што се (од стране аналитичара или алгоритамски) детерминише гранична вредност, при чему једној класи припадају све јединице посматрања чија је вредност посматране варијабле мање или једнака граничној вредности, а другој јединице чија је вредност већа од граничне вредности;
- у више класа, одређивањем интервала једнаких ширина, фреквенција итд;
- за квантитативну променљиву са малим бројем прекидних вредности, број класа може бити еквивалентан броју могућих дискретних вредности.

За избор најбоље поделе постоје многи критеријуми путем којих се, на сваком нивоу формирања стабла, проналази променљива која максимизира хомогеност или чистоћу зависне променљиве у свакој добијеној групи. Уколико је реч о класификационом стаблу (када је зависна варијабла по својој природи категоријска или је извршена трансформација нумеричке варијабле у категоријску), у најпопуларније критеријуме (мере) се убрајају: *Gini* коефицијент, мера ентропије, инфомациони добитак и хи-квадрат критеријум. Основна идеја која се налази иза ових критеријума је повећање хомогености новоформираних група (чворова) у погледу категорија зависне променљиве наспрам групе (чвора) над којом је извршена подела.

Ради објашњења наведних мера, нека се симболом q означи посматрани чвор стабла, симболом n број јединица посматрања у узорку за учење које припадају чвору q, симболом C_i класе зависне променљиве, где је $i = 1, 2, \ldots, h$.
За дати чвор q, *Gini* коефицијент[49] се дефинише на следећи начин:[50]

$$I_G(q) = I(q) = 1 - \sum_{i=1}^{h} p_i^2 , \qquad (11)$$

где је p_i релативна фреквенција i-те класе зависне променљиве у чвору q, која означава однос броја опсервација i-те класе и укупног броја опсервација у узорку за учење.

Максимална вредност ове мере, која указује на највећи степен нечистоће чвора, се постиже када су све опсервације равномерно распоређене по класама. Логично, минимална вредност *Gini* коефицијента, која је индикатор најмањег степена нечистоће, се постиже када све опсервације конкретне променљиве припадају једној класи. У случају бинарне поделе, *Gini* коефицијент се креће у интервалу од 0 (потпуно чист

[49] Овај коефицијент је, изворно као меру социјалне неједнакости у друштву, дефинисао италијански статистичар *Corrado Gini* (1884-1965).

[50] Као основа за дефинисање свих мера чистоће у математичкој нотацији коришћени су изрази према *Vercellis*-у (2009), с тим што су извршене извесне симболичке измене.

чвор: $p_1 = 0$ или $p_1 = 1$) до 0,5 (равномерна распоређеност опсервација по класама: $p_1 = p_2 = 1\text{-}0{,}5 = 0{,}5$), а при подели на три класе од 0 до 2/3. Дакле, са повећањем чистоће чвора, *Gini* коефицијент се смањује. Сходно томе, као најбоља променљива и најбоља тачка променљиве за поделу бирају се оне са најмањом вредношћу овог индекса.

Друга мера је ентропија[51] датог чвора q, која се одређује путем следећег израза:

$$E(q) = I(q) = -\sum_{i=1}^{h} p_i \log_2 p_i , \tag{12}$$

где $\log_2 p_i$ представља дуални логаритам релативне фреквенције, а према конвенцији важи: $0 \log_2 0 = 0$.

Мања вредност мере ентропије је индикатор веће чистоће чвора за дељење. Максимална вредност ове мере се добија када су опсервације равномерно распоређене по свим класама (једнаке релативне фреквенције свих класа унутар чвора q), а минимална када опсервације припадају једној класи и износи 0. При томе увек важи релација: $E(q) \geq 0$. У случају бинарне поделе, мера ентропије се креће у интервал од 0 ($p_1 = 1$ или $p_2 = 1$) до 1 ($p_1 = p_2$). Максимална вредност мере ентропије (као и *Gini* индекса) се постиже при релативној фреквенцији од 0,5.

Поред дефинисаних мера (не)хомогености, односно (не)чистоће у чвору q, дефинише се и мера којом се мери ефективност одређене независне променљиве у класификацији опсервација према категоријама зависне променљиве. У питању је мера која се назива информациони добитак (енгл. *information gain*) и према којој се, у сваком кораку формирања стабла, бира променљива са највећом вредношћу добитка.

За дефинисање ове мере нека се, на основу произвољне независне варијабле A, која се појављује у узорку за учење, изврши подела чвора q на K нових чворова q_1, $q_2,\ldots,q_K$, (сходно подели варијабле A) који садрже n_1, $n_2,\ldots$, n_K елемената, респективно. Нехомогеност новоформираних чворова дефинише се путем следећег израза:

$$I(q_1, q_2, \ldots, q_K) = \sum_{k=1}^{K} \frac{n_k}{n} I(q_k) , \tag{13}$$

где је $I(q_k)$ једна од претходно дефинисаних мера нечистоће одређена за новоформирани чвор k. Овако исказана нехомогеност нових подскупова је, у основи,

[51] У теорији информација, у свом научно-истраживачком раду, *Claude Shannon* (1916-1988) је изучавао вредност, то јест, информациони садржај одређеног саопштења (поруке). Полази се од тога да се реализације одређене активности остварују у домену скупа могућих исхода, при чему за сваки исход постоји одређена вероватноћа наступања. Након реализације активности и спознаје о томе у форми информације, ситуација у погледу могућих исхода је постала одређенија (или се на постављено питање може једнозначно одговорити или се смањио број могућих одговора), а тиме се смањила и неодређеност која је претходно постојала. Дакле, информација је утицала на смањење неодређености, односно, ентропије посматраног система (*Arsovski*, 2008, стр. 10-11).

очекивана вредност хетерогености или ентропије и представља пондерисану суму мера нечистоће свих новоформираних чворова, где улогу пондерационих фактора имају релативне фреквенције опсервација из чвора q које су смештене у кореспондирајући нови чвор k. При формирању стабла, циљ је минимизирати овај израз.

Сходно наведеном, информациони добитак се дефинише путем следећег израза:

$$\Delta(q_1, q_2, ..., q_K) = I(q) - I(q_1, q_2, ..., q_K) = I(q) - \sum_{k=1}^{K} \frac{n_k}{n} I(q_k), \tag{14}$$

и представља разлику између хетерогености или ентропије чвора q, $I(q)$, и очекиване (просечне) вредности нечистоће опсервација након што је извршена подела чвора q према могућим категоријама променљиве A. Дакле, информациони добитак је очекивано смањење ентропије или хетерогености које је узроковано познавањем вредности променљиве A. Другим речима, у питању је добитак у смислу хомогености и количине информација о вредности зависне променљиве, ако су познате вредности независне променљиве A. Једноставно, информациони добитак представља разлику ентропије пре гранања и ентропије након гранања над променљивом A.

Још један критеријум, који се користи у процедурама поделе хетерогене популације и смањења варијабилности дистрибуције зависне променљиве у новоформираним групама у поређењу са групом (чвором) над којом се врши подела, јесте χ^2 статистика и кореспондентна p-вредност. Хи-квадрат статистика, заснована на разлици између емпиријских и очекиваних фреквенција, као и претходне мере, одређује се за сваку независну променљиву и сваку могућу поделу. Променљива која има најмању p-вредност бира се као прва независна променљива за поделу, јер се налази у најјачој вези са зависном променљивом. Ако је p-вредност једнака или мања од унапред дефинисаног нивоа значајности α (прихвата се алтернативна хипотеза о зависности променљивих), врши се подела чвора (групе) користећи дату независну променљиву. У супротном, посматрани чвор се сматра завршним чвором, тако да се не спроводи даља подела (*Soldić-Aleksić*, 2009, стр. 131).

Уколико је, пак, зависна променљива квантитативна, а вредности које она узима у генерисаним чворовима нумеричке, уобичајени критеријуми за мерење чистоће и поделу регресионог стабла су: варијанса (просечно квадратно одступање од средине чвора) и статистика *Fisher*-овог F теста. Основна идеја која се налази иза ових критеријума је да варијанса вредности зависне променљиве у новоформираним чворовима (након поделе) мора бити мања од варијансе у чвору поделе. Другим речима, при формирању регресионог стабла, избор независних променљивих и подела

чворова се спроводи тако да се минимизира варијансе унутар класа, а максимизира између класа (*Tufféry*, 2011, стр. 331). Када су све вредности у чвору једнаке, варијанса, а самим тим и нечистоћа чвора, је нула.

У основи, процес дељења чворова се зауставља ако су испуњени одређени услови: на пример, све опсервације узорка за учење које припадају одређеном чвору имају исту вредност зависне променљиве (припадају истој класи тако да се ради о завршном чвору), или, постоји само једна опсервација у сваком чвору, или, недостају променљиве за гранање стабла. Статистички тестови се често користе за оцењивање статистичке значајности варијабли изабраних за поделу. Такође, увођење извесних додатних ограничења у форми критеријума за заустављање процеса раста стабла омогућава да се избегну негативни ефекти формирања стабла велике сложености (разгранатости). Ови критеријуми представљају скуп правила која се користе током развоја стабла одлучивања како би се утврдило да ли је неопходно креирати више грана и формирати нове чворове или дати чвор претворити у лист (*Vercellis*, 2009, стр. 250). Заправо, у питању су следећи критеријуми и њихове комбинације (*Rokach & Maimon*, 2010, стр. 157; *Tufféry*, 2011, стр. 316): ▸ дубина стабла је достигла одређену границу, ▸ број листова (а самим тим и правила) је достигао одређени максимум, ▸ број опсервација у завршном чвору је мањи од унапред одређеног броја испод којег се сматра да чвор не треба даље делити, ▸ даља подела било којег чвора ће резултирати креирањем једног или више нових чворова са бројем опсервација испод унапред дефинисане минималне величине чворова, ▸ квалитет стабла је адекватан, а даља подела не побољшава значајно квалитет стабла, и ▸ најбољи критеријум поделе није већи од дефинисане граничне вредности.

За постављање граница у развоју стабла одлучивања постоје два кључна разлога: (а) превише разгранато стабло прилично прецизно одражава специфичности опсервација у узорку за учење (укључујући шумове и екстремне вредности, које су својствене само тим подацима), али, истовремено, поседује мању способност генерализације, односно, превелика прилагођеност модела подацима за учење обезбеђује већу тачност класификације у том узорку, али узрокује велику класификациону грешку на подацима тест узорка, као и предиктивну грешку при примени креираног модела на потпуно новим подацима; и (б) превише разгранато стабло подразумева већи број листова стабла и формулисање дубоких класификационих правила, што, последично, смањује укупну интерпретабилност резултирајућег модела.

Један од добро познатих метода који се користи за решавање проблема превелике прилагођености стабла подацима у узорку за тренирање је метод скраћивања или подрезивања стабла (енгл. *pruning tree*), који се заснива на елиминисању статистички несигнификантних грана и редундантних информација. Разликују се два начина подрезивања стабла (*Gorunescu*, 2011, стр. 181): ▸ *a priori* детерминисање раста стабла (енгл. *prepruning*), које подразумева дефинисање правила за заустављање даљег раста стабла пре него што дође до постизања потпуне прилагођености, и ▸ *a posteriori* смањење формираног стабла (енгл. *postpruning*), које подразумева да се, након формирања стабла максималне комплексности, изврши елиминисање одређених грана све док се не постигне жељена сложеност стабла. Гране које треба елиминисати одређују се поређењем тачности оригиналног и редукованог стабла. Ако одстрањивање гране смањује тачност модела, тада се грана задржава у конфигурацији стабла. Обично се одређује више редукованих стабала, а затим међу њима бира оно које карактерише највећа тачност, то јест, најмања класификациона / предиктивна грешка.

Оба приступа имају своје предности и недостатке, али, генерално, изузетно је тешко унапред дефинисати сложеност стабла која ће обезбедити постизање жељене тачности класификације. У сваком случају, трагање за стаблом оптималне сложености захтева експериментисање на подацима који нису учествовали у формирању стабла и непрекидно мерење перформанси модела (односно, оцену тачности и вредновање других аспеката квалитета модела).

9.2.3. Примена метода стабло одлучивања

На темељу решавања претходно разматраних методолошких проблема развијени су многи алгоритми за формирање стабла одлучивања. Сходно томе, ови алгоритми се примарно разликују у погледу: начина одређивања променљивих за поделу (као и одређивања њихових вредности за поделу), редоследа дељења променљивих (једне или више истовремено), броја подела у сваком чвору (бинарна наспрам вишеструке поделе), критеријума за заустављање раста стабла и начина поткресивања стабла.

У групу главних алгоритама спадају (*Tufféry*, 2011, стр. 321):

- *C5.0*: овај алгоритам (као побољшана верзија основних *ID3* и *C4.5*[52] алгоритама) је погодан за истраживање свих типова независних променљивих, заснован на ентропији као мери чистоће чвора.

[52] Позната имплементација овог алгоритма је алгоритам *J48*, као део *DM* прогламског алата *Weka*.

- *CART* (акроним пуног назива - *Classification and Regression Trees*): овај алгоритам је погодан за истраживање свих типова независних променљивих и, у својој базичној верзији, генерише бинарна стабла (сваки чвор одлучивања има две гране), а избор најбоље поделе сваког чвора заснива се на *Gini* индексу;
- *CHAID* (акроним пуног назива - *Chi-Square Automatic Interaction Detection*): овај алгоритам користи χ^2 тест за дефинисање најзначајније променљиве сваког чвора и изворно је развијен за истраживање дискретних и квалитативних независних променљивих, мада већина софтвера у којима је имплементиран *CHAID* алгоритам има опцију и за аутоматску дискретизацију нумеричких променљивих.

Стабло одлучивања, као метод откривања законитости у подацима се интензивно користи у многим областима за проблеме класификације и предикције. Полазећи од тога да је основни задатак овог метода одређивање променљивих и њихових вредности које примарно детерминишу конкретну популацију или скуп појава (*Panian & Klepac*, 2003, стр. 305), може се успешно применити у следећим случајевима: за анализу склоности потрошача, класификацију подносилаца захтева за кредит према степену ризика, оцену квалитета услуга и нивоа сатисфакције корисника услуга, разврставање електронске поште као легитимне и исправне или бескорисне и неисправне, дефинисање основних карактеристика тржишних сегмената, идентификовање доминантних карактеристика корисника полисе животног осигурања и друго.

Метод стабло одлучивања је веома погодан за комбинацију са осталим методима откривања знања, тако да се може наћи у улози „суплемента, комлемента или супститута" за традиционалне статистичке анализе, *DM* алате и технике и недавно развијене мултидимензионалне форме извештавања и анализе у подручју пословне интелигенције (*de Ville*, 2006, стр. 1). На пример, у првим фазама *DM* истраживања, овај метод може обезбедити избор релевантних улазних варијабли за спровођење регресионе анализе и развој модела за предвиђање. Такође, резултате анализе груписања је често немогуће интерпретирати без даље и дубље анализе, а метод стабло одлучивања се показао веома успешним у анализи и дефинисању профила формираних група. Стога се ова два метода заједно користе у многим пословним апликацијама.

При избору адекватног метода за конкретни проблем из широке групе класификационих и предиктивних метода, потребно је имати у виду основне предности и недостатке метода стабла одлучивања, као једног од потенцијалних решења. Основни разлози практичне атрактивности овог метода односе се на следеће (*Tufféry*, 2011, стр. 327-328): ▸ концептуалну једноставност, разумљивост креираних модела и лаку

могућност интерпретације резултирајућих правила, ▸ одсуство захтева у погледу испуњености специфичних претпоставки о моделу расподеле вероватноћа случајних променљивих, ▸ флексибилност и могућност рада са свим типовима улазних променљивих, ▸ робустност у погледу недостајућих података, шума у подацима и екстремних вредности, и ▸ једноставност коришћења, укључујући и скромне захтеве у погледу рачунарских ресурса (мали капацитет меморије и разумно време израчунавања, чак и током фазе учења), као и брзо и ефикасно класификовање непознатих података на основу креираног модела.

Неки од основних недостатака метода стабло одлучивања односе се на следеће аспекте: ▸ даје мање тачне предикције за проблеме предвиђања континуираних вредности зависне променљиве, ▸ осетљивост на промене улазних података помоћу којих се тренира модел, тако да мале промене (укључујући и промене у дискретизацији променљивих) могу узроковати различите поделе и резултирати потпуно различитим конфигурацијама стабла, и ▸ у неким ситуацијама формирање стабла одлучивања може бити рачунски веома захтеван задатак, јер укључује избор не само најбоље променљиве за поделу и најбоље тачке за поделу променљиве, већ и подрезивање стабла, које затева формирање великог броја редукованих стабала, како би се, користећи податке узорка за валидацију, извршио избор најбољег стабла.

Важно је истаћи да се при формирању стабла одлучивања из базе података углавном узима мањи број променљивих него што је садржан у бази. При укључивању променљивих из базе треба бити јако обазрив, јер променљиве које су за разматрани проблем ирелевантне могу резултирати формирањем стабла са јако лошим перформансама у погледу класификационих и предиктивних могућности (екстремни примери таквих променљивих су матични број особа, поштански број и слично, при чему разврставање на основу ових променљивих и формирање једночланих класа је потпуно бескорисно решење).

Наведеним аспектима примене стабла одлучивања, свакако треба додати и чињеницу да је овај метод доступан у већини софтверских пакета за *DM* анализу (посебно у пакетима отвореног кода). Такође, савремени софтвери имају уграђене могућности за аутоматско партиционирање узорка података на део за тренирање, валидацију и тестирање, као и за оцену релевантности креираних класификационих и предиктивних модела. Заиста, с једне стране, софтверски алати омогућавају лако „руковање" моделом током целокупног процеса његовог развоја, али не треба занемарити улогу искуства, вештина и стручности аналитичара и њихово активно

учешће у подешавању критеријума и отвореним могућности кориговања резултата до којих је дошао алгоритам. Заправо, због постојања извесних недостатака и ограничења повезаних са применом овог метода, важно је истаћи значајну улогу аналитичара при коришћењу и прилагођавању параметара метода одређеном проблемском контексту. Сагласност изабраног метода и својстава конкретног проблема је кључ за успех *DM*-а.

9.3. Неуронске мреже

Вештачке неуронске мреже[53] (енгл. *artificial neural networks*) су, као што сам назив говори, метод вештачке интелигенције, настале као резултат реализације идеје да се у процесу обраде података симулирају неуронске мреже биолошких система (мозак и нервни систем човека). Фундаментална сазнања о биолошким моделима и понашању неуронских станица људског мозга (које функционишу по принципу активације) су искоришћена за дефинисање математичких формулација и метода, који своју примену налазе у пракси издвајања законитости из података. Неуронске мреже се успешно користе за креирање *DM* модела и спровођење предиктивних задатака класификације (у суштини, модели неуронских мрежа су класификатори) и предвиђања (које се може односити на темпоралне и нетемпоралне податке), али и за реализацију дескриптивних задатака груписања и визуелизације.

9.3.1. Концепт и структура неуронских мрежа

Хронолошки посматрано, почетак развоја неуронских мрежа везује се за имена *McCullock*-а (неуропсихолога) и *Pitts*-а (логичара), који су у свом раду, публикованом четрдесетих година XX века, под насловом „Логички рачун идеја карактеристичких за нервну активност", предложили једноставан математички модел по аналогији са функционисањем неурона, као основне ћелије нервног система живих бића (*Berry & Linoff*, 2004, стр.212-213; *Tufféry*, 2011, стр. 219). Њихов рад је имао снажан утицај на друге истраживаче и профилисање неуро-рачунарства, као гране вештачке интелигенције. Прва неуронска мрежа - перцептрон, развијена од стране *Rosenblatt*-а и заснована на *McCullock-Pitts*-овом неурону, појавила се 1958. год. До наглог пораста интересовања за ову област долази половином осамдесетих година XX века које је праћено одржавањем бројних научних конференција на тему неуронских мрежа са великим бројем учесника, објављивањем стручних публикација, оснивањем истраживачких центара и реализацијом едукативних програма из области неуронског

[53] У наставку текста користи се синтагма неуронске мреже, без придева вештачке.

рачунарства. Као последица наведеног развијени су бројни алгоритми за неуронске мреже који омогућавају успешно решавање практичних предиктивних и дескриптивних проблема. Данас, у *big data* ери, неуронске мреже се интензивно примењују и као део технологије за анализу друштвених мрежа, слика и говора.

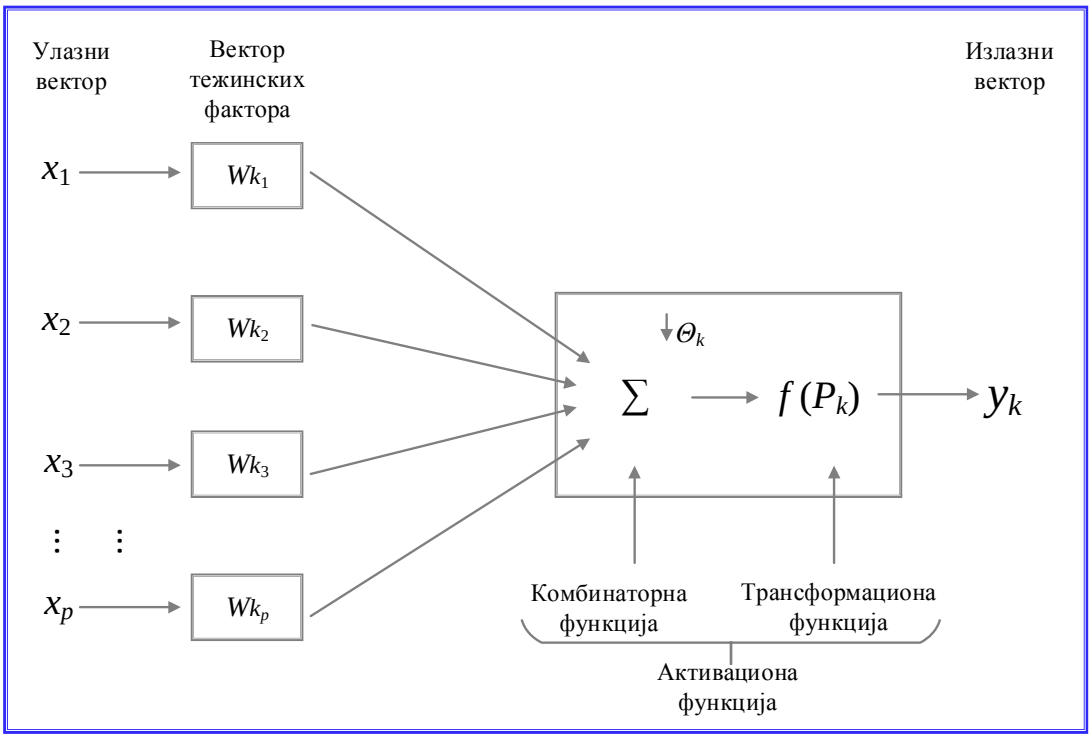
С обзиром на варијететност приступа у испитивању сличности у функционисању нервног и компјутерског сиситема од стране бројних експерата, јединствену дефиницију неуронских мрежа је врло тешко утврдити. Ипак, независно од различитих аспеката посматрања и третирања неуронских мрежа као нове информационе технологије, математичког модела, рачунарског система или колекције математичких техника, свака дефиниција полази од тога да је концепт вештачке неуронске мреже заснован на принципима функционисања биолошке неуронске мреже.

Људски мозак и целокупни нервни систем представљају биолошку неуронску мрежу састављену од 10^{11} нервних ћелија (неурона), које су међусобно повезане са око 10^{15} међусобних веза захваљујући којима се изводе процеси учења и размишљања (*Kantardžić*, 2011, стр. 200). Биолошки неурон је центар који прима и обрађује информације од других неурона, а затим шаље импулсе другим неуронима у мрежи. Аналогно биолошким мрежама, вештачке неуронске мреже су састављене од међусобно повезаних вештачких неурона. За појам вештачки неурон равноправно се користе и термини: процесни елемент, чвор или јединица. Вештачки неурон је јединица за обраду података (варијабли) која прима пондерисане улазне вредности, врши трансформацију примљених вредности према одређеној функцији и, коначно, резултира одређеним излазом (*Zekić-Sušac i drugi*, 2009, стр. 314). Дакле, сваки неурон прима улазе, процесира их и производи излаз. На Слици 19 представљен је модел вештачког неурона.

У математичкој нотацији, структуру неурона k чине следеће компоненте:

- $\boldsymbol{x} = [x_1, x_2, x_3, ..., x_p]$ – вектор улазних сигнала, то јест, вредности улазних варијабли, где је $i = 1, 2, ..., p$, тако да симбол x_{ik} репрезентује i-ти улаз k-тог неурона;
- $\boldsymbol{w}_k = [w_{k1}, w_{k2}, w_{k3}, ..., w_{kp}]$ – вектор тежинских фактора, који одражавају релативну јачину (важност) везе између сваке улазне величине и неурона (или у структури мреже између неурона), тако да симбол w_{ki} репрезентује тежински фактор неурона k повезан са i-тим улазом. Тежински фактори могу бити позитиван или негативан број, а имају исту функцију као синапсе код биолошког неурона: повезују излазе других неурона (аксоне) из околине посматраног неурона (који су истовремено улаз неурона k) са улазом функције сумирања;

- Σ – функција сумирања којом се одређује сума производа компоненти вектора улазних величина и вектора тежинских фактора;
- P_k (или, net_k) – пондерисана сума улазних вредности (пондери су тежински фактори), то јест, резултат функције сумирања као комбинаторне функције, са укљученим прагом активације неурона k, θ_k ;
- $f(P_k)$ – функција трансформације резултата функције сумирања;
- y_k – излаз k-тог неурона.



Слика 19: Структура вештачког неурона

Извор: Приказ аутора прилагођено према *Kantardžić* (2011, стр. 202)

Функционисање вештачког неурона одвија се на следећи начин (*Dalbelo Bašić i drugi*, 2008, стр. 8; *Kantardžić*, 2011, стр. 202):

- израчунавају се производи улазних вредности варијабли и кореспондирајућих тежинских фактора;
- добијени производи, као улазне променљиве комбинаторне функције (при избору облика комбинаторне функције постоји велика флексибилност, мада је најчешће у питању функција суме) се сабирају, а затим од добијене суме одузима праг θ_k , односно, симболички, пондерисана сума је:

$$P_k = x_1 w_{k1} + x_2 w_{k2} + \ldots + x_p w_{kp} - \theta_k = \sum_{i=1}^{p} x_i w_{ki} - \theta_k \,. \qquad (15)$$

Договорно, да би се поједноставио израз за пондерисану суму, праг θ_k се апсорбује увођењем додатног улаза са константном вредношћу $x_0 = 1$, који је повезан са телом неурона преко тежинског фактора $w_{k0} = -\theta_k$ (*Giudici & Figini*, 2009, стр. 77):

$$P_k = \sum_{i=0}^{p} x_i w_{ki}\,; \tag{16}$$

- добијени јединствени резултат се преусмерава, као улаз у функцију трансформације;
- применом функције трансформације на пондерисану суму (као јединствени резултат функције сумирања), P_k, добија се излаз k-тог неурона, y_k, односно:

$$y_k = f(P_k)\,. \tag{17}$$

Комбинаторна функција и функција трансформације заједно чине активациону функцију неурона (*Panian & Klepac*, 2003, стр. 315), мада многи аутори под активационом функцијом подразумевају само функцију трансформације. Комбинаторна функција генерише одређену комбинацију улаза и тежинских фактора у форми једне вредности, која се затим применом другог дела активационе функције (функције трансформације) трансформише у излазну вредност. Функције трансформације могу имати различите форме, као што су линеарна (која се у домену неуронских мрежа ретко користи), хиперболично-тангентна, а најчешће коришћена је сигмоидна (логистичка, *S*-функција), која гласи:

$$f(P_k) = \frac{1}{1+e^{-P_k}}\,. \tag{18}$$

Сходно наведеном, неуронска мрежа се може дефинисати као скуп међусобно повезаних улазно-излазних јединица за обраду података (неурона). Њено функционисање се заснива на софтверски[54] подржаном итеративном поступку учења и проналажењу везе између улазних и излазних варијабли модела на бази прошлих података, како би се за нове вредности улазних варијабли добила излазна вредност (категорија припадности или предвиђена вредност зависне променљиве).

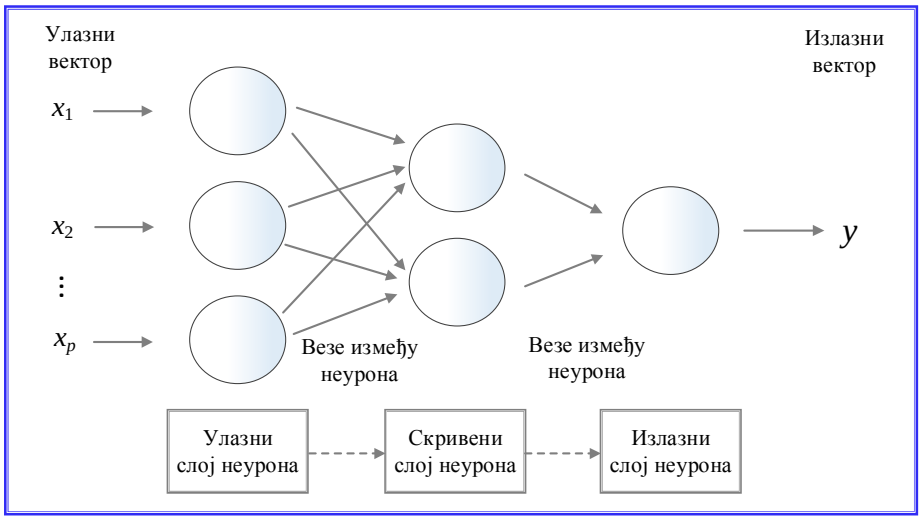
Структура сваке неуронске мреже, која се такође назива и архитектура или топологија мреже, односи се на организовање и међусобну повезаност неурона путем механизама тежинских фактора. Неурони су организовани у слојеве, при чему мрежа може имати по један улазни и излазни слој и један или више скривених слојева са различитим бројем јединица у сваком слоју. Слојеви мреже су у потпуности повезани

[54] Један од моћних, а истовремено приступачних и једноставних алата за коришћење и обликовање неуронских мрежа је *NeuroXL* софтвер. Дизајниран је као додатак за *Microsoft Excel*.

путем кореспондентних тежинских фактора, тако да сваки неурон одређеног слоја прима улазе од неурона претходног слоја и шаље своје излазе неуронима наредног слоја у структури мреже. Наиме, неурони одређеног слоја су повезани са неуронима у наредном слоју везама чија се јачина може мењати и које се користе за чување знања (*Kalinić & Marinković*, 2017, стр. 214).

Број неурона улазног слоја зависи од типа и броја независних варијабли у посматраном скупу података. Број скривених слојева, као и број неурона у сваком од њих дефинише и подешава корисник током процеса учења мреже, методом покушаја и грешке. Број неурона у скривеном слоју утиче на брзину обуке и прецизност креираног модела. Повећање броја неурона у скривеном слоју до одређеног нивоа доводи до повећања прецизности креираног модела. Међутим, након тога, свако даље повећање не само да не доводи до повећања прецизности, већ се, сходно добијању комплекснијих мрежа, смањује брзина обучавања и коришћења мреже (*Kalinić & Marinković*, 2017, стр. 214-215). Излазни слој може имати један или више неурона, зависно од типа проблема, односно задатка који се решава (*Larose*, 2005, стр. 132). Заправо, број неурона у излазном слоју једнак је броју зависних варијабли.

Структура неуронске мреже је најчешће састављена од три слоја (Слика 20). Улазни слој је једини слој који прима улазе (сигнале) из окружења. У скривеном слоју обрађују се информације и шаљу ка неурону излазног слоја. На трећем слоју добија се мрежни излаз. Овај процес се понавља у мрежи кроз онолико итерација (процес учења на подацима за тренирање) колико је потребно да се генерише излаз који је, уз унапред дефинисану толеранцију, најближи очекиваном излазу.



Слика 20: Трослојна неуронска мрежа

Извор: Приказ аутора прилагођен према *Vercellis* (2009, стр. 261)

Осим представљене структуре, у пракси се за сврхе ефикаснијег издвајања важних информације из улазног слоја, сходно карактеристикама разматраних проблема, дизајнирају и неуронске мреже сложенијих структура (са већим бројем скривених слојева, повратних петљи и слично). У основи, након утврђивања структуре мреже, активационих функција и спроведеног процеса учења, мрежа успешно обезбеђује решења проблема везаних за велике количине података, с тим што архитектура неуронске мреже управо детерминише комплексност процесирања.

9.3.2. Методологија неуронских мрежа

Значајна карактеристика неуронских мрежа је њихова способност да уче на бази искуства. У зависности од тога да ли се у процедури учења користи повратна веза околине или не, разликују се следеће парадигме (*Panian & Klepac*, 2003, стр. 317):

- надгледано учење: учење мреже се спроводи на случајевима у облику улаз-излаз, при чему су вредности излазних варијабли унапред задате, а разлике између очекиваног и стварног излаза се уклања методом повратне везе. Надгледано учење се користи углавном за решавање класификационих и предиктивних проблема. При томе, најчешће коришћена структура мреже надгледаног учења је вишеслојни перцептрон са пропагацијом грешке уназад (енгл. *feedforward back-propagation multilayer perceptron*).
- ненадгледано учење: учење мреже се спроводи без познавања вредности излазних варијабли и без појављивања повратне везе, тако да је нагласак је на самоорганизацији. У пословним применама за задатке груписања (сегментације) и визуелизације широко коришћена мрежа ненадгледаног учења је *Kohonen*-ова (самоорганизирајућа или самоадаптивна) мрежа формирања природних група (као излаза алгоритма) у структури улазних података. Ове мреже не користе функцију активације и немају скривене слојеве, већ само улазни и излазни слој.
- често се истиче и парадигма придружујућег појачавајућег учења (енгл. *reinforcement learning*), односно учења са критиком, која је евалуацијски оријентисана и у својој основи има повратну везу, али се ретко користи (углавном за проблем оптимизације током времена и адаптивно управљање) (*Bigus*, 1996, стр. 61).

Разматрања која следе су, пре свега, тангентана са поступком надгледаног учења и решавањем проблема класификације и предвиђања. Наиме, након дефинисања одређеног пословног класификационог или предиктивног проблема за који је, на основу прелиминарних разматрања његове природе, установљено да се може решити применом метода неуронских мрежа, приступа се имплементацији овог метода. Процес

имплементације обухвата следеће етапе (*Tufféry*, 2011, стр. 220-221): ▸ идентификовање улазних и излазних података, ▸ нормализација података, ▸ дизајнирање и успостављање мреже са погодном структуром, ▸ учење (обучавање, тренирање) мреже, ▸ тестирање мреже, ▸ примена (предиктивног или класификационог) модела генерисаног у процесу учења, и ▸ денормализација излазних података (укључујући и тумачење резултата).

Вештачке неуронске мреже су мреже вођене подацима, тако да квалитет резултирајућег модела зависи од квалитета (и количине) улазних података. Услед наведеног, након дефинисања улазних и излазних варијабли, избор података и њихова припрема су један од пресудних фактора за успешну примену овог метода. Подаци који се користе у неуронским мрежама морају се трансформисати у нумеричке износе одговарајућим поступцима који су детерминисани типом података. Такође, улазни подаци се морају и нормализовати. Сходно томе, њихове вредности су елементи {0,1} за бинарну променљиву или пак реални бројеви који се налазе у интервалу [0;1] или [-1;1], у зависности од ограничења активационе функције и примењених стратегија за трансформацију нумеричких варијабли и кодирање категоријских варијабли. (Више о претпроцесирању података за неуронске мреже видету у: *Larose*, 2005, стр. 130; *Shmueli et al*., 2005, стр. 163-164; *Tufféry*, 2011, стр. 223-224.)

Посебно питање у домену припреме података се односи на узорковање и поделу изабраног узорка на подузорак за тренирање мреже, подузорак за валидацију мрежа са различитим комбинацијама параметара и подузорак за тестирање мреже. При коришћењу старијих софтверских решења, поделу узорка на подузорке и однос између њихових величина углавном предлажу експерти, док новија решења, захваљујући развоју технологије, самостално раздвајају прикупљене податке на подузорке.

Критични корак у имплементационом процесу са доминантим импликацијама на резултат моделирања је детерминисање структуре неуронске мреже, као начина на који су организовани и повезани неурони у мрежи. За разликовање и класификовање структура неуронских мрежа постоје бројни критеријуми: број слојева, смер простирања, односно враћања информација од виших ка нижим слојевима, тип везе између улазних и излазних података, тип учења у мрежи, тип података, правила учења, активацијске функција, временске карактеристике итд.

Бројност критеријума и њихових комбинација, као и могућност подешавања параметара мреже резултирали су широким спектром потенцијалних структура и развојем читавог низа алгоритама неуронских мрежа. Уз унапред дефинисану структуру мреже, коначан облик мреже се одређује оптимизацијом параметара током

процеса учења. Изабрана структура мреже, такође, може се мењати током процеса тренирања мрежа кроз промену броја неурона у улазном, излазном или скривеном слоју или повећањем броја скривених слојева. Редизајнирање структуре мреже је хеуристички процес који се може одвијати у току самог процеса учења као последица добијања лоших резултата (*Klepac & Mršić*, 2006, стр. 52).

У основи, учење неуронских мрежа је процес који почиње додељивањем иницијалних вредности параметрима мреже (применом неког од правила учења или методом случајности). Након завршетка овог корака, истренирана мрежа се подвргава тестирању. Ако су резултати задовољавајући, мрежа се примењује у пракси, а уколико нису, процес се итеративно одвија променом вредности параметара све док се не добије излаз са задовољавајућим степеном тачности. Тако на пример, у процесу учења на бази алгоритма са пропагацијом грешке уназад, грешка која се одређује као разлика између стварног и очекиваног излаза, израчунава се и шаље уназад кроз систем мреже (од излазног према унутрашњим слојевима мреже). Итеративни поступак ширења грешке праћен је корекцијом параметара мреже у функцији минимизирања укупне грешке модела. Тежински фактори и вредности прага у неуронима се коригују (повећавају или смањују) обрнуто пропорционално величини грешке. Међутим, у спровођењу овог поступка треба бити обазрив, јер након одређеног броја итерација, вештачку неуронску мрежу је могуће претренирати тако да она губи својство генерализације, што резултира лошим (класификационим или предиктивним) резултатима процесирања преосталих података, односно података који нису учествовали у тренирању мреже.

Учење мреже (кроз итеративни поступак користећи различите параметре) се одвија на узорку за тренирање. Свака комбинација се тестира на подузорку за валидацију. Циљ је да се, од свих модела мреже чија је вредност грешке испод неког дозвољеног прага, пронађе онај модел који даје најбољи резултат на подузорку за валидацију. Наиме, на бази резултата подузорка за валидацију доноси се коначна одлука о најбољем моделу мреже за разматрани проблем. На крају, тако добијена мрежа се тестира на подацима који нису учествовали у дизајнирању мреже, а добијени резултат користи за оцену успешности мреже.

Суштински, циљ је „пропустити" кроз мрежу велики број случајева како би се обезбедио висок квалитет излаза. Мерила на основу којих се доноси одлука о најбољем моделу и оцени његове успешности зависи од типа проблема. У већини случајева за проблеме класификације се користи стопа погрешне класификације за сваку класу

појединачно и просечна класификациона грешка модела, док се за проблеме предвиђања најчешће користи просечна квадратна грешка, просечна грешка или просечна апсолутна грешка.

9.3.3. Примена неуронских мрежа

Неуронске мреже представљају метод вештачке интелигениције која се успешно примењује за решавање *DM* задатака у многим сегментима живота и бројним истраживачким подручјима. Уопштено, неуронске мреже омогућавају решавање таквог типа проблема код којих постоји одређене релације између предикторских (улазних) и зависних (излазних) варијабли, укључујући и присуство веома сложених веза у форми нелинеарних зависности (*Dalbelo Bašić i drugi*, 2008, стр. 9).

С обзиром на екстремну флексибилност и чињеницу да је емпиријски верификована њихова супериорност у односу на друге методе у решавању проблема код којих постоје веома сложене, како линеарне, тако и нелинеарне релације између варијабли (укључујући и проблеме са високим степеном флуктуација током времена), неуронске мреже се ефикасно примењују на широком спектру комплексних економских проблема у бројним областима, попут: финансијских апликација (предвиђање банкрота, утаја пореза, процена захтева за зајма, предвиђање солвентности, анализа монетарних трансакција, класификација акција у класе купити, продати, задржати или предвиђања кретања цена акција, обвезница, финансијских деривата и других хартија од вредности, класификација кредитних пласмана на повољне и неповољне, а клијената у групу високо или ниско ризичних корисника), анализе нових производа, процена персонала и кандидата за посао, предвиђања учинака запослених и слично. Такође, за потребе стратегијског планирања високообразовних институција, честа примена неуронских мрежа везује се за анализу и предвиђање успешности студирања кроз инкорпорирање читавог низа карактеристика студената, као улазних варијабли у процес моделирања.

Позитивни ефекти примене неуронских мрежа засновани су на низу њихових специфичних карактеристика, као што су (*Dalbelo Bašić i drugi*, 2008, стр. 9):

- погодност за моделирање сложених проблема који се одликују нелинеарним релацијама између великог броја променљивих;
- способност креирања односа између података који нису експлицитно задати, као и генерисања знања у контексту разматраног проблема кроз процес учења на бази примера (искуства);

- робустност, то јест инхерентна толерантност на неповољне услове функционисња мреже, попут присуства шумова у подацима, недостајућих података, екстремних вредности, оштећења дела мреже и прекида везе између неурона;
- адаптивност, односно поседовање уграђеног механизма за прилагођавање тежинских фактора непосредном окружењу, тако да се мрежа научена да функционише у једном окружењу може лако прилагодити променама у реалном времену;
- изузетно добре предиктивне перформансе.

С друге стране, увек треба размотрити и потенцијалне ризике и опасности које са собом носи примена неуронских мрежа, а које се односе на следеће аспекте (*Shmueli et al*., 2005, стр. 174):

- екстремна флексибилност и успешно учење захтевају, пре свега, да се обезбеди велика количина података, али, уколико подузорак за учење није довољно бројан[55], перформансе неуронске мреже биће лоше чак и када су релације између предиктора и зависне варијабле веома једноставне;
- и поред тога што поседују способност извођења општих закључака из података без деградирања перформанси модела чак и у случају недостајућих података, екстраполација представља озбиљан проблем, јер мреже не могу обезбедити валидна предвиђања изван интервала података на којем уче;
- висок степен апстрактности, јер корисницима не омогућавају стицање увида у начине, разлоге или правила подешавања тежинских фактора веза између вариајбли у процесу учења, већ само обезбеђују резултате у контексту циљева анализе;
- након одређеног броја итерација може доћи до пренаучености мреже и проналажења локалног оптималног решења, тако да она губи својство генерализације у смислу да даје добре резултате само на подацима подузорка за учење, док остале податке не обрађује успешно, што доводи до пада перформанси мреже;
- у погледу времена за тренирање и потребна израчунавања мреже су веома захтевне (нарочито при повећању броја улазних варијабли).

Приликом реализације *DM* задатака, у ланчаном процесу откривања знања из података, неуронске мреже се често комбинују са осталим методима (*Panian & Klepac,* 2003, стр. 327). Као разлог томе углавном се наводе скромне интерпретабилне могућности резултата обраде које карактеришу неуронске мреже. Коришћење додатних метода, попут неизразите логике, може знатно допринети побољшању

[55] У реализацији задатка класификације, у непосредној вези са овим проблемом је и питање накнадног узорковања како би се обезбедило довољно података за класу са најмањим учешћем.

квалитета интерпретације добијених излаза. Поред тога, неуронске мреже немају уграђен механизам за избор улазних варијабли, тако да потреба за пажљивим разматрањем и идентификацијом или ревидирањем кључних предиктора захтева комбинацију неуронских мрежа са анализом главних компоненти, класификационим или регресионим стаблом одлучивања. Поред наведених комбинација, у интегративном приступу решавању проблема, за остварење синергијских ефеката и добијање ефикасних предиктивних модела, могуће је неуронске мреже комбиновати и са генетским алгоритмом. Истовремено, ради побољшања тачности коначног модела, пожељно је на истом проблему, тестирати и упоредити резултате добијене не само применом више алгоритама и архитектура неуронских мрежа, већ и тестирати и упоредити резултате примене неуронских мрежа и других метода *DM* анализе.

Сумирањем претходно изнетог, најважније предности неуронских мрежа везују се за могућности откривања скривених нелинеарних правилности у подацима и њихову предиктивну моћ. Модели неуронских мрежа прилагођавају се променама улазних варијабли много лакше него, на пример, методи вишеструке регресионе анализе, тако да се сматрају посебно погодним у динамичким ситуацијама у којима је однос између зависне и предиктор променљивих предмет честих промена (*Lepojević & Janković-Milić,* 2008, стр. 109). Најслабија тачка, пак, односи се на апстрактност структуре мреже, као и на проблеме у настојању да се добију одговори на питања: шта се дешава у мрежи, или, који су разлози добијања одређених резултата из наученог модела. Услед тога неуронске мреже имају репутацију „црне кутије" (*Shmueli et al*., 2005, стр. 174). Међутим, критике упућене на рачун скромних могућности интерпретације могу се прихватити са становишта откривања и тумачења релација између неурона у самој мрежи и тумачења правила на основу којих су добијени излази, док, интерпретација добијених коначних резултата, посматрано из пословне перспективе и циљева анализе, не представља велики проблем (*Klepac & Mršić,* 2006, стр. 52).

9.4. Суштинска одређења осталих фреквентно коришћених *data mining* метода

Обзиром на бројност *DM* метода, практично је немогуће на једном месту направити потпуни преглед свих метода. Ипак, да би се презентовала широка апликативност *DM* приступа у анализи и решавању разноврсних задатака, у овом Потпоглављу, укратко су представљени суштински елементи изабраних популарних метода који су означени као група осталих фреквентно коришћених *DM* метода.

9.4.1. Асоцијативна анализа

Метод откривања асоцијативних правила (енгл. *association rules*) представља често коришћен метод за реализацију специфичног *DM* задатка усмереног на идентификовање интересантних фреквентних веза између елемената у репозиторијумима података, односно на проналажење фреквентних скупова (енгл. *frequent itemsets*) парова „атрибут - вредност атрибута". Законитости које су откривене путем алгоритама овог метода називају се асоцијативна правила (или, правила повезивања) и указују на то колико се често елементи одређеног скупа података појављују заједно. Реч је о дескриптивном, ненадгледаном методу груписања.

Асоцијативна анализа, као један од метода који се користи за решавање проблема код којих није дефинисана зависна варијабла, има широк распон примене. Типична примена односи се на анализу потрошачке корпе и откривање законитости о склоностима и комбинацијама производа приликом обављања куповина у малопродајним објектима. Такође, овај метод има изузетан потенцијал за анализу у домену едукативних, пословних, финансијских, *web* и медицинских података. Веома важно је истаћи да су асоцијативна правила погодна за моделирање категоријских варијабли номиналног типа. Уколико је реч о подацима нумеричког типа, пре примене овог метода неопходно је у фази претпроцесирања извршити њихову трансформацију у номиналне вредности. Асоцијативна правила не могу се користити за предикцију.

Асоцијативна анализа је повезана са употребом специфичне терминологије. Кључни термин је трансакција која се састоји од скупа елемената, то јест, ставки. Елемент (енгл. *item*) је појам који се односи на пар „атрибут - вредност атрибута". Скуп трансакција кореспондира са појмом скуп података. Подаци на које се примењују алгоритми асоцијативних правила су углавном организовани у форми трансакционих база података (Табела 2). Као јединица посматрања, трансакција представља ред у бази података. Алтернативно, база података може бити конвертована у бинарну матрицу у којој свака колона кореспондира са једним елементом, а сваки ред са једном трансакцијом, при чему свака ћелија садржи 1 или 0, у зависности од присуства или одсуства конкретног елемента у односној трансакцији (*Shmueli et al*., 2005, стр. 197).

Трансакције (односно редови) обично садрже различити број елемената, што представља значајну разлику у односу на матрицу података за потребе неких других метода моделирања. При томе, свака трансакција у скупу трансакција пружа информације о заједничком појављивању одређених комбинација елемената. На основу

тих информација формирају се табеле са бројем трансакција (односно, фреквенцијом) у којима се елементи појављују самостално, у пару са другим елементом или у комбинацији са већим бројем елемената, а које представљају основу за генерисање релевантних асоцијативних правила према одговарајућој алгоритамској процедури.

Табела 2: Једноставна форма трансакционих података и бинарне матрице

Ознака трансакције	Елементи трансакције
T_{10}	$\{e_1, e_2, e_5, e_6\}$
T_{30}	$\{e_2, e_4\}$
T_{89}	$\{e_2, e_3, e_5\}$
T_{125}	$\{e_1, e_2, e_4, e_5, e_6\}$
T_{150}	$\{e_1, e_2, e_3\}$

Ознака трансакције	Елементи					
	e_1	e_2	e_3	e_4	e_5	e_6
T_{10}	1	1	0	0	1	1
T_{30}	0	1	0	1	0	0
T_{89}	0	1	1	0	1	0
T_{125}	1	1	0	1	1	1
T_{150}	1	1	1	0	0	0

Најпознатији алгоритам за генерисање асоцијативних правила је *a priori* алгоритам (*Shmueli et al*., 2005, стр. 196). Начелно, генерисање асоцијативних правила је итеративни процес који омогућава проналажење фреквентних скупова парова „атрибут - вредност атрибута". У том контексту, кључна идеја овог алгоритма је да се, најпре, формирају фреквентни (под)скупови који се, састоје од по једног елемента, а затим, рекурзивно, фреквентни подскупови који садрже два и више елемената. У свакој наредној итерацији, елиминишу се кандидати парова који не задовољавају одговарајуће критеријуме (односно, не понављају се у довољном броју трансакција), тако да се подскупови формирају комбиновањем само елементата подскупова који су у претходним итерацијама означени као фреквентни. Заправо, идеја асоцијативних правила је да се испитају сва могућа правила повезаности између елемената у облику „ако-онда", али уз избор оних правила која су, сходно дефинисаним критеријумима, индикатори релевантне повезаности између елемената.

Основни концепти путем којих су представљена суштинска својства метода асоцијативних правила, формално се могу представити на следећи начин (*Cios et al*., 2007, стр. 292): Нека је E скуп k елемената, то јест, $E = \{e_1, e_2,\ldots, e_k\}$, а D скуп трансакција, где за сваку трансакцију T, која садржи макар један елемент (није празан скуп), важи релација $T \subseteq E$. Нека су A и B два дијунктна скупа која се састоје од једног или већег броја елемената. Трансакција T садржи A ако и само ако важи релација $A \subseteq T$. Асоцијативно правило се, као импликација (која значи истовремено догађање, а не узрочност), представља у форми $A \Rightarrow B$, где је $A \subset E$, $B \subset E$ и $A \cap B = \varnothing$.

Међутим, сва правила генерисана путем конкретног алгоритма нису подједнако значајна. Управо, кључни проблем у овом процесу моделирања је издвојити из базе података статистички значајна асоцијативна правила (*Giudici & Figini*, 2009, стр. 92). За евалуацију интересантности и релевантности резултирајућих правила, поред субјективних разматрања од стране корисника који поседују одређена знања о анализираном скупу података, најчешће се примењују мера подршке (енгл. *support*) и мера поузданости (енгл. *confidence*) правила. Заправо, асоцијативним правилима су придружене одговарајуће вероватноће.

На основу мере подршке одређује се да ли је неки (под)скуп фреквентан или не и елиминишу се правила која су мање фреквентна. Подршка скупа A (са једним или групом елемената) је релативна фреквенција, односно вероватноћа његовог појављивања у случајно одабраној трансакцији и добија се као однос између броја трансакција у којима се појављују елементи тог скупа, N_A, и укупног броја трансакција, N. Односно, симболички:

$$\text{подршка}\{A\} = \frac{\text{број трансакција које садрже } A}{\text{укупан број трансакција}} = \frac{N_A}{N}. \tag{19}$$

Консеквентно, подршка за правило $A \Rightarrow B$ је вероватноћа заједничког појављивања A и B, а добија се као однос броја трансакција које садрже A и B, $N_{A \Rightarrow B}$, и укупног броја трансакција у скупу D. Симболички, подршка за правило $A \Rightarrow B$ је:

$$\text{подршка}\{A \Rightarrow B\} = \frac{N_{A \Rightarrow B}}{N} = P(A \cup B). \tag{20}$$

Мера поузданости за правило $A \Rightarrow B$ дефинише се као однос између броја трансакција које садрже A и B и броја трансакциај које садрже само A. Другим речима, поузданост је релативна фреквенција, односно условна вероватноћа да транскција која садржи A, такође садржи и B, $P(B \mid A)$. Симболички, поузданости за правило $A \Rightarrow B$ је:

$$\text{поузданост}\{A \Rightarrow B\} = \frac{N_{A \Rightarrow B}}{N_A} = \frac{N_{A \Rightarrow B} / N}{N_A / N} = \frac{\text{подршка}\{A \Rightarrow B\}}{\text{подршка}\{A\}} = P(B \mid A). \tag{21}$$

Правила која задовољавају кориснички дефинисане вредности минималних прагова подршке и поузданости су релевантна правила и називају се јаким асоцијативним правилима (*Cios et al.*, 2007, стр. 293). Интерпретација асоцијативних правила везује се за %, тако да мера подршке правила показује % трансакција у скупу

свих трансакција које садрже и *A* и *B*, док мера поузданости правила показује који % од свих оних трансакција које садрже *A*, такође садрже и *B*.

Као и сваки други метод, метод откривања асоцијативних правила има своје позитивне и негативне стране (*Giudici & Figini*, 2009, стр. 95). Кључне предности се односе на једноставност и интерпретабилност генерисних правила, док су недостаци повезани са временом егзекуције алгоритма и трошковима анализе. С обзиром да софтверски пакети из великих база генеришу мноштво правила, након завршетка рада алгоритма, неопходно је да аналитичари, поред респектовања критеријума минималне подршке и поузданости, додатним анализама смање њихов број идентификовањем и елиминисањем бесмислених и редундантних правила.

9.4.2. *Bayes*-ови методи

У анализама заснованим на *DM* приступу, *Bayes*-ови методи се често користе у својству *DM* метода за решавање предиктивних и класификационих задатака. Ови методи припадају групи пробабилистичких статистичких метода, тако да њихова примена подразумева одређивање вероватноће неког догађаја уз услов да се догодио неки други догађај или више догађаја истовремено. У основи, претпостављају да се знање о неком догађају представља вероватноћом појаве тог догађаја. Као што сам назив говори, ова група метода је засновна на *Bayes*-овој теореми (формули), чији је основни концепт условна вероватноћа.

Уколико се посматрају два догађаја, *A* и *B*, вероватноћа остварења догађаја *A* под условом да се већ остварио догађај *B* назива се условна вероватноћа. У математичкој нотацији, условна вероватноћа представљена симболом $P(A \mid B)$, по правилу се дефинише на следећи начин:

$$P(A \mid B) = \frac{P(AB)}{P(B)}, \tag{22}$$

под условом да догађај *B* не представља немогућ догађај, односно вероватноћа његове реализације је $P(B) > 0$. Симболом $P(AB)$ представљена је вероватноћа остварења и догађаја *A* и догађаја *B*. Такође, ако је вероватноћа реализације догађаја *A*, $P(A) > 0$, условна вероватноћа $P(B \mid A)$ је:

$$P(B \mid A) = \frac{P(AB)}{P(A)}. \tag{23}$$

На основу израза (22), $P(AB)$ се може изразити као:

$P(AB) = P(A \mid B)\, P(B)$. (24)

Лева страна израза (24) је симетрична за A и B, док симетрија за десну страну није очигледна. Полазећи од израза (23), $P(AB)$ се може представити и на следећи начин:

$P(AB) = P(B \mid A)\, P(A)$. (25)

На основу наведеног произлази следећа релација:

$P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$, (26)

тако да се условна вероватноћа дефинише путем израза:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \quad (27)$$

што представља први облик *Bayes*-ове теореме (*Kadane*, 2011, стр. 89).[56]

Нека се догађај B представи путем следећег израза:

$$B = AB \cup \bar{A}B \quad (28)$$

Пошто су AB и $\bar{A}B$ међусобно искључиви догађаји (дисјунктни), вероватноћа реализације догађаја B може се представити следећом релацијом:

$P(B) = P(AB) + P(\bar{A}B) = P(B \mid A)\, P(A) + P(B \mid \bar{A})\, P(\bar{A})$. (29)

На основу израза (28) и (29) изводи се други облик *Bayes*-ове теореме, који гласи:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})}. \quad (30)$$

Нека су $A_1, A_2, \ldots, A_k$ дисјунктни скупови чија унија представља потпун простор S. Генерализацијом израза (28), добија се:

$$B = \cup_{i=1}^{k} A_i B \quad (31)$$

при чему су скупови A_iB дисјунктни. Консеквентно:

$$P(B) = \sum_{i=1}^{k} P(A_i B) = \sum_{i=1}^{k} P(B \mid A_i)P(A_i). \quad (32)$$

Генерализацијом израза (30), дефинише се трећи облик *Bayes*-ове теореме:

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P_i(A_i)}. \quad (33)$$

[56] За представљање *Bayes*-ове тероеме у потпуности је коришћен претходно референциран литературни извор, тако да се у наставку неће поново наводити.

У оквиру *Bayes*-ове теореме разликују се *a priori* и *a posteriori* вероватноће. *A priori* вероватноће су иницијалне, унапред познате вероватноће случајних догађаја које одржавају степен расположивих информација, пре прикупљања емпиријских података. Све вероватноће независних догађаја који се међусобно искључују су *a priori* вероватноће. *A posteriori* вероватноће су вероватноће одређене након спроведеног експеримента и прикупљених додатних информација. Ове вероватноће представљају *a priori* вероватноће кориговане на основу информација из узорка и посматрања мноштва случајних догађаја, тако да се називају и емпиријским вероватноћама.

Сходно наведеном, *DM* методи засновани на *Bayes*-овој теореми комбинују претходно и ново знање (у форми вероватноће) о класама зависне и независних варијабли, при чему је ново знање добијено на основу података за учење. Израчунавање тражених условних вероватноћа непосредном применом *Bayes*-ове теореме могуће је у ситуацијама када су условне вероватноће последица директног међусобног утицаја између варијабли. Међутим, за одређивање тражених вероватноћа у ситуацијама када условне зависности одражавају веома сложене односе између већег броја ланчано повезаних варијабли и њихових категорија развијене су *Bayes*-ове мреже. Реч је о графичкој структури за представљање условних вероватноћа и веза између скупова варијабли које су укључене у анализу. Заправо, *Bayes*-ове мреже се дефинишу као „графички модели који показују пробабилистичке релације засноване на условним вероватноћама између скупова варијабли" (*Klepac & Mršić,* 2006, стр. 54).

Bayes-ове мреже се графички приказују као повезани чворови који се састоје од табела условних вероватноћа. Помоћу табела (које се као елементи унутрашње структуре мреже конструишу за сваку варијаблу) израчунавају се заједничке вероватноће. Чворови у мрежи репрезентују варијабле, док гране (као линије које спајају чворове) репрезентују пробабилистичку зависност између кореспондентних варијабли (*Ben-Gal*, 2008). Дакле, случајна зависност између појединих варијабли представљена је структуром чворова. Конкретно, грана од чвора X_i до чвора X_j означава да вредност коју узима варијабла X_j зависи од вредности коју узима варијабла X_i, или, другим речима, варијабла X_i утиче на варијаблу X_j. У овом случају чвор X_i назива се родитељски чвор, а чвор X_j дете чвора X_i. Бројем родитељ варијабли које претходе конкретној варијабли детерминисана је димензионалност њене табеле вероватноће. Варијабле које немају родитеље називају се корени и, у констелацији одабраних варијабли, на њих не утиче ниједна варијабла.

Суштинска карактеристика *Bayes*-ових мрежа односи се на способност учења, које је засновано на вероватноћи. У поступку дизајнирања мреже најпре се врши одабир скупа варијабли релевантних са становишта разматраног проблема, а затим успостављају везе између варијабли уважавајући њихове узрочно-последичне односе. Сходно томе, поред статистичког знања, за креирање квалитетног модела потребна су и одговарајућа експертска знања из конкретног подручја. Генерално, на основу експертског знања одређују се везе између чворова, смерови тих веза и процењује потенцијални значај веза између одабраних варијабли (*Klepac & Mršić,* 2006, стр. 56).

У домену економских појава и процеса постоји велики број група предиктивних и класификационих проблема који се могу решавати применом ове групе метода, попут, сегметације тржишта, процене кредитног ризика, процене понашања клијената итд. Важна напомена се односи на чињеницу да анализом обухваћене варијабле морају узимати категоријске или дискретне вредности, с тим што је могуће укључити и варијабле са прекидним и континуираним нумеричким вредностима, дефинисањем оптималног броја категорија или дискретних вредности.

9.4.3. Регресиона анализа

Природа многих (варијабилних) појава указује на њихову међусобну повезаност. Посебан допринос откривању суштине и законитости међусобних веза и утицаја две или више појава везује се за статистичка истраживања и примену, вероватно најпознатијег и најчешће коришћеног статистичког параметарског метода, регресионе (и корелационе) анализе. У основи, сврха регресионе анализе је идентификовање функционалне форме (облика) међузависности варијација појава (променљивих). Сходно томе, ова група метода има своју улогу у решавању *DM* проблема.

Опште је познато да се приликом истраживања међусобних веза две променљиве примењују методи просте (линеарне и нелинеарне) регресионе (и корелационе) анализе, а у случају посматрања више променљивих методи вишеструке (линеарне и нелинеарне) регресије (и корелације). Такође, примена било којег метода регресионе анализе захтева идентификовање променљиве која има улогу зависне, а која улогу независне (или независних променљивих). Како помоћу регресионих метода није могуће идентификовати узрочно-последичне везе између променљивих, у статистичкој терминологији најчешће се уместо термина независна, користи термин објашњавајућа променљива (*Lovrić,* 2009, стр. 334-335). Будући да је мултидимензионалност инхерентно својство *DM* проблема и задатака, у наставку следи кратак осврт на

суштинска одређења стандардног модела вишеструке линеарне регресије, као и специфичности његове примене у *DM* окружењу.

Вишеструка линеарна регресија се користи за одређивање линеарне везе између нумеричке зависне променљиве Y и скупа од две или више нумеричких објашњавајућих променљивих, означених са X_m (где је m = 1, 2, ..., p). Општи облик вишеструког линеарног регресионог модела популације гласи:

$$Y = \beta_O + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon, \quad (34)$$

где су: β_0, ..., β_p, регресиони параметри модела, а ε стохастички члан.

Структура дефинисаног модела се састоји из два адитивна дела: детерминистичког и стохастичког. Детерминистички део представља линеарну функцију вишеструке регресије и показује просечан утицај објашњавајућих променљивих на зависну променљиву, а стохастички члан је случајна грешка и одражава ефекте случајних фактора и фактора који нису обухваћени моделом. Пошто се ретко располаже подацима скупа, закључивање се спроводи на подацима узорка, а регресиони параметри се могу оцењивати употребом бројних метода. Један од најчешће коришћених је метод најмањих квадрата чија је идеја да се на основу података узорка одреде најбоље могуће оцене непознатих регресионих парметара минимизирањем суме квадрата вертикалних одступања (резидуала) емпиријских вредности од моделираних вредности.

Оптимално оцењивање регресионих параметара подразумева да буду испуњене одређене претпоставке регресионог модела. При томе, неке од претпоставки се проверавају само у фази претпроцесирања података, док друге могу бити оцењене тек након креирања регресионог модела. Због изузетног значаја који наведно питање има са аспекта интерпретације модела и квалитета формулисаних закључака, у сваком релевантном статистичком извору могу се наћи таксативно наведене и објашњене претпоставке за валидно спровођење вишеструке линеарне регресије.

У начелу, циљ регресионе анализе је креирање адекватног регресионог модела који ће омогућити да се, најпре, разуме, објасни и интерпретира веза између зависне и објашњавајућих променљивих, а затим да се, на основу идентификоване функционалне релације, оцене и предвиде вредности зависне променљиве за одабране вредности објашњавајућих променљивих. Заправо реч је о експланаторним и предиктивним циљевима регресионе анализе, који се реализују кроз процесе експланаторног и предиктивног моделирања (*Shmueli*, 2010).

У основи циљ експланаторног моделирања је максимизирање количине информација о релацијама које постоје између анализираних појава у популацији, а које су претпостављене узимајући у обзир теоријске поставке о конкретном проблему у кореспондентној истраживачкој области. Наиме, у контексту регресије, у питању је тестирање хипотеза о зависној променљивој као функцији објашњавајућих променљивих. Предиктивно моделирање је, пак, процес примене статистичких модела или *DM* алгоритама за сврхе предвиђања нових или будућих опсервација. Реч је о предвиђању вредности зависне променљиве за нове вредности објашњавајућих варијабли.[57] Мада се модел вишестуке регресије користи за реализацију оба циља, постоје ставови да је класичан статистички приступ фокусиран на експланаторни, а *DM* приступ чешће на предиктивни циљ (*Shmueli et al.*, 2005, стр. 122-123).

Суштинска разлика између ова два приступа односи се на чињеницу да су подаци који се користе за креирање регресионог модела у класичној, статистичкој имплементацији метода вишеструке линеарне регресије истовремено и подаци за оцену поузданости модела. При томе се закључци и кључне информације о релевантним везама у скупу података изводе на основу ограничене количине расположивих података узорка. Насупрот томе, *DM* приступ се, по правилу, везује за огромну количину података, тако да се прилагођавање модела и одређивање оцена регресионих параметара спроводи на делу података за учење, док се оцењивање перформанси модела спроводи на новим подацима који нису коришћени за израчунавање оцена параметара. Добар експланаторни модел може имати ниску предиктивну прецизност, као што и модел који је слабо прилагођен подацима може имати добра предиктивна својства. Због ових разлога, важно је дефинисати циљ анализе пре моделирања.

У процесу развоја модела вишеструке линеарне регресије, генерално, а посебно у *DM* окружењу, веома важно питање се односи на избор подскупа предиктор варијабли које ће ефективно објаснити највећи део варијабилитета зависне променљиве. Једноставније речено: колико објашњавајућих променљивих треба укључити у модел? Или, посматрано из другог угла, формулисано питање се односи на смањење димензионалности модела (смањење броја објашњавајућих променљивих), али под претпоставком да карактеристике модела не буду нарушене. Одговор на постављено

[57] У истом раду, ауторка *Shmueli* наводи и дескриптивно моделирање као трећи тип моделирања развијен од стране статистичара. Овај тип моделирања се мање ослања на теоријске постулате и усмерен је на сумаризацију и компактно представљање структуре података. У контексту регресије, прилагођени регресиони модел може бити дескриптивног карактера уколико се примарно користи за стицање увида и знања о вези између зависне и објашњавајућих променљивих, а не доминантно за анализу зависне променљиве у функцији објашњавајућих променљивих, односно за закључивање и предвиђање.

питање је комплексан задатак, уз препоруку укључивања што је могуће мањег броја објашњавајућих променљивих које су заиста битне са аспекта зависне променљиве. У литератури је формулисано више метода за смањење димензионалности модела вишеструке линеарне регресије. Суштински, они су засновани на једном од следећа два приступа: ▸ први, подразумева потпуну претрагу простора свих комбинација објашњавајућих променљивих и избор најбољег подскупа, то јест, најбоље прилагођеног регресионог модела, и ▸ други, подразумева постепено елиминисање и / или укључивање променљивих у модел. Утицај велике количине података на својства регресионог модела (превасходно са аспекта димензионалности) и вредност статистика за вредновање његовог квалитета дискутован је Потпоглављу 11.4.

Вишеструка линеарна регресија, као поступак надгледаног моделирања, са нумеричком зависном варијаблом, се примењује у многим *DM* ситуацијама, пре свега, за реализацију задатака оцењивања и предвиђања. Међутим, често зависна променљива није нумеричка, већ квалитативна. У таквим случајевима за описивање релација између категоријске зависне променљиве (најчешће бинарне) и скупа објашњавајућих променљивх (које могу бити и категоријске и нумеричке) примењује се метод логистичке регресије. Суштински, применом овог метода, чији се параметри оцењују методом максималне веродостојности, одређује се (предвиђа) дискретна вредност (категорија) и припадајућа вероватноћа зависне променљиве, а на основу познатих вредности једне или више објашњавајућих променљивих.

Са становишта циљева анализе логистичка регресија је доста слична са вишеструком регресионом и дискриминационом анализом. Међутим, у односу на сродне мултиваријационе технике, логистичка регресија је заснована на мање рестриктивним претпоставкама, а самим тим је и флексибилнија у домену примене. У *DM* окружењу, логистичка регресија представља често изабрани метод за решавање проблема припадности сваке опсервације одређеној групи (класификација) или предвиђања вероватноће за одређену дискретну вредност зависне променљиве.

9.4.4. Дискриминациона анализа

Дискриминациона анализа је мултиваријациони статистички метод чија је основна сврха да се оцени веза између једне категоријске зависне променљиве и скупа нумеричких независних променљивих. Суштински, дискриминациона анализа омогућава раздвајање (дискриминацију) различитих група и, према одређеним критеријумима, разврставање (алокацију) опсервација у унапред дефинисане групе

(категорије) зависне променљиве. У случајевима када зависна променљива узима две категорије реч је дискриминационој анализи за две групе, а у случајевима више категорија, тада се говори о дискриминационој анализи за више група.

Применом дискриминационе анализе утврђује се разлике између, на пример, група лојалних и нелојалних клијената одређене трговине, сталних и повремених корисника одређених услуга, купаца марке производа А, Б или Ц, успешних и неуспешних предузећа, а затим исте објашњавају на основу скупа независних варијабли укључених у анализу. Слично томе, подносиоци кредитних захтева, на основу сличности њихових ставова (изражених на Ликертовој скали) или социодемографских и психолошких карактеристика са карактеристикама особа које су претходно испуниле или не своје обавезе према банци, могу се класификовати у категорије ниско, средње или високо ризичних клијената. Наведени примери јасно илуструју потенцијал примене дискриминационе анализе у многим проблемским *DM* ситуацијама како за дескрипцију разлика између група, тако и за реализацију предиктивних задатака класификације јединица посматрања чија је група припадања позната или предикције група припадања оних јединица посматрања које нису учествовале у креирању дискриминационог модела.

Полазећи од претходно наведене примарне сврхе дискриминационе анализе, у аналитичком контексту могуће је идентификовати следеће њене циљеве:

- одређивање дискриминационих функција (као линеарнх комбинација независних променљивих) које обезбеђују раздвајање опсервација према дефинисаним групама (односно, категоријама зависне променљиве) тако да се минимизира укупна вероватноћа погрешне класификације, или, другим речима, максимизира релативни однос варијабилитета између група и варијабилитета унутар група;
- дефинисање процедура за класификацију јединица посматрања (појединаца, предузећа, производа итд.) у групе на основу њихових дискриминационих скорова и формулисаних класификационих правила (одређених на основу изведене дискриминационе функције);
- утврђивање да ли постоје статистички значајне разлике између дефинисаних група у погледу просечних вредности посматраних независних варијабли како би се утврдила њихова дискриминациона моћ и допринос у креирању дискриминационог модела, а самим тим и идентификовале независне варијабле које имају највећи утицај на разлике између група;

- спровођење класификације по групама како опсервација које су учествовале у креирању модела, тако и нових јединица посматрања, а на основу познатих вредности независних променљивих, и, с тим у вези, оцењивање класификационе и предиктивне прецизности креираног модела.

Као и у случају сваког другог метода, дискриминациона анализе се заснива на провери испуњености бројних статистичких и имплементационих претпоставки. Кључне статистичке претпоставке су: једнодимензионална и мултидимензионална нормалност независних варијабли, нормалност варијабли по групама, постојање статистички значајне линеарне везе између независних варијабли и одсуство проблема мултиколинеарности, одсуство једнодимензионалних и мултидимензионалних екстремних вредности и хомогеност матрица варијанси и коваријанси по групама. Имплементационе претпоставке су тангентне са проблемом величине узорка, величином група у саставу узорка, као и са поделом узорка на подузорак за анализу (који се користи за оцењивање дискриминационе функције) и тестни подузорак (који се задржава и користи за валидацију оцењеног дискриминационог модела)[58].

Дискриминациона анализа се заснива на одређивању дискриминационог модела, састављеног од једне или више дискриминационих функција, односно, линеарних комбинација независних променљивих које на најбољи начин врше раздвајање јединица посматрања између унапред дефинисаних група. За креирање класификационог модела и алокацију јединица посматрања користе се, потпуно равноправно, следећа два приступа, која се разликују са становишта начина одређивања критеријума за дискриминацију (односно правила за класификацију): ▸ први, заснован на одређивању дискриминационих *Z* функција, и ▸ други, заснован на одређивању *Fisher*-ових линеарних класификационих функција.

Суштину спровођења поступка дискриминационе анализе према првом приступу чини издвајање статистички значајних (једне или више) дискриминационих *Z* функција на основу којих се израчунавају *Z* скорови по свакој јединици посматрања, просечне вредности дискриминационих *Z* скорова по групама (то јест, центроиди група), као и тачке пресека (енгл. *cutting Z values*) за разврставање јединица посматрања у одговарајуће групе. Дискриминациона *Z* функција (на основу које се, у форми *Z* скора, одређује вредност за сваку јединицу посматрања, $i = 1,2,..., n$), представља се у форми следеће линеарне једначине:

[58] О претпоставкама за дискриминациону анализу видети у: *Sharma* (1996); *Hair et al*. (2010).

$Z_j = b_{0j} + b_{1j}X_1 + b_{2j}X_2 + ... + b_{pj}X_p$, где је: (35)

Z_j – Z скор j-те дискриминационе функције (за j = 1, 2,…, k), за свако i = 1,2,…, n,

X_m – вредност m-те независне варијабле, за m = 1, 2,..., p,

b_{0j} – оцењена вредност параметра одсечка j-те дискриминационе функције,

b_{mj} – оцењена вредност коефицијента j-те функције за m-ту независну променљиву,

при чему се дискриминациони коефицијенти оцењују тако да се постигне максимална разлика између група путем максимизирања количника суме квадрата одступања између група и суме квадрата одступања унутар група.

Оцењивање коефицијената дискриминационе функције спроводи се применом једног од следећих метода: ▸ директан метод, који подразумева истовремено укључивање свих независних променљивих у модел без обзира на дискриминациону моћ коју оне појединачно имају, и ▸ метод корак по корак, заснован на постепеном укључивању предиктора у модел сходно њиховом потенцијалу да допринесу раздвајању елемената две групе (*Đorđević i drugi*, 2011, стр. 127). Након оцењивања дискриминационих коефицијената, неопходно је, применом одговарајућих тестова, проверити статистичку значајност добијене дискриминационе функције и оценити статистичку значајност доприноса изабраних варијабли објашњењу разлика између група.

Код просте дискриминационе анализе засноване на двема групама довољно је пронаћи једну дискриминациону функцију која раздваја опсервације на две групе, док се код вишегрупне дискриминационе анализе одређује више дискриминационих функција, с тим што не морају све бити статистички значајне. Начелно, број дискриминационих функција, k, одговара минимуму од следећа два броја: број независних променљивих, p, или број група (категорија, модалитета) зависне променљиве, g, умањен за један (*Soldić-Aleksić & Chroneos Krasavac*, 2009, стр. 171). Односно, симболички, $k = min\ (p; g - 1)$.

Уколико групе нису исте величине, оптимална тачка пресека (раздвајања) између било које две групе одређује се као пондерисани просек центроида група, применом следећег израза:

$$Z_{cs} = \frac{n_A\bar{Z}_B + n_B\bar{Z}_A}{n_A + n_B}, \quad (36)$$

где симболи имају следеће значење:

Z_{cs} – оптимална тачка пресека између група *A* и *B*,

n_A и n_B – број јединица посматрања у групама *A* и *B*, респективно,

$\overline{Z}_A$ и $\overline{Z}_B$ – центроиди група *A* и *B*, респективно.

Уколико су групе једнаке величине, оптимална тачка пресека је једноставно просек два центроида, односно симболички:

$$Z_{cs} = \frac{\overline{Z}_A + \overline{Z}_B}{2}. \tag{37}$$

Правило за класификацију гласи: уколико је $Z_i < Z_{cs}$, тада јединица посматрања припада групи *A*, а уколико је $Z_i > Z_{cs}$, тада јединица посматрања припада групи *B*.

Према другом приступу, погоднијем у случају када је $k \geq 2$, класификационе функције, познате под називом *Fisher*-ове линеарне дискриминационе функције, као одговарајуће линеарне комбинације анализом обухваћених независних променљивих, креирају се за сваку од дефинисаних група. За разлику од првог приступа, број класификационих функција код овог приступа једнак је броју категорија зависне променљиве. Класификација јединица посматрања спроводи се израчунавањем класификационог скора за сваку јединицу посматрања по свакој од *Fisher*-ових функција, утврђених за појединачне групе зависне променљиве. Заправо, број класификационих скорова по свакој јединици посматрања једнак је броју класификационих функција, односно броју група. Конкретна јединица посматрања распоређује се унутар групе којој одговара највећа вредност класификационог скора.

Поступак разврставања јединица посматрања у оквиру оба приступа према дефинисаним критеријумима се итеративно понавља док све јединице посматрања не буду класификоване у једну од дефинисаних група. Резултати класификационе процедуре се приказују у форми класификационе матрице, која представља примарно средство за оцену класификационе и предиктивне прецизности креираног дискриминационог модела, односно одређивање пропорције тачно класификованих јединица посматрања, као кључног показатеља успешности модела. Наравно, дискриминациона анализа може бити на истим подацима спроведена применом оба приступа (уколико је број статистички значајних дискриминационих функција мањи од два), с тим што ће се интерпретација резултата и формулисање закључака заснивати на оном моделу дискриминације који даје боље резултате (прецизнију класификацију).

10. *DATA MINING* ВРЕМЕНСКИХ СЕРИЈА

Многи *DM* проблеми укључују временске аспекте, а најчешћа форма представљања временских података су временске серије. У том смислу, анализа временских серија у *DM* окружењу представља специфичну област истраживања у којој је примена бројних техника и метода прилагођена временској природи података. Полазећи од наведеног, у овом Поглављу је указано на кључна одређења анализе временских серија у *DM* окружењу, као и типичних задатака ове анализе. Посебно је истакнут значај истраживања сличности временских серија и представљена идеја која се налази у основи *SAX* алгоритма за редукцију димензионалности оригиналних података и идентификовање темпоралних законитости.

10.1. Концепт и задаци *data mining*-а у анализи временских серија

Решење проблема који укључују временску димензију у условима велике количине података резултирало је концептом, који се назива *data mining* у анализи временских серија (енгл. *Time Series Data Mining - TSDM*). *TSDM* концепт се може определити као адаптивна и / или иновативна примена: ▸ принципа и метода класичне *DM* анализе, ▸ традиционалних метода анализе временских серија (базираних на декомпозицији временских серија), и ▸ специјално дизајнираних алгоритама и метода за анализу велике количине темпоралних података, са сврхом идентификовања карактеристика и предвиђања стохастичких временских серија.

Наиме, независно од примењеног приступа, суштински, циљеви анализе временских серија су повезани са откривањем корисних законитости у структури временских серија и предвиђањем будућих вредности посматраних појава. Пошто су циљеви класичног и *DM* приступа у анализи временских серија у суштини исти, поставља се питање разлике између ова два приступа. Кључна разлика се превасходно односи на огромну количину података која је настала као последица *IT* револуције. Заправо, у односу на традиционалну анализу, *TSDM* се спроводи над знатно већом количином података или над знатно већим бројем временских серија. Као последица тога, у *TSDM* апликацијама, аутоматско моделирање је једино могући поступак који се може применити за идентификовање комплексних карактеристика високо димензионалних скупова података представљених у форми временских серија.

Временска серија се дефинише као низ реалних вредности о посматраној појави које су сложене хронолошким редоследом у сукцесивним, једнаким временским периодима (година, квартал, месец, недеља, дан, сат ...). У контексту разматрања која

следе, корисно је временску серију дефинисати као секвенцу вредности посматране појаве у функцији времена. Симболички, временска серија као секвенцијални скуп уређених n парова података се представља на следећи начин:

$$Y = \{(y_1, t_1), (y_2, t_2), \ \ldots, (y_n, t_n)\}, \tag{38}$$

где је (за $i = 1, 2, ..., n$): y_i – реална вредност појаве, t_i – време, (y_i, t_i) – пар података.

Суседни делови временске серије, који се састоје од одређеног броја суседних тачака називају се сегменти временске серије и, такође, сами по себи, представљају временску серију. Сегменте репрезентује једна репрезентативна вредност или модел генерисан на подацима сегмента (*Chundi & Rosenkrantz*, 2009, стр. 1753). У том случају, временска серија Y, дужине n, може се представити у виду низа k непреклапајућих, узастопних сегмената. Или, симболички: $Y = (S_1; S_2, ..., S_\kappa)$, при чему се сваки сегмент, S_k, састоји од одређеног броја парова података, (y_i, t_i), односно тачака s_{ik}, где i показује редослед тачке у временској серији, а k припадност конкретном сегменту. С друге стране, скуп већег броја временских серија конституише базу података временских серија (енгл. *Time Series DataBase - TSDB*).

Класификација временских серија може се извршити са становишта различитих критеријума. У овој дисертацији, наводи се, сходно типу мерне скале, подела временских серија на нумеричке и симболичке. Нумеричка временска серија се дефинише као низ хронолошки уређених нумеричких вредности, а симболичка временска серија обухвата хронолошки уређене номиналне или ординалне вредности. У оба случаја реч је о низу кореспондентних парова података за сваку временску јединицу серије или о хронолошком низу вредности које репрезентују сегменте временске серије (на пример, низ аритметичких средина или репрезентативних симбола сегмената серије). За потребе анализе, спроводе се различити поступци трансформације нумеричких у симболичке временске серије, и обратно.

У релевантној литератури, која се односи на истраживање временских серија применом *DM* приступа, истичу се следећи, суштински и методолошки повезани, типични *TSDM* задаци (*Keogh & Kasetty*, 2003, стр. 350; *Mörchen*, 2006, стр. 23–43; *Keogh*, 2011, стр. 339): претпроцесирање, истраживање сличности, класификација, груписање, откривање аномалија, сегментација и предвиђање. Наведени задаци имају низ сличности, али и специфичности у поређењу са кореспондентним *DM* задацима у анализи нетемпоралних података. Уз важну напомену да се остварење циљева анализе временских серија у решавању реалних проблема путем *DM* апликација заснива,

углавном, на међусобном комбиновању наведених задатака (*Lovrić et al*., 2012, стр. 1607), у наставку текста следи кратак опис типичних *TSDM* задатака:

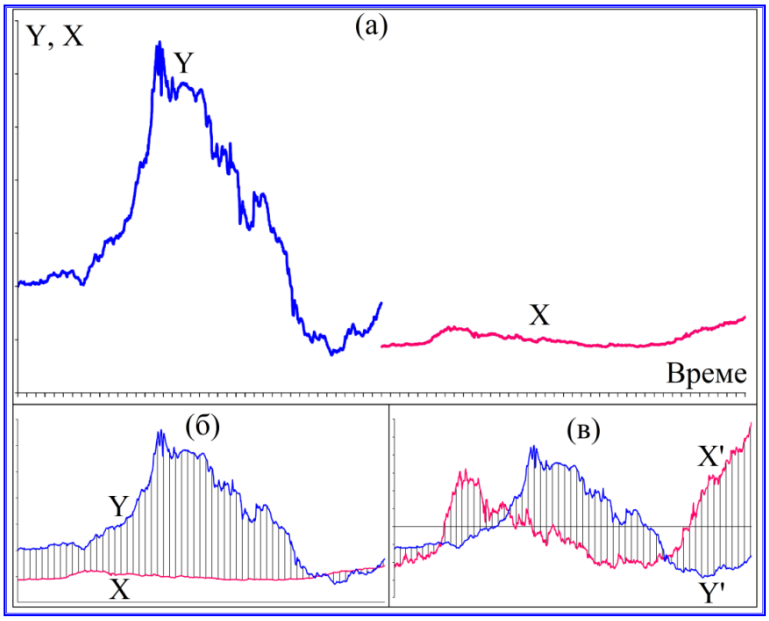
- Претпроцесирање: у обезбеђењу квалитета података, као пресудног фактора за успешну анализу, неопходно је идентификовати и отклонити или редуковати недостатке оригиналних података.
- Истраживање сличности (енгл. *similarity search*): подразумева истраживање сличности у понашању временских серија у бази података (или подсеквенци у временским серијама) коришћењем одговарајућих мера сличности (или одстојања).
- Груписање (енгл. *clustering*): подразумева формирање природних група временских серија у бази података, али на начин да су временске серије унутар сваке групе међусобно сличне, а да се временске серије које припадају различитим групама међусобно разликују.
- Класификација (енгл. *classification*): подразумева креирање класификационог модела за позиционирање сваке временске серије, сходно њеним карактеристикама, у једну од две или више претходно дефинисаних класа.
- Откривање аномалија (енгл. *anomaly detection*): односи се на идентификовање оних делова временске серије којима је својствено понашање знатно различито од очекиваног, уобичајеног обрасца понашања.
- Сегментација (енгл. *segmentation*): обезбеђује редукцију димензионалности поделом временске серије на интерно хомогене делове (или сегменте), али тако да је њихов број знатно мањи од броја података у оригиналној временској серији.
- Предвиђање (енгл. *forecasting*): подразумева да се на на основу дате временске серије која садржи n тачака, одреди њена будућа вредност у времену $n+1$.

Заједничка компонента за већину *TSDM* задатака је истраживање сличних образаца понашање скривених у огромним количинама сирових података кроз процес који обухвата следеће кораке: ▸ приказивање временских серија у форми погодној за даљу анализу, ▸ дефинисање мера сличности између временских серија, и ▸ примена адекватног *TSDM* метода и реализација конкретног задатка сходно природи истраживачког проблема (*Antunes & Oliveira,* 2001). Наиме, истраживање сличности доприноси остварењу следећих користи: ▸ идентификовање временских серија са сличним обрасцима понашања током времена претраживањем *TSDB*-а, и ▸ у својству помоћне или подржавајуће активности у многим апликацијама омогућава реализацију осталих *TSDM* задатака. У основи већине *TSDM* задатака налази се концепт сличности.

10.2. Истраживање сличности и прикази временских серија

Вероватноћа да две временске серије током истог периода имају потпуно исте вредности је врло мала. Имајући то у виду, трагање за идентичним временским серијама је практично бескорисно, тако да добро спроведено истраживање и откривање сличних образаца у понашању временских серија добија све већи практични значај.[59] Консеквентно, истраживање сличности временских серија постаје подручје које привлачи посебну пажњу истраживача из *TSDM* области.

Највећи број приступа за истраживање сличности предложених у литератури односи се на проналажење сличности која је заснована на облику (енгл. *shape-based similarity*). Суштински, проблем сличности се може интерпретирати на следећи начин: ако су дате две временске серије, *Y* и *X*, испитивање њихове сличности подразумева дефинисање и одређивање функције сличности (или функције различитости) између њих. При томе, две временске серије су сличне уколико имају сличан облик, односно сличан образац понашања.



Слика 21: Приказ идеје концепта сличности

Слика 21 илуструје базичну идеју концепта сличности. Визуелно, између представљених временских серија постоји велика удаљеност (Слика 21а) која је последица њихових различитих вредности на *y* оси, као и померања дуж *x* осе, а не само резултат стварне разлике у облику посматраних кривих. Стога, пре анализе

[59] У контексту економских истраживања, упити којима се идентификује сличност временских серија треба да омогуће идентификовање: ▸ компанија са сличним тенденцијама у кретањима индикатора раста, ▸ производа са сличним кретањем продаје (количински и / или вредносно), ▸ акција са сличним кретањем цена, ▸ значајних варијација контролисане карактеристике квалитета процеса и слично.

сличности и квантификовања степена сличности израчунавањем конкретне мере сличности / одстојања, неопходно је, спровести одговарајући поступак трансформације и елиминисати утицај поменутих вертикалних и хоризонталних померања дуж ординате и апцисе (Слика 21б и 21в). На тај начин трансформациони процеси (као што су скалирање, померање, нормализација и одређивање покретних просека) омогућавају да се открије реални степен сличности између две временске серије (*Milanović & Stamenković*, 2011б, стр. 336).

Истраживање сличности може бити спроведено путем потпуног (енгл. *whole matching*) или парцијалног (енгл. *subsequence matching*) упаривања временских серија (*Ratanamahatana et al*., 2010, стр. 1056-1057; *Kontaki et al*., 2005). У првом случају полази се од претпоставке да су све временске серије које су предмет поређења (енгл. *candidate time series* или *target sequences*) исте дужине као и упитна временска серија, то јест, временска серија са којом се врши поређење (енгл. *query time series* или *user's sequence*). При томе, упитна временска серија се упарује са сваком временском серијом у бази, у циљу идентификовања и издвајања оних серија које су њој најсличније. Парцијално упаривање се користи када је упитна временска серија краћа од временских серија које су предмет поређења. Полазећи од упитне временске серије, *Y*, и дуже временске серије, *X*, циљ је идентификовати делове у временској серији *X*, који најбоље одговарају серији *Y*. Заправо, трага се за оним деловима временских серија (енгл. *subsequences*) који су најсличнији са упитном, краћом, временском серијом.

Истраживање сличности, као комплексан *DM* задатак, заснива се на квантификовању сличности / одстојања путем одговарајућих мера. Генерално о мерама блискости је дискутовано у Потпоглављу 9.1. Када је реч о мерама сличности временских серија, у литератури се могу пронаћи њихове бројне класификације према различитим критеријумима (*Mörchen*, 2006, стр. 23-27; *Ratanamahatana et al*., 2010, стр. 1056-1057; *Vlachos et al*., 2004, стр. 67-100). При избору конкретне мере у одређеној ситуацији за утврђивање сличности између временских серија полази се од карактеристика временских серија које се пореде (као што су, тип временских серија, дужина временских серија, присуство екстремних вредности и шумова, расположиво знање истраживача о структури података и слично).

Најчешће коришћена мера за утврђивање сличности између две нумеричке временске серије јесте Еуклидска мера удаљености (или нека њена изведена форма).[60]

[60] Преко 80% публикованих стручних радова засновано је на коришћењу Еуклидског одстојања за истраживање сличности између објеката (*Soon & Lee*, 2007, стр. 31).

У методолошком смислу, ова мера је заснована на поређењу и утврђивању разлике између вредности у *i*-тој тачки једне временске серије и вредности у *i*-тој тачки друге временске серије. Сходно томе, израз 1, прилагођен временској природи података може се записати у следећем облику:

$$D(Y,X)=\sqrt{\sum_{i=1}^{n}\left(y_i-x_i\right)^2}\ , \tag{39}$$

где су симболима *Y* и *X* представљене две временске серије једнаке дужине, *n*[61].

У непосредној, концепцијској и методолошкој повезаности са мерама сличности и, генерално, задатком истраживања сличности су одговарајући методи приказа временских серија (енгл. *representations of time series*). По правилу, временске серије карактерише висока димензионалност тако да директан рад са сировим подацима је веома незахвалан из перспективе ефикасности складиштења, преноса и процесирања података. Отуда, основна идеја приказивања временских серија у облику другачијем од оригиналног садржана је у намери да се путем одговарајућих апроксимативних форми, кроз редукцију димензионалности, изврши адекватна припрема оригиналних података за даље процесирање у циљу екстракције основних карактеристика и законитости високо димензионалних серија уз минимални губитак релевантних информација. Дакле, прикази временских серија, методолошки, представљају трансформације путем којих се временске серије високе димензионалности из оригиналног простора пресликавају у редуковани простор нижег реда димензионалности.

У литератури су предложени бројни методи за приказивање временских серија, попут: *DFT* (енгл. *Discrete Fourier Transform*), *DWT* (енгл. *Discrete Wavelet Transform*), *APCA* (енгл. *Adaptive Piecewise Constant Approximation*), *SVD* (енгл. *Singular Value Decomposition*), *PAA* (енгл. *Piecewise Aggregate Approximation*), *PLA* (енгл. *Piecewise Linear Approximation*) итд.[62] Генерално, избор конкретног приказа је тесно повезан и знатно детерминисан конкретним *TSDM* задатком, циљевима истраживања сличности, као и карактеристикама посматраних временских серија.

Посебна категорија приказа јесу симболички прикази. Симболизацијом се временска серија поједностављује, а тиме и омогућава лакше идентификовање

[61] За анализу сличности временских серија неједнаке дужине, једна од најпопуларнијих мера је динамичко искривљење времена (енгл. *Dynamic Time Warping - DTW*). За разлику од Еуклидског одстојања, ова мера се заснива на нелинеарном упаривању тачака (једна тачка једне серије може бити упарена са низом суседних тачака друге серије). Међутим, даља дискусија у овом правцу је изван оквира овог рада.

[62] Детаљне информације о хијерархијској класификацији, као и опис појединих категорија приказа временских серија, предложених за потребе подршке реализацији *TSDM* задатака видети у: *Keogh* (2011, стр. 341); *Ratanamahatana et al*. (2010, стр. 1065-1073).

темпоралних законитости уз пратеће повећање ефикасности нумеричких израчунавања. Од бројних симболичких метода, као моћан метод за редукцију димензионалности и откривање информација скривених у временским серијама издваја се метод познат под називом симболичка агрегатна апроксимација (енгл. *Symbolic Aggregate ApproXimation - SAX*), предложен од стране *Lin et al*. (2003).

10.3. Редукција димензионалности временских серија применом *SAX* алгоритма

У *TSDM* анализи, симболизација се дефинише као процес конвертовања реалних вредности података оригиналне временске серије у серију симбола. Суштина симболизације, као претпроцесне активности, огледа се у дискретизацији вредности временске серије путем одређеног броја симболичких вредности, чиме се креира нова, симболичка серија која представља предмет даљег процесирања.

Независно од метода коришћеног за трансформацију сирових података у симболичке низове, са становишта евалуације перформанси примењеног симболичког метода, неопходно је указати на значај следећа три параметра (као и њихових међусобних, конфликтних међуодноса): величина алфабета, губитак информација и стопа компресије (*Sant' Anna & Wickström*, 2011, стр. 2223).

Број симбола који се користе за трансформацију временске серије назива се величина алфабета, односно величина скупа симбола (*Daw & Finney*, 2003). Хронолошки уређена комбинација коришћеног скупа симбола резултира низом који представља израз конкретне временске серије у форми речи као симболичке секвенце. У најједноставнијем, бинарном случају, реч се формира комбинацијом само два могућа симбола [0 и 1] и тада је величина алфабета 2. С обзиром да је број коришћених симбола (много) мањи од броја различитих вредности у оригиналној серији, симболизација је увек праћена извесним губитком информација. Стога, истраживач неизбежно доноси одлуку о величини алфабета која директно утиче на карактеристике и квалитет симболичког приказа серије.

Да би се идентификовале смислене информације о динамици појаве током времена, неопходно је временску серију поделити на одређени број узастопних, интерно хомогених делова (сегмената) и сваком од њих доделити одговарајући симбол у зависности од тога којем региону тај део припада. Према томе, битна компонента процеса симболизације је сегментација временске серије. Путем сегментације се постиже компресија података, односно редукује величина посматраних скупова података и иманентно присутна висока димензионалност. Исход спроведене

компресије се може оценити коришћењем различитих мера. Једна од њих јесте стопа компресије који се може дефинисати на следећи начин (*Salomon & Motta*, 2010, стр. 12): ▸ као однос између величине резултирајућег низа симбола и величине полазног низа података, и ▸ као однос између броја сегмената редуковане серије и броја података оригиналне временске серије.

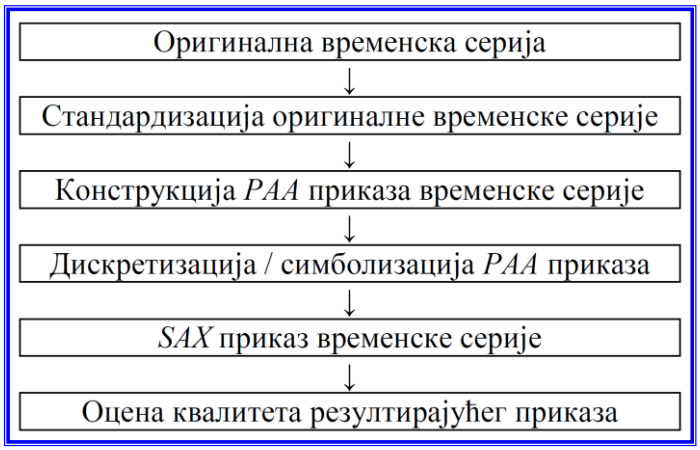
По правилу, за уочавање значајних темпоралних образаца (уколико постоје), при одређивању броја сегмената, а самим тим и дужине симболичке секвенце, експертска процена истраживача има доминантну улогу, при чему је неопходно водити рачуна о *trade-off* ефекту између броја сегмената и прецизности новоформираног симболичког приказа, односно степена губитка информација (већи број сегмената→мања редукција димензионалности→мањи губитак информација, и обратно).

На основу изнетих разматрања, симболизација се у формалном смислу може дефинисати на следећи начин: Временска серија Y представља секвенцу n реалних вредности у конкретном временском тренутку или интервалу времена. Симболизација представља поступак којим се дата временска серија Y дели на k делова, а затим сваком делу додељује конкретни симбол из претходно дефинисаног алфабета. На тај начин се секвенца n оригиналних података трансформише у секвенцу k симбола, при чему је, по правилу, $k < n$ (углавном, $k << n$).

Табела 3: Ознаке коришћене за конструкцију *SAX* приказа и њихово значење

Ознака	Значење
Y	оригинална временска серија: $Y = y_1, y_2, ..., y_n$
Y'	стандардизована серија: $Y' = y'_1, y'_2, ..., y'_n$
$\bar{Y}'$	*PAA* приказ серије: $\bar{Y}' = \bar{y}'_1, \bar{y}'_2, ..., \bar{y}'_k$
$\hat{Y}'$	*SAX* приказ серије: $\hat{Y}' = \hat{y}'_1, \hat{y}'_2, ..., \hat{y}'_k$
n	величина (број података) временске серије Y
k	број *PAA* сегмената (*PAA* коефицијената)
α	величина алфабета (на пример, за $\{a, b, c, d\}$, $\alpha = 4$)
S_h	h-ти симбол алфабета S, $h = 1, 2, ..., \alpha$ (на пример: за $\alpha = 4$, $S_1 = a$, $S_2 = b$, $S_3 = c$, $S_4 = d$)
β_i	преломне тачке ($i = 1, ..., \alpha$-1)

За потребе презентовања идеје *SAX* приступа, на Слици 22 представљени су кораци поступка трансформације оригиналне временске серије у симболичку *SAX* апроксимацију, док је у Табели 3 представљена листа коришћених ознака.



Слика 22: Кораци у конструкцији *SAX* приказа

SAX метод је заснован на парцијалној агрегатној апроксимацији (*PAA* приказ) уз претпоставку да су временске серије нормално дистрибуиране. Пошто се *SAX* метод примењује на стандардизованој серији, иницијални корак у поступку креирања *SAX* приказа је трансформација оригиналне временске серије путем процеса *z*-стандардизације у њену стандардизовану форму, са аритметичком средином 0 и стандардном девијацијом 1, симболички: [$Y \rightarrow Y'$: N(0;1)]. Наиме, пре трансформације временских серија у симболе, неопходно је извршити њихову нормализацију. *Keogh & Kasetty* (2003, стр. 361) управо истичу значај нормализације и наводе да истраживање сличности (и консеквентно, реализација осталих *TSDM* задатака) без нормализације нема смисла. Полазећи од наведеног, спровођење наредних корака засновано је на својствима стандардизованог нормалног распореда (*Stamenković i drugi*, 2012).

У наставку процеса, димензионалност временске серије се смањује путем парцијалне агрегатне апроксимације стандардизоване серије (односно, *PAA* приказа). Како је *SAX* приказ изведен из *PAA* приказа, следи кратак преглед ове, са аспекта израчунавања, једноставне и ефикасне методе, коју су, независно једни од других, предложили *Keogh et al*. (2000) и *Yi & Faloutsos* (2000, стр. 385-394). Путем *PAA* приказа, временска серија, *Y'*, дужине *n*, се сегментира на *k* узастопних делова, једнаке дужине, при чему се за сваки формирани сегмент израчунава аритметичка средина применом следећег израза:

$$\bar{y}'_j = \frac{k}{n} \sum_{i=\frac{n}{k}(j-1)+1}^{\frac{n}{k}j} y'_i \text{ , за } i = 1, 2, ..., n, \text{ и } j = 1, 2, ..., k \quad . \qquad (40)$$

Серија формираних *k* средина (које се називају и *PAA* коефицијенти) представља нови, редуковани приказ временске серије. Заправо, временска серија *Y'* дужине *n* се

апроксимира серијом $\overline{Y}'$ дужине k, при чему је $k = n$. Оптимална вредност параметра k се налази између екстремних случајева: када је $k = 1$, временска серија се пресликава у њену просечну вредност, док у случају $k = n$, временска серија се не трансформише, тако да се редукција димензионалности не постиже. Релација n / k означава број података од којих се састоји сваки сегмент, а релација k / n је стопа компресије.

Након формирања *PAA* приказа, спроводи се дискретизација (односно, симболизација) која подразумева конверзију нумеричке временске серије *PAA* коефицијената у низ симбола. Процес дискретизације започиње одређивањем величине алфабета, односно броја симбола који ће се користити за симболичко приказивање временске серије. Величина алфабета је арбитрарно одређена целобројна вредност која се обележава симболом α, при чему увек важи релација $\alpha > 2$. При томе, сваки симбол има подједнаку и независну вероватноћу појављивања. Наведено произилази из чињенице да је временска серија у првом кораку стандардизована, [Y': N(0;1)]. Сходно величини алфабета, на графику нормалног распореда могуће је одредити α–1 преломних тачака, β_i (за $i = 1,..., \alpha$–1), које ће обезбедити α једнаких површина испод нормалне криве, при чему се свакој површини додељује одговарајући симбол, S_h (за $h = 1, 2,..., \alpha$). Површина испод нормалне криве између две суседне преломне тачке β_i (од β_i до β_{i+1}) једнака је $1 / \alpha$, а β_0 и β_α су дефинисани као $-\infty$ и $+\infty$, респективно (*Lin et al.*, 2003, стр. 5). Дакле, број преломних тачака зависи од величине алфабета, а вредност (локација) сваке преломне тачке одређује се коришћењем таблице стандардизованог нормалног распореда. У емпиријском делу дисертације, сходно потребама истраживања у контексту прве дефинисане проблемске ситуације, у Табели 7 дате су вредности преломних тачака за величину алфабета од 3 до 12, одређене коришћењем таблице стандардизованог нормалног распореда.

Након утврђивања преломних тачака, сваком од *PAA* коефицијената ($\overline{y}'_j$, за $j = 1, 2,..., k$) додељује се одговарајући *SAX* симбол ($\hat{y}'_j$, за $j = 1, 2,..., k$), сходно интервалу између две узастопне преломне тачке (површина испод нормалне криве) којем сваки појединачни *PAA* коефицијент припада, на следећи начин: ▸ сви *PAA* коефицијенти који су испод најмање преломне тачке (β_1) означавају се симболом „*a*”, (S_1); ▸ сви коефицијенти који су једнаки или већи од најмање преломне тачке (β_1), а уједно и мањи од друге, наредне најмање преломне тачке (β_2), означавају се симболом „*b*”, (S_2) итд. Наведене релације се могу представити путем израза: $\beta_{j-1} \le \overline{y}'_i < \beta_j$. Дакле, *PAA* вредности се трансформишу у симболе коришћењем табеле преломних тачака.

Спој ових симбола се назива *SAX* реч, симболички: $\hat{Y}' = \hat{y}'_1, \hat{y}'_2, ..., \hat{y}'_k$. На овај начин је, у форми *SAX* речи, дефинисан *SAX* приказ временске серије. Број симбола у *SAX* речи је детерминисан бројем сегмената, а њен садржај, за једну исту временску серију, може се разликовати у зависности од изабране величине алфабета.

Процесом конверзије оригиналне временске серије у симболичке репрезенте редукује се димензионалност оригиналних података. Јасно је да у овом процесу долази до губитка извесних информација, тако да се поставља питање прецизности *SAX* апроксимације. Прецизност симболичког апроксимативног приказа је врло тешко квантификовати. Један од начина је одређивање грешке апроксимације, коју *Wei et al.* (2008, стр. 359) називају грешка *SAX* реконструкције (енгл. *SAX reconstruction error*). Ова грешка се дефинише на следећи начин: за временску серију, *Y'*, и њену *SAX* реч, *Ŷ'*, грешка *SAX* апроксимације, E_{SAX}, је квадратни корен суме квадрата одступања сваког податка временске серије, $y_i^{'}$, од одговарајуће средње вредности, μ_h (за h=1, 2,..., α), која површину сваког симбола испод нормалне криве (дефинисане двема суседним преломним тачкама) дели на два дела једнаке вероватноће. Симболички:

$$E_{SAX} = \sqrt{\sum_{i=1}^{n} \left(y_i^{'} - \mu_h \right)^2} \text{ , за } h = 1, 2, ..., \alpha. \qquad (41)$$

При одређивању и избору конкретне μ_h вредности, у односу на коју се мери одступање конкретног оригиналног податка, неопходно је водити рачуна о симболу којим је тај податак представљен (то јест, сегменту којим је тај податак обухваћен), као и чињеници да се површина сваког симбола дели на два дела једнаке вероватноће путем кореспондентне μ_h вредности. Стога су ове вредности еквивалентне преломним тачкама за дуплирану величину алфабета коришћеног за конструкцију *SAX* приказа.

Поред редукције димензионалности и оцене прецизности *SAX* апроксимације за конкретну временску серију, поставља се питање употребе овог приказа у истраживању сличности и компарацији две временске серије (*Milanović et al.*, 2012), а самим тим и питање дефинисања одговарајуће мере сличности / одстојања, као и њеног коришћења у реализацији *TSDM* задатака.

У поступку формирања *SAX* приказа, поређење две временске серије може се спровести на: ▸ стандардизованим временским серијама, ▸ *PAA* апроксимативним формама, и ▸ *SAX* апроксимативним формама.

За две стандардизоване временске серије Еуклидско одстојање између њих се дефинише путем следећег израза:

$$D(Y', X') = \sqrt{\sum_{i=1}^{n} (y_i' - x_i')^2}\,, \tag{42}$$

где *X'* и *Y'* представљају стандардизоване форме оригиналних временских серија. Дакле, Еуклидско одстојање је квадратни корен из суме квадрата одстојања (разлике) сваког пара кореспондентних тачака стандардизованих података.

Одстојање између два *PAA* приказа дефинише се као квадратни корен из суме квадрата разлике између сваког пара кореспондентних *PAA* коефицијената, пондерисан са квадратним кореном из стопе компресије, односно, симболички:

$$D(\bar{Y}', \bar{X}') \equiv \sqrt{\frac{n}{k}} \sqrt{\sum_{j=1}^{k} (\bar{y}_j' - \bar{x}_j')^2}\,. \tag{43}$$

Након трансформације *PAA* коефицијената у симболички приказ, одстојање између две *SAX* речи, односно, две симболичке временске серије дефинише се као квадратни корен из суме квадрата одстојања између сваког пара симбола, пондерисан са квадратним кореном из стопе компресије, односно, симболички:

$$MINDIST(\hat{Y}', \hat{X}') \equiv \sqrt{\frac{n}{k}} \sqrt{\sum_{j=1}^{k} [dist(\hat{y}_j', \hat{x}_j')]^2}\,, \tag{44}$$

где је $dist(\hat{y}_j', \hat{x}_j')$ одстојање између упарених симбола две временске серије.

За одређивање одстојања између упарених симбола два *SAX* приказа, $dist(\hat{y}_j', \hat{x}_j')$, користи се специјално дизајнирана табела одстојања упарених симбола, сходно величини алфабета, типа $\alpha \times \alpha$. Заправо, број колона (*c*) и редова (*r*) у табели једнак је величини алфабета, а вредности у пољима табеле одређују се путем следећег израза (*Lin et al*., 2003, стр. 6):

$$\text{поље табеле}_{r,c} = \begin{cases} 0, \text{ ако је} \mid r - c \mid \ \leq 1 \\ \beta_{\max(r,c)-1} - \beta_{\min(r,c)} \text{ у осталим случајевима} \end{cases}. \tag{45}$$

Одстојање између два симбола се очитава у пресеку реда и колоне табеле који одговарају упареним симболима, при чему се редови односе на симболе једне, а колоне на симболе друге временске серије. У емпиријском делу дисертације, сходно међурезултатима истраживања у контексту прве дефинисане проблемске ситуације, у Табели 9 дате су вредности одстојања између два симбола за величину алфабета упитне временске серије која је изабрана као оптимална. Заправо, табела преломних тачака и табела одстојања формирају се за сваку величину алфабета.

У основи, као најважнија карактеристика *SAX* приказа наводи се да представљена мера одстојања између две *SAX* речи обезбеђује доњу граничну вредност Еуклидске

мере одстојања кроз две етапе: у првој етапи, према корацима у креирању *SAX* приказа, добија се *PAA* мера одстојања која представља доњу граничну вредност Еуклидске мере; у другој етапи добија се *MINDIST* одстојање које је мање од *PAA* мере одстојања, тако да сходно својству транзитивности, *SAX* метод резултира (мањим) одстојањем које представља доњу границу Еуклидског одстојања између оригиналних временских серија (Доказ ове констатације видети у: *Lin et al*., 2007, стр. 118–122). Наведене релације се могу представити путем израза: $D(Y', X') \geq D(\bar{Y}', \bar{X}') \geq MINDIST(\hat{Y}', \hat{X}')$.

11. МЕТОДОЛОШКИ ОКВИРИ ЗА ОЦЕЊИВАЊЕ КАРАКТЕРИСТИКА *DATA MINING* МОДЕЛА

Оцењивање откривеног знања је једна од битних компоненти *DM* процеса. У зависности од карактеристика конкретног *DM* проблема и примењених метода процесирања, разликују се мере, критеријуми и методи за оцењивање резултирајућих предиктивних и дескриптивних *DM* модела. У том контексту, у овом Поглављу пажња је фокусирана на разматрање како општих аспеката оцењивања квалитета креираних модела, тако и на специфичности које се односе на вредновање, најпре, резултата примене дескриптивних метода анализе груписања, а затим и предиктивних метода.

11.1. Проблем оцењивања и избора модела

Оцењивање [63] карактеристика креираних модела и вредновање делова знања издвојеног из података представља важан задатак у *KDD* фази постпроцесирања. Стриктно речено, у стандардизованом *KDD* процесу, оцењивање карактеристика модела је између фаза моделирања и примене развијеног модела. Међутим, не треба превидети да се оцењивање не односи само на поузданост и валидност коначног исхода процеса моделирања, већ представља интегрални део сваког сегмента процеса креирања модела и *DM* анализе. Из наведеног произлази да је коришћење поузданих критеријума и метода за оцењивање квалитета различитих аспеката *DM* модела током целог процеса откривања знања од суштинског значаја у контексту корисности модела за разматрани пословни проблем транспонован у конкретни *DM* задатак.

[63] Оцењивање, вредновање, валидација и евалуација (енгл. *evaluation*) су термини који се користе као синоними при одређивању квалитета било којег *DM* модела и односе се на оцењивање карактеристика модела путем одговарајућих мера за оцењивање. Такође, у случају предиктивних метода, укључује и подешавање параметара у итеративном процесу учења у функцији побољшања карактеристика модела и избор најприхватљивијег (сходно карактеристикама) модела или алгоритма. Овај вид оцењивања спроводи се током поступка спровођења одговарајуће процедуре, као и на крају процеса при оцени коначних резултата. Оцењивање у наведеном контексту треба разликовати од оцењивања параметара скупа (енгл. *estimation*) које се базира на одређивању одговарајућих статистика узорка. Наравно да и мере за евалуацију квалитета *DM* модела могу представљати статистичке оцене.

Суштински, типичну *DM* ситуацију карактерише генерисање више модела. Стога се појављује потреба за оцењивањем и компарацијом њихових својстава. Заправо, непходност оцењивања квалитета модела је резултат:

- бројности *DM* метода и алгоритама и чињенице да се за реализацију истог *DM* задатка може користити више метода;
- различитих комбинација (конфигурација) и промена параметара у оквиру једног метода које могу довести до битно различитих резултата;
- решавања проблема засновано на комбинованом приступу, односно примени два или више метода;
- примене разноврсних трансформационих процеса који доводе до формирања различитих скупова / узорака података за моделирање, а тиме и различитих модела;
- настојања да се утврди да ли откривене законитости у форми једног или више модела имају пословни смисао и практични значај у процесу пословног одлучивања.

Генерално, оцењивање карактеристика модела може бити усмерено на (*Hastie et al*., 2009, стр. 196): оцењивање карактеристика различитих модела и избор једног као најбољег у односу на разматрани проблем и оцењивање предиктивне тачности изабраног модела приликом његове примене на подацима који нису учествовали у формирању модела. У том смислу, да би се обезбедила валидна интерпретација добијених резултата моделирања, неопходно је при оцењивању водити рачуна о следећим аспектима: статистичкој и практичној значајности модела, подели расположивих података за потребе развоја модела и корисности софтверских излаза.

Статистичка значајност модела подразумева примену различитих статистичких критеријума у свим фазама *KDD* процеса у функцији развоја модела, као и приликом компарације модела и избора коначног решења. Практична значајност модела се односи на константну проверу смисла и поузданости конкретног *DM* решења у разматраном (пословном) сценарију. Модел који је са становишта статистичких критеријума најбољи није и нужно најбољи у погледу практичне корисности, и обрнуто. Начелно, обе, и статистичка и практична значајност детерминишу одговор на питање који модел изабрати у одређеној ситуацији.

Оцењивање карактеристика модела је непосредно повезано са питањем поделе података на узорак за учење, валидацију и тестирање. С обзиром да се циљеви оцењивања на нивоу ових узорака разликују, при интерпретацији резултата треба водити рачуна о којем узорку је реч.

Софтверски алати који се користе за примену *DM* метода на подацима углавном генеришу мноштво излазних резултата. Међутим, нису сви излази подједнако корисни за разматрани реални проблем. Стога је важно, најпре, проценити који софтверски излази (нумеричке информације) су релевантни за дефинисану проблемску ситуацију, а затим путем прикладног презентовања одабраних резултата (укључујући и визуелизацију) крајњим корисницима (који нису упознати са техничким детаљима коришћених *DM* метода) обезбедити разумевање истих.

Оцењивањем карактеристика модела треба добити одговоре на следећа питања (*Berry & Linoff*, 2004, стр. 78): ▸ колико је модел тачан; ▸ колико добро модел репрезентује емпиријске податке, или, колико је добро модел прилагођен подацима; ▸ колико је модел поуздан са становишта предвиђања; ▸ колико је модел разумљив и интерпретабилан.

Начини добијања одговора на ова питања примарно зависе од типа креираног модела за реализацију конкретног *DM* задатка. У основи, разликују се критеријуми и методи за оцену дескриптивних и предиктивних *DM* модела. Услед одсуства зависне променљиве (и њених познатих вредности), као и чињенице да се у процесу развоја модела не издваја део података за тренирање модела, знатно је већи проблем оценити моделе који су резултат примене ненадгледаних метода учења (*Cios et al*., 2007, стр. 471). Наиме, објективне критеријуме за оцену дескриптивних модела је тешко дефинисати, због чега искуство, логика и здрав разум аналитичара и експерта из одређеног подручја доминирају при оцени квалитета дескриптивних модела. Другим речима, у целом процесу трагања за дескриптивним моделом и одговором на питање да ли су и у ком степену откривене законитости заиста релевантне, примарна улога припада експертском знању и одлукама.

За оцењивање предиктивних модела развијени су бројни објективни приступи сходно типу зависне променљиве. Заправо, приступи за оцењивање карактеристика модела који генеришу нумеричке предикције (зависна променљива је нумеричка) се разликују од приступа за оцењивање класификационих модела који се користе за алокацију опсервација према категоријама зависне променљиве и предикцију номиналне категорије (или вероватноће класе) којој припадају нове опсервације (зависна променљива је категоријска).

Процес моделирања, а тиме и вредновања својстава модела није линеаран, већ итеративан, интерактиван и креативан процес. Врло ретко ће иницијално истрениран модел заиста бити најбоље решење (*Nisbet et al*., 2009, стр. 85). Наиме, често се

оцењивањем карактеристика модела откривају извесни проблеми који се могу решити додатним претпроцесирањем и новим процесирањем података кроз варирање параметара модела или промену алгоритма моделирања. Ове измене треба да допринесу побољшању прецизности, корисности, разумљивости, као и других својстава модела. Како не постоји опште правило које ће „рећи" када је модел довољно добар, одговор се обезбеђује, сходно специфичностима конкретног проблема, кроз итеративни процес и серију активности моделирања, евалуације и побољшања.

11.2. Оцењивање дескриптивних модела

Експлоративни, односно дескриптивни проблеми и циљеви су један од највећих *DM* изазова који се често решавају применом метода анализе груписања. Оцењивање резултата примене овог метода, базираног на парадигми ненеадгледаног учења, је знатно сложенији проблем у односу на оцењивање надгледаних метода. Услед наведеног, одговор на питање колико је резултат груписања добар, најтешње је повезан са конкретним подручјем и проблемом истраживања. У суштини, оцена квалитета груписања је проблемско и контроверзно питање (*Rokach*, 2010, стр. 273). Ипак, и поред одсуства стриктних процедура, постоје бројни индикатори и поступци који омогућавају адекватну процену међурешења и коначног решења анализе груписања.

Евалуација груписања објеката се односи на оцену оправданости спровођења поступка груписања на посматраним подацима и оцену квалитета резултата добијених применом овог метода (*Han et al*., 2012, стр. 484). Добра процедура груписања треба да: омогући откривање структура присутних у подацима, обезбеди детерминисање оптималног броја група, резултира јасно диференцираним групама и групама које остају стабилне када су присутне и мале промене у подацима, ефикасно процесира не само велике количине података, већ и, ако је потребно, различите типове променљивих (*Tufféry*, 2011, стр. 242). Кључни задаци евалуације груписања укључују: ▸ оцену оправданости груписања података, ▸ одређивање оптималног броја група, ▸ мерење различитих аспеката значајности формираних група, ▸ интерпретацију добијеног модела груписања, и ▸ верификацију добијених резултата.

Сваки алгоритам груписања ће резултирати одређеним групама независно од тога да ли подаци заиста суштински садрже смислене и разумљиве законитости. Међутим, модел груписања, то јест, формирана структура је валидна само ако основано није производ случајности. Због тога је, пре примене алгоритма, неопходно установити да

ли постоји природна тенденција података ка груписању која ће довести до формирања група са одговарајућин значењем.

Витално питање у анализи груписања се односи на одређивање оптималног броја група. Начелно, статиситчки је оправдано да се процес груписања прекине када почне спајање група између којих постоји велика удаљеност или у случају када почне процес раздвајања на групе које нису много удаљене. Мада не постоје јака правила у погледу могућих решења овог проблема, разликује се више корисних приступа и критеријума, хеуристичког и формалног карактера, који обезбеђују корисне смернице за одређивање броја група. Неки од тих приступа су:

- Број група може бити сугерисан и унапред дефинисан од стране менаџера и аналитичара уз поштовање теоријских, концептуалних, практичних и / или логичних разлога. На пример, ако су, на једној страни, за потребе истраживања потребне четири групе, а алгоритам је резултирао разврставањем података у три групе, или пак, на другој страни, унапред детерминисани број група не кореспондира са стварном дистрибуцијом објеката, тада се групе и границе између њих одређују арбитрарно.
- Код хијерархијске процедуре груписања, проблем дефинисања броја група се може решити путем вредности мера одстојања при којој се две групе удружују у једну, и то праћењем: апсолутних вредности мера одстојања, прираштаја одстојања или суме квадрата одстојања (грешака). Информација о апсолутној вредности мере одстојања може се добити на основу дендрограма хијерархијске структуре, тако да се његовим пресецањем на одређеном нивоу одстојања доноси одлука о броју група. Нагли скок мере одстојања указује на знатно смањење повезаности и повећање различитости између две групе. Број група у кораку који је претходио том кораку постаје оптималан. Такође, ако се уместо мере одстојања као критеријум оптималног избора користи прираштај или сума квадрата одстојања, драстична промена њихових вредности сугерише да је оптималан број група одређен у претходном кораку.
- При избору опималног броја група код нехијерархијског груписања разматра се однос између укупног варијабилитета унутар група (укупна сума квадрата одстојања свих објеката од центроида група којима припадају) и варијабилитета између група (укупна сума квадрата одстојања центроида сваке групе од глобалног центроида, као центра гравитације свих опсервациаја). Обично се овај количник и кореспондирајући број група представљају графички, тако да тачка на дијаграму где долази до нагле промене вредности количника указује на оптималан број група. За нехијерархијску процедуру *k*-средина је специфично да се вредност овог количника не добија као

међурезултат током алгоритамског поступка груписања. Вредност наведеног односа одређује се на крају процеса за дефинисани број група, а затим итеративни процес започиње са новим бројем група. Као оптимални броја група код нехијерархијског груписања може се користити исход агломеративног хијерархијског груписања.

- Релативна величина група и расподела објеката по групама може бити значајан фактор при избору одговарајућег броја група. Наиме, као резултат примене алгоритма могу настати групе које садрже само један елемент (што може бити последица реално присутне екстремне вредности). Такође, у последњим корацима груписања могуће је да дође до спајања група које садрже велики број елемената са знатно удаљеним појединачним објектима (при чему треба бити обазрив, јер велика удаљеност може бити узрокована и грешком насталом приликом уноса података). Стога, уколико не постоје теоријска и логична оправдања за добијене резултате груписања, неопходно је формулисати и имплементирати старатегију за елиминацију оваквих проблема, укључујући и понављање поступка груписања.

У оцењивању различитих аспеката квалитета група и груписања издвајају се следећа три приступа (а у оквиру сваког од њих развијени су бројне мера) (*Jain & Dubes*, 1988, стр. 161):

- екстерно вредновање се базира на оцени степена подударности формираних група са детерминисаним групама на бази искуства, постојећег знања експерата о подацима и слично, коришћењем екстерних критеријума и мера, попут мера ентропије;
- интерно вредновање се заснива на мерењу релевантности структуре груписања узимајући у обзир само информације из самих података, коришћењем интерних критеријума и мера;
- оцена резултата груписања заснована на релативним критеријумима и мерама подразумева поређење идентификованих структура група које су резултат примене различитих метода груписања или истог алгоритма, али са различитим вредностима улазних параметара.

За разлику од екстерних мера, које своју примену налазе у избору најбољег алгоритма груписања и поређењу резултирајућих структура, интерне мере имају ширу употребу, јер осим наведеног, користите се и за избор оптималног броја група (*Liu et al*., 2010, стр. 911). Како у многим практичним сценаријима екстерне информације нису расположиве, интерне мере постају једини начин вредновања груписања. Имајући у виду речено, као и бројност предложених мера за оцену квалитета груписања, у

наставку ових разматрања издвајају се неке од интерних мера које су засноване на суми квадрата одступања и концептима компактности и сепарације.

Уопштено, путем интерних мера, груписање се оцењује тако што се испитује интерна хомогеност (компактност или кохезија унутар група) и екстерна хетерогеност (међусобна раздвојеност или сепарација група) (*Liu et al*., 2010, стр. 911-912; *Rendón et al*., 2011, стр. 27-28). Да би се дефинисали наведени концепти, нека се посматра n објеката који су разврстрани у скуп од k група: $C=(C_1, C_2, ..., C_k)$, где је $h = 1, 2,..., k$.

Компактност групе се односи на блискост објеката у једној групи. Као индикатор компактности сваке групе C_h уобичајено се користе сума квадрата одступања свих опсервација унутар конкретне групе од њеног центроида, W_h. На основу овог индикатора одређује се варијанса која се даље користи за израчунавање бројних мера компактности групе. Генерално, нижа варијанса указује на бољу компактност. Такође, компактност се може оценити и путем максималног или просечног одстојања између парова објеката унутар исте групе.

На основу компактности група, могуће је дефинисати укупну компактност скупа група C (*Vercellis,* 2009, стр. 313). Уколико се симболом *coh* (C_h) означи компактност групе C_h, тада се укупна компактност (где је : $C_h \in C$) дефинише на следећи начин:

$$coh(C) = \sum_{h=1}^{k} coh(C_h). \tag{46}$$

Уколико се као индикатор компактности користи сума квадрата одступања опсервација групе од њеног центроида, тада се укупна компактност дефинише као:

$$coh(C) = W = \sum_{h=1}^{k} W_h. \tag{47}$$

Са аспекта интерне хомогености, решење добијено применом једног метода груписања је боље од резултирајућег исхода другог метода, ако је индикатор његове укупне компактност мањи.

Сепарација се односи на испитивање одстојања једне групе од сваке друге групе у скупу C. Широко коришћени индикатори сепарције између било које две групе су одстојање између центара група (репрезентативних вредности) или минимално одстојање од свих одстојања између парова објеката две групе. Уколико се симболом $sep(C_i, C_j)$ представи сепарација између било које две групе из скупа C (где је: C_i и $C_j \in$ C), тада се укупна сепарација дефинише путем следеће релације:

$$sep(C) = \sum sep(C_i, C_j). \tag{48}$$

Уколико се као индикатор сепарације користи сума квадрата одступања средина група од кореспондирајуће опште средине, тада се укупна сепарација дефинише путем следећег израза, односно:

$$sep(C) = B. \tag{49}$$

Са аспекта екстерне хетерогености, решење добијено применом једног метода груписања је боље од решења добијеног применом другог метода ако је индикатор његове укупне сепарације већи (боље је разграничење између група).

Поред парцијалних оцена компактности и сепарације (на бази одстојања унутар и између група), посебна мера, која укључује оба аспекта квалитета груписања је коефицијент силуете (енгл. *silhouette coefficient*). Ова мера се може одредити за ниво: ▸ појединачних објеката, ▸ група, и ▸ укупног резлтата груписања. У том контексту, нека се посматра група C_i, чији је елемент, између осталих, објекат i. Такође, нека су C_1, C_2,..., C_{k-1} суседене групе, различите од групе C_i.

Коефицијент силуете за сваки објекат групе дефинише се на следећи начин (*Rousseeuw*, 1987, стр. 56):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{50}$$

Елементи у изразу, илустровани на Слици 23, имају следеће значење:

- $a(i)$ – представља меру компактности за групу C_i, а одређује се као просечно одстојање између објекта i и осталих објеката који припадају истој групи као објекат i, односно групи C_i. Уколико је ова мера мања, група је компактнија. Заправо, вредност $a(i)$ је просечна дужина свих линија унутар групе C_i.
- $b(i)$ – се дефинише као просечно одстојање између објекта i у групи C_i и објеката у најближој суседној групи. Сходно наведеном, вредност $b(i)$ се одређује према процедури која садржи следеће кораке:

1) израчунати одстојања између објекта i и свих објеката у другим групама,

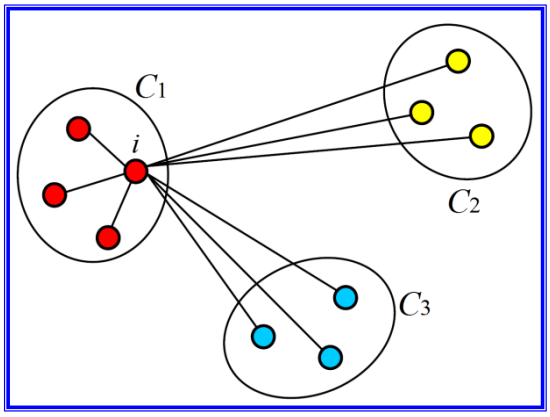
2) за сваку групу израчунати просек ових одстојања, и

3) од свих просечних вредности одстојања група, израчунатих на начин описан у корацима 1 и 2, пронаћи минималну вредност. Група коју карактерише најмање просечно одстојање назива се најближом суседном групом објекта i. Уколико је вредност $b(i)$ већа, објекат i је јасније одвојен, односно удаљенији од осталих група.

Заправо, просечно одстојање $b(i)$ одређује се на основу дужина линија које крећу од објекта i у групи C_1 ка објектима у преосталим групама (у овом случају C_2 и C_3).

После одређивања просечног одстојања између објекта i и ових група, бира се минимална вредност просечног одстојања.

- $max\{a(i), b(i)\}$ – већа вредност од двеју претходно дефинисаних вредности.



Слика 23: Приказ елемената укључених у израчунавање коефицијента силуете
Извор: *Rousseeuw* (1987, стр. 55)

Вредност коефицијента силуете варира између –1 и 1, односно, $-1 \leq s(i) \leq 1$. При тумачењу његових могућих вредности истичу се и образлажу три типичне ситуације:

- Уколико је група C_i компактна, то значи да између њених елемената постоји мала удаљеност, што доводи до тога да је вредност $a(i)$ мали позитиван број. Слично, ако је група C_i добро одвојена од њених суседа, $b(i)$ ће бити много већи позитиван број, такав да је $max\ \{a(i);\ b(i)\} = b(i)$, а $b(i) - a(i) \approx b(i)$. Дакле, уколико објекат i припада хомогеној групи, добро раздвојеној од свих њених суседа, тада је $s(i) \approx 1$.

- Насупрот томе, нека се претпостави да објекат i путем примењене процедуре груписања није добро класификован. Тада следи да је $a(i)$ велики позитиван број, док је $b(i)$) мали позитиван број. У том случају, $max\ \{a(i);\ b(i)\} = a(i)$, а $b(i) - a(i) \approx -a(i)$, имплицирајући $s(i) \approx -1$.

- Коначно, уколико у подацима постоји случајна структура, а не инхерентна структура у форми законитости, може се очекивати да ће $a(i)$ и $b(i)$ имати исту вредност (односно, најбоља група за објекат i суштински није боља од друге најбоље групе за исти објекат), имплицирајући $s(i) \approx 0$. Дакле, ако је $a(i) = b(i)$, или ако је објекат i једини објекат у групи C_i, вредност коефицијента силуете износи 0.

Генерално, позитивна вредност коефицијента силуете произлази из односа $a(i) < b(i)$. Уколико је вредност коефицијента силуете близу 1, то значи да је објекат i добро класификован. Негативна вредност индицира да је просечно одстојање $a(i)$ веће од минималног просечног одстојања $b(i)$. Ако је његова вредност близу –1, то значи да је

објекат *i* ближи објектима у некој другој групи, него објектима у истој групи и тада је природније објекат *i* пребацити из групе C_i у најближу суседну групу.

Вредност коефицијента силуете за једну групу одређује се као просечна вредност индивидуалних коефицијената силуете објеката који су њени елементи, док се за укупан резултат груписања просечна вредност коефицијената свих посматраних објеката, која се назива просечна ширина силуете, израчунава према следећој формули:

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(i). \qquad (51)$$

Да би се коефицијент сулуете могао користити за оцену валидности груписања пожељно је да постоје најмање две групе. У случају да је објекат *i* једини објекат у групи C_i, просечна вредност коефицијента силуете за ту групу је једнака 0 (ова вредност је арбитрарно дефинисана као неутрална вредност). Са становишта оцене квалитета укупног резултата груписања, односно коначног модела груписања, просечна вредност коефицијента силуета од –1, указује да је модел изузетно лошег квалитета, а вредност 1 указује на добро формирану структуру група и модел изврсног квалитета. Такође, са становишта дефинисања броја група, онај број за који се постиже максимална вредност $\bar{s}$, проглашава се оптималним. Коефицијенти силуете могу бити графички илустровани путем дијаграма силуете (*Vercellis,* 2009, стр. 314).

Једна од интерних мера за оцену квалитета груписања је мера која изражава пропорцију укупног варијабилитета који се може приписати варијабилитету између група (*Tufféry*, 2011, стр. 245). Ова мера се заснива на декомпоновању укупне суме квадрата одступања (*T*) и критеријуму смањења (суме квадрата) одступања унутар група (*W*) и повећања (суме квадрата) одступања између група (*B*). Симболички, може се изразити путем следећег количника:

$$R_g^2 = 1 - \frac{W}{T} = \frac{B}{T}. \qquad (52)$$

Вредност R_g^2 варира од 0 до 1, то јест, $R_g^2 \in [0;1]$. Уколико постоји само једна група $R_g^2 = 0$, а уколико је број група једнак броју објеката, тада је $R_g^2 = 1$, тако да са повећањем броја група, расте вредност R_g^2, и обрнуто. Генерално, уколико је вредност R_g^2 близу 1, конкретан резултат груписање је добро решење, јер су објекти унутар група међусобно хомогенији (мања вредност *W*), а групе добро раздвојене (већа вредност *B*). Сходно томе, уколико R_g^2 тежи 0, релевантност груписања се смањује. Међутим, то не значи да вредност овог коефицијента треба максимизирати. Заправо, његова максимална вредност произлази из максималног броја група (које садрже један

елемент), односно, $k = n$, што свакако не представља оптимално решење процеса груписања. Вредност R_g^2 треба да буде ближа 1, али без много група. Стога, (максимална) вредност R_g^2 не може бити једини критеријум за дефинисање оптималног броја група и, уопште, оцену квалитета груписања. Ипак, са становиште ове мере, опште правило за вредновање валидности груписања гласи:

- при поређењу финалних структура груписања, уколико је вредност R_g^2 већа, кореспондирајуће решење груписања је боље решење, јер су објекти у истим групама хомогенији[64], а групе боље раздвојене;
- при одређивању броја група, нагли пад вредности R_g^2 указује да је дошло до спајања две групе које се међусобно знатно разликују, тако да се број група који претходи овој наглој промени проглашава оптималним.

У непосредној вези са овом мером је мера која представља прираштај одстојања унутар група одређен након спајања две групе, а добија се као разлика између R_g^2 одређеног након новог спајања и R_g^2 одређеног у претходној итерацији. Нагли пораст вредности ове мере указује да је извршено спајање хетерогених група, тако да је број група који претходи овом наглом повећању оптималано решење.

Уобичајена мера која прати R_g^2 је псеудо-F мера[65]. Уколико је број група k, а број објеката n, тада ће псеудо-F мера бити:

$$\text{псеудо-}F = \frac{B/(k-1)}{W/(n-k)}, \qquad \text{или,} \qquad \text{псеудо-}F = \frac{R_g^2/(k-1)}{(1-R_g^2)/(n-k)}. \tag{53}$$

У начелу, вредност ове мере се смањује са смањењем броја група k, јер одступање унутар група постепено расте, а између група опада. Псеудо-F мера се користи за одређивање оптималног броја група тако што се прате промене њене вредности из корака у корак. Оног тренутка када дође до спајања две међусобно знатно различите групе, доћи ће и до велике промене одступања унутар и између група, а тиме и до наглог смањења њене вредности. Број група који непосредно претходи овој промени сматра се оптималним. Како је реч о мери којом се оцењује квалитет груписања са аспекта раздвајања између група, њена висока вредност указује на добро груписање. Ова мера није погодна за оцену квалитета једноструког хијерархијског повезивања.

Пошто је сврха анализе груписања формирање смислених група, веома је важно испитати валидност резултирајућег модела груписања са становишта могућности

[64] При томе треба водити рачуна о потенцијалној опасности формирања група са малим бројем елемената која проистиче из релације да се са повећањем броја група хомогеност група повећава, а величина група смањује.

[65] Префикс „псеудо" се уводи због аналогије ове мере са статистиком F теста у једнофакторској анализи варијансе, али не следи F дистрибуцију (*Tufféry*, 2011, стр. 246).

логичне и разумљиве интерпретације добијених група у контексту дефинисаних циљева анализе. Да би се групе интерпретирале, приступа се испитивању њихових карактеристика путем (*Shmueli et al*., 2005, стр. 220): ▸ одређивања кључних статистика сваке групе по свакој коришћеној варијабли у анализи; ▸ додатног испитивања формираних група према варијаблама које нису коришћене за груписање; ▸ одређивања (уколико постоји) доминантне карактеристике у свакој групи како би се на основу ње извршило декларисање, то јест, додељивање одређеног имена или ознаке свакој групи (на пример, Група поборника здраве хране). На основу наведених испитивања карактеристика група и њиховог упоређивања дефинише се профил сваке групе, као један од начина интерпретације резултата груписања, уз коришћење адекватних визуелних приказа.

Да би се установило да ли генерисане групе кореспондирају са стварно присутним законитостима у подацима, при оцењивању резултата груписања неопходно је укључути проверу стабилности добијених решења. Услед недостатка строго дефинисаних процедура за верификацију резултата, истраживачи често посежу за *ad hoc* процедурама, као што су: ▸ примена различитих метода груписања уз коришћење различитих мера блискости на истим подацима и упоређивање добијених резултата; ▸ случајна подела расположивих података на два дела и спровођење поступка груписања посебно за сваки део, а затим упоређивање резултата, дефинисање профила група и њихово испитивање; ▸ спровођење поступка груписања са мањим бројем варијабли и упоређивање са резултатима добијеним коришћењем свих варијабли путем којих је дефинисан профил јединица посматрања; ▸ спровођење низа итеративних процеса груписања са различитим бројем група како би се, одредио оптимални број група; ▸ спровођење низа итеративних процеса нехијерархијског груписања коришћењем различитог поретка опсервација све док се не постигне стабилност резултата.

Сумирајући наведена разматрања, јасно је да не постоје јединствени критеријуми, синтетички показатељи или „златна" правила за дефинисање доброг груписања. Међутим, било би неоправдано услед наведеног умањити улогу и важност претходно илустрованих методолошких оквира за оцену квалитета груписања.

11.3. Оцењивање класификационих модела

Оцењивање квалитета класификационих модела се односи на мерење њихове ефикасности, односно способности да правилно класификују елементе (објекте, опсервације) одређеног скупа / узорка података у претходно дефинисане класе

(категорије) и предвиде класу за нове податке. При поређењу и оцењивању класификационих *DM* резултирајућих модела треба узети у обзир следеће аспекте њиховог квалитета (*Han et al*., 2012, стр.369):

- предиктивну тачност: односи се на способност модела да тачно предвиди класу за нове или претходно непознате податке;
- брзину: односи се на компјутерско време потребно за развијање (укључујући и тестирање) модела;
- робустност: односи се на способност модела да тачно предвиди класу за податке који садрже шум, недостајуће или погрешне вредности;
- скалабилност: односи се на способност ефикасне конструкције модела за велике количине података;
- разумљивост: односи се на то колико је модел јасан и разумљив (углавном из перспективе субјективне процене корисника).

Иако класификациони модели могу бити оцењени према различитим критеријумима, тачност се користи као најчешћи евалуациони параметар. Тачност модела дефинише се као пропорција података који су тачно класификовани коришћењем формираног модела. У непосредној вези са својством тачности је концепт грешке. Под грешком се подразумевају погрешно класификоване опсервације.

Примарни извор за разумевање и одређивање мера за оцену квалитета класификационих модела је класификациона матрица. Матрица садржи сумарне информације о правилно и неправилно класификованим подацима на основу поређења унапред познате класе са класом одређеном од стране модела. Класификациона матрица за бинарни класификациони проблем са класама C_1 и C_2 (на пример, купац / није купац), чији редови и колоне кореспондирају са реалним (емпиријским) и предиктивним (моделираним) класама, приказана је у Табели 4. За конкретни проблем, издваја се једна класа као циљна или класа од интереса (C_1), а подаци деле на позитивне и негативне исходе те класе. На главној дијагонали матрице налазе се правилно класификоване опсервације, а садржај ћелија ван дијагонале указује на број погрешно класификованих опсервација. Заправо, могући исходи класификације су:

- f_{TP}: је број опсервација које припадају класи C_1, а које је класификациони модел заиста класификовао као чланове те класе (стварно позитивни исходи);
- f_{FP}: је број опсервација које припадају класи C_1, а које је класификациони модел неправилно класификовао као чланове класе C_2 (лажно негативни исходи);

- f_{FN}: је број опсервација које припадају класи C_2, а које је класификациони модел неправилно класификовао као чланове класе C_1 (лажно позитивни исходи);
- f_{TN}: је број опсервација које припадају класи C_2, а које је класификациони модел заиста класификовао као чланове те класе (стварно негативни исходи).

Табела 4: Класификациона матрица

Класе		Предиктивне класе	
		C_1 (+)	C_2 (–)
Емпиријске класе	C_1 (+)	f_{TP}	f_{FP}
	C_2 (–)	f_{FN}	f_{TN}

Наведене четири категорије исхода класификације користе се за израчунавање и разумевање многих евалуационих мера (*Witten et al*., 2011, стр. 157-177; *Shmueli et al*, 2005, стр. 49-50). Неке од мера за евалуацију квалитета класификационих модела су представљене у наставку текста.

Тачност (T) представља пропорцију успешно класификованих опсервација и одређује се као однос броја исправно класификованих опсервација (позитивних и негативних) и укупног броја опсервација. У директној вези са тачношћу модела је укупна грешка (G) модела, која представља однос броја погрешно класификованих опсервација и укупног броја опсервација. Односно, симболички:

$$T = \frac{f_{TP} + f_{TN}}{f_{TP} + f_{FP} + f_{FN} + f_{TN}} \qquad vs \qquad G = \frac{f_{FP} + f_{FN}}{f_{TP} + f_{FP} + f_{FN} + f_{TN}} \tag{54}$$

Поред ове две мере, често се у практичним ситуацијама нарочито у случајевима неравномерне дистрибуције јединица посматрања по класама и појаве ретких класа, користе комплементарни парови зависних мера за оцењивање модела, попут: сензитивност – специфичност, прецизност – одзив, позитивне предиктивне вредности – негативне предиктивне вредности.

Сензитивност (S^+), која се назива и стварно позитивна стопа, мери тачност у позитивним исходима класе од интереса, док специфичност (S^-) (која се назива и стварно негативна стопа), мери тачност у негативним исходима класе од интереса. Односно, симболички:

$$S^+ = \frac{f_{TP}}{f_{TP} + f_{FP}} \qquad vs \qquad S^- = \frac{f_{TN}}{f_{FN} + f_{TN}}. \tag{55}$$

Прецизност (P) мери пропорцију стварно позитивних исхода на следећи начин: од предиктивних исхода које је модел означио као позитивне, одређује се учешће

стварно позитивних, док одзив (*O*) такође мери пропорцију стварно позитивних исхода, али на следећи начин: од стварно позитивних исхода одређује се учешће исхода који су означени као позитивни и од стране модела. Дакле, одзив је мера тачно предвиђених позитивних исхода. Односно, симболички:

$$P = \frac{f_{TP}}{f_{TP} + f_{FP}} \qquad vs \qquad O = \frac{f_{TP}}{f_{TP} + f_{FN}}. \tag{56}$$

Позитивна предиктивна вредност (*PPV*) и негативна предиктивна вредност (*NPV*) су мере засноване на принципима прецизности, при чему се путем прве, мери тачност у позитивним исходима (односно, позитивној предикцији), а путем друге, тачност у негативним исходима (односно, негативној предикцији) путем следећих израза:

$$PPV = \frac{f_{TP}}{f_{TP} + f_{FN}} \qquad vs \qquad NPV = \frac{f_{TN}}{f_{TN} + f_{FN}}. \tag{57}$$

Мера перформанси модела која комбинује прецизност и одзив у један број позната је под називом *F*-мера (или *F*-скор) и израчунава се путем следећег израза:

$$F - \text{мера} = \frac{2 \times \text{прецизност} \times \text{одзив}}{\text{прецизност} + \text{одзив}} = \frac{2 \cdot f_{TP}}{2 \cdot f_{TP} + f_{FP} + f_{FN}} \tag{58}$$

Наведене мере се не односе само на класификационе проблеме са две класе, већ се користе и за класификационе проблеме са већим бројем класа, представљањем истих као серије бинарних проблема. За сваки мултикласни проблем, издваја се једна класа као циљна, а скуп података дели на позитивне и негативне исходе циљне класе, где категорији негативних исхода припадају исходи свих осталих класа осим циљне.

Једно од важних питања при креирању и оцењивању не само класификационих, већ, генерално свих предиктивних модела јесте начин поделе узорка посматраних података на подузорке и избегавање опасности од превелике прилагођености модела подацима, о чему је дискутовано у Потпоглављу 6.4. Генерално, уколико одређену ситуацију карактерише довољно велика количина података, најбоље је извршити случајну поделу података на три дела: узорак за тренирање модела, валидацију и тестирање, док се примена изабраног модела и оцена укупне грешке изабраног модела спроводи на новим подацима који нису учествовали у формирању модела. Осим ове поделе, уобичајено је, нарочито, у ситуацијама када не постоји довољно података за поделу на три дела, спровести двостепену методологију класификације и одвојити део података за тренирање (развој модела), а део за тестирање модела.

У зависности од тога да ли је циљ одређивање степена прилагођености модела подацима за учење, подешавање параметара и избор најбољег модела или генерализација модела и примена на потпуно новим подацима, тачност / грешка се може одредити на подацима узорка за тренирање, валидацију или тестирање. Међутим, за потребе евалуације коначног, изабраног класификационог модела морају се користити подаци који нису учествовали у процесу учења модела. Заправо, оцењивање тачности / грешке модела на бази података за учење није добар индикатор квалитета модела, већ само омогућава да се процени степен прилагођености модела подацима. Тачност модела изведеног на бази узорка за тренирање се оцењује на узорку за тестирање. Тест подаци се никада не користе за подешавање параметара модела.

За добијање оцене (енгл. *estimate*) стварне грешке, као и оцена осталих евалуационих мера класификационог модела користе се бројни методи (*Vercellis,* 2009, стр. 228-230; *Witten et al*., 2011, стр. 148-158).

У случају *holdout* метода, односно метода задржавања, на који је већ индиректно указано у претходним дискусијама о подели података, расположиви подаци се на случајан начин распоређују на две независне партиције. У пракси је уобичајено да се за сврхе извођења модела у узорак за учење алоцира више података, а тачност модела се оцењује на мање бројном, тестном узорку. Будући да је оцена тачности класификационог модела заснована на тестном узорку, она може бити прецењена или потцењена у односу на стварну грешку модела. Претпоставке за добијање поуздане оцене грешке модела применом овог метода су подела јединица посматрања базирана на једноставној процедури случајног узорковања, расположивост довољне количине података и избалансиран распоред података по класама. Уколико у структури расположивих података класе нису равномерно дистрибуиране, тада се узорковањем које је засновано на концепту стратификације обезбеђује да пропорције свих класа у узорцима за тренирање и тестирање буде приближно једнаке њиховим реалним пропорцијама у расположивим подацима, а формирани узорци буду репрезентативни. Такође, метод поновљеног случајног узорковање (енгл. *repeated random sampling method*) је варијанта *holdout* метода, у којој се *holdout* процедура понавља више пута. У свакој итерацији случајним путем (по могућству уз стратификацију) се бира узорак за тренирање модела, а преостали део података представља тест узорак. Укупна грешка модела се одређује као аритметичка средина грешака из спроведених итерација.

За метод унакрсне валидације (енгл. *cross validation method*) је карактеристично да се, најпре, оригинални скуп посматраних података случајно дели у *k* дисјунктних

делова приближно једнаке величине, а затим, се спроводи вишеструко понављање (кроз *k* итерација) процеса евалуације: у свакој итерацији, бира се једна партиција за тестирање, а унија свих осталих партиција, (*k*–1), за тренирање модела. Тренирање и тестирање се спроводе исти број пута. На крају процедуре, укупна тачност се израчунава као аритметичка средина *k* индивидуалних тачности. У практичним ситуацијама се обично преферира већи број итерација у циљу добијања робустнијих оцена тачности класификационих модела. Типична форма унакрсне валидације (која се углавном користи у *DM* софтверима) је 10-острука унакрсна валидација, где се оригинални скуп дели на 10 партиција. Генерално, метод унакрсне валидације је рачунски врло захтеван, али може дати добре резултате и у условима малог броја опсервација, јер се базира на њиховом максималном искоришћавању.

У условима мале количине података преферирани евалуациони методи су метод изостављања једног случаја (енгл. *leave-one-out*) и *bootstrap* метод. У првом случају реч је о методу заснованом на логици унакрсне валидације, с тим што је у питању *n*-тострука унакрсна валидација, где је *n* број јединица посматрања у посматраном узорку. У свакој од *n* итерација, класификациони модел се формира на *n*–1 случајева, а затим тестира на једној преосталој јединици. Метод изостављања једног случаја је рачунски захтеван метод у којем изостаје потреба за случајним узорковањем. Обезбеђује изузетно прецизне резултате и максималну искоришћеност доступних података. За разлику од *holdout* метода и метода унакрсне валидације, *bootstrap* метод је заснован на статистичкој процедури узорковања са понављањем.

Коначно, веома важан аспект оцењивања класификационих модела се односи на компарацију два (или више) модела и избор најбољег модела у одређеној ситуацији. Наиме, уколико је генерисано више класификационих модела поставља се питање који је модел најбољи за посматрани класификациони проблем. Може изгледати логично да је најбоље бирати модел који се карактерише са најмањом стопом погрешних класификација, али не треба превидети чињеницу да се модели примењују на подацима који нису учествовали у њиховом креирању. Услед наведеног, често је избор заснован на спровођењу одговарајућих статистичких тестова за утврђивање статистичке значајности разлике између израчунатих индикатора квалитета модела.

11.4. Оцењивање модела нумеричке предикције

Оцењивање квалитета резултирајућих модела нумеричке предикције базира се на разматрању истих критеријума као и приликом оцењивања класификационих модела:

тачност, брзина, робустност, скалабилност и разумљивост. Такође, базични принцип, који подразумева да се поуздано оцењивање перформанси модела заснива на коришћењу независних података који нису употребљени приликом креирања модела, подједнако је валидан као при оцењивању класификационих модела. Међутим, како је зависна варијабла нумеричка, мере квалитета модела и начин њиховог одређивања се знатно разликује од ситуација када је зависна варијабла категоријска.

За потребе разматрања овог проблема, неопходно је јасно истаћи разлику између предиктивне тачности и прилагођености модела емпиријским подацима (по аналогији са предиктивним и експланаторним моделирањем). С обзиром да се нумеричка предикција најчешће повезује са регресионом анализом, најједноставније је ову разлику објаснити на примеру регресионог модела. Заправо, у регресионој анализи циљ је пронаћи регресиони модел који се најбоље прилагођава емпиријским подацима. Након што су оцењени регресиони параметри, а тиме и дефинисани регресиони модел, поставља се питање колико је модел заиста успешан и колико добро репрезентује емпиријске податке. Познате мере за оцену репрезентативности регресионог модела и степена његове прилагођености стварним подацима су коефицијент детерминације и стандардна грешка регресије, подржани анализом резидуала. Регресиони модел који добро репрезентује емпиријске податке у узорку за тренирање може се користити за предвиђање зависне променљиве. Модел који је најбоље прилагођен подацима у узорку за тренирање не мора бити и модел који ће обезбедити високу тачност предвиђања. Стога, наведене мере нису добри показатељи квалитета модела са становишта његове успешности у предвиђању нових случајева.

За мерење квалитета нумеричке предикције постоји неколико алтернативних мера. У циљу њиховог дефинисања нека су симболом y_i означене реализоване, а симболом $\hat{y}_i$ предвиђене вредности зависне променљиве *Y*. Мере предиктивне тачности одређују се на основу предиктивне грешке, која представља разлику између реализоване / емпиријске и предиктивне вредности. Симболички, за опсервацију *i*, грешка предвиђања, као и њене апсолутна и квадратна вредност су, респективно:

$$e_i = y_i - \hat{y}_i, \qquad |e_i| = |y_i - \hat{y}_i|, \qquad e_i^2 = (y_i - \hat{y}_i)^2 . \tag{59}$$

На бази ових релација изведене су следеће мере[66] (статистике) за оцену тачности нумеричке предикције (*Witten et al*., 2011, стр. 180-181; *Shmueli et al*, 2005, стр. 68):

[66] Ове мере су означене симболима који представљају акрониме њихових назива на енглеском језику: *Mean Absolute Error*, *Mean Error*, *Mean Absolute Percentage Error*, *Root Mean Squared Error* и *Sum of Squared Errors*, респективно.

$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$, средња апсолутна грешка; (60)

$ME = \frac{1}{n}\sum_{i=1}^{n}e_i$, просечна грешка; (61)

$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{y_i}\right)\cdot 100$, средњи апсолутни процентуални износ грешке; (62)

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}$, корен средње квадратне грешке; (63)

$SSE = \sum_{i=1}^{n}e_i^2$, сума квадрата грешака (одступања). (64)

Сходно подели посматраних података, одређивање свих мера се спроводи на узорку за валидацију и / или узорку за тестирање, односно подацима који нису коришћени за избор предиктор варијабли и оцену параметара предиктивног модела. Кључни недостатак примене ових мера односи се на чињеницу да су под утицајем екстремних вредности. Поред наведних мера базираних на грешки у предвиђању, интерсантна мера за вредновање квалитета модела нумеричке предикције је коефицијент корелације између стварних и предвиђених вредности.

Свака од претходно наведених мера може се користити за компарацију два (или више) модела. У зависности од тога који су подаци коришћени за њихово израчунавање, утврђена разлика између конкретних мера два модела указује како на боље прилагођени модел, тако и на модел који има већу предиктивну моћ.

При интерпретацији наведених мера треба бити јако обазрив у смислу података на основу којих су исте израчунате: заправо, при оцењивању перформанси експланаторних модела, оцена регресионих параметара и мерење ваљаности регресионог модела (односно, његове прилагођености емпиријским подацима на бази разлике између стварних и оцењених, моделираних вредности зависне променљиве) се спроводи на узорку за учење, док се перформансе предиктивних модела оцењују путем предиктивне прецизност мерене на новим подацима који нису учествовали у креирању модела (а на бази разлике између стварних и предвиђених вредности зависне променљиве).

У наставку овог дела излагања апострофира се, у литератури ретко истраживани проблем примене регресионих модела за реализацију *DM* задатака, а који се односи на утицај велике количине података на ефикасност и ефективност ових модела. Заправо,

поставља се питање какве су импликације велике количине података на: (а) вредност стандардних статистика које се користе у процесу развоја и избора одговарајућег регресионог модела, и (б) димензионалност модела, односно укључивање објашњавајућих варијабли у регресиони модел или њихово искључивање из модела.

Посматрано из угла математичке нотације и (опште познатих) формула за израчунавање стандардних статитика за оцену регресионог модела (односно, *F* статистике, коефицијента детерминације, коригованог коефицијента детерминације, просечне суме квадрата одступања, *t* статистике), велика количина података узрокује следеће тенденције (*Hill et al*., 2004, стр. 241-243):

- Вредност *F* статистике, којом се тестира статистичка прихватљивост регресионог модела, са повећањем величине узорка такође се повећава, као и број степени слободе, док се критична вредност теста смањује. Као резултат тога, нулта хипотеза (која гласи да је коефицијент детерминације једнак нули, односно, вредност свих регресионих коефицијената је једнака нули) биће веома лако одбачена. Консеквентно, *F* статистика губи своју улогу коју има у стандардној процедури избора адекватног регресионог модела.
- Кориговани коефицијент вишеструке детерминације је статистика којом се мери да ли је повећање објашњеног варијабилитета зависне променљиве услед укључивања нове објашњавајуће променљиве у модел вредно и оправдано с обзиром на губитак броја степени слободе. Међутим, са повећањем величине узорка, вредност коригованог коефицијента вишеструке детерминације конвергира ка вредности некоригованог коефицијента вишеструке детерминације. Као последица тога, доводи се у питање смисао примене коригованог коефицијента.
- Просечна квадратна грешка је статистика чијим смањењем се генерално повећава предиктивна моћ модела, а која добија на значају посебно у ситуацијама када је из модела искључена нека од релевантних објашњавајућих променљивих. Значајно повећање величине узорка има за резултат да се вредност средње квадратне грешке приближава вредности варијансе, и такође постаје неосетљива на додавање или елиминисање из модела једне по једне варијабле.
- Статистика *t* теста се користи за тестирање хипотеза о статистичкој значајности регресионих коефицијената и детерминисање утицаја појединачних објашњавајућих варијабли на зависну варијаблу, а самим тим утиче на димензионалност коначног модела (у смислу укључивања или елиминисања појединих

објашњавајућих променљивих у модел). Са значајним повећањем величине узорка, оцена регресионог коефицијнта постаје изузетно прецизна, а њена стандардна грешка постаје веома мала. Као резултат тога, нулта хипотеза (која гласи да одговарајућа објашњавајућа променљива не утиче на зависну променљиву, односно да је одговарајући регресиони параметар у основном скупу једнак нули) ће бити лако одбачена, чиме ефективност *t* статистике у тестирању статистичке значајности регресионих коефицијената и мерењу статистичке значајности утицаја појединих објашњавајућих променљивих на зависну променљиву постаје дискутабилна.

Логистичка регресија се, у односу на вишеструку линеарну регресију, заснива на другачијим теоријским претпоставкама, методима за оцену параметара и тестовима за оцену статистичке значајности регресионих коефицијента. Упркос томе, анализа утицаја велике количине података на кретање вредности, ефикасност и ефективност кључних статистика за избор одговарајућег логистичког регресионог модела (статистике одступање, псеудо коефицијента детерминације и *Wald* статистике) упућује на исте закључке као код вишеструке линеарне регресије (*Hill et al*., 2004, стр. 243-246).

Имајући у виду наведено, неспорно, развој регресионог модела у условима великих количина података је изузетно отежан. Међутим, то не значи да регресиона анализа у *DM* окружењу нема практични значај, већ само јасно упућује на неопходност реализације *DM* задатака кроз форму колаборативних пројекта свих учесника у *DM* процесу са доминантном улогом статистички образованих истраживача. Осим тога, не постоји метод који може бити ефикасан у раду са десетинама милиона записа (и знатно више) услед чега се и развијају нове технологије за обраду, складиштење и, генерално, управљање количинама података у тим размерама.

Део IV

ЕМПИРИЈСКО ИСТРАЖИВАЊЕ: АНАЛИЗА ПРОБЛЕМСКИХ СИТУАЦИЈА ПРИМЕНОМ *DATA MINING* ПРИСТУПА

12. РЕАЛИЗАЦИЈА DATA MINING ЗАДАТАКА У КОНТЕКСТУ ДЕФИНИСАНЕ ПРОБЛЕМСКЕ СИТУАЦИЈЕ 1

Полазећи од теоријских постулата изложених у претходним Поглављима и узимајући у обзир чињеницу да многи *DM* проблеми укључују временске аспекте, у овом Поглављу је представљено емпиријско истраживање засновано на анализи берзанских података у форми временских серија берзанских индекса, као кључних показатеља динамике финансијског тржишта. У том смислу, конципиран је и приказан методолошки оквир истраживања базиран на интегрисаној имплементацији *SAX* алгоритма и хијерархијске агломеративне процедуре груписања у функцији развоја модела сличности и формирање хомогених група одабраних берзи на основу временских серија референтних берзанских индекса. Поред наведеног, добијени резултати примењене методологије су на одговарајући начин приказани и дискутовани.

12.1. Идентификовање проблемске ситуације

Савремене тржишне економије су високо монетизоване и стога трансакције са новцем и хартијама од вредности (на основу којих се пласира или набавља новац или капитал) представљају механизме који обезбеђују значајне изворе финансирања економског раста. Наиме, финансијски систем је битан сегмент економског система, тако да је опште признат став по којем између привредног развоја и развијености финансијског система постоји висок степен међузависности. *De facto*, финансијска тржишта, као кључни елемент финанасијског система, имају велики утицај на вођење монетарне и економске политике како развијених, тако и земаља у развоју.

Нераскидиви део финансијког тржишта сваке тржишно оријентисане економије је берзанско тржиште (односно, берза). Као облик организовања финансијског тржишта, берзе представљају финансијске институције у којима овлашћена лица обављају трговину одређеним берзанским материјалом, на утврђеном месту[67] и по усвојеним правилима (*Dugalić & Štimac*, 2009, стр. 20). Савремено берзанско пословање обухвата куповину и продају бројних финансијских инструмената, али, без сумње, по броју обављених трансакција доминирају акције као власничке хартије од вредности. У том процесу трговања, кроз односе понуде и тражње, формирају се цене финансијских инструмената.

[67] Треба нагласити да је развој компјутерске технологије и електронског трговања узроковао редефинисање места као елемента дефиниције берзе, које се не може више третирати као једна локација или сала за трговање, већ се берзанска трговина одвија кроз информациони систем берзе и електронске платформе за трговање.

За анализу берзанског пословања и доношење одлука о куповини и продаји акција упоредо коегзистира више показатеља промена цена акција, међу којима посебно место припада берзанским индексима. Ови индекси не одређују како ће се цене акција мењати у будућности, већ узимају у обзир све промене у прошлости, изражене у терминима базних индекса. У том контексту, берзански индекси који региструју ценовне промене у остварeним берзанским трансакцијама, одражавају реалне односе који владају на берзи, али су и показатељ развоја привреде у целини.

У настојањима да инвеститори обезбеде информације о перформансама појединих финансијских инструмената, или портфолија финансијских инструмената, развијене су серије берзанских индекса (*Jakšić*, 2016, стр.174). Оне служе (а) као израз трговања у статичком смислу, за одређивање стања на тржишту, и (б) као основа за утврђивање динамике тржишта, то јест, *ex post* анализу кретања цена акција. Наиме, серије берзанских индекса се формирају у циљу унапређења процеса информисања учесника на берзи и (инвестиционе) јавности, јачања транспарентности берзанских догађања и обезбеђења упоредивости података са различитих берзанских тржишта.

У основи, берзански индекси, као показатељи промена цена хартија од вредности којима се тргује на берзи, својим кретањем треба јасно да сигнализирају шта се дешава на финансијском тржишту. Сходно томе, уколико се анализа берзанских индекса прошири укључивањем и других кључних економских показатеља, могуће је, у зависности од дефинисаног територијалног обухвата анализе, идентификовати релевантне релације између различитих индекса и сектора економије унутар једне економије, као и релације које постоје на регионалном или светском нивоу.

Однос између берзанског тржишта и економије представља честу тему политичких и истраживачких дебата и дискусија. Заправо, аналитичари покушавају да, анализом историјских података, идентификују законитости о кретању тржишта, као и обрасце који се понављају у различитим фазама раста и пада тржишта, а затим да на основу њих објасне актуелна и предвиде будућа кретања не само на финансијском тржишту, већ и у економији уопште.

Савремено берзанско пословање практично је немогуће замислити без компјутеризације и нових *ICT* решења, које је, самим тим, праћено генерисањем и складиштењем велике количине података о трговању на берзи. Због тога се као важан проблем појављује проналажење ефикасних начина за сумирање и визуелизацију великих количина берзанских података који ће омогућити добијање корисних информација о понашању берзанског тржишта. Појава *DM* алата и софистицираних

технологија база података обезбедила је основу за релативно једноставно и лако решавање овог проблема. Заправо, *DM* приступ може бити коришћен за различите аспекте анализе берзанског тржишта, као што су откривање тенденција у ценовним променама финансијских инструмената, дефинисање профила компанија које послују на берзи, управљање ризиком портфолија, предвиђање берзанских индекса и слично.

Последњих година забележен је пораст истраживачког интересовања за издвајање значајних информација и законитости из берзанских података путем *DM* приступа. Наведену констатацију илуструје повећани број објављених радова у којима су разматрани различити аспекти пословања на берзанском тржишту уз примену *DM* метода. Због инхерентних својстава динамичности, комплексности и стохастичности берзанских тржишта, у знатном броју публикација превасходно је тангирана проблематика предвиђања берзанских трендова. Одлични прикази успешних имплементација *DM* метода, не само у реализацији задатка предвиђања берзанских трендова, већ и анализи осталих аспеката берзанског пословања, уз детаљне прегледе постојеће литературе, могу се пронаћи у следећим радовима: *Nanda et al*. (2010); *Wu et al*. (2014); *Aghabozorgi & Teh* (2014); *Peker et al*. (2017).

У принципу, *DM* приступ може бити примењен за било који тип берзанских података (нумеричких, просторних, темпоралних, мултимедијалних). Из перспективе доношења квалитетних одлука од стране свих учесника на тржишту, у методолошком смислу један од највећих истраживачких изазова односи се на анализу историјских података представљених у форми берзанских временских серија. Анализа временских серија, генерално, је повезана са откривањем корисних образаца и законитости у структури временских серија, као и предвиђањем будућих вредности посматране појаве. Класична статистичка и економетријска анализа временских серија се широко примењују у различитим истраживањима повезаним са анализама берзанског тржишта. Међутим, ови традиционални методи су базирани на строгим претпоставкама (попут линеарности) које је врло тешко испунити када су у питању динамичке берзанске временске серије. За разлику од њих, *TSDM* апликације имају знатно мање ограничења и могу се успешно применити у идентификовању комплексних карактеристика и предвиђању стохастичких, непериодичних, нелинеарних и хаотичних временских серија које често укључују и структурне ломове, типичне за берзанска тржишта.

Полазећи од чињеница: (а) да су берзански индекси најквалитетнији показатељи динамике берзанског тршишта, и (б) да велика количина складиштених података у форми временских серија берзанских индекса садржи скривене законитости које могу

имати значајан предиктивни потенцијал у контексту доношења инвестиционих одлука, емпиријско истраживање у овом делу дисертације, засновано на методологији интегрисане имплементације *SAX* алгоритма и хијерархијске агломеративне процедуре груписања, је усмерено на развој модела сличности и формирање хомогених група одабраних берзи на основу временских серија референтних берзанских индекса.

Сходно наведеном, предмет овог истраживања је 14 водећих берзанских индекса одабраних европских земаља. Основни циљ истраживања састоји се у утврђивању сличности и класификацији / груписању берзи одабраних земаља у интерно хомогене и екстерно хетерогене групе према вредностима берзанских индекса у анализом обухваћеном периоду. Генерално, сврха спроведеног истраживања садржана је у обезбеђивању иновативног проширења методолошког приступа у анализи берзанских података, који треба да омогући идентификовање законитости у форми модела сличности формираних група, као корисног извора за разумевања глобалних финансија и објашњење, у литератури потврђеног, става да кретања (цена акција) на једној берзи у једној земљи утичу на кретања на другим берзама у другим земљама или у региону.

Сходно дефинисаном и изложеном предмету, као и постављеним циљевима истраживања, формулисане су следеће истраживачке хипотезе, које ће бити предмет провере у овом емпиријском делу дисертације:

(H–1_{12}): Складиштење велике количине берзанских података условљава потребу за применом иновативних методолошких приступа у анализи берзанских тржишта.

(H–2_{12}): Берзе земаља Централне и Југоисточне Европе, које су чланице бивше Југославије, карактеришу се високим степеном међусобне сличности са становишта тенденција у кретању вредности њихових берзанских индекса.

12.2. Методолошки аспекти истраживања

У складу са претходно датим општим напоменама о берзанском пословању, као и, у том контексту, образложеним предметом, дефинисаним циљем и прецизираним истраживачким хипотезама, дизајниран је оквир емпиријског истраживања за идентификовање тржишта са најсличнијим тенденцијама у кретању вредности берзанских индекса представљених у форми дневних временских серија. Кључне фазе емпиријског истраживања су: ▸ избор података (дефинисање временског и просторног обухвата и извора података), ▸ претпроцесирање сирових, оригиналних података, ▸ обрада претпроцесираних података, и ▸ приказ и интерпретација добијених резултата. Свака фаза је детаљно елаборирана у наставку текста овог Поглавља.

За потребе реализације планираног истраживања коришћене су финансијске временске серије које се односе кретање вредности (водећих) индекса берзанских тржишта одабраних земаља Централне и Југоисточне Европе. Одабране берзе и берзански индекси представљени су у Табели 5. У анализу је укључен и регионални индекс који израђује Будимпештанска берза са званичним називом централноевропски „*blue chip*" индекс (енгл. *Central European Blue Chip Index – CETOP*). Реч је о бенчмарк индексу у чију корпу су укључене акције компанија са највећом тржишном вредношћу и прометом, а којима се тргује на берзама следећих земаља: Мађарска, Чешка Република, Пољска, Словачка, Словенија, Хрватска и Румунија (*https://www.bse.hu/*).

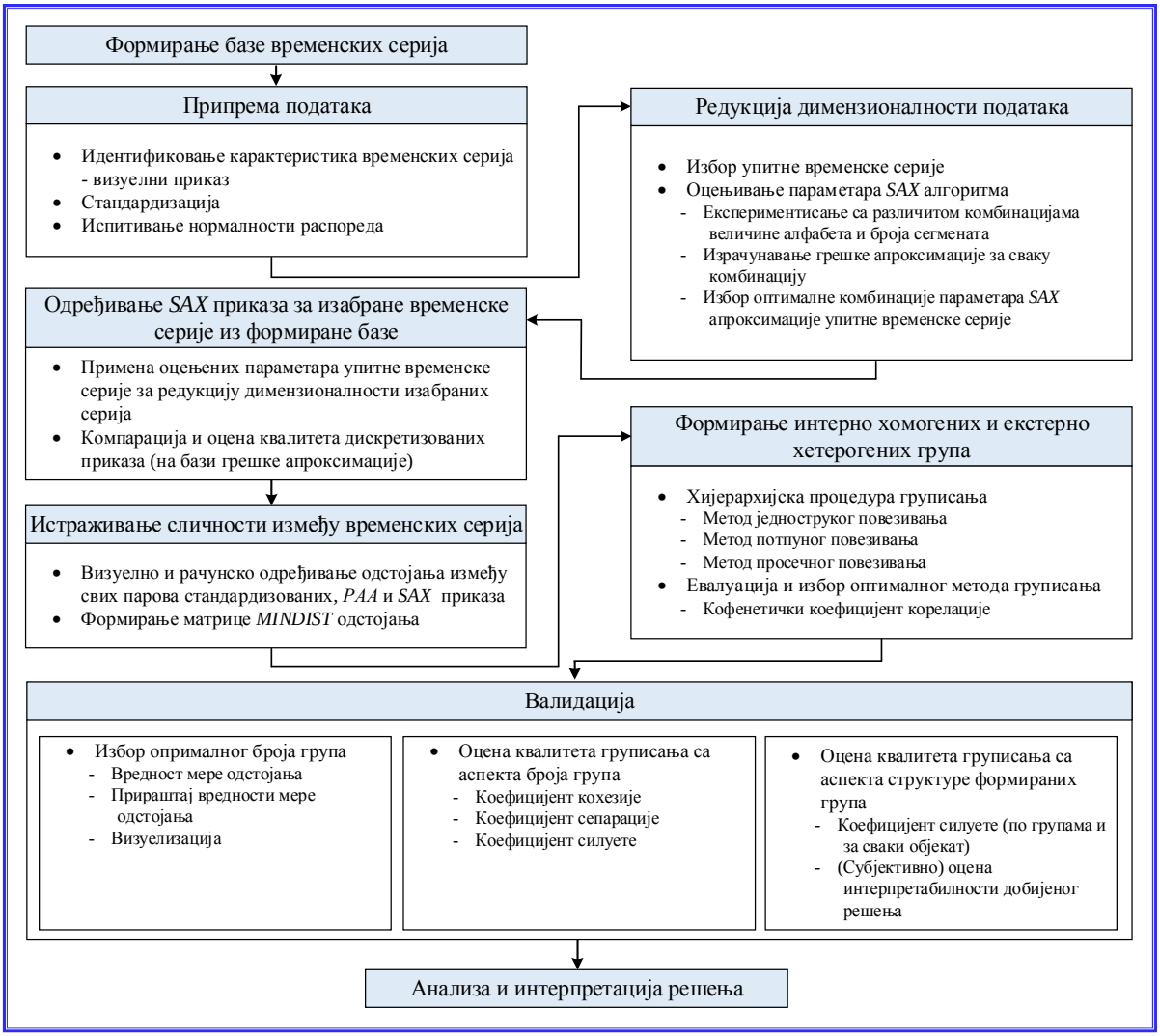
Табела 5: Берзанска тржишта и берзански индекси коришћени у истраживању

Берза	Држава	Индекс	Сајт берзе	Ознака
Београдска	Србија	BELEX-15	*http://www.belex.rs/*	*SE*-01
Загребачка	Хрватска	CROBEX-10	*http://zse.hr/*	*SE*-02
Црногорска	Црна Гора	MONEX	*http://www.mnse.me/*	*SE*-03
Сарајевска	БиХ Федерација	SASX-10	*http://www.sase.ba/*	*SE*-04
Бањалучка	Република Српска	BIRS	*https://www.blberza.com/*	*SE*-05
Македонска	Македонија	MBI-10	*http://www.mse.mk/*	*SE*-06
Љубљанска	Словенија	SBITOP	*http://www.ljse.si/*	*SE*-07
Букурештанска	Румунија	BET	*http://www.bvb.ro/*	*SE*-08
Софијска	Бугарска	SOFIX	*http://www.bse-sofia.bg/*	*SE*-09
Словачка (у Братислави)	Словачка	SAX (BR)	*http://www.bsse.sk/*	*SE*-10
Будимпештанска	Мађарска	BUX	*https://www.bse.hu/*	*SE*-11
Варшавска	Пољска	WIG-20	*https://www.gpw.pl/*	*SE*-12
Прашка	Чешка Република	PX	*https://www.pse.cz/*	*SE*-13
Будимпештанска	Мађарска	CETOP	*https://www.bse.hu/*	*SE*-14

Анализом су обухваћене серије дневних података посматране варијабле у периоду од (прве половине) 2010. до (прве половине) 2017. године, укључујући 1800 дана трговања на свакој од одабраних берзи. Подаци су обезбеђени из секундарних извора, односно, електронских база са званичних сајтова посматраних берзи.[68] За статистичку обраду података коришћени су стандардни алати: статистички програмски пакет за друштвене науке *IBM SPSS* верзија 20.0, док су остала неопходна (табеларна) израчунавања спроведена у програму *Excel* 2007, као део пакета *Microsoft Office*. Осим тога, за редукцију димензионалности временских серија и њихову трансформацију у *SAX* секвенце (речи), са пратећим графичким илустрацијама, дизајниран је специјални програм написан у програмском језику *Java*, који је назван *MMS Statistics*.

[68] Посебно се наглашава да је коришћена база временских серија формирана на основу података преузетих са званичних сајтова берзи током вишегодишњег периода. Разлог томе се налази у чињеници да у реалном времену (у тренутку посете сајту), системи за дистрибуцију података неких берзи ограничавају расположивост података у смислу дужине њиховог временског обухвата.

Свеобухватан методолошки оквир за развој модела сличности тенденција берзанских индекса са током спроведеног истраживања представљен је на Слици 24, где се јасно уочавају два главна потпроцеса: ▸ први, који се односи на одређивање сличности заснованој на редукованим *SAX* приказима и одређивању компатибилних мера одстојања, и ▸ други, који обухвата примену различитих метода хијерархијске агломеративне процедуре формирања хомогених група посматраних објеката.



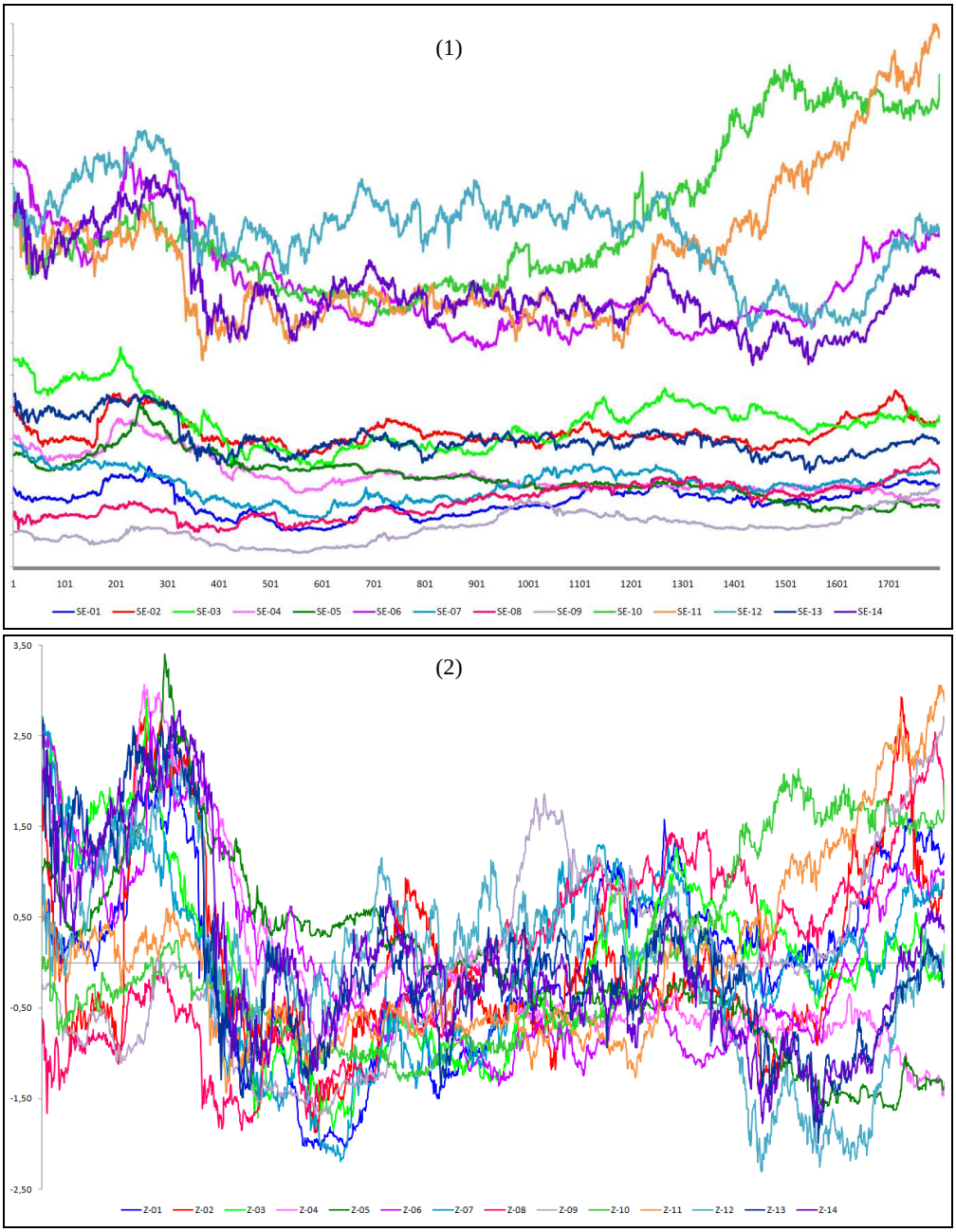
Слика 24: Шематски приказ концептуално-методолошког оквира истраживања 1

12.3. Резултати истраживања

У наставку текста се, сходно конципираном методолошком оквиру истраживања, презентују, евалуирају и анализирају резултати обраде података, а који су непосредо повезани са дефинисаним истраживачким хипотезама.

12.3.1. Карактеристике временских серија у формираној бази

За реализацију планираног истраживања сличности временских серија одабраних берзанских индекса и обезбеђења адекватних улазних података за имплементацију *SAX* алгоритма, у оквиру иницијалне припреме података спроведене су следеће активности: подешавање дужине серија (будући да се овом приликом истраживање сличности спроводи путем потпуног упаривања временских серија које полази од претпоставке да су све временске серије исте дужине) и представљање серија путем погодних графичких приказа (да би се обезбедиле кључне информације о њиховим својствима).



Слика 25: Оригиналне временске серије (1) и њихове стандардизоване форме (2)

Извор података: формирана *MMS* база

Слика 25 илуструје дневно кретање анализом обухваћених берзанских индекса у посматраном периоду. На основу овог приказа јасно се уочава да су све генерисане временске серије високо димензионалне нестационарне серије са променљивим варијабилитетом, што се манифестује различитим облицима кривих и различитим тенденцијама у појединим сегментима кривих. Берзански индекси се исказују у различитим јединицама (индексним поенима или неким вредносним категоријама, уз примену различитих методологија за њихово израчунавање), што утиче на вредности и позицију кривих на ординати, тако да овај визуелни приказ директно указује и на неопходност стандардизације података. Да би се обезбедила упоредивост временских серија у функцији истраживања и мерења (квантификовања) њихове сличности потребно је спровести стандардизацију истих. Визуелни приказ стандардизованих (међусобно упоредивих) форми временских серија посматране варијабле представљен је, такође, на Слици 25. Начелно, стандардизација података је претпроцесна активност која је инкорпорирана у *SAX* процедуру трансформације великих, високо димензионалних скупова података у приказе нижег реда димензионалности.

12.3.2. Резултати примене *SAX* алгоритма

Пошто су у формираној *MMS* бази обезбеђене временске серије берзанских индекса једнаке дужине и извршена визуелизације њихових карактеристика, наредна истраживачка активност је усмерена на детерминисање оптималне комбинације параметара *SAX* приказа. За потребе формирања редукованих и међусобно упоредивих приказа свих елемената базе, временска серија *BELEX*-15 је издвојена и означена као упитна серија на основу које су оцењени *SAX* параметри.

У складу са процедуром примене *SAX* алгоритма, представљеној у Поглављу 10, извршена је, сходно комбинацији кључних параметара[69], конструкција већег броја *SAX* приказа временске серије *BELEX*-15. Заправо, формирано је 15 редукованих приказа као резултат комбинација следећих величина алфабета и броја сегмената серије: α = [4, 5, 6] и k = [10, 12, 15, 18, 20]. Сходно броју сегмената и дужини серије од 1800 података, експериментисање је спроведено уз следеће величине сегмената: n_k = [180, 150, 120, 100, 90]. За оцену квалитета сваке комбинације параметара, односно резултирајућег апроксимативног приказа коришћена је грешка апроксимације, E_{SAX}. У Табели 6

[69] Избор броја сегмената и величине алфабета примарно је детерминисан карактеристикама података. Временске серије са усклађеним обрасцима понашања могу се представити са мањим бројем сегмената, док редукција временских серија које карактеришу фреквентне промене образаца понашања захтева већи број сегмената како би се идентификовале критичне промене. У погледу другог параметра, резултати бројних студија указују да вредности α = 3 или α = 4 представљају добар избор за већину база временских серија (*Lin & Li*, 2009, стр. 467).

представљен је скуп свих комбинација параметара и кореспондентних грешака апроксимације.

Табела 6: Величина грешке апроксимације за различите комбинације *SAX* параметара

Величина алфабета	Величина сегмената (број сегмената)					Просечна грешка
	n_k=180 (k=10)	n_k=150 (k=12)	n_k=120 (k=15)	n_k=100 (k=18)	n_k=90 (k=20)	
α = 4	19,880	20,146	20,965	19,737	19,470	20,040
α = 5	19,511	19,070	20,086	18,348	18,352	19,073
α = 6	19,606	18,943	19,342	16,900	17,130	18,384
Просечна грешка	19,666	19,386	20,131	18,328	18,317	/

Анализом резултата емпиријског оцењивања (Табела 6) може се запазити да је најмања просечна грешка, при варирању броја сегмената k, добијена за величину алфабета α = 6 (E_{SAX} = 18,384). Посматрано по колонама, при варирању величине алфабета α, уочава се незнатна разлика између просечних грешака апроксимације за k = 18 и k = 20 (E_{SAX} = 18,328 и E_{SAX} = 18,317, респективно). Такође, добијени резултати сугеришу да се најмања грешка апроксимативног приказа постиже за комбинацију параметара α = 6 и k = 18 (E_{SAX} = 16,9). Сумирањем изнетих разматрања, ова комбинација параметара се означава оптималним избором. Интересантно је указати на чињеницу да овим избором стопа компресије износи 0,01, односно оригинална серија n димензионалности (n = 1800) се редукује у серију k димензионалности (k = 18).

Ток трансформације оригиналне временске серије *BELEX*-15 у *SAX* апроксимацију за α = 6 и k =18 представљен је на Слици 26. Према алгоритамској процедури, најпре је извршена стандардизација временске серија тако да је њена аритметичка средина 0, а стандардна девијација 1. Након тога, поделом серије на 18 сегмената једнаке величине и израчунавањем аритметичке средине по сваком сегменту, редукована је димензионалност серије и обезбеђен њен приказ у форми *PAA* апроксимације. У наредном кораку *PAA* приказ је дискретизован. За те сврхе, с обзиром да је величина алфабета 6, одређено је 5 преломних тачака. На тај начин генерисано је α области једнаке површине испод нормалне криве и свакој од њих додељен одговарајући симбол, при чему су вероватноће симбола међусобно једнаке. Сходно потребама овог истраживања, у Табели 7 представљене су вредности преломних тачака за α = [3,...,12], а њихове вредности су утврђене коришћењем таблице стандардизованог нормалног распореда. Наиме, за трансформацију *PAA* приказа у симболе коришћене су преломне тачке које одговарају величини алфабета 6, док су за одређивање грешке апроксимације коришћене преломне тачке које одговарају

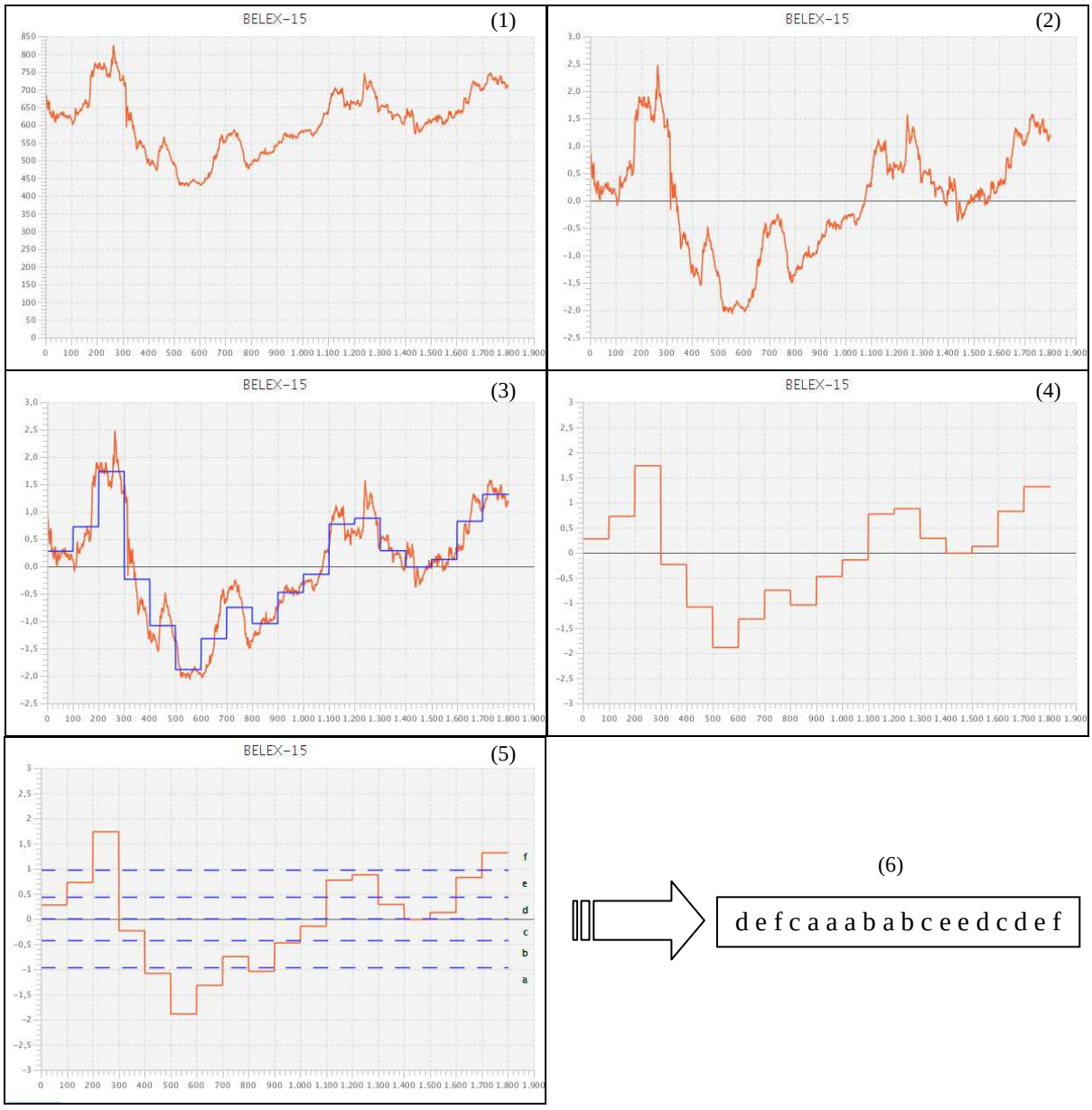
дуплираној величини алфабета. *PAA* коефицијенти су трансформисани у симболе на следећи начин: ▸ сви *PAA* коефицијенти чија је вредност мања од -0,97 означени су симболом „*a*”, ▸ сви *PAA* коефицијенти чија је вредност једнака или већа од -0,97, а истовремено и мања од -0,43 означени су симболом „*b*”, ▸ сви *PAA* коефицијенти чија је вредност једнака или већа од -0,43, а истовремено и мања од 0 означени су симболом „*c*”, ▸ сви *PAA* коефицијенти чија је вредност једнака или већа од 0, а истовремено и мања од 0,43 означени су симболом „*d*”, ▸ сви *PAA* коефицијенти чија је вредност једнака или већа од 0,43, а истовремено и мања од 0,97 означени су симболом „*e*”, и ▸ сви *PAA* коефицијенти чија је вредност једнака или већа од 0,97 означени су симболом „*f*”. Спајањем ових симбола добијена је *SAX* апроксимација, односно оригинална временска серија *BELEX*-15 од 1800 података је трансформисана у *SAX* реч „*defcaaababceedcdef*” (за $\alpha = 6$), која има 18 карактера.

Табела 7: Вредности преломних тачака

Преломне тачке (β_i)	Величина алфабета (α)									
	3	4	5	6	7	8	9	10	11	12
β_1	-0,43	-0,67	-0,84	-0,97	-1,07	-1,15	-1,22	-1,28	-1,34	-1,38
β_2	0,43	0	-0,25	-0,43	-0,57	-0,67	-0,76	-0,84	-0,91	-0,97
β_3		0,67	0,25	0	-0,18	-0,32	-0,43	-0,52	-0,60	-0,67
β_4			0,84	0,43	0,18	0	-0,14	-0,25	-0,35	-0,43
β_5				0,97	0,57	0,32	0,14	0	-0,11	-0,21
β_6					1,07	0,67	0,43	0,25	0,11	0
β_7						1,15	0,76	0,52	0,35	0,21
β_8							1,22	0,84	0,60	0,43
β_9								1,28	0,91	0,67
β_{10}									1,34	0,97
β_{11}										1,38

Аналогно поступку који је представљен за временску серију *BELEX*-15, одређене су *SAX* апроксимације осталих временских серија садржаних у формираној бази, коришћењем вредности параметара изабраних у конструкцији *SAX* приказа временске серије *BELEX*-15. Сумарни резултати трансформације 14 (нумеричких) временских серија берзанских индекса у (симболичке) *SAX* секвенце, представљени су у Табели 8.

Будући да не постоји прецизно дефинисани критеријум у смислу која је вредност грешке апроксимације индикатор квалитетне апроксимације, значајно је приметити да се добијене вредности грешака апроксимација знатно не разликују од грешке серије *BELEX*-15, односно серије на основу које су детерминисани *SAX* параметри. Наведено упућује на закључак да, са становишта квалитета приказа, резултирајуће *SAX* апроксимације свих серија у бази представљају задовољавајућа решења.



Слика 26: Различите форме приказа кретања вредности индекса *BELEX*-15

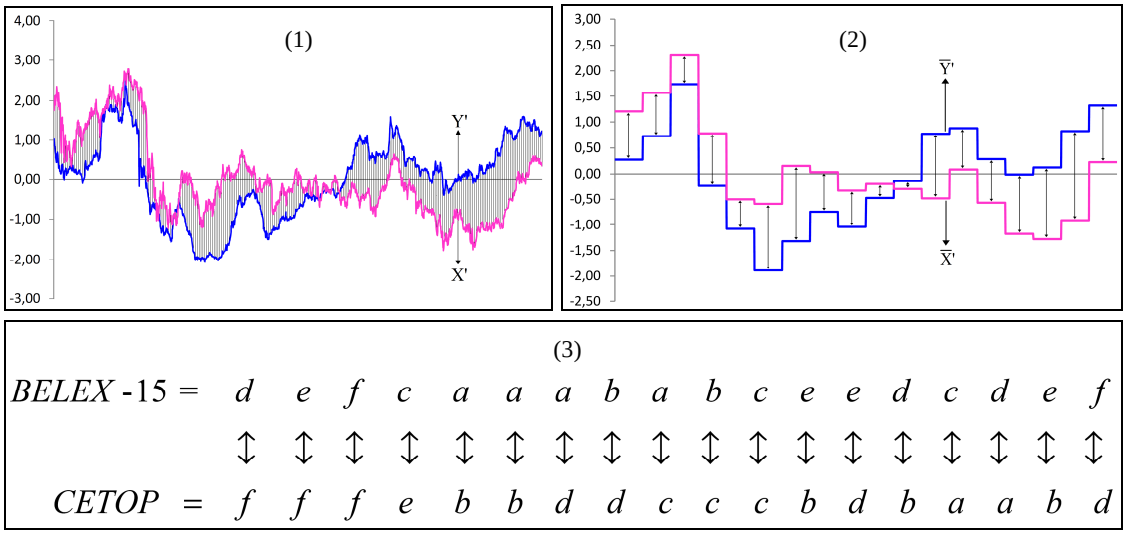
(1) Оригинална врменска серија; (2) Стандардизована временска серија;
(3) Стандардизована серија и *PAA* приказ; (4) *PAA* приказ;
(5) Преломне тачке и дискретизација; (6) *SAX* реч

С обзиром да су обезбеђени међусобно упоредиви прикази, спроведено је упаривање свих апроксимација и израчунавање одстојања између сваког пара *SAX* секвенци. На Слици 27 демонстриран је поступак одређивања одстојања између две *SAX* речи, и то временских серија *BELEX*-15 и *CETOP*. Истовремено је потврђено да мера одстојања између два симболичка низа, креирана путем *SAX* алгоритма, представља доњу границу Еуклидског одстојања између две стандардизоване оригиналне серије. Том приликом је имплементирана *MINDIST* функција уз коришћење одговарајуће статистичке табеле одстојања између упарених симбола два

SAX приказа. У питању је Табела 9, димензија 6×6, чија су поља одређена на бази Табеле 7 и израза 45. У основи, овај поступак је примењен на све комбинације парова *SAX* приказа, а резултирајуће *MINDIST* мере одстојања представљене су путем матрице одстојања типа 14×14 у Табели 10.

Табела 8: *SAX* прикази генерисани за „оптималне" вредности параметара

Берзански индекс	*SAX* реч (α = 6 и k =18)	Грешка апроксимације
BELEX-15	„ *d e f c a a a b a b c e e d c d e f* "	16,900
CROBEX-15	„ *c d f d b a b d c b c c c c b c f f* "	23,876
MONEX	„ *f f f d a a a b a b b d e e d c c d* "	14,910
SASX-10	„ *f f f f d b c d c c b b b b b b b a* "	15,476
BIRS	„ *e f f f e d e d d c b c c b b a a a* "	15,107
MBI-10	„ *f e f f d c b b a a b b b a b b d e* "	16,078
SBITOP	„ *f f e c a a a a b c e e e e c c c d e* "	15,222
BET	„ *a b c a a a a a b c d e e f f e e e f* "	14,112
SOFIX	„ *b b c b a a a b d e f e e d c d f f* "	14,427
SAX-BR	„ *c c c c b b a a b b b b d e f f f f* "	11,148
BUX	„ *d d d b b b b b b b b b c d e f f f* "	16,935
WIG-20	„ *d f f d c c e d d e d d d b a a a c* "	16,762
PX	„ *f f f c b b c c c d c c d c b a b c* "	20,041
CETOP	„ *f f f e b b d d c c c b d b a a b d* "	20,131



Слика 27: Визуелизација одређивања одстојања између два приказа временских серија

(1) Еуклидско одстојање између две стандардизоване серије, $n_1 = n_{14} = 1800$, $D(Y',X')$=45,691;

(2) Одстојање између два *PAA* приказа, $k_1 = k_{14} = 18$, $D = (\bar{Y}', \bar{X}') = 43{,}374$;

(3) Одстојање између два *SAX* приказа за $\alpha = 6$, $MINDIST(\hat{Y}', \hat{X}') = 22{,}544$

Табела 9: Статистичка табела за *MINDIST* функцију, α = 6

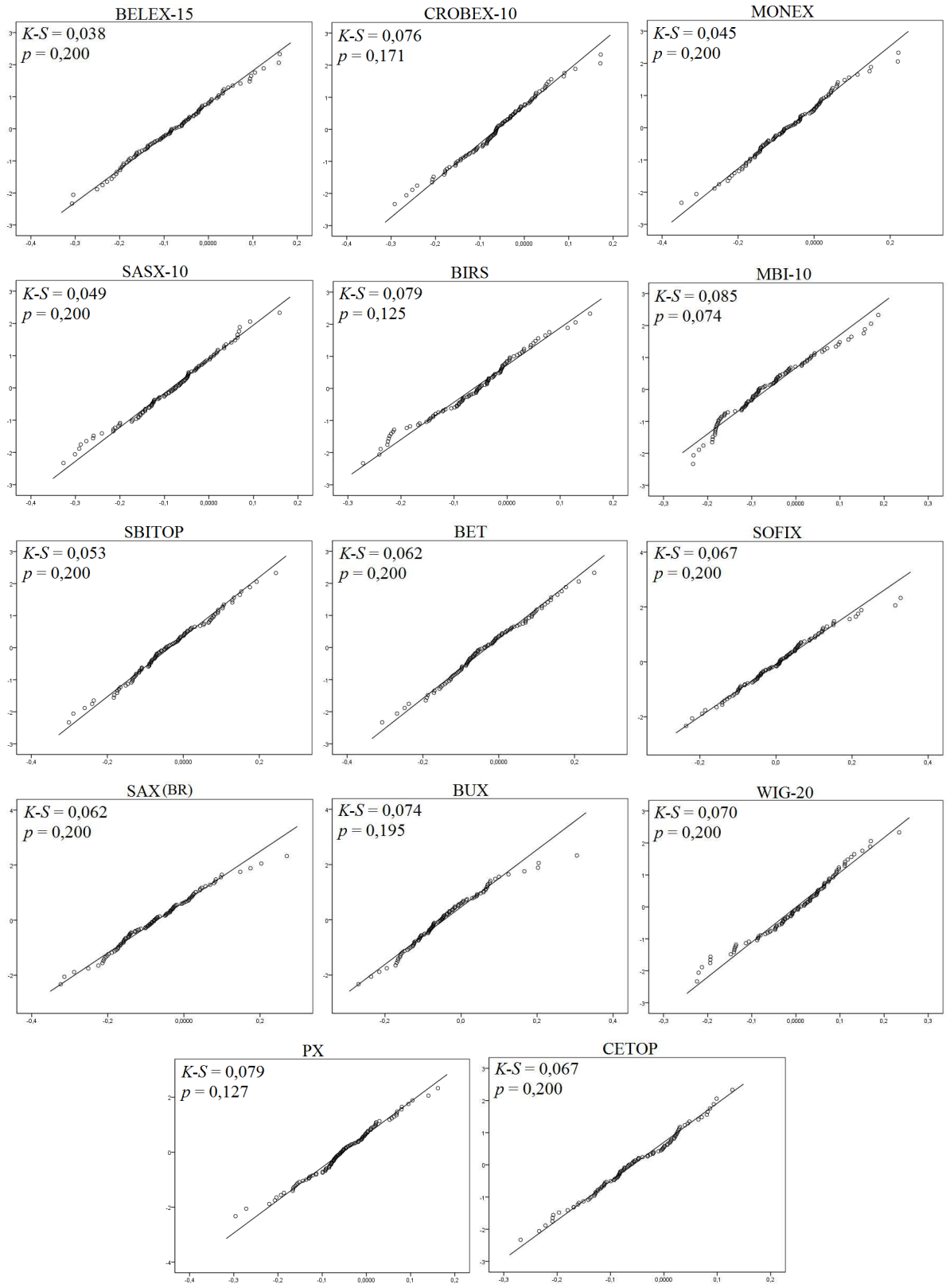
Симболи	„*a*"	„*b*"	„*c*"	„*d*"	„*e*"	„*f*"
„*a*"	0	0	0,54	0,97	1,4	1,94
„*b*"	0	0	0	0,43	0,86	1,4
„*c*"	0,54	0	0	0	0,43	0,97
„*d*"	0,97	0,43	0	0	0	0,54
„*e*"	1,4	0,86	0,43	0	0	0
„*f*"	1,94	1,4	0,97	0,54	0	0

У контексту валидне имлементације *SAX* алгоритма, важно је истаћи да су аутори алгоритма спровели обимне анализе различитих скупова података временских серија и потврдили да, генерално, подаци (стандардизованих) временских серија следе нормалан распоред. На основу тога, у емпиријским демонстрацијама *SAX* алгоритма, претпоставка о нормалности података се углавном подразумева. Чак и уколико се у конкретним ситуацијама провером ове претпоставке констатује да за одређене серије у бази постоји одступање од нормалног распореда, ефикасност алгоритма неће бити у посебној мери угрожена, а тиме и његова тачност. Тачност алгоритма је загарантована својствима *MINDIST* мере у симболичком простору, која представља доњу границу одстојања између стандардизованих серија (*Lin et al.*, 2007, стр. 114).

Табела 10: *MINDIST* одстојање између *SAX* приказа

Ознака	*SE*-01	*SE*-02	*SE*-03	*SE*-04	*SE*-05	*SE*-06	*SE*-07	*SE*-08	*SE*-09	*SE*-10	*SE*-11	*SE*-12	*SE*-13	*SE*-14
SE-01	0													
SE-02	9,199	0												
SE-03	8,764	18,698	0											
SE-04	30,486	27,263	21,887	0										
SE-05	37,612	33,177	29,853	6,081	0									
SE-06	22,517	16,639	21,432	17,068	22,490	0								
SE-07	6,903	17,377	9,615	27,350	33,425	22,306	0							
SE-08	20,127	24,299	30,794	48,195	50,911	45,116	27,406	0						
SE-09	21,182	18,831	31,494	42,374	46,705	38,198	21,832	6,903	0					
SE-10	17,601	22,288	23,394	39,685	45,835	30,901	24,299	15,934	22,038	0				
SE-11	13,010	15,303	18,981	34,370	40,979	25,807	19,143	20,056	21,576	0,000	0			
SE-12	29,541	26,568	26,737	13,414	10,667	24,676	24,196	40,723	35,231	43,644	36,878	0		
SE-13	19,662	21,471	11,956	12,658	15,303	17,901	12,706	36,373	31,812	35,214	29,093	6,903	0	
SE-14	22,544	19,901	18,906	11,449	13,658	12,279	19,389	42,282	35,015	39,253	31,850	8,133	4,300	0

Ипак, независно од тога што се нормалност распореда подразумева, за сврхе овог истраживања, провером претпоставке о облику распореда временских серија у оквиру формиране *MMS* базе, откривено је да немају све временске серије нормалан распоред. При томе, експериментисање је спроведено коришћењем стандардизованих (1800) података, применом метода тестирања статистичких хипотеза и графичког метода (у форми дијаграма нормалне вероватноће). Из тог процеса издвојено је следеће запажање: тестирањем статистичких хипотеза о нормалности распореда на основу 1800 података, за сваку серију добијена *p*-вредност једнака је нули. Тиме је потврђена, у литератури често истицана, констатација о неефикасности процедуре тестирања хипотеза за случај изразито велике количине података. Заправо, у условима велике количине података, наспрам тестирања хипотеза, визуелизација података (у овом случају у форми дијаграма нормалне вероватноће) се, показала знатно кориснијом за потребе откривања облика распореда података.



Слика 28: Оцена облика узорачког распореда аритметичких средина временских серија

Полазећи од наведеног, а у циљу откривања распореда аритметичких средина узорака и валидности примене резултата Централне граничне теореме (без обзира на распоред основног скупа), спроведена је процедура која је обухватила следеће кораке: ▸

за сваку временску серију у бази, коришћењем генератора случајних бројева, извршен је избор 100 великих узорака[70] (и то тако да сваки узорак садржи апроксимативно 100 опсервација (елемената), што је компатибилно са изабраном величином сегмената), ▸ за сваки узорак израчунате су аритметичке средине, ▸ по сваком формираном низу аритметичких средина, спроведено је испитивање нормалности распореда аритметичких средина. Резултати испитивања нормалности дати су на Слици 28.

Као што се може приметити, очекивано, свих 14 распореда аритметичких средина тежи нормалном распореду (што потврђује приближно права линија на графицима нормалне вероватноће). Упоредо са визуелизацијом проблема нормалности распореда аритметичких средина, представљени су и резултати тестирања нулте хипотезе о нормалности (применом *Kolmogorov-Smirnov*-ог теста), јер логика тестирања хипотеза добија свој смисао будући да је (случајним) узорковањем постигнута редукција података. Полазећи од вредности *Kolmogorov-Smirnov*-ог теста нормалности распореда (на Слици 28 означен као *K–S*), хипотеза о нормалности распореда посматране варијабле (вредност берзанских индекса) се одбацује за сваку посматрану временску серију (H_0: Посматрана варијабла следи нормалан распоред), будући да су кореспондентне *p*-вредности мање од дефинисаног ризика грешке I врсте, $\alpha = 0{,}05$. Валидност и научна заснованост примењене процедуре оцене нормалности распореда у домену *SAX* алгоритма, има потпуно оправдање у чињеници да су *PAA* коефицијенти по својој природи аритметичке средине.

12.3.3. Резултати примене анализе груписања

Мерење сличности између објеката представља веома важан сегмент истраживања у многим *DM* апликацијама. Сходно томе, након истраживања сличности између временских серија у посматраној бази, које је засновано на одређивању и компарацији мере одстојања компатибилне са *SAX* приказом, у наставку анализе је спроведена класификација берзанских тржишта у одређени (унапред непознати) број група, које се карактеришу интерном хомогеношћу и екстерном хетерогеношћу према тенденцијама у кретању вредности референтних берзанских индекса.

Полазећи од резултирајуће матрице *MINDIST* одстојања, у складу са представљеним концептуално-методолошким оквиром истраживања, примењени су

[70] Овом приликом се истиче да је за потребе анализе могао бити изабран мањи број узорака и мањи број елемената по узорку. Извршени избор представља последицу чињенице да је визуелни приказ у форми дијаграма нормалне вероватноће, као средства за формулисање закључака о моделу распореда података, поузданији ако је анализом обухваћена већа количина података.

различити методи хијерархијске агломеративне процедуре груписања, и то методи једноструког, потпуног и просечног повезивања. У основи, груписање 14 временских серија (односно њихових симболичких приказа дужине $k = 18$) базирано је на матрици одстојања (сличности / различитости), с тим што су, у зависности од конкретног метода, коришћена минимална, максимална или просечна одстојања, као критеријуми удруживања у групе.

С обзиром да је експериментисање извршено уз примену наведена три метода, поставља се питање избора најбољег решења.

Испитивањем степена квантитативног слагања између одговарајућих елемената оригиналне и изведених матрица одстојања (при којем је извршено удруживање свих разматраних парова објеката) за свако решење добијено применом наведених метода хијерархијског груписања, утврђене су вредности кофенетичког коефицијента корелације, као показатеља степена квалитета појединачних решења проблема груписања (Табела 11). За потребе даље анализе, решење проблема груписања применом метода просечног повезвања идентификовано је као најквалитетније у поређењу са резултатима осталих метода, будући да је највећа вредност кофенетичког коефицијента ($r = 0,757$) управо његова карактеристика.

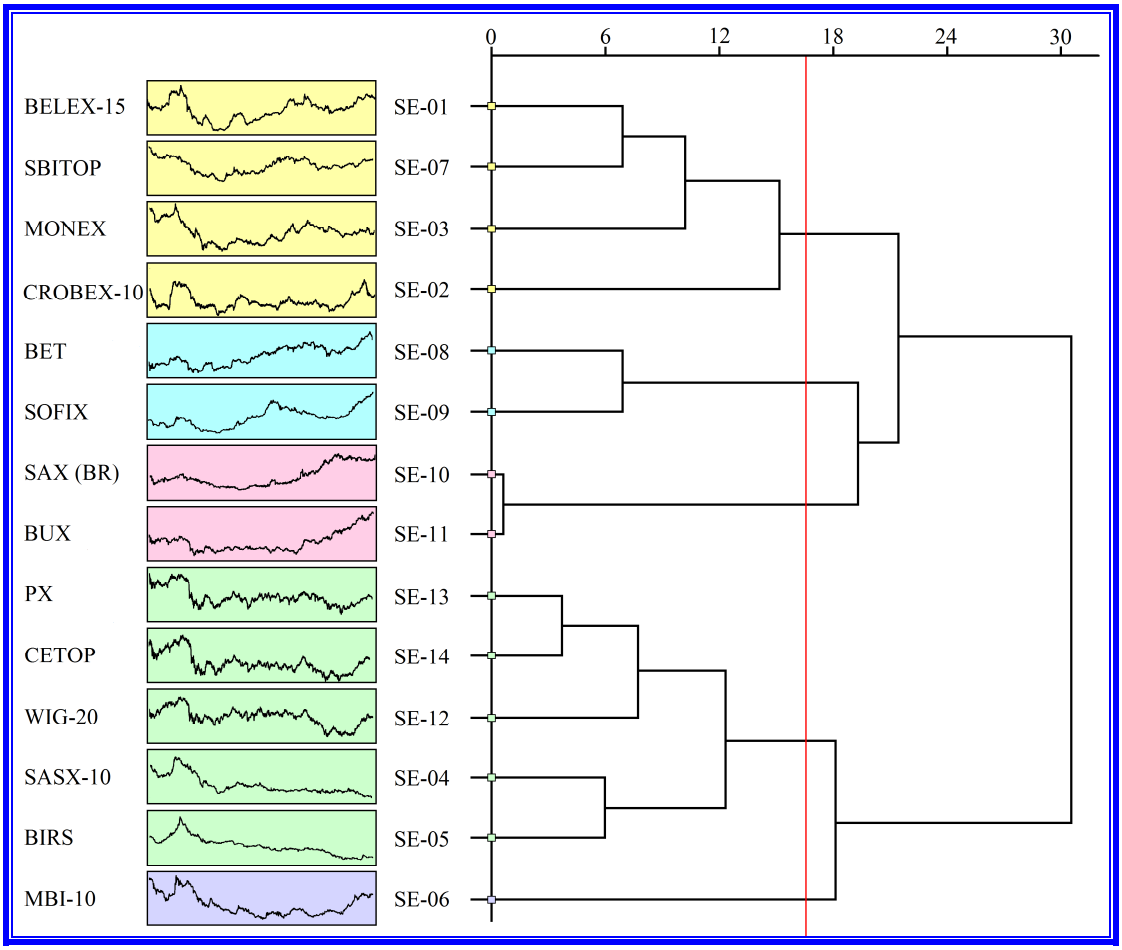
Табела 11: Вредности кофенетичког коефицијента за примењене методе груписања

Методи хијерархијског груписања	Кофенетички коефицијент (r)
Метод једноструког повезивања	0,631
Метод потпуног повезивања	0,336
Метод просечног повезивања	0,757

Визуелна интерпретација поступка и резултата спровођења хијерархијског агломеративног груписања изабраних берзи и вредности одговарајућих 14 водећих индекса, који је заснован на имплементацији метода просечног повезивања и *MINDIST* одстојања, представљени су у форми дендрограма на Слици 29. Очигледно, поступак груписања започиње са 14 група са по једним објектом, након чега се, сходно критеријуму просечног одстојања, врши удруживање (повезивање) у веће групе, тако да се у коначној форми добија једна група са инкорпорираним свим објектима. Заправо, из итерације у итерацију величина група се повећава, а њихов број смањује.

С обзиром да поступак хијерархијске агломерације не полази од унапред дефинисаног броја група, за потребе идентификовања оптималног броја група, анализиране су вредности уобичајених критеријума за доношење одлуке овог типа, и

то, мере одстојања при којој се врши удруживање група и прираштај мере одстојања између група за сваки корак у поступку груписања. Стога је, као комплемент графичком приказу агломерације у форми дендрограма, дата Табела 12 која указује не само на редослед удруживања берзи или група берзи према вредностима берзанских индекса у посматраном периоду, већ садржи и кореспондентне вредности претходно наведених критеријума за избор оптималног броја група.



Слика 29: Дендрограм - метод просечног повезивања

Анализом вредности мере одстојања између група у сукцесивним корацима итеративног поступка хијерархијског удруживања берзи и група берзи, уочава се, растућа тенденција вредности ове мере, што је логична и природна последица укрупњавања хијерархијске структуре и удруживања међусобом, у већој мери, различитих група. При томе се у првим корацима итеративног процеса (изузимајући други корак и тренутак формирања 12 група) бележи спорији раст вредности мере одстојања, а од осмог корака (и тренутка формирања 6 и мање од 6 група) долази до незнатног убрзавања овог пораста и значајнијег повећања прираштаја вредности мере одстојања, с тим што се не може констатовати да је реч о великом или драстичном

повећању вредности. Такође, забележено је још једно успоравање раста вредност мере одстојања, које је, последично, праћено и падом прираштаја мере одстојања у једанаестом кораку (и тренутку формирања 3 групе). Поред наведеног, (изузимајући последњи итеративни корак) изузетно је важно приметити да је највећи прираштај вредности мере одстојања (3,792) остварен у тренутку формирања 4 групе (десети корак). Сходно томе, број група детерминисан у претходном кораку представља оптимално решење. Такође, визуелна верификација изнете констатације је представљена на Слици 29, тако што је пресек дендрограма, према правилу, извршен при броју група који непосредно претходи кораку у итеративној процедури удруживања у којем су настале значајне промене вредности анализираних критеријума оптималности. Заправо, као оптимално решење спроведене хијерархијске процедуре груписања издваја се решење са 5 група берзи.

Табела 12: Редослед удруживања берзи (берзанских индекса)

Редни број корака	Формирање група		Број група	Вредност одстојања	Прираштај одстојања
Иницијални корак	[*SE*-01]; [*SE*-02]; [*SE*-03]; [*SE*-04]; [*SE*-05]; [*SE*-06]; [*SE*-07]; [*SE*-08]; [*SE*-09]; [*SE*-10]; [*SE*-11]; [*SE*-12]; [*SE*-13]; [*SE*-14]		14	/	/
1	[*SE*-10]	[*SE*-11]	13	0,000	/
2	[*SE*-13]	[*SE*-14]	12	4,300	4,300
3	[*SE*-04]	[*SE*-05]	11	6,081	1,781
4	[*SE*-08]	[*SE*-09]	10	6,903	0,822
5	[*SE*-07]	[*SE*-01]	9	6,903	0,000
6	[*SE*-13; *SE*-14]	[*SE*-12]	8	7,518	0,615
7	[*SE*-01; *SE*-07]	[*SE*-03]	7	9,190	1,672
8	[*SE*-04; *SE*-05]	[*SE*-12; *SE*-13; *SE*-14]	6	12,858	3,688
9	[*SE*-01; *SE*-07; *SE*-03]	[*SE*-02]	5	15,091	2,233
10	[*SE*-04; *SE*-05; *SE*-12; *SE*-13; *SE*-14]	[*SE*-06]	4	18,883	3,792
11	[*SE*-08; *SE*-09]	[*SE*-10; *SE*-11]	3	19,901	1,018
12	[*SE*-01; *SE*-02; *SE*-03; *SE*-07]	[*SE*-08; *SE*-09; *SE*-10; *SE*-11]	2	21,874	1,973
13	[*SE*-01; *SE*-02; *SE*-03; *SE*-07; *SE*-08; *SE*-09; *SE*-10; *SE*-11]	[*SE*-04; *SE*-05; *SE*-06; *SE*-12; *SE*-13; *SE*-14]	1	31,333	9,453

Оцена валидности добијеног решења представља веома важан аспект анализе груписања, при чему се одређивање оптималног броја група сматра фундаменталним проблемом валидности груписања. Како не постоји аутоматски начин и универзални општеприхваћени формални критеријум за идентификовање најбољег решења, уведени су и разматрани опциони показатељи за оцену квалитета исхода груписања. Наиме, статистичка евалуација валидности и квалитета издвојеног решења груписања, које подразумева алокацију берзи у 5 група, спроведена је поређењем вредности следећих статистичких показатеља: коефицијента кохезије, сепарације и силуете. Табела 13

садржи вредности наведених коефицијената, које су одређене за решења анализе груписања са 2, 3, 4, 5 и 6 група.

Табела 13: Вредности коефицијената за оцену квалитета решења груписања

Коефицијенти	Број група				
	6	5	4	3	2
Коефицијент кохезије	31,821	41,174	56,304	65,62	56,17
Коефицијент сепарације	299,521	359,939	108,784	50,773	11,956
Коефицијент силуете	0,462	0,489	0,354	0,407	0,449

Као што се може запазити, значајан пораст вредности коефицијента кохезије (који је, у овом случају, исказан путем максималног одстојања између парова објеката унутар исте групе) забележен је код решења са четири групе. Алтернативно, значајан пад вредности коефицијента сепарације (који је, у овом случају, мерен путем минималног одстојања између парова објеката двеју група) забележен је, такође, код решења са четири групе. Према овим коефицијентима као оптимало решење се издваја оно које непосредно претходи израженој промени у кретању њихових вредности. Осим тога, решење анализе груписања са 5 група се одликује највећом вредношћу коефицијента силуете ($\overline{s} = 0,489$). Заправо, анализа ових вредности потврђује претходно изведени закључак да је са аспекта интерне хомогености и екстерне хетерогености, у поређењу са осталим могућим и тестираним исходима анализе груписања, класификација берзи у 5 група најповољније решење. Распоред берзи по групама утврђеним на бази вредности берзанских индекса представљен је у Табели 14.

Осим пружања информација о броју група и квалитету укупног резултата груписања, одређивањем коефицијената силуете за сваку формирану групу и сваки објекат унутар група добијене су додатне информације које су омогућиле стицање увида у хомогеност појединих група и успех у позиционирању (класификацији) појединих објеката. На основу анализа вредности коефицијената силуете по групама, примећено је да у контексту интерне хомогености и екстерне хетерогености најбољу позицију има група 3 ($\overline{s}_3 = 1,00$), док су остале групе, у опадајућем низу вредности овог коефицијента, рангиране према следећем редоследу: Група 2 ($\overline{s}_2 = 0,65$), Група 4 ($\overline{s}_4 = 0,42$) и Група 1 ($\overline{s}_1 = 0,40$), док структуру Групе 5 чини једна берза, тако да је коефицијент силуете једнак нули. Истовремено, запажено је да не постоји забележена негативна вредност појединачних коефицијената силуте по берзама. Наведено имплицира да примењена процедура груписања није резултирала погрешно

класификованим берзама, с тим што постоје случајеви са мање повољним вредностима коефицијената, попут *CROBEX* индекса Хрватске берзе ($s_{crobex} = 0,09$), што је узроковало и најмању вредност коефицијента силуете за Групу 1.

Табела 14: Распоред берзи према формираним групама

Ознака групе	Број елемената у групи	Ознака берзе и изабраног берзанског индекса
Група 1	4	[*SE*-01], [*SE*-02], [*SE*-03], [*SE*-07]
Група 2	2	[*SE*-08], [*SE*-09]
Група 3	2	[*SE*-10], [*SE*-11]
Група 4	5	[*SE*-04], [*SE*-05], [*SE*-12], [*SE*-13], [*SE*-14]
Група 5	1	[*SE*-06]

Генерално, проналажење решења анализе груписања јесте хеуристички проблем. Коректно коришћење и правилна интерпретација резултата у сваком кораку примене анализе груписања (али, уз висок ниво обазривости услед чињенице да анализа груписања резултира одређеним групама чак и у случају одсуства било какве структуре у подацима хомогених база) поседује значајан потенцијал за откривање структура у подацима које се не могу идентификовати применом других стандардних метода.

12.3.4. Анализа креираног модела сличности

У реализованом емпиријском истраживању, у оквиру којег је инкорпорирано креирање *DM* апликације за трансформацију нумеричких у симболичке податке, идентификоване су законитости скривене у историјским подацима о вредностима берзанских индекса, и то у форми утврђивања степена сличности у понашању између парова финансијских временских серија и формирања (оквирних) група сличних финансијских временских серија.

На основу анализе резултата представљених у Табели 10 уочавају се законитости које се односе на одређене аспекте сличности у понашању током времена посматраних серија, а које могу бити третиране као опште или парцијалне (то јест, карактеристичне за сваку од посматраних серија у компарацији са осталим временским серијама). Полазећи од тога да мања мера одстојања представља индикатор веће сличности (односно, мање различитости), кључне уочене законитости су: ▸ највећа сличност постоји између кретања *SAX* (*BR*) индекса Словачке берзе и *BUX* индекса Будимпештанске берзе; ▸ најмања сличност се односи на кретање *BIRS* индекса Бањалучке берзе и *CETOP* индекса Будимпештанске берзе; ▸ кретање *BELEX*-15 индекса Београдске берзе је најсличније са кретањем *SBITOP* индекса Љубљанске и

CROBEX-10 индекса Загребачке берзе, док је најмање слично са кретањем *SASX*-10 индекса Сарајевске и *BIRS* индекса Бањалучке берзе. Велики значај имплементиране иновативне процедуре за редукцију димензионалности података је директна последица чињенице да је исте или сличне законитости (истина, изразито комплексне како у методолошком, тако и суштинском смислу) тешко, практично, чак, и немогуће извести на основу визуелног приказа велике количине оригиналних и стандардизованих података временских серија у бази (Слика 25) или, пак, на основу процедура које не укључују редукцију димензионалности. Заправо, *ICT* развој и непрекидно и убрзано повећање количине прикупљених података условили су раст тражње за иновативним, ефикасним и ефективним методолошким приступима и алатима за анализу берзанских података, тако да се, посматрано из перспективе дефинисаних истраживачких хипотеза, може констатовати да је прва хипотеза (H-1_{12}) потврђена.

На темељу утврђене сличности (одстојања) између парова финансијских временских серија, формиране су групе берзи према кретању вредности изабраних берзанских индекса. Анализом структуре формираних група идентификују се извесне специфичности сваке од њих са становишта чланства конкретне државе којој берза припада у Европској унији (ЕУ) и привредне развијености држава којима припадају берзе у формираним групама. Као индикатор развијености, у регионалним оквирима, у истраживању се користи просечна вредност БДП-а *per capita* (према паритету куповне моћи, исказана у *USD*) у периоду од 2010-2017. године за сваку земљу у компарацији са просечном вредношћу БДП-а *per capita* за све анализом обухваћене земље у истом периоду (који износи 20006 *USD*).[71]

Полазећи од представљеног распореда берзи и кореспондентних берзанских индекса у оквиру формираних група (Табела 14), може се уочити да су у оквиру групе која је означена као Група 1, сходно вредностима референтних берзанских индекса, класификоване берзе више од половине земаља бивше Југославије, и то Београдска, Љубљанска, Монтенегро и Хрватска берза. Реч је о групи берзи земаља од којих две имају статус чланица ЕУ (Словенија и Хрватска), а две статус земаља кандидата за чланство са започетим преговорима (Србија и Црна Гора). Са становишта привредне развијености, Словенију (у износу од 29981 *USD*) карактерише знатно већа просечна вредност БДП-а *per capita* од просечне вредности свих посматраних држава, Хрватска се карактерише просеком (у износу од 21496 *USD*) који је не у непосредној близини

[71] За израчунавање просечних вредности БДП-а *per capita* за све анализиране земље, у периоду од 2010. до 2017. године, коришћена је база података Светске банке (*http://data.worldbank.org/*).

просечне вредности са којом се врши поређење, док се Србија и Црна Гора одликују знатно нижим просечним вредностима БДП-а *per capita* (у износу од 13444 *USD* и 15161 *USD*, респективно).

У оквиру Групе 2, сходно тенденцијама у кретању *BET* и *SOFIX* индекса, удружиле су се Румунска и Бугарска берза. У питању су берзе држава које су истовремено (2007. године) укључене у европске интеграционе процесе. За обе државе важи релација да је просечна вредност БДП-а *per capita* мања од просечне вредности истог показатеља за све анализом обухваћене земље у разматраном периоду. У поређењу са Србијом и Црном Гором (као државама које су алоциране у Групу 1), реч је о вредностима, које су, иако мање, знатно ближе просеку посматране групе земаља.

Групу 3 формирају Словачка и Мађарска берза, које се, имајући у виду вредност мере удаљености при којој се врши удруживање, одликују изузетно сличним тенденцијама у кретању *SAX* и *BUX* индекса. Наведене државе су своје чланство у ЕУ стекле 2004. године и одликују се просечном вредношћу БДП-а *per capita* која је знатно већа (нарочито у случају Словачке у износу од 27754 *USD*) од просечне вредности БДП-а *per capita* за све анализом обухваћене земље.

Даљом анализом формираних група може се уочити да су у оквиру групе 4 позиционирane две, међусобом јасно издиференциране, подгрупе индекса: ▸ прва, којом су обухваћени *CETOP* (бенчмарк индекс), *PX* (Чешка Република) и *WIG*-20 (Пољска), и ▸ друга, којом су обухваћени *SASX-10* и *BIRS* (као индекси берзи (неразвијеног) финансијског тржишта Босне и Херцеговине). Интересантно је уочити да су у Групи 4 (али у различитим подгрупама) алоцирани *CETOP* и *BIRS* индекс као пар који карактерише највећа вредност *MINDIST* одстојања, односно највећа разлика у погледу кретања њихових вредности током времена. Генерално, Група 4 је, са аспекта степена економске развијености земаља и финансијских тржишта у њеном саставу, најхетерогенија група. У структури прве подгрупе (изузимајући *CETOP* индекс који одражава кретање тржишних цена акција компанија којима се тргује на више берзи) налазе се берзе држава ЕУ које су своје чланство стекле 2004. године, а карактеришу се изразито високим просечним вредностима посматраног индикатора привредне развијености. Овој подгрупи припада берза Чешке Републике, која од свих држава анализираног географског подручја бележи највећу просечну вредност БДП-а *per capita* (30929 *USD*) у посматраном периоду. За разлику од прве, у другој подгрупи су берзе држава потенцијалних кандидата за чланство у ЕУ, које имају најнижу просечну

вредност БДП-а *per capita* (10727 *USD*) у оквирима посматраног територијалног и временског обухвата.

Коначно, групу 5 чини само *MBI*-10 индекс Македонске берзе. Македонија је држава која има статус кандидата за чланство у ЕУ (без започетих перговора). У погледу привредне развијености, одликује се просечном вредношћу БДП-а *per capita* у посматраном периоду у износу од 12872 *USD*, која је знатно мања од просечне вредности БДП-а *per capita* изабране групе земаља.

У овом делу квалитативне анализе формираних група посебно треба апострофирати чињеницу да су све земље обухваћене истраживањем започеле процес транзиције и структурних промена (укључујући приватизацију, инстистуционалне реформе и стварање финанасијских система за одржавање тржишно оријентисних економија) деведесетих година XX века и да су, за разлику од остатка европског континента, као земље релативно слабије економске моћи, у знатно већем степену биле изложене утицају глобалних дешавања. Као што је већ истакнуто, неке од ових земаља (пре свега земље Централне и Источне Европе) су успешно спровеле реформе и остариле солидан економски раст, а тиме обезбедиле и улазак у ЕУ. С друге стране, у стварању тржишно оријентисаних економија и спровођењу важних реформи, неке од њих, нарочито земље Југоисточне Европе, напредовале су знатно спорије и још увек нису део јединственог тржишта ЕУ. Потпуно је природно да су (спроведене, али и актуелне) динамичке промене праћене и оживљавањем функција берзанског тржишта и успостављањем берзанског трговања. Упркос започетим реформама, ипак, појединачна тржишта одабраних земаља (нарочито земаља Југоисточне Европе) карактеришу се недовољном величином, ниском ликвидношћу, недостатком већег броја компанија велике тржишне капитализације, ниском разноврсношћу хартија од вредности и других финансијских инструмената, а самим тим и недовољном атрактивношћу из угла глобалних инвеститора.[72] Сходно томе, њихово укрупњавање (и регионално повезивање) представља неопходан услов за даљу реализацију стратегија економског развоја. У том контексту, полазећи од утврђене сличности у кретању вредности берзанских индекса, а у скалду са реализацијом идеје о колаборацији и регионалном

[72] У циљу потврде изнете констатације и стицања оквирног увида у перформансе берзанских тржишта посматраних земаља, наводи се неколико илустративних релација. Према извештајима Светске банке за 2016. годину, Варшавска берза, као највећа берза на простору обухваћеним овим истраживањем, одликује се тржишном капитализације (у % од БДП-а) у износу од 29,42%, што је знатно испод просечне стопе на светском нивоу (79,62%). Пажњу завређује и поређење са нивоом истог индикатора у САД (146,86%), Француској (87,48%) или Русији (48,48%) (*https://www.theglobaleconomy.com*). Према истом извору, стопа тржишне капитализације за Мађарску берзу износи 17,93%, а Словеначку 11,77%, док позицију српског берзанског тржишта за 2016. годину (детерминисану на основу података са сајта Београдске берзе) репрезентује стопа тржишне капитализације чија је вредност око 12%.

повезивању малих и недовољно атрактивних берзи, могуће је обезбедити значајне погодности са становишта смањења трошкова истраживања тржишта и алокације слободних финансијских средстава глобалних инвеститора у приносне инвестиционе алтернативе.

Полазећи од представљених резултата и изнетих разматрања, треба запазити да се више од половине берзи (али не све) земаља чланица бивше Југославије налази у истој групи према вредностима референтних берзанских индекса у анализираном периоду, што упућује на закључак да се друга истраживачка хипотеза не може у потпуности сматрати потврђеном. Међутим, неопходно је истаћи да су оправдана очекивања према којима ће проширење анализе и укључивање осталих индикатора развијености одређеног берзанског тржишта обезбедити да иницијално дефинисана друга истраживачка хипотеза буде у потпуности потврђена. Заправо, логично је да ће кроз посматрање више аспеката стања на берзанском тржишту доћи до удруживања Сарајевске, Бањалучке и Македонске берзе, као неразвијених берзи, или у оквиру посебне групе или у оквиру подрупе земаља Централне и Југоисточне Европе које су чланице бивше Југославије.

У основи, добијени резултати спроведеног истраживања недвосмислено показују да је примењени методолошки приступ од велике користи за истраживање сличности високо димензионалних економских / финансијских феномена и, конкретно, за класисификацију берзи према вредностима водећих берзанских индекса, укључујући одговарајуће (потенцијалне) економске импликације у форми усмеравања одлука инвеститора и дефинисања могућих праваца регионалног повезивања појединачних берзанских тржишта. У складу са наведеним разматрањем, а узимајући у обзир и ограничења реализованог истраживања, будућа истраживачка настојања ће иницијално бити усмерена на проширење анализе у овом домену берзанског пословања, и то кроз следећа три правца: ▸ спровођење компаративне анализе резултата истраживања сличности добијених применом различитих мера сличности, ▸ идентификовање промена у структури формираних група током времена, и ▸ укључивање нових варијабли у анализу (као што су учешће тржишне капитализације акција и вредности промета акција у друштвеном бруто производу, коефицијент обрта за тржиште акција, волатилност индекса цена акција итд.) уз повећање просторног обухвата. Такође, са методолошког аспекта посматрано, презентовано истраживање потврђује ограниченост и непрактичност примене традиционалних методолошких оквира у анализи динамичког берзанског тржишта и високо димензионалних берзанских података. Стога

је представљени иновативни, концептуално-методолошки приступ, који чини основу спроведеног емпиријског истраживања, могуће имплементирати и додатно тестирати при анализи и других аспеката берзанског пословања, као и осталих сличних високо димензионалних феномена у подручју економије.

Коначно, на основу претходног разматрања различитих аспеката процеса развоја модела сличности, могуће је закључити следеће: ▸ прво, примењени оквир истраживања одликује се изразитом методолошком комплексношћи; ▸ друго, неструктурираност дефинисане проблемске ситуације и заснованост истраживања на подацима захтевали су, у функцији ефикасног управљања обрадом и анализом података, коришћење више софтверских алата уз висок степен познавања статистичког начина размишљања; ▸ треће, да би решење било које проблемске ситуације до којег се долази применом *DM* приступа у анализи података имало практични смисао и значај, од конципирања пословног проблема до контекстне интерпретације решења, неопходна је адекватна подршка експерата са компентецијама у домену конкретног подручја (што у склопу овог истраживања значи да дубље и детаљније елаборирање добијених међурезултата и коначних резултата захтева ангажовање експерата из области финансијских тржишта).

13. РЕАЛИЗАЦИЈА *DATA MINING* ЗАДАТАКА У КОНТЕКСТУ ДЕФИНИСАНЕ ПРОБЛЕМСКЕ СИТУАЦИЈЕ 2

Имајући у виду значај развоја услужних делатности у савременој економији, као и чињеницу да су односи са корисницима услуге централна тачка свих менаџмент и маркетинг активности у домену услужног пословања, у овом Поглављу је представљено емпиријско истраживање усмерено на мерење сатисфакције корисника услуге једног услужног предузећа. Сходно томе, након дефинисања проблемског контекста, конципиран је и приказан методолошки оквир истраживања, базиран на комбинованој примени *DM* метода са инкорпорираним елементима статистичког начина размишљања. Поред наведеног, такође су, логично, приказани и анализирани резултати реализованог истраживања, формулисани и дискутовани закључци о предмету истраживања, уз осврт на ограничења и будуће правце истраживања.

13.1. Идентификовање проблемске ситуације

У савременим условима пословања, ефикасан економски систем није могуће замислити без развијеног сектора услуга. Наиме, услуге заузимају централно место у

структури економских активности, а самим тим добијају огроман (и континуирано растући) значај у сфери тржишне конкуренције. Неспорно, услужни послови и услужна еволуција покрећу сваку економију.

Посматрано из перспективе пословне економије, развој услужне делатности условио је промене у начину пословања предузећа и подстакао њихову транзицију од произвођачке логике ка моделу услужног предузећа, са фокусом на повезаност са купцем (потрошачем / корисником / клијентом), као централном тачком свих менаџмент и маркетинг активности. Предузећа која су прихватила услужну револуцију као неминовност, проналазе изворе за успех и профит тамо где друга предузећа не успевају. У супротном, уколико се занемари ова потреба и, последично, недовољно инвестира у услуге, конкретно предузеће се излаже великом ризику да ће постати жртва конкуренције. Једноставно, свако предузеће да би опстало, расло и развијало се мора постати услужно предузеће (*Senić & Senić*, 2008, стр. 44-45).

За предузећа која су усмерена ка купцу, сатисфакција купца је и циљ и главни фактор његовог (пословног и финансијског) успеха (*Senić*, 2000, стр. 42), а самим тим и есенцијална компонента дугорочног опстанка. Уопштено узевши, сатисфакција, као концепт који је осамдесетих година XX века постао веома популаран у маркетинг и менаџмент литератури, је повезана са осећањем задовољства или разочарања особе које се јавља као резултат поређења уочене перформансе и очекивања. Сходно одступању запажених и очекиваних нето користи утврђује се висина (ниво) сатисфакције, при чему се разликују следеће две типичне ситуације: ако испоручена услуга испуњава очекивања, корисник је задовољан, док у противном, ако испоручена услуга не испуњава очекивања, корисник је незадовољан. Заправо, задовољство корисника је претпоставка стварања лојалних корисника. Међутим, пословна филозофија предузећа не треба да буде оријентисана на максимирање сатисфакције купца, већ на постизање високог нивоа сатисфакције уз истовремено обезбеђење барем прихватљивог нивоа сатисфакције осталим стејкхолдерима у предузећу.

Кључна детерминанта вредности услуге која опредељује позицију предузећа на тржишту, обезбеђује веће тржишно учешће и профит и, сходно томе, доприноси сатисфакцији (како екстерних, тако и интерних) купаца јесте квалитет услуге. За разлику од првобитног схватања концепта квалитета у контексту филозофије управљања квалитетом и његовог повезивања са произвођачким спецификацијама и стандардима, дефинисање квалитета услуге у менаџмент и маркетинг литератури примарно се односи на квалитет који запажају купци, односно корисници услуге.

Другим речима, циљ пружања квалитетних услуга је сатисфакција корисника услуге. Суштински, сатисфакција купца и квалитет услуге имају заједничке елементе, али је, генерално сатисфакција шири концепт, при чему се се квалитет услуге односи конкретно на атрибуте услуге (*Wilson et al.*, 2008, стр. 78).

Евидентно је да бројни фактори врло брзо могу довести задовољног купца у стање незадовољства. Будући да је мерење квалитета услуга добар начин да се одреди да ли су услуге добре или не, а тиме, индиректно, одреди и степен сатисфакције купца, потпуно је разумљиво што у академској литератури постоји велики број емпиријских истраживања о релацијама између сатисфакције купца и квалитета услуге и, генерално, развијених приступа за праћење и мерење сатисфакције купца. Међутим, увидом у релевантну литературу, са методолошког становишта, може се запазити парцијална примена појединих метода, као и изузетно ниска варијететност у погледу примењених метода, при чему углавном доминирају методи дескриптивне статистичке анализе, анализа главних компоненти, регресиона анализа и методи статистичке контроле процеса (односно, контролне карте за оцену квалитета услужних процеса).

ICT прогрес је омогућио прикупљање и складиштење (брже и прецизније) огромних количина података у свим доменима, а самим тим и података о потребама и очекивањима корисника услуге. Истовремено, акумулирање великих количина података праћено је развојем одговарајућих софтверских решења за њихово процесирање и, последично, значајним помаком у анализи података о корисницима услуге. Заправо, долази до објављивања све већег броја истраживачких радова у којима су презентовани нови приступи за потребе евалуације различитих аспеката сатисфакције засновани на комбинованој примени више метода и раду са великом количином података о купцима (енгл. *customer data*). Уосталом, прве примене *DM* приступа у анализи података везују се, управо, за управљање односима са купцима.

Узимајући у обзир: (а) да је сатисфакција купца изразито варијабилна категорија коју је веома тешко квантификовати с обзиром да се њена оцена базира на субјективном доживљају корисника услуге, и (б) да је квалитет услуге, као детерминанта сатисфакције купца, генерално мултидимензионални концепт, који се сходно различитим проблемским контекстима, описује и представља путем различитих димензија (варијабли), емпиријско истраживање у овом делу дисертације, у методолошком смислу засновано на комбинованој примени надгледаних и ненадгледаних *DM* метода, усмерено је на квантификовање и испитивање утицаја

одабраних обележја квалитета услужне понуде на степен сатисфакције корисника услуга у домену ресторатерства.

Сходно наведеном, предмет овог истраживања је сатисфакција корисника услуге на основу различитих обележја квалитета услужне понуде једног градског ресторана. Примарни циљ истраживања је презентација конципираног методолошког приступа за проблем добијања (откривања) релевантних информација о сатисфакцији корисника утврђеној на основу испитивања њихових ставова о обележјима квалитета услуге. Секундарни циљ је формирање сегмената / група корисника услуга у ресторану према нивоу сатисфакције на бази обележја квалитета услуге и одређивање профила сваке групе. Суштински, спроведено истраживање треба да обезбеди идентификовање законитости скривених у великој количини података о корисницима услуге, а које су, из угла доносилаца пословних одлука, менаџера и маркетинг менаџера, релевантне за добијање одговора на бројна питања у вези са разумевањем детерминанти и формулисањем корисних смерница за побољшање нивоа потрошачке сатисфакције.

На основу дефинисаног предмета и постављених циљева, формулисане су следеће истраживачке хипотезе (као претпоставке о могућем решавању проблема), чија је истинитост (кроз прихватање или одбацивање) проверена у овом емпиријском делу дисертације:

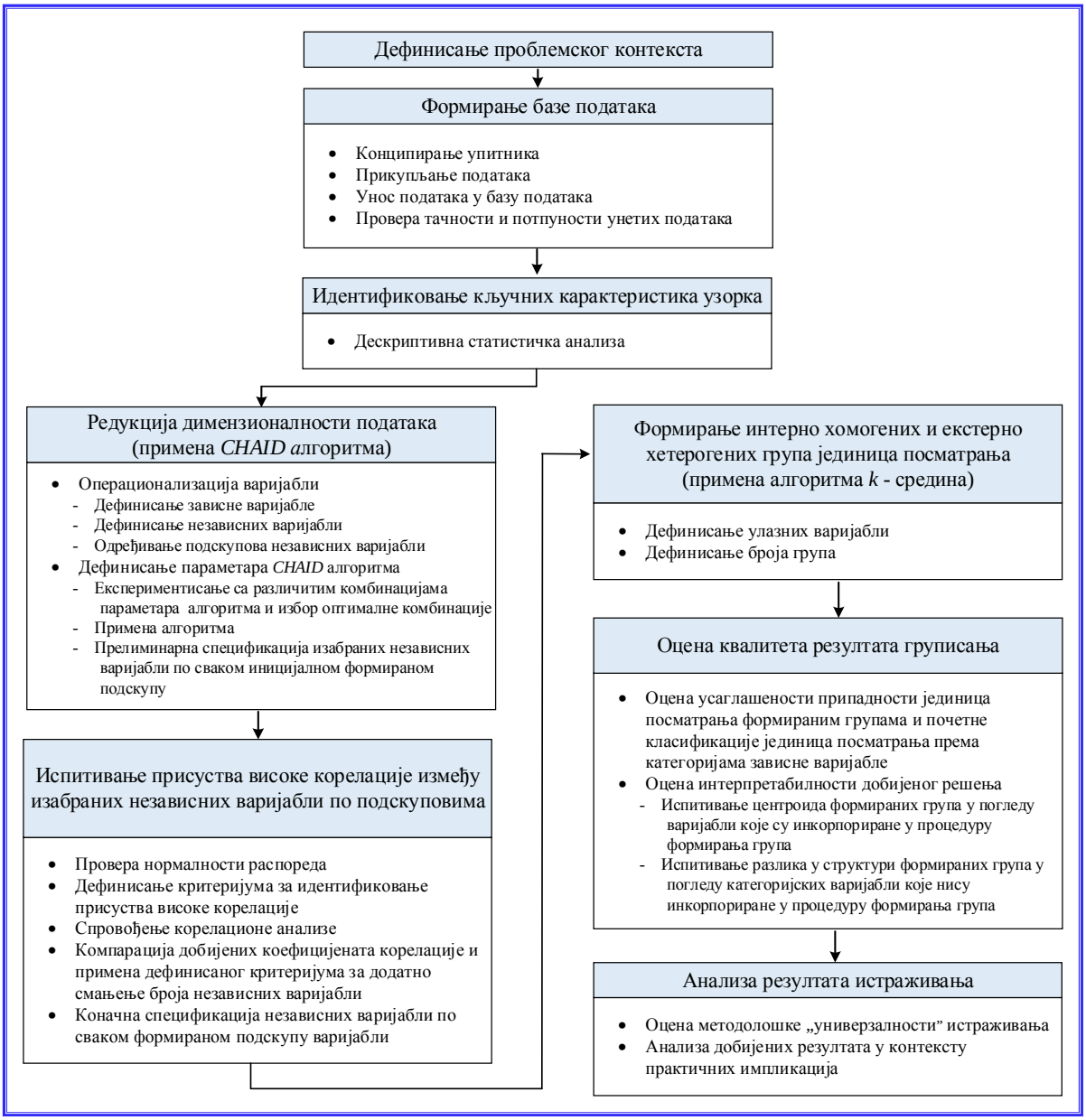
(H-1_{13}): Комбинованом применом надгледаних и ненадгледаних *DM* метода може се унапредити анализа (велике количине) података о корисницима услуге и, сходно томе, обезбедити идентификовање релевантних детерминанти и откривање сета (скривених) законитости које су повезане са сатисфакцијом корисника у различитим областима услужне делатности.

(H-2_{13}): Применом *DM* приступа у анализи података могуће је, на основу разлика / сличности у погледу просечног степена сатисфакције корисника услуге појединим обележјима квалитета услужне понуде, идентификовати интерно хомогене и екстерно хетерогене тржишне сегменте посетилаца ресторана, с једне стране, и испитати и установити (не)постојање разлика између тако формираних група у погледу, како традиционалних социодемографских, тако и осталих специфичних критеријума за сегментацију, с друге стране.

13.2. Методолошки аспекти истраживања

Полазећи од претходно представљених релација између квалитета услуге и сатисфакције корисника услуге, и сходно томе, дефинисаног предмета, прецизираних

циљева и формулисаних истраживачких хипотеза, дизајниран је концептуално-методлошки оквир емпиријског истраживања за генерисање модела који омогућава да се: ▸ идентификује утицај одабраних елемената квалитета услужне понуде на сатисфакцију корисника услуга у домену ресторатерског пословања, и ▸ спроведе сегментација тржишта како би се идентификовале релативно хомогене групе посетилаца ресторана према њиховим ставовима у вези са (изабраним) аспектима квалитета услуге. Редослед предвиђених и реализованих активности по појединим фазама истраживања, а које су дискутоване у наставку овог Поглавља, приказан је на Слици 30, где се могу издвојити три главна потпроцеса: анкетно истраживање, редукција димензионалности података и формирање хомогених група објеката.



Слика 30: Шематски приказ концептуално-методолошког оквира истраживања 2

Емпиријско истраживање усмерено на мерење сатисфакције и утврђивање ставова посетилаца једног градског ресторана о квалитету услужне понуде, спроведено је анализом примарних података, прикупљених применом анкетног истраживања.

Заправо, за наведену сврху формиран је упитник чију структуру, између осталог, чине питања која су повезана са социодемографским карактеристикама испитаника, као и њиховим посетама ресторанима. При конципирању овог дела упитника и одабиру карактеристика испитаника узете су у обзир оне карактеристике које се често појављују као предмет разматрања у истраживачким студијама сличног карактера. [73]

Такође, упитник садржи и 30 посебно формулисаних тврдњи (питања) које се односе на одређене аспекте квалитета услуге у оквиру ресторатерског пословања. Примарна сврха овог дела упитника је да се утврди да ли су и у којој мери испитаници задовољни квалитетом одабраних аспеката (у методолошком смислу варијабли) ресторанске услужне понуде. За мерење (не)сагласности учесника у анкети у погледу сваке од наведених тврдњи коришћена је *Likert*-ова скала са седам нивоа интензитета. Испитаници су свој став о свакој тврдњи изражавали избором (заокруживањем) једне од понуђених вредности, уз респектовање значења следећих релација „вредност - став": 1 - У потпуности се не слажем; 2 - У великој мери се не слажем; 3 - У малој мери се не слажем; 4 - Неутралан (индиферентан) став; 5 - У малој мери се слажем; 6 - У великој мери се слажем; 7 - У потпуности се слажем.

Анкетно истраживање је спроведено током маја, јуна и јула 2017. године међу посетиоцима једног реномираног градског ресторана на подручју града Крагујевца.[74] Узорак за истраживање је формиран случајним избором посетилаца ресторана, односно

[73] Избор и адаптација тврдњи које су укључене у упитник извршени су на основу увида у релевантну литературу и резултате реализованих студија у контексту испитивања ставова корисника услуга уопштено узевши, а посебно у области ресторатерства. Наиме, услед чињенице да је оцена квалитета услуге у функцији мерења сатисфакције корисника услуге изразито комплексна, настала је потреба за формулисањем стандардизованог оквира за мерење субјективних оцена корисника услуге у погледу различитих аспеката квалитета понуде у услужним делатностима, укључујући и ресторатерство. Сходно томе, развијен је већи број модела упитника као мерних инструмената за оцену квалитета услужне понуде. При обликовању овог дела упитника, коришћени су следећи (општи и специјализовани) модели упитника: ▸ *SERVQUAL* модел, који су, осамдесетих година XX, креирали *Parasuraman et al*. (1988) за оцену квалитета у услужном сектору, а који се може успешно користити и за праћење квалитета услуге током времена, компарацију резултата са конкурентима, оцењивање појединачних аспеката услужне понуде и мерење општег задовољства купца у погледу испоручене услуге; ▸ *DINESERV* модел, који су *Stevens et al*. (1995) специјално дизајнирали за мерење квалитета услуга у области ресторатерства; ▸ *DINESCAPE* модел, који је, као специјализован упитник, креирао *Ryu* (2005), такође, за процену перцепције посетилаца у погледу димензија укупног квалитета услужне понуде ресторана; ▸ *CFFRSERV* модел, који су креирали *Tan et al*. (2014) иницијално за испитивање квалитета услуге у кинеским ресторанима брзе хране.

[74] Важно је напоменути да, овом приликом, пилот истраживање (односно прелиминарно тестирање упитника) није спроведено, пре свега, због чињенице да је садржај упитника обликован на основу модела упитника чија је валидност већ више пута проверена и потврђена у пракси. Проблеми који су се јављали током прикупљања података су уобичајени проблеми повезани са анкетним истраживањима, тако да одређени број потенцијлних респондената није прихватио да учествује у давању одговора на питања у упитнику. Упркос томе, обезбеђено је 500 валидно попуњених упитника.

корисника услужне понуде. Упитници су циљној групи испитаника достављени лично (у штампаној верзији). Прикупљени подаци су кодирани и обрађени помоћу статистичког пакета *IBM SPSS* верзија 20.0, док су остала неопходна (табеларна) израчунавања спроведена у програму *Excel* 2007, као део пакета *Microsoft Office*.

Пошто је извршено прикупљање, а затим и формирање одговарајуће базе података, подаци из упитника су прелиминарно оцењени коришћењем метода дескриптивне статистичке анализе, укључујући и оцену поузданости интерне конзистентности тврдњи у делу упитника који се односи на аспекте квалитета услуге.

Након спроведене дескриптивне анализе, утврђен је степен (укупне) сатисфакције сваког испитаника као просечна вредност ставова испитаника по основу свих 30 тврдњи (варијабли), а затим извршено и груписање сродних тврдњи у 5 подскупова, које одражавају основне димензије ресторанске услужне понуде. У наставку анализе, применом метода мултиваријационе статистичке анализе - *CHAID* стабла одлучивања над свакој од формираних група варијабли, постигнута је редукција димензионалности података и извршено издвајање оних тврдњи о квалитету услужне понуде које су у најјачој интеракцији са утврђеним степеном сатисфакције. У циљу реализације задатка формирања релативно хомогених група посетилаца ресторана, односно сегментације тржишта, на редукованом подскупу варијабли примењена је нехијерархиска процедура груписања заснована на алгоритму *k*-средина. За потребе компарације и детаљнијег разумевања ставова испитаника по формираним групама, примењена је метода тестирања статистичких хипотеза, пре свега, у контексту социодемографских карактеристика испитаника и њихових посета ресторанима. Коначно, идентификоване законитости су на одговарајући начин презентоване и протумачене.

13.3. Резултати истраживања

У наставку текста су, сходно конципираном методолошком оквиру истраживања, приказани и анализирани неки од резултата обраде података, при чему је њихов избор извршен са становишта директне повезаности са дефинисаним истраживачким хипотезама.

13.3.1. Карактеристике узорка

Спровођењем анкетног истраживања добијени су подаци о социодемографским карактеристика испитаника, укључујући и податке о карактеристикама које су директно повезане са њиховим посетама ресторанима. У наставку текста су

представљени резултати анализе прикупљених података у форми сумарних карактеристика узорка од 500 посетилаца посматраног ресторана. Заправо:

- у погледу полне структуре учешће испитаника мушког (51,6%) и женског пола (48,4%) је приближно уједначено;
- већина испитаника припада старосној категорији од 36 до 45 година (28,2%), затим следе категорије испитаника од 26 до 35 година и од 46 до 55 година које имају исту процентуалну заступљеност у структури узорка (22,4%), као и категорије до 25 година и од 56 до 65 година (са учешћем од 12%), а 3% укупног броја анкетираних посетилаца је старије од 65 година;
- у погледу образовне структуре највећи број испитаника је означио да има високо образовање (41%), затим, са приближно истим учешћем, следе испитаници са завршеном средњом школом (39,8%), испитаници који поседују више образовање чине 16,4% укупног броја испитаника, док њих 2,8% има основно образовање;
- у структури испитаника доминира група која одлази у ресторанске посете са пријатељима (43,4%), а такође се уочава и велика група испитаника која најчешће породично одлази у ресторане (35%), док са колегама или сами одлази 16,2% и 5,4% укупног броја испитаника, респективно;
- у погледу програма здраве исхране, преко половине испитаника се изјаснило да су поборници здраве исхране (52,6%), 32,6% испитаника делимично поштује начела здраве исхране, а 14,8% испитаника не води рачуна о начелима здраве исхране;
- интересантно је да приближно ⅔ испитаника (74%) радо проба „специјалитете ресторанских кућа”, док 26% није склоно „експериментисању”.

Поред анализе претходних карактеристика, у Табели 15 су приказане основне статистике ставова (оцена) испитаника за сваку од 30 тврдњи које се односе на различите аспекте квалитета услужне понуде у ресторану. На основу представљених резултата може се уочити да су испитаници генерално исказали релативно висок ниво сатисфакције квалитетом услуге, с обзиром да је у случају 28 тврдњи забележена просечна вредност оцене изнад 5 (и то у интервалу од 5,09 до 5,67), док је само у случају 2 тврдње забележен умерен ниво сатисфакције са просечном вредношћу оцене мањом од 5, али већом од 4. Посетиоци су најзадовољнији аспектима квалитета који се односе на чистоћу сале за ручавање (5,67) и прибора за послуживање (5,60), док су најлошије оцењени аспекти квалитета који се односе на специјалне - наменске јеловнике (4,75) и цену (4,99).

Анализом дисперзије оцена испитаника по свакој од наведених тврдњи, може се запазити да је најмања вредност стандардне девијације израчуната за тврдњу која је означена симболом X_6 (Штампани материјали су визуелно привлачни и у складу са имиџом ресторана), што указује на највећи степен слагања ставова испитаника у погледу сатисфакције овим аспектом квалитета услуге. С друге стране, највећа вредност стандардне девијације је израчуната за тврдњу која је означена симболом X_{30} (Ресторанска понуда укључује специјалне - наменске јеловнике), што указује на највеће варијације ставова испитаника у погледу сатисфакције овим аспектом квалитета услуге. Наведене закључке у погледу варијабилитета (односно хомогености) ставова потврђују и израчунати коефицијенти варијације (Табела 15 – колона V), при чему је за тврдњу X_6 карактеристична најмања, а за тврдњу X_{30} највећа вредност коефицијента варијације.

13.3.2. Резултати редукције димензионалности података

Како је анкетом обухваћен велики број тврдњи о квалитету услуге (које из статистичке перспективе представљају варијабле у процесу моделирања), а чије мноштво може угрозити откривање релевантних законитости, у наставку истраживања спроведна је редукција димензионалности података заснована на примени метода стабло одлучивања, уз избор *CHAID* алгоритма за креирање одговарајућих модела. Сходно тој сврси, као и самој природи *CHAID* алгоритма, најпре је извршена операционализација зависне и независних (објашњавајућих, предиктор) варијабли.

Као зависна варијабла, *Y*, дефинисан је степен (ниво) сатисфакције корисника услуге. Вредности ове варијабле су одређене кроз следеће кораке: на основу 30 варијабли (од X_1 до X_{30}) које одражавају све анкетом (анализом) обухваћене аспекте квалитета услуге, по сваком испитанику, израчуната је просечна вредност исказаних ставова, а затим извршено кодирање на начин да су све просечне вредности одговора мање од 4 означене кодом 1 (који указује на низак ниво сатисфакције), просечне вредности 4 и више од 4, а мање од 5 су означене кодом 2 (који указује на умерен ниво сатисфакције, укључујући и неутралан - индиферентан став: „нити задовољан нити незадовољан”), док просечне вредности 5 и више од 5 су означене кодом 3 (који указује на висок ниво сатисфакције). Овим поступком је извршена трансформација нумеричке у номиналну варијаблу (са три модалитета), тако да се креирање *CHAID* модела може базирати на χ^2 критеријуму поделе. Наиме, зависна варијабла је категоријска и узима три вредности.

Табела 15: Кључне статистике ставова корисника о обележјима квалитета услуге

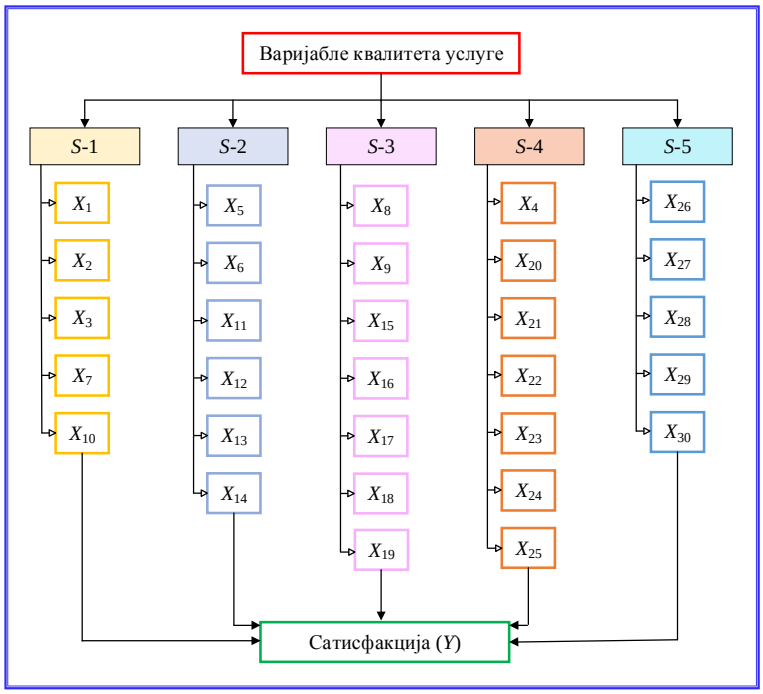
Тврдња (Варијабла)	Симбол	$\bar{x}$	s	V (у %)
Прилаз ресторану је приступачан и лепо уређен.	(X_1)	5,09	1,54	30,27
Декоративни елемени у унутрашњости ресторана су у складу са имиџом ресторана.	(X_2)	5,48	1,30	23,68
Ресторан има визуелно атрактиван простор за ручавање.	(X_3)	5,40	1,33	24,66
Особље ресторана је прикладно и уредно обучено.	(X_4)	5,51	1,28	23,29
Штампани материјали повезани са понудом ресторана су разумљиви.	(X_5)	5,48	1,23	22,52
Штампани материјали су визуелно привлачни и у складу са имиџом ресторана.	(X_6)	5,51	1,11	20,18
Сала за ручавање је пространа.	(X_7)	5,36	1,28	23,81
Сала за ручавање је чиста.	(X_8)	5,67	1,19	20,99
Преостали делови ресторанског простора намењени за госте су чисти и уредни.	(X_9)	5,47	1,36	24,79
Ресторански намештај је удобан.	(X_{10})	5,36	1,34	25,07
Аудио-визуелни ефекти у ресторану су пријатни и опуштајући.	(X_{11})	5,21	1,35	25,82
Осветљење у ресторану доприноси пријатној атмосфери.	(X_{12})	5,33	1,29	24,18
Температура у ресторану је оптимална.	(X_{13})	5,36	1,24	23,13
Декорација на столу је визуелно допадљива.	(X_{14})	5,21	1,40	26,77
Прибор за послуживање је чист и квалитетан.	(X_{15})	5,60	1,30	23,13
Особље ресторана пружа услуге поуздано и доследно имиџу ресторана.	(X_{16})	5,36	1,32	24,61
Сервирана храна је у складу са извршеном наруџбином.	(X_{17})	5,36	1,48	27,65
Особље ресторана брзо отклања пропусте настале приликом пружања услуге.	(X_{18})	5,24	1,45	27,70
Цене у ресторану су прикладне.	(X_{19})	4,99	1,60	32,08
Особље ресторана брзо опслужује госте.	(X_{20})	5,21	1,44	27,59
Особље ресторана је спремно да испуни специјални захтев госта.	(X_{21})	5,16	1,44	27,86
Особље ресторана је љубазно и улива поверење госту.	(X_{22})	5,36	1,37	25,59
Особље ресторана располаже потребним знањем о понуди ресторана.	(X_{23})	5,19	1,40	26,91
Особље ресторана је компетентно, професионално и међу собом координирано.	(X_{24})	5,24	1,40	26,65
Особље ресторана посвећује довољно пажње сваком госту.	(X_{25})	5,13	1,47	28,60
Ресторан нуди разноврстан избор јела.	(X_{26})	5,42	1,25	23,00
Храна у ресторану је укусна.	(X_{27})	5,46	1,48	27,08
Храна у ресторану је свежа.	(X_{28})	5,56	1,30	23,44
Изглед и декорација послужене хране су визуелно привлачни.	(X_{29})	5,26	1,41	26,76
Ресторанска понуда укључује специјалне-наменске јеловнике.	(X_{30})	4,75	1,73	36,40

Напомена: За сваку варијаблу обухваћену анализом (од X_1 до X_{30}) симбол $\bar{x}$ означава просечну вредност ставова испитаника, *s* стандардну девијацију, а *V* коефицијент варијације.

Приликом специфицирања независних варијабли, извршено је груписање скупа 30 посматраних варијабли у сродне подскупове које одражавају следеће (кључне) димензије ресторанске услужне понуде: (1) погодност локације, уређеност и опремљеност простора, (2) атмосферу и квалитет брошура, (3) поузданост услуживања и уредност, (4) професионалност услужног особља и (5) квалитет хране. При томе су, сходно наведеној категоризацији, уведене кореспонденте ознаке за сваки подскуп од *S*-1 до *S*-5, респективно. На Слици 31 је представљена структура подскупова из угла варијабли које их чине, уз коришћење одговарајућих симбола од X_1 до X_{30} из Табеле 15. Над сваким од формираних подскупова варијабли примењен је *CHAID* алгоритам како би се постигла редукција димензионалности података и извршило издвајање оних

варијабли квалитета услужне понуде (по сваком подскупу) које су у најјачој интеракцији са зависном варијаблом, односно утврђеним степеном сатисфакције.

Узимајући у обзир начин функционисања *CHAID* алгоритма, након спецификације варијабли, у оквиру примењеног статистичког програма, дефинисани су преостали елементи и параметри неопходни за његово покретање. Заправо, за испитивање зависности између варијабли одабран је *Pearson*-ов χ^2 тест, одређене су уобичајене вредности ризика грешке за поделу чворова и удруживање категорија независних варијабли од 0,05[75], уз аутоматско прилагођавање *p*-вредности (за решавање проблема погрешног одбацивања нулте хипотезе у контексту вишеструке компарације). Недостајуће вредности нису идентификоване нити у једној варијабли. У контексту валидације модела примењен је 10-оструки метод унакрсне валидације. Након спроведеног експериментисања са различитим вредностима параметара алгоритма и увида у прелиминарне резултате, дефинисани су коначни критеријуми за одређивање граница раста стабла одлучивања (односно, заустављање рекурзивног дељења посматраних података), и то: одабрана дубина стабла је три нивоа испод почетног (кореног) чвора, а минимални број јединица посматрања у унутрашњим чворовима (подгрупама, сегентима) 50 за „родитељске" и 15 за „дечије" чворове.



Слика 31: Независне варијабле по подскуповима и релације са зависном варијаблом

[75] У питању је стандардна гранична вредност која се користи код χ^2 теста за испитивање зависности између две вариајбле и проверу нулте и алтернативне хипотезе, које гласе: ▸ H_0: Између посматраних варијабли не постоји статистички зачајна веза, и ▸ H_1: Између посматраних варијабли постоји статистички значајна веза.

CHAID процедура са идентичним параметрима примењена је на све подскупове варијабли. Сумарни приказ кључних елемената примене алгоритма представљен је у Табели 16 и састоји се од 2 основна дела: први се односи на дефинисање захтеваних елементе у погледу рада алгоритма, а други на део резултата примене алгоритма. Резултирајућа *CHAID* стабла одлучивања (у форми графичког приказа) представљена су на Сликама 32, 33, 34, 35 и 36. Мада је примарна сврха креирања *CHAID* модела усмерена на смањење броја варијабли, ипак, у контексту сагледавања њихових карактеристика важно је указати и на одређене аспекте постигнуте класификационе прецизности. Заправо, за сваки модел је забележена висока укупна прецизност (односно, стопе успешних класификација јединица посматрања су 84,6%, 83,4%, 89,8%, 89,2% и 87, респективно по посматраним групама варијабли), што упућује на закључак да у случају када су познати ставови неког посетиоца ресторана у смислу наведених независних променљивих, ризик његове погрешне класификације по питању нивоа укупне сатисфакције је низак.

Осим примене *CHAID* алгоритма, у контексту реализације задатка редукције димензионалности разматрана је и ситуација у којој се појављује проблем високе корелације између независних варијабли, који често узрокује стварање редундантних информација, указујући на чињеницу да су у анализу укључене неке ирелевантне варијабле. Стога су у циљу додатне редукције димензионалности (посматрано из угла броја коришћених варијабли), резултати примене *CHAID* алгоритма употпуњени откривањем постојања високе корелације између независних варијабли. Уопштено узевши, један од начина за откривање ове појаве је корелациона анализа и израчунавање коефицијената корелације за све парове (применом *CHAID* процедуре издвојених, статистички значајних) независних варијабли (број издвојених варијабли је 21), као и парцијалних коефицијената корелације између сваке од њих и зависне варијабле *Y* (која је исказана као метричка варијабла, односно, представља просечну вредност ставова испитаника о различитим аспектима квалитета услуге). У складу са наведеним, примењена је следећа двоетапна процедура:

- у првој етапи, извршена је примена метода визуелзације и формулисање закључака о нормалности распореда посматраних - издвојених варијабли, и
- у другој етапи, извршено је, најпре, одређивање корелационе матрице, а затим и анализа њених елемената из угла додатног критеријума за смањење броја варијабли.

Табела 16: Сумарне карактеристике *CHAID* модела

		Подскупови варијабли				
		S-1	*S*-2	*S*-3	*S*-4	*S*-5
Спецификација модела	Зависна варијабла	Ниво сатисфакције – *Y*: низак, умерен и висок				
	Независне варијабле	$X_1; X_2; X_3; X_7; X_{10}$	$X_5; X_6; X_{11}; X_{12}; X_{13}; X_{14}$	$X_8; X_9; X_{15}; X_{16}; X_{17}; X_{18}; X_{19}$	$X_4; X_{20}; X_{21}; X_{22}; X_{23}; X_{24}; X_{25}$	$X_{26}; X_{27}; X_{28}; X_{29}; X_{30}$
	Валидација модела	Унакрсна оцена модела				
	Max дубина стабла	3				
	Min број случајева у „родитељским" чворовима	50				
	Min бројслучајева у „деци" чворовима	15				
Резултати моделирања	Укључене независне варијабле	$X_3; X_{10}; X_7$	$X_{13}; X_5; X_{11}; X_{12}; X_{14}$	$X_{16}; X_9; X_{17}; X_{15}$	$X_{22}; X_{21}; X_{23}; X_{24}; X_4$	$X_{28}; X_{26}; X_{27}; X_{30}$
	Број чворова	16	18	16	18	15
	Број завршних чворова	10	12	10	11	10
	Дубина стабла	3	3	3	3	3

Заправо, на основу графичког приказа ставова испитаника по свакој варијабли у форми хистограма фреквенција и анализом одговарајућих показатеља спроведена је анализа облика распореда фреквенција. Том приликом, у случају свих посматраних варијабли забележен је негативно асиметричан распоред, при чему се вредности мере асиметрије, α_3, крећу у интервалу од –0,992 (за варијаблу X_{27}) до –0,329 (за варијаблу X_{26}). Наведено јасно указује на чињеницу да су ставови са ознакама (вредностима) 4, 5, 6 и 7 доминантни ставови посетилаца ресторана о разматраним аспектима квалитета услужне понуде. Истовремено, при мерењу облика распореда фреквенција у погледу заобљености, за 19 варијабли вредност показатеља заобљености, α_4, је мања од 0 (креће се у интервалу од –1,138 за варијаблу X_{22} до –0,048 за варијаблу X_{24}), тако да су њихови распореди фреквенција више заобљени у односу на теоријски модел нормалног распореда, док је за две варијабле (X_{23} и X_{27}) вредност овог показатеља већа од 0, што упућује на закључак да и њихови распореди фреквенција одступају од егзактно нормалног распореда, с тим што је у питању мања заобљеност од заобљености код нормалног распореда. Будући да универијациона нормалност није постигнута, не може се очекивати да је заједнички, дводимензионални распоред случајних променљивих нормалан. Услед тога, при мерењу степена међусобне зависности између анализом обухваћених, а путем *CHAID* процедуре издвојених, варијабли користи се непараметарска мера корелације – *Spearman*-ов коефицијент корелације ранга (r_s). Сходно томе, полазећи од вредности коефицијента корелације ранга, дефинисан је додатни критеријум за елиминисање варијабли из даље анализе, који гласи: уколико

између две независне варијабле постоји висока корелација, тада из анализе искључити једну од њих, и то ону која има мањи степен повезаности са зависном варијаблом. Како су у статистичкој пракси градирање и интерпретација вредности коефицијента корелације ранга сагласни са градацијом и интерпретацијом вредности коефицијента просте линеарне корелације, за граничну вредност која сигнализира високу корелацију узета је апсолутна вредност коефицијента корелације ранга од 0,8.

Сходно претходно наведеном, у даљем тексту се представљају резултати реализације задатка редукције димензионалности по сваком од формираних подскупова сродних варијабли.

Резултати примене *CHAID* процедуре у случају првог подскупа указују да креирани модел садржи, унутар три нивоа дубине стабла, укупно 16 чворова, од којих се издваја 10 завршних чворова. Истовремено, од укупно 5 иницијално специфицираних независних варијабли у оквиру подскупа *S*-1, у коначан модел је укључено 3, док преостале 2 нису статистички значајне са становишта повезаности са нивоом укупне сатисфакције.

Као што се може запазити на Слици 32, на првом нивоу стабла се налази варијабла X_3, што значи да је реч о варијабли која из групе *S*-1 има највећу моћ у подели, разликовању и класификацији посетилаца ресторана на три групе са становишта сатисфакције квалитетом услуге. (Статистичка значајност варијабле X_3, као првог дискриминатора, одређена је, уз ниво значајности α = 0,05, на основу следећих вредности: χ^2 = 391,546, v = 8 и p-вредност = 0,000). За поделу родитељских чворова формираних у претходном кораку, у оквиру другог нивоа стабла одлучивања статистички значајна варијабла је X_{10}, док је у оквиру трећег нивоа дубине стабла одлучивања издвојена варијабла X_7.

Анализом коефицијента корелације ранга између свих комбинација парова варијабли X_3, X_7 и X_{10}, нису идентификовани случајеви високе корелације, јер су све вредности коефицијената мање од дефинисане граничне вредности 0,8.

Резултати примене *CHAID* процедуре у случају другог подскупа варијабли указују да креирани модел садржи, унутар три нивоа дубине стабла, укупно 18 чворова, од којих се издваја 12 завршних чворова. Истовремено, од укупно 6 иницијално специфицираних независних варијабли у оквиру подскупа *S*-2, *CHAID* алгоритам је издвојио 5 статистички значајних варијабли, а једна варијабла је искључена из даљих разматрања.

Као што је на Слици 33 представљено, на првом нивоу стабла се налази варијабла X_{13}, што значи да је реч о варијабли која из групе *S*-2 има највећу моћ у подели, разликовању и класификацији посетилаца ресторана на три групе са становишта сатисфакције квалитетом услуге. (Статистичка значајност варијабле $X_{13,}$ као првог дискриминатора, одређена је, уз ниво значајности α = 0,05, на основу следећих вредности: χ^2 = 439,434, ν = 8 и *p*-вредност = 0,000). За поделу подгрупа формираних у претходном кораку, у оквиру другог нивоа стабла одлучивања, као статистички значајне варијабле су издвојене X_5 и X_{11}, док су у оквиру трећег нивоа дубине стабла одлучивања издвојене варијабле X_{12} и $X_{14.}$

Сходно дефинисаном критеријуму за идентификовање високе корелације између парова независних варијабли X_5, X_{11}, X_{12}, X_{13} и X_{14} уочено је присуство овог проблема између варијабли X_{12} и X_{13} ($r_{X_{12}X_{13}}$ = 0,80). Како је вредност коефицијента корелације ранга између варијабли X_{13} и Y (r_{X13Y} = 0,82) већа од вредности истог коефицијента између варијабли X_{12} и Y ($r_{X_{12}Y}$ = 0,81), за потребе даље анализе из модела се накнадно елиминише варијабла X_{12}.

Резултати примене *CHAID* процедуре у случају трећег подскупа варијабли указују да креирани модел садржи, унутар три нивоа дубине стабла, укупно 16 чворова, од којих се издваја 10 завршних. Истовремено, од укупно 7 иницијално специфицираних независних варијабли у оквиру групе *S*-3, издвојено је 4 статистички значајне варијабле, а преостале 3 су елиминисане из коначног модела.

Као што је илустровано на Слици 34, на првом нивоу стабла се налази варијабла X_{16}, што значи да је реч о варијабли која из подскупа *S*-3 има највећу моћ у подели, разликовању и класификацији посетилаца ресторана на три групе са становишта сатисфакције квалитетом услуге. (Статистичка значајност варијабле $X_{16,}$ као првог дискриминатора, одређена је, уз ниво значајности α = 0,05, на основу следећих вредности: χ^2 = 522,607, ν = 6[76] и *p*-вредност = 0,000). За поделу подгрупа формираних у претходном кораку, у оквиру другог нивоа стабла одлучивања, као статистички значајне варијабле су издвојене X_9 и X_{17}, док је у оквиру трећег нивоа дубине стабла издвојена варијабла $X_{15.}$

Приликом провере присуства проблема високе корелације између независних варијабли које су остале у моделу након примене *CHAID* процедуре, утврђено је да између варијабли X_{16} и X_{17} постоји, сходно дефинисаном критеријуму, висока

[76] С обзиром да све независне варијабле имају исти број модалитета, промена броја степени слободе (ν) указује на својство овог алгоритма да се током процесирања података поред редукције броја независних варијабли остварује и редукција њихових категорија, спајањем категорија које се статистички значајно не разликују.

корелација ($r_{X_{16}X_{17}}$ = 0,83). Међутим, како је вредност коефицијента корелације ранга између варијабли X_{16} и Y ($r_{X_{16}Y}$ = 0,88) већа од вредности истог коефицијента између варијабли X_{17} и Y ($r_{X_{17}Y}$ = 0,85), за потребе даље анализе из модела се накнадно елиминше и варијабла X_{17}.

Резултати примене *CHAID* процедуре у случају четвртог подскупа варијабли указују да креирани модел садржи, унутар три нивоа дубине стабла, укупно 18 чворова, од којих се издваја 11 завршних. Истовремено, од укупно 7 иницијално специфицираних независних варијабли у оквиру подскупа *S*-4, издвојено је 5 статистички значајних варијабли, а преостале 2 су елиминисане из коначног модела.

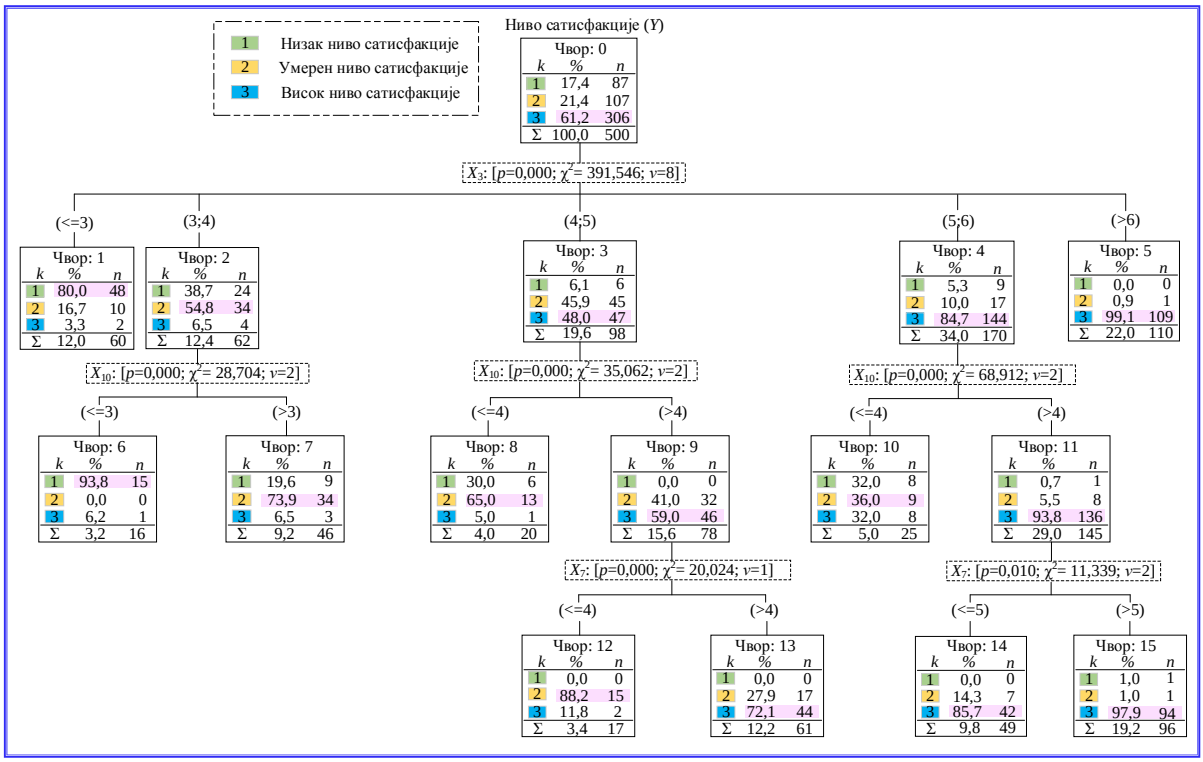
Као што се може уочити на Слици 35, на првом нивоу стабла се налази варијабла X_{22}, што значи да је реч о варијабли која из групе *S*-4 има највећу моћ у подели, разликовању и класификацији посетилаца ресторана на три групе са становишта сатисфакције квалитетом услуге. (Статистичка значајност варијабле $X_{22,}$ као првог дискриминатора, одређена је, уз ниво значајности α = 0,05, на основу следећих вредности: χ^2 = 506,029, ν = 6 и *p*-вредност = 0,000). За поделу подгрупа формираних у претходном кораку, у оквиру другог нивоа стабла одлучивања, као статистички значајне варијабле су издвојене X_{21}, X_{23} и X_{24}, док је у оквиру трећег нивоа дубине стабла издвојена варијабла $X_{4.}$

Након примене *CHAID* процедуре на варијабле у подскупу *S*-4, приликом провере присуства проблема високе корелације између независних варијабли које су остале у моделу, утврђено је да између варијабли X_{21} и X_{22} постоји, сходно дефинисаном критеријуму, висока корелација ($r_{X_{21}X_{22}}$ = 0,84). Међутим, како је вредност коефицијента корелације ранга између варијабли X_{22} и Y ($r_{X_{22}Y}$ = 0,89) већа од вредности истог коефицијента између варијабли X_{21} и Y ($r_{X_{21}Y}$ = 0,86), за потребе даље анализе из модела се накнадно елиминише и варијабла X_{21}. Такође, вредност коефицијента корелације између варијабли X_{23} и X_{24} је изнад дефинисане граничне вредности ($r_{X_{23}X_{24}}$ = 0,82), а како је вредност коефицијента корелације ранга између варијабли X_{24} и Y ($r_{X_{24}Y}$ = 0,86) већа од вредности истог коефицијента између варијабли X_{23} и Y ($r_{X_{23}Y}$ = 0,84), за потребе даље анализе у моделу се задржава X_{24}, а елиминише X_{23}.

Коначно, резултати примене *CHAID* процедуре у случају петог подскупа варијабли указују да креирани модел садржи, унутар три нивоа дубине стабла, укупно 15 чворова, од којих се издваја 10 завршних. Истовремено, од укупно 5 иницијално специфицираних независних варијабли у оквиру подскупа *S*-5, издвојено је 4 статистички значајне варијабле, а 1 варијабла је елиминисана из коначног модела.

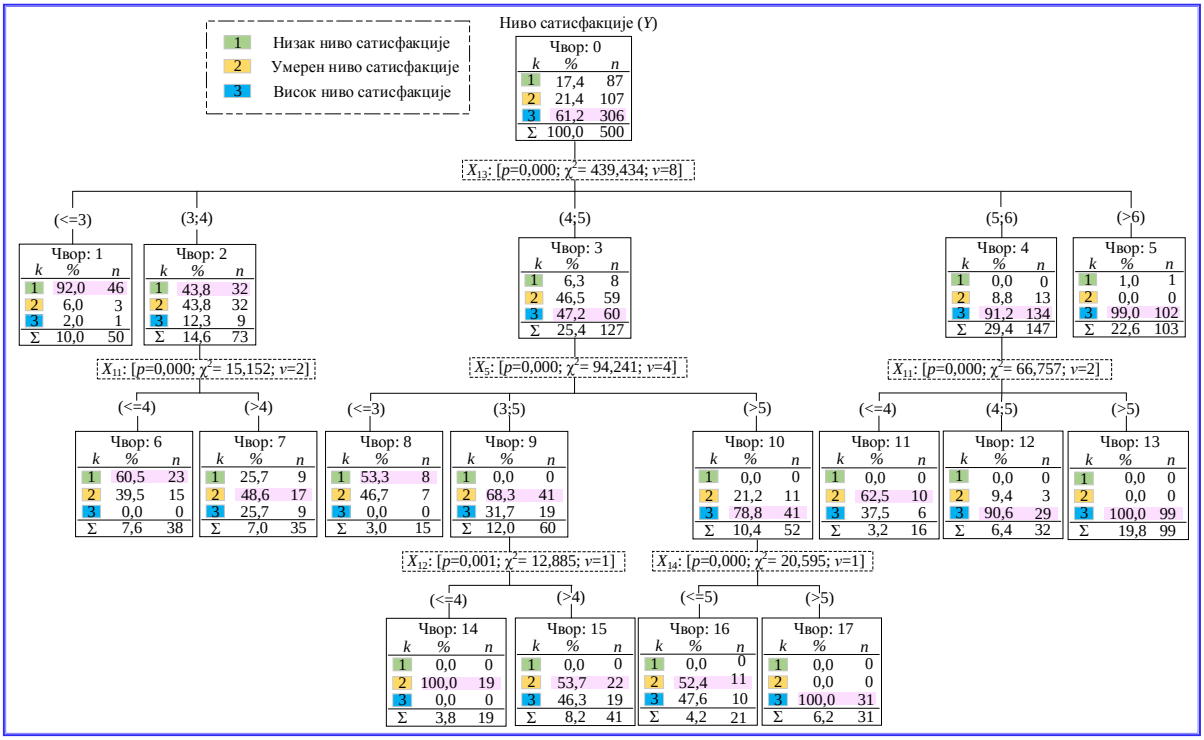
Као што се може уочити на Слици 36[77], на првом нивоу стабла се налази варијабла X_{28}, што значи да је реч о варијабли која из подскупа *G*-5 има највећу моћ у подели, разликовању и класификацији посетилаца ресторана на три групе са становишта сатисфакције квалитетом услуге. (Статистичка значајност варијабле $X_{28,}$ као првог дискриминатора, одређена је, уз ниво значајности α = 0,05, на основу следећих вредности: χ^2 = 499,022, v = 8 и p-вредност = 0,000). За поделу подгрупа формираних у претходном кораку, у оквиру другог нивоа стабла одлучивања, као статистички значајне варијабле су издвојене X_{26} и X_{27}, док је у оквиру трећег нивоа дубине стабла издвојена варијабла $X_{30.}$

Након примене *CHAID* процедуре на варијабле у подскупу *G*-5, приликом провере присуства проблема високе корелације између независних варијабли које су остале у моделу, утврђено је да између пара X_{27} и X_{28} постоји, сходно дефинисаном критеријуму, висока корелација ($r_{X_{27}X_{28}}$ = 0,87). Међутим, како је вредност коефицијента корелације ранга између варијабли X_{28} и Y ($r_{X_{28}Y}$ = 0,86) већа од вредности истог коефицијента између варијабли X_{27} и Y ($r_{X_{27}Y}$ = 0,85), за потребе даље анализе из модела се накнадно елиминише варијабла X_{27}.

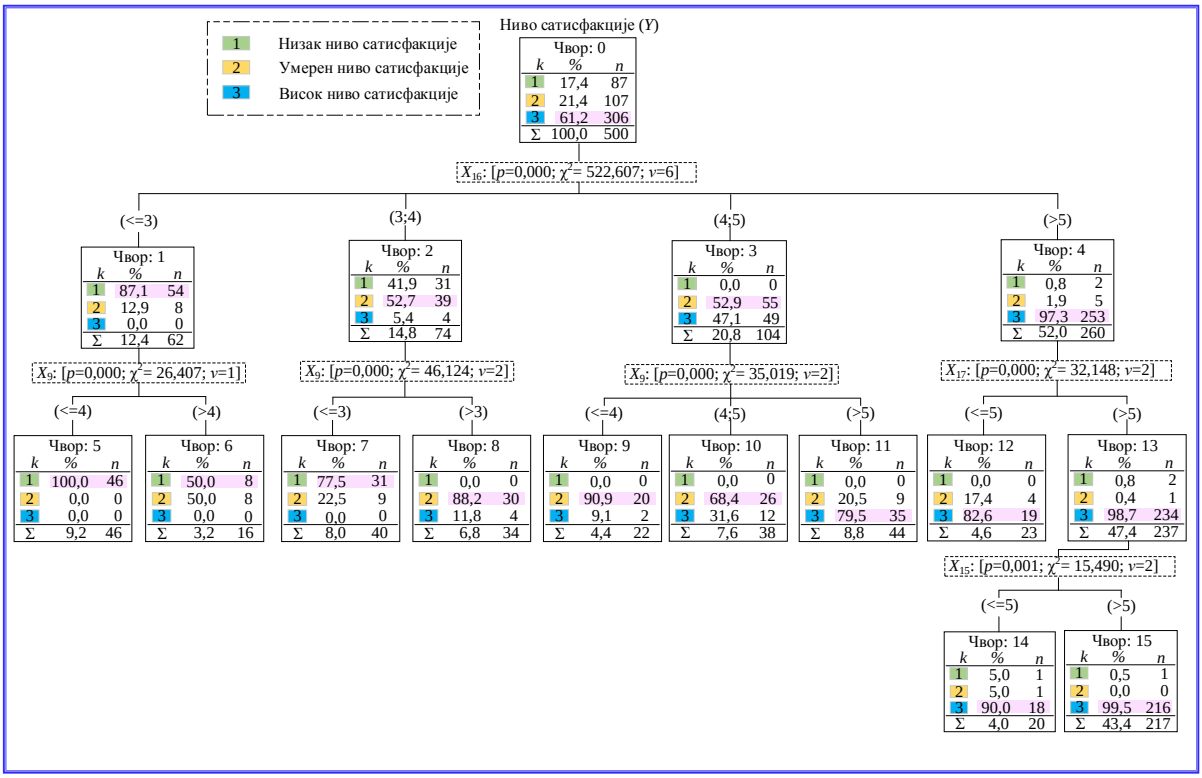


Слика 32: *CHAID* стабло одлучивања *S*-1

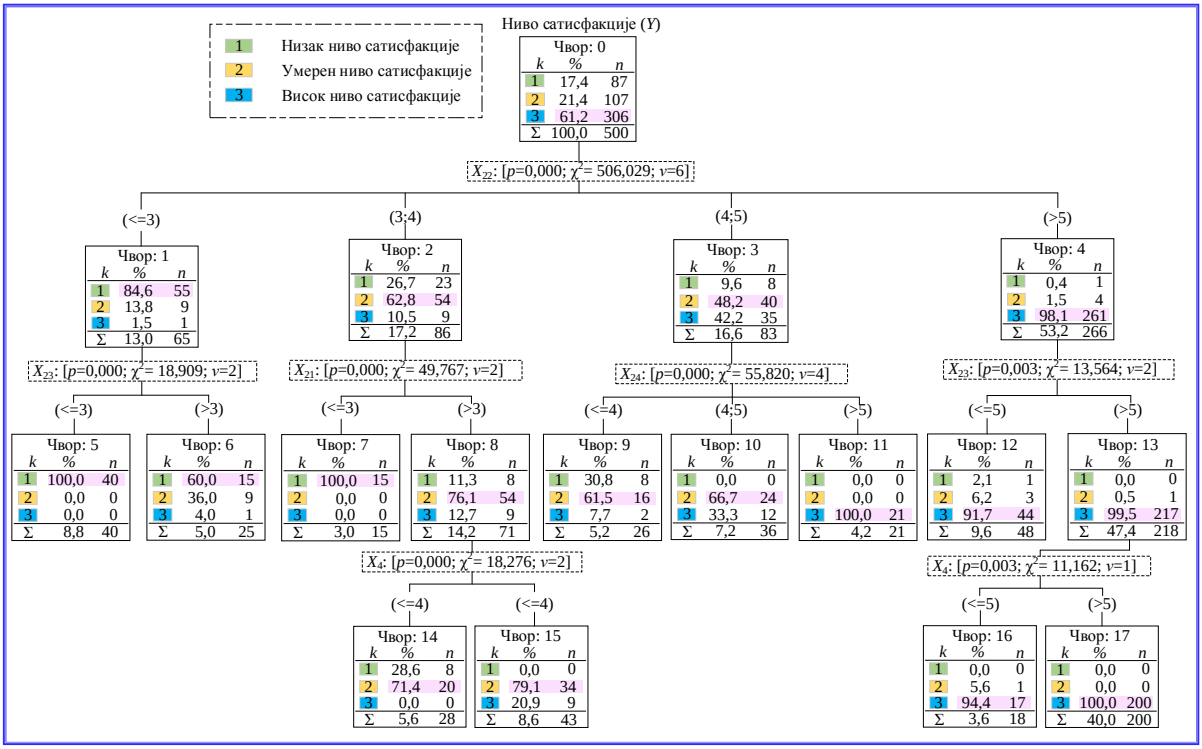
[77] Напомена: На графичким илустрацијама резултирајућих стабала одлучивања (Слике 32, 33, 34, 35 и 36), симболи χ^2, p, и v имају следеће значење: χ^2 = вредност статистике теста, p = прилагођена p-вредност и v = број степени слободе.



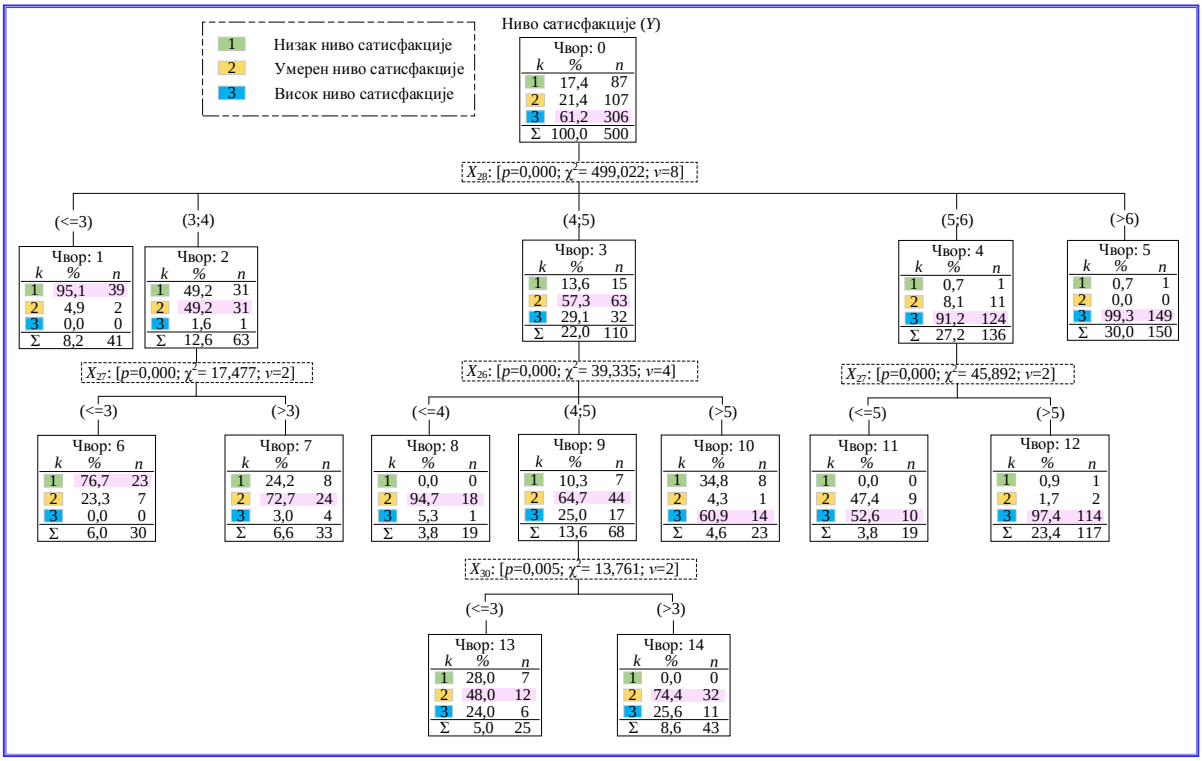
Слика 33: *CHAID* стабло одлучивања *S*-2



Слика 34: *CHAID* стабло одлучивања *S*-3



Слика 35: *CHAID* стабло одлучивања *S*-4



Слика 36: *CHAID* стабло одлучивања *S*-5

Дакле, на основу вредности χ^2 критеријума (уз кореспондентну *p*-вредност) и резултата провере присуства високе корелације између парова независних варијабли,

изабран је подскуп оних варијабли које представљу улазне елементе за даље истраживање и примену нехијерархијске процедуре груписања.

13.3.3. Резултати примене анализе груписања

Претходно презентована процедура, базирана на примени *CHAID* и корелационе анализе, омогућила је издвајање варијабли које највише могу да допринесу откривању група корисника услуге, генерално, међусобно различитих са становишта перцепције (ресторанске) услужне понуде. С обзиром да се анализа груписања сматра једном од изузетно ефективних метода управо за сврхе идентификовања и разумевања група корисника услуге које карактерише слично понашање (у овом случају у погледу (не)задовољства карактеристикама квалитета услуге), а које се, истовремено, међусобно разликују, на подскупу издвојених варијабли, спроведена је нехијерархијска процедура груписања заснована на примени метода (алгоритма) *k*-средина.

Методологија *k*-средина подразумева да се најпре одреде (иницијални) центри одређеног, претходно дефинисаног, броја група, а затим да се сви објекти чија је удаљеност од центра једне групе мања од унапред постављеног критеријума групишу у оквиру те групе. Реч је о итеративном поступку груписања и одређивања центара група који траје све док се сваком од обрађених слогова по јединицама посматрања не додели одговарајућа ознака групе. Такође, на крају, ова процедура резултира и коначним центрима група, који, суштински, представљају центроиде група (односно, вектор аритметичких средина за варијабле на основу којих је спроведено груписање, а које су, логично, одређене узимањем у обзир објеката који припадају конкретној групи).

Сходно наведеном, у Табелама 17 и 18 су представљени сумарни резултати примене метода *k*-средина за груписање посетилаца посматраног ресторана према изабраним варијаблама (које одражавају ставове посетилаца о одређеним карактеристикама квалитета услуге), при чему је унапред дефинисан број од 3 групе. Важно је запазити, на бази одстојања између формираних група, да је група 3 оштро разграничена од групе 1, док код група 1 и 2 не постоји толико изражено разграничење као у случају групе 3. Такође, за сврхе правилне интерпретације добијених резултата обраде података и креираног модела сегментације тржишта указује се и на следеће: ставови испитаника на основу којих су формиране групе су исказани на истој, седмостепеној интервалној скали. Услед тога, израчунавање аритметичке средине вектора аритметичких средина по свакој групи има смисла, тако да се у функцији те вредности може одредити просечан ниво сатисфакције групе. Заправо, раније

елабориран начин одређивања модалитета варијабле Y и њихово превођење у категорије 1, 2 и 3 (односно низак, умерен и висок степен сатисфакције, респективно), може се потпуно оправдано применити и при интерпретацији формираних група у погледу просечног нивоа сатисфакције.

Анлизом груписања, очигледно, детерминисана је и нова (категоријска) варијабла (X_{40}) која указује на припадност сваке јединице посматрања конкретној групи. Полазећи од тога, у Табели 19 је представљена структура формираних група према социодемографским карактеристикама испитаника и карактеристикама које су повезане са њиховим посетама ресторанима. Консеквентно, у даљу анализу је укључено и разматрање следећег истраживачког питања: да ли између наведених карактеристика (односно варијабли: X_{31}, X_{32}, X_{33}, X_{34}, X_{35} и X_{36}) и припадности испитаника конкретној групи (X_{40}) постоји статистички значајна веза. За добијање одговора на постављено питање коришћен је метод тестирања статистичких хипотеза. Како су све посматране варијабле категоријске, у функцији испитивања наведене међузависности, имплементиран је непараметарски χ^2 тест независности два обележја.

Табела 17: Одстојање између центроида формираних група

Група	1	2	3
1	0		
2	5,028	0	
3	10,361	5,679	0

Табела 18: Центроиди формираних група

Варијабле	Број групе (X_{40})		
	1 (n_1 = 109)	2 (n_2 = 129)	3 (n_3= 262)
X_3	3,64	5,10	6,27
X_7	3,61	5,38	6,08
X_{10}	3,72	4,89	6,28
X_5	3,91	5,21	6,26
X_{11}	3,74	4,54	6,15
X_{13}	3,72	5,07	6,19
X_{14}	3,50	4,66	6,20
X_9	3,54	5,22	6,39
X_{15}	3,86	5,40	6,42
X_{16}	3,72	4,77	6,32
X_4	3,60	5,36	6,37
X_{22}	3,65	4,64	6,47
X_{24}	3,72	4,52	6,23
X_{26}	4,06	4,74	6,32
X_{28}	4,06	4,97	6,49
X_{30}	2,97	3,84	5,93
„Просек центроида”	3,69	4,89	6,27

Табела 19: Карактеристике испитаника по групама

Назив и симбол варијабле	Модалитети	Група (X_{40})					
		1		2		3	
		f_i	%	f_i	%	f_i	%
Пол (X_{31})	мушки	66	60,6	73	56,6	119	45,4
	женски	43	39,4	56	43,4	143	54,6
Старост (X_{32})	до 25	9	8,3	8	6,2	43	16,4
	26-35	9	8,3	21	16,3	82	31,3
	36-45	27	24,8	46	35,7	68	26,0
	46-55	38	34,9	31	24,0	43	16,4
	56-65	17	15,6	22	17,1	21	8,0
	више од 65	9	8,3	1	0,8	5	1,9
Ниво образовања (X_{33})	основно	7	6,4	1	0,8	6	2,3
	средње	32	29,4	72	55,8	95	36,3
	више	24	22,0	9	7,0	49	18,7
	високо	46	42,2	47	36,4	112	42,7
С ким најчешће одлазите у ресторан? (X_{34})	са пријатељима	28	25,7	50	38,8	139	53,1
	са породицом	57	52,3	47	36,4	71	27,1
	са колегама	8	7,3	24	18,6	49	18,7
	сами	**16**	14,7	8	6,2	3	1,1
Да ли сте поборник здраве исхране? (X_{35})	да	67	61,5	67	51,9	129	49,2
	не	9	8,3	27	20,9	38	14,5
	делимично	33	30,3	35	27,1	95	36,3
Да ли радо пробате специјалитете ресторана? (X_{36})	да	76	69,7	96	74,4	198	75,6
	не	33	30,3	33	25,6	64	24,4

Напомена: Симболом f_i је представљен број испитаника по модалитетима варијабли.

На основу резултата спроведеног тестирања (односно, вредности парцијалних статистика χ^2 теста независности и придружених p-вредности, за сваки посматрани пар варијабли), уз ниво значајности теста $\alpha = 0{,}05$, може се закључити следеће: ▸ у случају варијабли X_{31} ($\chi^2 = 8{,}789$, $\nu = 2$. p-вред. = 0,012), X_{32} ($\chi^2 = 56{,}546$, $\nu = 8$; p-вред. = 0,000), X_{33} ($\chi^2 = 18{,}944$, $\nu = 4$; p-вред. = 0,010), X_{34} ($\chi^2 = 61{,}279$, $\nu = 6$; p-вред. = 0,000) и X_{35} ($\chi^2 = 11{,}088$; $\nu = 4$; p-вред. = 0,026) постоји довољно емпиријских доказа за одбацивање нулте хипотезе и прихватање алтернативне хипотезе да између сваке од наведених варијабли и варијабле X_{40} постоји статистички значајна веза; ▸ у случају варијабле X_{36} ($\chi^2 = 1{,}384$, $\nu = 2$, p-вред. = 0,501) не постоји довољно доказа за одбацивање нулте хипотезе, тако да међусобна зависност варијабли X_{36} и X_{40} није статистички значајна.[78]

13.3.4. Анализа креираног модела груписања

Спроведено емпиријско истраживање, у основи засновано на примени *CHAID* анализе и анализе груписања у условима велике количине података, резултирало је

[78] Током спровођења поступка тестирања хипотеза и провере испуњености претпоставки на којима се заснива валидна примена χ^2 теста, у ситуацијама појављивања сувише малих теоријских (очекиваних) фреквенција у ћелијама табеле контигенције (вредности мање од 5) вршено је спајање редова (колона) којима оне припадају са суседним редом (колоном). Мале очекиване фреквенције су се појавиле у случају варијабли X_{32} и X_{33}.

формирањем три групе испитаника према њиховим ставовима о карактеристикама квалитета услуге. Упоређивањем центроида (Табела 18) и социодемографских (Табела 19) формираних група испитаника, у наставку се, између осталог, укратко представљају профили сваке од њих.

У оквиру прве групе класификовано је 109 посетилаца посматраног ресторана, који представљају 21,8% величине расположивог узорка. Реч је о групи која, у поређењу са другим групама, има најмањи „просек центроида" ($\overline{\overline{x}}_1 = 3{,}69$). Како је, сходно уведеном критеријуму за категоризацију, $\overline{\overline{x}}_1 < 4$, ова група се може назвати група испитаника са ниским (просечним) нивоом сатисфакције. Осим тога, од 16 варијабли на основу којих је спроведена анализа груписања, доминирају варијабле (укупно 14) чија је забележена аритметичка средина ($\overline{x}_i$) мања од 4. Истовремено, за 2 варијабле аритметичка средина незнатно одступа од 4 што је индикатор неутралног, индиферентног става посетилаца који припадају овој групи у погледу карактеристика квалитета услуге означених симболима X_{26} и X_{28}. У погледу варијабли из дела упитника који се односи на социодемографска обележја испитаника и њихове посете ресторанима, структуру ове групе углавном чине испитаници који поседују следеће карактеристике: мушки пол (60,6%), старост од 46 до 55 година (34,9%), високо образовања (42,2%), најчешће са породицом одлазе у ресторане (52,3%) и прихватају начела здраве исхране (61,5%).

Друга група обухвата 129 испитаника, који представљају 25,8% величине расположивог узорка. Просечна вредност вектора аритметичких средина ове групе износи $\overline{\overline{x}}_2 = 4{,}89$. Како је $4 \leq \overline{\overline{x}}_2 < 5$, ова група се, генерално, може назвати група испитаника са умереним (просечним) нивооом сатисфакције. Детаљнија анализа показује да су испитаници ове групе у случају 8 карактеристика квалитета исказали свој став који је еквивалентан са категоријом умерен ниво сатисфакције. Такође, у оквиру ове групе треба запазити да је аритметичка средина 7 варијабли еквивалентна са категоријом висок степен сатисфакције (уз напомену да је реч о релативно високом нивоу будући да се просечан ниво задовољства креће у интервалу од 5,07 до 5,40), док је у случају варијабле X_{30}, штавише, аритметичка средина еквивалентна са категоријом низак степен сатисфакције. У погледу варијабли из дела упитника који се односи на социодемографска обележја испитаника и њихове посете ресторанима, у структури ове групе доминирају испитаници који поседују следеће карактеристике: мушки пол

(56,6%), старост од 36 до 45 година (35,7%), средње образовања (55,8%), најчешће са пријатељима одлазе у ресторане (38,8%) и прихватају начела здраве исхране (51,9%).

Трећа група је највећа и обухвата 262 посетиоца, што чини 52,4% посматраног узорка. Ову групу карактерише највећа вредност „просека центроида" ($\bar{\bar{x}}_3$ = 6,273). С обзиром да је $\bar{\bar{x}}_3 \geq 5$, може се констатовати да је у питању група испитаника са високим (просечним) нивоом сатисфакције. У структури 16 посматраних варијабли, скоро све варијабле (укупно 15) су од стране испитаника ове групе, у просеку, оцењене оценом изнад 6. У погледу варијабли из дела упитника који се односи на социодемографска обележја испитаника и њихове посете ресторанима, у структури ове групе углавном доминирају испитаници који поседују следеће карактеристике: женски пол (54,6%), старост од 26 до 35 година (31,3%), високо образовања (42,7%), најчешће са пријатељима одлазе у ресторане (53,1%) и прихватају начела здраве исхране (49,2%).

Имајући у виду претходне закључке, основано се може тврдити да је добијено решење дефинисане проблемске ситуације у форми креираног модела сегментације тржишта у ресторатерском пословању логично, смислено и интерпретабилно. Не улазећи, овом приликом, у приказ бројних формалних процедура и практичних инструкција за оцењивање валидности формираних група, ради оцене квалитета креираног модела извршена је компарација зависне варијабле *Y* -укупан ниво сатисфакције (и њених категорија 1, 2 и 3) са новом варијаблом која је изведена у процесу моделирања – припадност испитаника одређеној групи, а чије су категорије у погледу значења са становишта нивоа сатисфакције сагласне са категоријама варијабле *Y*. Сумарни приказ извршене компарације представљен је у Табели 20.

Као што се може уочити, из угла нивоа сатисфакције који је дефинисан узимањем у обзир 30 иницијалних варијабли, идентификовано је 87 испитаника чији је просечан ниво сатисфакције означен као низак, 107 испитаника чији је просечан ниво сатисфакције означен као умерен (укључујући и неутралан став у смислу „нити задовољан нити незадовољан") и 306 испитаника чији је просечан ниво сатисфакције означен као висок. Након редукције димензионалности и избором 16 независних варијабли за спровођење метода анализе груписања, креиран је модел који је успешно разврстао 407 (главна дијагонала у Табели 20) од 500 испитаника у посматраном узорку. Заправо, модел је успешно класификовао 81,4% од укупног броја посетилаца у узорку, што је, неспорно, између осталог, индикатор доброг избора варијабли за формирање група. Посматрано по категоријама, модел груписања је алоцирао 22

испитаника више у групу ниског нивоа сатисфакције, такође, 22 више у групу умереног нивоа сатисфакције и, последично, група испитаника високог нивоа сатисфакције је мања за 44 испитаника у односу на класификацију према варијабли *Y*. Мада, генерално, најбоља мера перформанси резултирајућег модела није сирова прецизност, већ његова корисност и успешност у остваривању примарне сврхе због које је и формиран у домену решења одређеног проблема, ипак се мора констатовати да креирани модел поседује високу прецизност у разврставању опсервација по категоријама.

Табела 20: Компарација распореда испитаника према нивоу сатисфакције и групама

Варијабле		Укупан ниво сатисфакције			Σ
		1	2	3	
Припадност групи	1	79	30	0	109
	2	7	75	47	129
	3	1	2	253	262
Σ		87	107	306	500

Полазећи од резултата спроведеног истраживања на подацима расположивог узорка, могу се издвојити, између осталих, следеће законитости у погледу степена сатисфакције посетилаца ресторана квалитетом услужне понуде:

- највећи број испитаника припада категорији испитаника са високим степеном сатисфакције квалитетом услуге;
- у свим групама испитаника, формираним сходно њиховим ставовима у вези са карактеристикама квалитета услуге, најлошије је вреднован онај аспект ресторанске понуде који се односи на укључивање специјалних - наменских јеловника;
- у трећој формираној групи испитаника, испитаници су исказали висок степен слагања у погледу свих тврдњи о квалитету услуге, с тим што је најбоље оцењен онај аспект ресторанске понуде који се односи на свежину хране;
- у другој формираној групи испитаника, просечна вредност ставова испитаника варира од ниске до високе сатисфакције, с тим што преовладава група карактеристика квалитета за које је просечна вредност става еквивалентна умереном степену сатисфакције, при чему је најбоље оцењен аспект квалитета и чистоће прибора за послуживање;
- у првој формираној групи испитаника, испитаници су исказали низак степен слагања у погледу готово свих тврдњи о квалитету услуге, изузев две тврдње које се односе на одређене аспекте квалитета хране;
- од укупног броја испитаника који најчешће сами одлазе у ресторан, скоро 60% припада првој формираној групи, то јест, групи незадовољних;

- у првој групи, то јест, групи најмање задовољних испитаника, доминирају испитаници који најчешће у ресторан одлазе са породицом, док у трећој групи, то јест, групи највише задовољних испитаника доминирају испитаници који најчешће у ресторан одлазе са пријатељима итд.

Поред наведених, постоји још читав низ откривених законитости чије би експлицитно формулисање употпунило слику о тржишним сегментима корисника услуге. Другим речима, из угла менаџмента ресторана, практична корисност резултата (добијених реализацијом предложеног концептуално-методолошког оквира истраживања) се огледа у чињеници да обезбеђују детаљно упознавање група корисника у погледу њихових различитих карактериситка. Сходно томе, спроведено истраживање пружа основу за дефинисање јасних смерница и предузимање конкретних акција у правцу побољшања квалитета услуге и, консеквентно, нивоа сатисфакције корисника услуге конкретног услужног предузећа.

Сумирањем изнетих разматрања, посматрано из перспективе дефинисаних истраживачких хипотеза, може се констатовати да су обе хипотезе потврђене. Међутим, са реализованим истраживањем су повезана и извесна ограничења, која треба узети у обзир приликом интерпретације резултата. Наиме, формулисани закључци су валидни и поуздани, али се односе искључиво на податке расположивог узорка, тако да сваки покушај да се исти генерализују захтева повећање величине узорка, као и временског периода прикупљања података. Са променама ових аспеката истраживања стекао би се додатни увид у приказане резултате. Заправо, узимајући у обзир наведена ограничења, будућа истраживачка настојања иницијално ће бити усмерена на дубљу анализу добијених резултата (укључивањем нових метода) и репликацију спроведеног поступка у циљу обезбеђења континуираног праћења и компарације идентификованих резултата. Такође, полазећи од својства флексибилности конципираног методолошког оквира, будућим истраживањима биће обухваћена и његова имплементација и додатно тестирање при анализи високо димензионалних феномена у проблемским контекстима других услужних области.

Генерално, мерење (кватификовање) квалитета услуге и сатисфакције корисника услуге је важно питање за већину организационих ентитета у савременом окружењу. Заправо, без података, као резултата мерења, нема разумевања разматраних феномена, а самим тим ни потенцијалних побољшања. Сходно томе, оправдано се може констатовати да, без података нема анализе, а без анализе нема *feedback*-а и корективне акције.

ЗАКЉУЧАК

Концептуализацијом садржаја дисертације кроз (а) синтетизовање истраживачког материјала и грађе многих теоретичара и практичара и (б) конципирање и спровођење оригиналног емпиријског истраживања, отворена су бројна теоријско-методолошка питања и апликативне дилеме и осветљени кључни аспекти извођења законитости из економских података коришћењем *DM* приступа у анализи података. Несумњиво, на тај начин су аргументована, већ у уводном делу ове докторске дисертације наговештена, својства комплексности, изузетног значаја и актуелности разматране проблематике. У оквиру обрађених делова, као конститутивних елеметата и међусобно повезаних целина структуре дисертације, у оквиру којих је организовано излагање материје, изнети су одређени ставови по разним питањима у вези са дефинисаним предметом истраживања. Ипак, резимирајући резултате спроведеног истраживања, могуће је формулисати и генералне закључне констатације, чију валидност треба, пре свега, самерити у констелацији са формулисаним хипотезама и дефинисаним циљевима дисертације. Заправо, прецизирани сет циљева и хипотеза примарно опредељује и детерминише природу и контекст закључака овог рада.

Резултати спроведеног истраживања недвосмислено указују на следеће:

✓ Једна од највећих промена у сфери савремене економије, иницирана, али и вођена развојем информационо-комуникационих технологија, односи се на растућу улогу података као извора знања, раста и значајног потенцијала за генерисање економске вредности из перспективе њихових корисника. Наиме, дигитална револуција је омогућила стварање и складиштење огромне количине разноврсних података. Последично, организациони ентитети су постали преплављени подацима који сами по себи немају вредност (већ представљају трошак) са становишта решавања конкретних проблема, тако да се за врло кратко време појавило питање начина издвајања вредних информација и корисног знања из обиља расположивих података, а које се суштински може изразити кроз следећу формулацију: Како искористити предности поседовања података. Полазећи од упозорења, која се често истичу у литератури, да ће, без радикалних промена у начину процесирања, велике количина података постати пасивне архиве, у раду је констатовано да су, поред препознавања и схватања значаја података за побољшање различитих аспеката пословања, неопходна и одговарајућа усклађивања стварних потреба и аналитичких могућности за њихово процесирање и анализу. Стога је, у контексту могућих одговора за решење проблема

несклада између расположиве количине података и степена њихове искоришћености, апострофирана улога *DM*-а, као мултидисциплинараног приступа и методолошког оквира за разумевање, претраживање, обраду и анализу великих количина података.

✓ Валидна интерпретација и примена *DM*-а у анализи података претпоставља, између осталог, решавање концептуалних и језичко-терминолошких непрецизности и недоумица. Имајући у виду варијететност критеријума и аспеката посматрања, *DM* је опредељен као: ▸ вишеетапни, интерактиван, итеративан и креативан процес (а не једнократна активност) проналажења корисних и иновативних информација и знања у великим количинама података; ▸ централни потпроцес ширег процеса откривања знања из података; ▸ скуп компјутерски подржаних метода и алгоритама дизајнираних за претраживање и анализу велике количине података у циљу проналажења законитости у подацима; и ▸ мултидисциплинарна научно-истраживачка и апликативна област, која се, као производ компјутерске ере, континуирано мења и развија. Такође, истраживањем су идентификовани и издвојени следећи елементи релевантни са становишта правилне интерпретације *DM* концепта: велика количина података, смислене законитости, подручје примене и заснованост на процесном приступу у разматрању конкретних феномена.

✓ За реализацију *DM* процеса и, шире, процеса откривања знања из података неопходна су одговарајућа експертска знања и кореспондентне професионалне улоге учесника у процесу. Заправо, процес откривања законитости из података једна особа не може самостално да спроведе. Успешна решења захтевају тимски рад и ангажовање стручњака из домена апликативног подручја, информационих система и анализе података. Истина, без аутоматског процесирања немогуће је спровести *DM* анализу, али због постојања потенцијалне опасности да се интерпретација читавог концепта сведе на примену софтверских алата за откривање значајних информација и знања из података, посебно се истиче чињеница да пресудну улогу у спровођењу *DM* анализе има људски фактор, као и да су софтверска решења, такође, незаобилазни, али ипак помоћни алати, који аутоматски не решавају *DM* проблеме и задатке. Генерално, може се констатовати да успешна реализација *DM* процеса, као форма пројектног задатка, зависи од бројних фактора, при чему сарадњи између свих учесника са одговарајућим одговорностима, компетенцијама и вештинама припада посебно место.

✓ Свеобухватно разумевање комплексности *DM*-а представља битан услов за апликативну успешност при решавању задатака у реалним проблемским контекстима, чиме се афирмише неопходност изучавања и сагледавања позитивних и негативних

страна примене *DM*–а из различитих перспектива. *DM* је идентификован и препознат као значајан концепт и технологија за претварање великих количина података у вредне законитости, а самим тим је постао опште прихватљив као средство за остварење циљева у бројним подручјима. Међутим, *DM* апликације истовремено са собом носе бројне изазове и потенцијалне негативне импликације са становишта друштва, организације и појединаца. Свест о различитим типовима могућих заблуда и грешака и њиховим консеквенцама, односно, алтернативно, о истинама и реалностима везаним за *DM* примену знатно смањује ризик од неуспеха при реализацији пројекатних *DM* задатака.

✓ Могућности примене смислених методолошких поступака над подацима, ток саме анализе и квалитет добијених резултата детерминисани су карактеристикама података, као основном компонентом *DM* приступа у откривању законитости о анализираним феноменима. Заправо, полазећи од тога да су подаци сирови материјал за генерисање релевантних информација, који, самим тим, директно опредељује вредност идентификованих законитости, у дисертацији је актуелизован и укратко изложен проблем посматрања карактеристика података кроз призму категоризације (типологије), начина оргазнизовања и квалитета података. На основу наведених разматрања установљено је следеће: ▸ дијапазон дозвољених (оправданих) рачунских и логичких операција и метода, које је могуће применити у конкретној ситуацији зависи од скупа (то јест, домена) могућих вредности анализом обухваћених обележја јединица посматрања, као и нивоа обухвата података; ▸ појава велике количне и изузетно комплексних форми података (у виду не само бројева, већ и знакова, симбола, слика и звукова) условила је изналажење нових облика логичког представљања и повезивања, односно организовања података у рачунарским меморијама, чиме је створена основа за ефикасно спровођење целокупног поступка извођња и обезбеђења законитости из података у правом тренутку и облику погодном за крајње кориснике; и ▸ ефикасно управљање квалитетом података захтева институционализовање процеса (са пратећим процедурама) који спречавају настанак података лошег квалитета, укључујући и успостављање адекватносг система мерења димензија квалитета, као скупа карактеристика података, попут доступности, потпуности, тачности, разумљивости и слично.

✓ Реализација *DM* циљева заснива се на успешном комбиновању и спровођењу сета појединачних задатака до чијих се решења долази применом изабраних кореспондентних метода претпроцесирања, процесирања и постпроцесирања података.

Наиме, сваки задатак откривања знања из података се може решити помоћу неколико различитих методолошких поступака, као што и један метод може бити употребљен за реализацију више задатака. У том смислу, како не постоји универзални поступак за решавање конкретних задатака, избор адекватне комбинације метода представља прави истраживачки изазов у домену целог процесу откривања знања из података. Такође, развијени софтверски алати отварају велике могућности комбиновања метода што додатно доприноси истраживачкој атрактивности овог питања. Сходно широком опсегу метода који налазе примену у *DM* окружењу, поред питања избора покренуто је и питање правилне употребе одређених метода у решавању конкретних проблемских ситуација током сваке фазе процеса откривања знања. Заправо, претпоставка квалитетне анализе податка засноване на *DM* приступу је познавање кључних одређења метода (претпоставки и услова коришћења, предности и недостатака) и разматрање њихове апликативности у констелацији са карактеристикама конкретног проблема. Фокусирањем истраживачке пажње на представљање и анализу основних и специфичних својстава изабраних методолошких поступака / метода из група означених као методи претпроцесирања, процесирања и постпроцесирања, дошло се до следећег закључка: избор методолошких поступака треба да буде базиран на разматрању карактеристика одређене проблемске ситуације (преточене у методолошке оквире) наспрам карактеристика потенцијалних методолошких решења, што захтева широк распон стручног знања о својствима метода у циљу сагледавања могућности за њихову валидну примену у конкретној ситуацији.

✓ У оквиру широке групе *DM* метода (преузетих из других дисциплина и прилагођених *DM* окружењу или иницијално развијених за откривање знања у великим скуповима података), сходно чињеници да практично ниједну фазу процеса откривања знања из података није могуће реализовати без употребе статистике и статистичког начина размишљања, значајна улога припада статистичким методима. При томе, посебно треба истаћи да коришћење статистичких метода у *DM* окружењу не умањује њихова статистичка својства нити их декларише као искључиво *DM* методе. И док је у прошлости између статистичара и *DM* аналитичара постојалао узајамно игнорисање или пак упућивање критика са негативним конотацијама једних на рачун других, у последње време се апострофира потреба њиховог повезивања у функцији не само решавања посматраних проблема и дефинисаних задатака, већ развоја и статистичког и *DM* приступа у анализи података. Успешност у имплементацији *DM*-а у будућности ће критично зависити од прилагођавања статистичких метода *DM* окружењу и

могућности (које су детерминисане, пре свега знањем и умећем истраживача) њихове интеграције у методолошки контекст *DM* приступа. Из другог угла посматрано, *DM* иницира и обезбеђује изузетне могућности за развој нових методолошких решења у домену статистике. Да би се потенцијал њихове интеграције заиста искористио неопходна су обострана прилагођавања уз извесне модификације базичних парадигми и оперативних принципа оба приступа у анализи података, јер као што *DM* неће бити ефективан у откривању знања из података без статистичког размишљања, тако и статистика без елемената *DM*-а неће бити успешна у раду са великим скуповима података.

✓ Примена *DM* метода у конкретној ситуацији, при решавању конкретних задатка, резултира одговарајућим моделом (или, моделима) из података, који има (имају) свој животни циклус, односно трајање. Наиме, са протоком времена од креирања до употребе модела долази до деградирања његових перформанси. Временска зависност и лимитираност модела значи да креирани модел који је изабран у конкретној ситуацији, сходно одговарајућим критеријумима, временом губи на поузданости. Наведено јасно указује на неопходност непрекидног праћења функционалности модела и његовог прилагођавања новим околностима и новим (ажурираним) подацима и упућује на потребу креирање нових и побољшаних верзија модела. При томе, ново моделирање се може базирати на већ примењеној или, пак, некој новој комбинацији метода.

✓ *DM* приступ у анализи података је ефикасан и ефективан методолошки оквир за откривање законитости о појавама, процесима и феноменима у области економије, пословне економије и менаџмента. Полазећи од ове констатације, изведене на основу увида у релевантан теоријски материјал, дефинисане су две конкретне проблемске ситуације и, у складу са тим, спроведена одговарајућа емпиријска истраживања заснована на (а) берзанским подацима у форми временских серија берзанских индекса и (б) анкетним подацима о корисницима услуге:

- У оквиру првог реализованог емпиријског истраживања испитана је и анализирана сличност и извршена класификација одабраних берзи земаља Централне и Југоисточне Европе на основу кретања вредности водећих берзанских индекса. Анализом су обухваћене временске серије дневних података посматране варијабле, укључујући 1800 дана трговања на свакој од одабраних берзи. У методолошком смислу, основу овог истраживања чини процедура интегрисане имплементације *SAX* алгоритма и различитих метода хијерархијског агломеративног груписања, путем које

је, најпре, креиран модел сличности (уз одређивање *MINDIST* мера одстојања, као мера компатибилних са *SAX* приказом временских серија), а затим извршено формирање хомогених група одабраних берзи на основу временских серија референтних берзанских индекса. За статистичку обраду података поред коришћења стандардних статистичких алата, коришћен је и специјално развијен програм за трансформацију нумеричких временских серија у њихове симболичке секвенце, односно *SAX* речи. Посебно се истиче чињеница да су у сваком кораку истраживања, заснованог на *DM* приступу, инкорпорирани елементи статистичке логике и, сходно току подацима вођене анализе, примењени одговарајући статистички критеријуми и методи. Генерално, на основу резултата спроведеног истраживања могуће је констатовати да презентовани методолошки оквир обезбеђује генерисање корисних темпоралних законитости о кретању берзанских индекса, с тим што је исти могуће применити при анализи података који се односе како на друге аспекте функционисања берзи, тако и на високо димензионалне феномене сличних својстава у другим областима.

- У оквиру другог реализованог емпиријског истраживања извршено је мерење, а затим и анализа сатисфакције корисника услуге на основу различитих обележја (варијабли) квалитета услужне понуде у домену ресторатерског пословања, а на примеру једног ресторана. За потребе овог истраживања подаци су обезбеђени анкетирањем случајно одабраних посетилаца ресторана, односно корисника услужне понуде. У методолошком смислу, основу испитивања сатисфакције корисника услуге чини процедура базирана на комбинованој примени *DM* приступа са инкорпорираним статистичким методима и елементима статистичког начина размишљања, а која укључује: ▸ имплементацију *CHAID* алгоритма, којом је постигнута редукција димензионалности података и извршено издвајање оних варијабли квалитета услужне понуде које се налазе у најјачој интеракцији са степеном сатисфакције; ▸ сходно дефинисаном критеријуму, додатно смањење броја варијабли на основу испитивања постојања високе корелације између независних варијабли које су применом *CHAID* алгоритма задржане у моделу; ▸ имплементацију алгоритма *k*-средина, којом су, на основу редукованог подскупа варијабли квалитета услуге, формиране релативно хомогене групе посетилаца ресторана, односно спроведена сегментација тржишта; и ▸ примену метода тестирања статистичких хипотеза за потребе компарације и детаљнијег испитивања ставова испитаника по формираним групама, то јест, тржишним сегментима. Суштински, добијени резултати истраживања су јасно потврдили да је предложени методолошки оквир врло ефикасно средство за

идентификовање законитости скривених у великој количини података о корисницима услуге, а које су релевантне са становишта пословног одлучивања и формулисања праваца деловања у циљу побољшања степена сатисфакције корисника услуге. Наравно, наведена констатација недвосмислено сугерише да се иста методологија успешно може применити и у контексту разматрања сатисфакције корисника услуге у домену других услужних области.

Имајући у виду наведене закључке, који су изведени на основу резултата истраживања теоријског материјала и спроведених оригиналних емпиријских студија, са правом се може констатовати да су потврђене полазне истраживачке хипотезе (*H*-1, *H*-2 и *H*-3), а које су директно повезане са дефинисаним примарним циљем и специфицираним сетом посебних циљева ове дисертације. Поред ових закључака, кључни допринос дисертације односи се на демистификацију феномена *DM*-а и аргументовано образложену и потврђену неопходност и оправданост примене *DM* приступа у економским истраживањима и анализи економских података. Као посебан допринос истраживања реализованих у дисертацији истиче се расветљавање односа између статистике и *DM*-а као научних дисциплина и утврђивање концептуалних сличности и разлике између статистичког и *DM* приступа у анализи података.

На крају, неспорна је чињеница да свако истраживање има извесна ограничења која треба узети у обзир при критичком осврту на резултате актуелног истраживања (у циљу продубљивања изучавања конкретног феномена и разматрања могућности за добијање побољшаних и нових, а не оспоравања постојећих резултата истраживања), али и при конципирању праваца будућих истраживања. С тим у вези, на основу увида у ограничења овог истраживања, у домену будућих истраживачких активности аутора, профилисана су два потенцијална усмерења:

❖ Будући да је *DM* изузетно комплексна истраживачка материја, али и широка научна и апликативна област (која има бројне подобласти), потпуно је јасно да свеоукупни обухват свих релевантних теоријско-прагматичних аспеката није могуће постићи у оквиру једног научно-истраживачког рада у форми докторске дисертације. Питања која су кроз предложену структуру отворена, а кроз садржај дисертације елаборирана, је не само корисно, већ и неопходно продубити, а добијене резултате надограђивати континуираним праћењем промена и изучавањем научне литературе, теоријских расправа и резултата спроведених истраживања у пракси. Заправо, свако питање је могуће издвојити, а самим тим и обухватити као предмет посебних истраживања. Услед наведеног, а сходно афинитетима аутора, као фокус даљих

истраживања издваја се питање инкорпорирања статистике као начина размишљања и рада у *DM* окружење.

❖ У склопу реализованих емпиријских студија, које су интегрални део ове дисертације, потврђени су потенцијал примењивости и корисност *DM* приступа у анализи економских података и моделирању реалних проблемских ситуација у економији, пословној економији и менаџменту, али су истовремено, на основу добијених резултата, сугерисане и будуће смернице у погледу примене и унапређења предложених концептуално-методолошких оквира. Осим ових, већ наведених, предлога, такође, интересантно је осмислити и спровести истраживање које ће указати на могућности и спремност коришћења (у датим околностима) *DM* приступа за анализу података у (свакодневној) пракси српских предузећа и државних институција. Суштински, кроз опсежна анкетна прикупљања релевантних информација, треба установити да ли, изван оквира академских, институтских и агенцијских истраживања, постоје услови за практично коришћење потенцијала *DM* анализе у функцији стицање знања из података и, последично, метафорички речено, постизања видљивих резултата уз помоћ „невидљиве" имовине.

ЛИТЕРАТУРА

Ackoff, R. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16: 3-9.

Aghabozorgi, S. & Teh, Y.W. (2014). Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applications*, 41(4): 1301–1314.

Allee, V. (1997). 12 Principles of Knowledge Management. *Training & Development*, 51(11): 71-74.

Andritsos, P. (2002). Data Clustering Techniques.
Доступно на: *ftp://www.cs.toronto.edu/public_html/public_html/cs/ftp/pub/reports/csri/443/depth.pdf*

Antunes, C. & Oliveira, A. (2001). Temporal Data Mining: An Overview. In: *Proceedings of the Workshop on Temporal Data Mining at the 7th International Conference on Knowledge Discovery and Data Mining*: 1-13. ACM Press.

Arsovski, Z. (2008). *Informacioni sistemi*. Kragujevac: Ekonomski fakultet Univerziteta u Kragujevcu.

Azevedo, A. & Santos, M.F. (2008). KDD, SEMMA and CRISP-DM: A Parallel Overview. У: *Proceedings of the IADIS European Conference Data Mining* 2008: 182-185.

Azzalini, A. & Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*. New York: Oxford University Press, Inc.

Babić, V. (2012). *Uvod u menadžment*. Kragujevac: Ekonomski fakultet Univerziteta u Kragujevcu.

Baicoianu, A. & Dumitrescu, S. (2010). Data Mining Meets Economic Analysis: Opportunities and Challenges. *Bulletin of the Transilvania University of Braşov*, 3(52), Series V: Economic Sciences: 185-192.

Bal, M., Bal, Y. & Demirhan, A. (2011). Creating competitive advantage by using data mining technique as an innovative method for decision making process in business. У: *Proceedings of the Annual Conference on Innovations in Business & Management*, UK.
Доступно на: *https://www.researchgate.net/publication/220449370_Creating_Competitive_Advantage_by_Using_Data_Mining_Technique_as_an_Innovative_Method_for_Decision_Making_Process_in_Business*

Barnet, V. & Lewis, T. (1994). *Outliers in statistical data*. London: John Wiley and Sons.

Batini, C. & Scannapieca, M. (2006). *Data Quality: Concepts, Methodologies, and Techniques*. Berlin Heidelberg: Springer-Verlag.

Batista, G.E.A.P.A. & Monard, M.C. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence*, 17(5/6): 519-533

Ben-Gal, I. (2008). Bayesian Networks. In: Ruggeri, F., Faltin, F. & Kenett, R. (Eds), *Encyclopedia of Statistics in Quality & Reliability*. John Wiley & Sons.
Доступно на: *https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470061572.eqr089*

Ben-Gal, I. (2010). Outlier detection. In: Maimon, O. & Rokach, L. (Eds), *Data mining and knowledge discovery handbook*, 2nd edition: 117-130. New York: Springer Science +Business Media, LLC.

Benjamini, Y. & Leshno M. (2005). Statistical Methods for Data Mining. In: Maimon, O. & Rokach, L. (Eds), *Data mining and knowledge discovery handbook*, 2nd edition: 523-541. New York: Springer Science+Business Media, LLC.

Berry, M. & Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management*, 2nd edition, USA: John Wiley & Sons Ltd.

Bigus, J.P. (1996). *Data Mining with Neural Networks: Solving Business Problems – from Application Development to Decision Support*. USA: McGraw-Hill.

Bole, U., Popovič, A., Žabkar, J., Papa, G. & Jaklič, J. (2015). A case analysis of embryonic data mining success. *International Journal of Information Management*, 35(2): 253-259

Brown, E.D. (2014). *Drowning in data, starved for information*.
Доступно на: *http://ericbrown.com/drowning-in-data-starved-for-information.htm*

Bruh, I. & Famili, A. (2000). Postprocessing in Machine Learning and Data Mining. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA.
Доступно на: *http://www.kdd.org/exploration_files/KDD2000PostWkshp.pdf*

Cavanillas, J.M., Curry, E. & Wahlster, W. (2016). The Big Data Value Opportunity. In: Cavanillas, J.M., Curry, E. & Wahlster, W. (Eds) *New horizonts for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Date in Europe*: 3-11. Switzerland: Springer International Publishing AG.

Chandola, V., Banerjee, A. & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3): 15:1–15:58.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 – Step-by-step data mining guide*. The CRISP-DM (Cross–Industry Standard Process for Data Mining) consortium.
Доступно на: *http://www.crisp-dm.org/CRISPWP-0800.pdf*

Chundi, P. & Rosenkrantz, D. (2009). Segmentation of Time Series Data. In: Wang, J. (Ed), *Encyclopedia of Data Warehousing and Mining*: 1753-1758. New York: Information Science Reference.

Cios, K., Pedrycz, W., Swiniarski, R.W. & Kurgan, L. (2007). *Data mining: A Knowledge Discovery Approach*. New York: Springer Science+Business Media, LLC.

Clark, D. (2004). *Understanding and Performance*.
Доступно на: *http://www.nwlink.com/~donclark/performance/understanding.html*

Чубукова, И.А. (2006). *Data Mining*.
Доступно на: *https://www.intuit.ru/studies/courses/6/6/info*

Dalbelo Bašić, B., Čupić, M. & Šnajder, J. (2008). *Umjetne neuronske mreže*. Zagreb: Fakultet elektrotehnike i računarstva, Zavod za elektroniku, mikroelektroniku, računalne i inteligentne sustave.

Dalbelo-Bašić, B. (2011). Distance Measures. In: Lovrić, M. (Ed) *International Encyclopedia of Statistical Science*: 397-398. Berlin: Springer-Verlag.

Dalkir, K. (2011). *Knowledge Management in Theory and Practice*, 2nd edition. Cambridge, MA: Massachusetts Institute of Technology.

Dash, M. & Liu, H. (1997). Feature Selection for Classification. *Intelligent Data Analysis*, 1(3): 131-156.

Dasu, T. & Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. New York (USA): John Wiley and Sons Ltd.

Daw, C.S., Finney, C.E.A. & Tracy, E.R. (2003). A review of symbolic analysis of experimental data. *Review of Scientific Instruments*, 74(2): 915-930.
Доступно на: *https://pdfs.semanticscholar.org/4060/a718533fcc021b6aa9ad0de7e673af748934.pdf*

Deal, J. (2013). The ten most common data mining business mistakes. *Elder Research Paper*. Доступно на: *https://www.elderresearch.com/company/resource-center/white-papers*.

de Ville, B. (2006). *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*. USA: SAS Institute Inc.

Dugalić, V. & Štimac, M. (2009). *Osnove berzanskog poslovanja*, IV izdanje. Beograd: Stubovi kulture.

Đorđević, V., Lepojević, V. & Janković-Milić, V. (2011). *Primena statističkih metoda u istraživanju tržišta*. Niš: Ekonomski fakultet Univerziteta u Nišu.

Đuričin, D. & Janošević, S. (2009). Strategijska analiza ljudskih resursa. *Ekonomske teme*, XLVII(1): 1-46.

Efron, B. & Tibshirani, R. (1991). Statistical Data Analysis in the Computer Age. *Science*, New Series, 253(5018): 390-395.

English, L. (1999). *Improving Data Warehouse and Business Information Quality*. New York: John Wiley & Sons.

European Commission (EC), (2014). *Communucation from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. Brussels: EC.
Доступно на: *http://ec.europa.eu/environment/eussd/pdf/SustainableBuildingsCommunication.pdf*

Evans, M., Dalkir, K. & Bidian, C. (2014). A Holistic View of the Knowledge Life Cycle: The Knowledge Management Cycle (KMC) Model. *The Electronic Journal of Knowledge Management*, 12(2): 85-97.

Everitt, B.S. (2006). *The Cambridge Dictionary of Statistics*, 3rd edition. UK: Cambridge University Press.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3): 37-54.

Feelders, A.J. (2002). Data Mining in Economic Science. In: Meij, J. (Ed), *Dealing with the data flood*: 166-175. The Hague (Netherlands): STT/Beweton.

Filzmoser, P., Ruiz-Gazan, A. & Thomas-Agnan, C. (2014). Identification of local multivariate outliers. *Statistical Papers*, 55(1): 29-47.

Friedman, J.H. (1997). Data Mining and Statistics: What's the Connection?. *Computing Science and Statistics*, 29(1): 3-9.

Ganesh, S. (2002). Data Mining: Should it be Included in the 'Statistics' Curriculum?. In: *Proceedings of the 6th International Conference on Teaching Statistics*, South Africa.
Доступно на: *http://www.stat.auckland.ac.nz/~iase/publications/1/3l4_gane.pdf*

Gibert, K., Sànchez-Marrè, M. & Codina, V. (2010). Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation. In: *Proceedings of the International Congress on Environmental Modelling and Software*, Ottawa, Canada.
Доступно на:
http://www.iemss.org/iemss2010/papers/S23/S.23.03.Choosing%20the%20Right%20Data%20Mining%20Technique%20-%20Classification%20of%20Methods%20and%20Intelligent%20Recommendation%20-%20MIQUEL%20SANCHEZ-MARRE.pdf

Giudici, P. & Figini, S. (2009). *Applied Data Mining: For Business and Industry*, 2nd edition. Chichester (UK): John Wiley & Sons Ltd.

Giudici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry*. Chichester (UK): John Wiley & Sons Ltd.

Gorunescu, F. (2011). *Data Mining*: *Concepts, Models and Techniques*. Berlin Heidelberg: Springer-Verlag.

Hair, Jr.J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2010). *Multivariate Data Analysis*, 7th edition. NewYork: Pearson Prentice Hall.

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: concepts and techniques*, 3rd edition. Waltham (USA): Morgan Kaufmann Publishers.

Hand, D.J. (1998). Data Mining: Statistics and More?. *The American Statistician*, 52(2): 112-118.

Hand, D.J. (1999a). Why Data Mining is More than Statistics Writ Large. *Bulletin of the International Statistical Institute*, 1: 433-436.

Hand, D.J. (1999б). Statistics and Data Mining: Intersecting Disciplines. *SIGKDD Explorations*, 1(1): 16-19.

Hand, D.J. (2009). Mining the past to determine the future: Problems and possibilities. *International Journal of Forecasting*, 25(3): 441–451.

Hand, D.J., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. UK: MIT Press.

Hand, D.J., Blunt, G., Kelly, M.G. & Adams, N.M. (2000). Data Mining for Fun and Profit. *Statistical Science*, 15(2): 111-131.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. USA: Springer.

Haug, A., Zachariassen, F. & van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, 4(2): 168-193.

Hill, C.M., Malone, L.C. & Trocine, L. (2004). Data Mining and Traditional Regression. In: Bozdogan, H. (Ed), *Statistical Data Mining and Knowledge Discovery*: 223-249. Boca Raton (Florida): CRC Press LLC.

Hodge, V. & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2): 85–126.

Jain, A. K. &. Dubes, R.C. (1988). *Algorithams for Clustering Data*. NJ: Prentice Hall.

Jakšić, M. (2016). *Finansijsko tržište – instrumenti i institucije*. Kragujevac: Ekonomski fakultet Univerziteta u Kragujevcu.

Janošević, S., Senić, R., Stefanović, Ž., Arsovski, Z. & Šolak, Nj. (1999). *Menadžment ukupnog kvaliteta*. Kragujevac (Srbija): Ekonomski fakultet Univerziteta u Kragujevcu.

Jifa, G. (2013), Data, Information, Knowledge, wisdom and meta-synthesis of wisdom-comment on wisdom global and wisdom cities. *Procedia Computer Science*,17:713-719

Jin, R., Breitbart, Y. & Muoh, C. (2009). Data Discretization Unification. *Knowledge and Information Systems*, 19(1): 1-29.

Juran, J.M. & Gryna, F. (1993). *Quality Planning and Analysis*, 3rd edition. New York: McGraw-Hill, Inc.

Kahn, B.K., Strong, D.M. & Wang, R.Y. (2002). Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*, 45(4): 184-192.

Kadane, J.B. (2011). Bayes' Theorem. In: Lovrić, M. (Ed) *International Encyclopedia of Statistical Science*: 89-90. Berlin: Springer-Verlag.

Kalinić, Z. & Marinković, V. (2017). Određivanje relativnog uticaja pojedinih faktora na prihvatanje mobilne trgovine primenom neuronskih mreža. *Poslovna ekonomija*, X(II): 206-223.

Kamel, M. (2009). Data Preparation for Data Mining. In: Wang, Y. (Ed), *Encyclopedia of Data Warehousing and Mining*, 2nd edition: 538 -543. New York: Information Science Reference (IGI Global).

Kantardžić, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms,* 2nd edition. New Jersey: John Wiley & Sons, Inc.

Kennedy, J. (2014). *The data driven economy*.
Доступно на: *https://www.uschamberfoundation.org/sites/default/files/Joe%20Kennedy%20Article.pdf*

Keogh, E. (2011). Data Mining Time Series Data. In: Lovrić, M. (Ed), *International Encyclopedia of Statistical Science*: 339-342. Berlin: Springer - Verlag

Keogh, E,. Chakrabarti, K,. Pazzani, M. & Mehrotra, S. (2000). Dimensionality reduction for fast similarity search in large time series databases. *Journal of Knowledge and Information Systems*, 3(3): 263–286.

Keogh, E. & Kasetty, S. (2003). On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery*, 7(4): 349–371.

Khabaza, T. (2010). *Nine Laws of Data Mining*.
Доступно на: *http://khabaza.codimension.net/index.htm*

Klepac, G. & Mršić, L. (2006). *Poslovna Inteligencija kroz poslovne slučajeve*. Zagreb: Lider Press & TIM Press.

Knobbe, A., Crémilleux, B., Fürnkranz, J. & Scholz, M. (2008). From Local Patterns to Global Models: The LeGo Approach to Data Mining. In: *Proceedings of the ECML/PKDD-08 Workshop* (*LeGo*-08), Antwerp, Belgium.
Доступно на: *https://www.kiminkii.com/publications/lego.pdf*

Kontaki, M., Papadopoulos, A. & Manolopoulos, Y. (2005). Similarity Search in Time Series Databases.
Доступно на: *http://delab.csd.auth.gr/papers/IDEA05kpm.pdf*

Kovačić, Z. (1994). *Multivarijaciona analiza*. Beograd: Ekonomski fakultet Univerziteta u Beogradu.

Kumar, V. & Minz, S. (2014). Feature Selection: A literature Review. *Smart Computing Review*, 4(3): 211-229.

Kuonen, D. (2005). Is Data Mining for Gold 'Statistical déjà vu'?. *CRM Zine*, 53.
Доступно на: *http://www.statoo.com/en/publications/articleDMStat4CRMToday.pdf*

Kurgan, L.A. & Musilek P. (2006). A Survey of Knowledge Discovery and Data Mining Process Models. *The Knowledge Engineering Review*, 21(1): 1-24.

Lallich, S., Teytaud, O. & Prudhomme, E. (2006). Statistical inference and data mining: false discoveries control. In: Rizzi, A. & Vichi, M. (Eds), *Compstat 2006 – Proceedings in Computational statistics*: 325-336. Physica-Verlag HD.

Larose, D.T. (2005). *Discovering knowledge in data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, Inc.

Lee, Y.W., Pipino, L.L., Funk, J.D. & Wang, R.J. (2006). *Journey to Data Quality*. Cambridge (UK): MIT Press.

Lepojević, V & Janković-Milić, V. (2008). Using Data Mining Techniques in Market Research. *Economic themes*, XLVI (4): 101-115.

Liew, A. (2007). Understanding Data, Information, Knowledge and their Inter-Relationships. *Journal of Knowledge Management Practice*, 8(2).
Доступно на: *http://www.tlainc.com/articl134.htm*

Liew, A. (2013). DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. *Business Management Dynamics*, 2(10): 49-62.

Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of the* 8*th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*: 2–11.

Lin, J., Keogh, E., Wei, L. & Lonardi, S. (2007). Experiencing SAX: a Novel Symbolic Representation of Time Series. *Data Mining and Knowledge Discovery*, 15(2):107-144.

Lin, J. & Li, Y. (2009). Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation. In: *Proceedings of the International Conference of Scientific and Statistical Database Management*: 461-477. Berlin: Springer.

Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. In: *Proceedings of the 10th IEEE International Conference on Data Mining*: 911-916.

Liu, Q. (2014). *The Application of Exploratory Data Analysis in Auditing*. PhD Dissertation. Newark (New Jersey): Rutgers, The State University of New Jersey.

Liu, H., Hussain, F. & Tan, C.L. (2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery*, 6(4): 393–423.

Loshin, D. (2011). Evaluating the Business Impacts of Poor Data Quality. *Information Quality Journal*.
Доступно на: *http://dataqualitybook.com/kii-content/BusinessImpactsPoorDataQuality.pdf*

Lovrić, M. (2009). *Osnovi statistike*. Kragujevac: Ekonomski fakultet Univerziteta u Kragujevcu.

Lovrić, M., Komić, J. & Stević, S. (2017). *Statistička analiza: Metodi i primjena,* II izdanje. Banja Luka: JU Narodna i univerzitetska biblioteka Republike Srpske.

Lovrić, M., Milanović, M. & Stamenković, M. (2012). Time series data mining: similarity search and its application to the stock indices in the region. *TTEM*, 7(4): 1605-1614.

Lukić, J. (2013). „Istraživač podataka"– zanimanje za 21. vek. У: *Proceedings of the IX Symposium of Business and Science SPIN'*13 *– New Industralization, Reengineering and Sustainability*: 325–332. Beograd: Fakultet organizacionih nauka.

Lyman, P. & Varian, H.R. (2003). *How much Information*. Technical report, UC Berkeley.
Доступно на: *http://www.sims.berkeley.edu/research/projects/how-much-info-2003/*

Magnani, M. (2004). Techniques for Dealing with Missing Data in Knowledge Discovery Tasks.
Доступно на: *http://magnanim.web.cs.unibo.it/index.html*

Maletic, J.I. & Marcus, A. (2010). Data Cleansing: A Prelude to Knowledge Discovery. In: Maimon, O. & Rokach, L. (Eds), *Data mining and knowledge discovery handbook*, 2nd edition: 19-31. New York: Springer Science+Business Media, LLC.

Maimon, O. & Rokach, L. (2010). Introduction to Knowledge Discovery and Data Mining. In: Maimon, O. & Rokach, L. (Eds), *Data mining and knowledge discovery handbook*, 2nd edition: 1-14. New York: Springer Science+Business Media, LLC.

Mariscal, M., Marban, O. & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2): 137–166.

Marr, B. (2015). *Big Data–Using smart big data, analytics and metrics to make better decisions and improve performance*. UK: John Wiley & Sons.

Marsh, R. (2005). Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management. *Database Marketing & Customer Strategy Management*, 12(2): 105–112.

McAfee, A. & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90(10): 60-68.

McKinsey Global Institute (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company.
Доступно на: *https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation*

Mikut, R. & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery*, 1(5): 431-443.

Milanović, M. & Stamenković, M. (2011a). Data Mining in Time Series. *Economic Horizons*, 13(1): 5-25.

Milanović, M. & Stamenković, M. (2011б). Istraživanje sličnosti u dubinskoj analizi vremenskih serija. У: *Zbornik radova XXXVIII Simpozijuma o operacionim istraživanjima* (*SYM-OP-IS* 2011): 335-338. Zlatibor: Ekonomski fakultet u Beogradu.

Milanović, M., Stamenković, M. & Đurić, Z. (2012). Dimensionality Reduction of Time Series Data based on SAX Representation. In: *Proceedings of the 2nd International Scientific Conference "Contemporary Issues in Economics, Business and Management – EBM 2012"*: 629-641. Kragujevac: Faculty of Economics, University of Kragujevac.

Mörchen, F. (2006). *Time Series Knowledge Mining*. PhD Dissertation. Marburg (Germany): Philipps-University Marburg.
Доступно на: *http://www.mybytes.de/papers/moerchen06tskm.pdf*

Motoda, H. & Liu, H. (2002). Feature Selection, Extraction and Construction. In: *Proceedings of the Communication of IICM* (*Institute of Information and Computing Machinery*, Taiwan), Volume 5: 67-72.
Доступно на: *http://www.ar.sanken.osaka-u.ac.jp/~motoda/papers/fdws02.pdf*

Myatt, G. & Johnson, W. P. (2014). *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, 2nd ed. New Jersey: John Wiley & Sons.

Nanda, S.R. Mahantly, B. & Tiwari, M.K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, 37: 8793-8798.

Neha D. & Vidyavathi, B.M. (2015). A Survey on Applications of Data Mining using Clustering Techniques. *International Journal of Computer Applications*, 126(2): 7-12.

Nisbet, R., Elder, J. & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. USA: Academic Press (Elsevier Inc.).

Organisation for Economic Co-operation and Development (OECD), (1996). *The Knowledge-based Economy*. Paris: OECD.
Доступно на: *https://www.oecd.org/sti/sci-tech/1913021.pdf*

Organisation for Economic Co-operation and Development (OECD), (2013). Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data". *OECD Digital Economy Papers*, No. 222, Paris: OECD Publishing.

Panian, Ž. & Klepac, G. (2003). *Poslovna inteligencija*. Zagreb: Masmedia.

Panian, Ž., Pejić Bach, M., Mršić, L., Brešić, B., Kockar, I., Jaković, B., Obradović, M., Kanižaj, T., Žmirak, Z., Oreščanin, D. & Karaga, L. (2007). *Poslovna inteligencija: Studije slučajeva iz hrvatske prakse*. Zagreb: Narodne novine.

Parasuraman, A., Zeithaml, V.A. & Berry, L.L. (1988). Servqual: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1): 12-40.

Peker, S., Aktan, B. & Tvaronavičienė, M. (2017). Clustering in key G-7 stock market indices: an innovative approach. *Маркетинг і менеджмент інновацій*, (1): 300-310.

Pipino, L. L., Lee, Y. W. & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4): 211-218.

Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco, California, (USA): Morgan Kaufmann Publishers.

Ramzan, S, Zahid, M.F. & Ramzan, S. (2013). Evaluating Multivariate Normality: A Graphical Approach. *Middle-East Journal of Scientific Research*, 13(2): 254-263.

Ratanamahatana, C.A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M. & Das, G. (2010). Mining time series data. In: Maimon, O. & Rokach, L. (Eds), *Data mining and knowledge discovery handbook*, 2nd edition: 1056-1057. New York: Springer.

Redman, T.C. (2001). *Data quality: the field guide*. Boston: Digital Press.

Redman T.C. (2013). Data's credibility problem. *Harvard Business Review*, 91(12): 84-88.

Rendón, E, Abundez, I, Arizmendi, A. & Quiroz, E.M. (2011). Internal versus External Cluster Validation Indexes. *International Journal of Computers and Communications*, 5(1): 27-34.

Roberts, J. & Armitage, J. (2008). The ignorance economy. *Prometheus:Critical Studies in Innovation*, 26(4): 335-354.

Rockart, J.F. (1979). Chief executives define their own needs. *Harvard Business Review*, 57(2): 81–93.

Rokach, L. & Maimon, O. (2008). *Data mining with decision trees:theory and applications*. Singapore: World Scientific Publishing Co.

Rokach, L. & Maimon, O. (2010). Classification Trees. In: Maimon, O. & Rokach, L. (Eds), *Data mining and knowledge discovery handbook*, 2nd edition: 157-174. NewYork: Springer Science+Business Media, LLC.

Romesburg, H.C. (2011). Cluster Analysis: An Introduction. In: Lovrić, M. (Ed) *International Encyclopedia of Statistical Science*: 262-265. Berlin: Springer-Verlag.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53-65.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 33(2): 163–180.

Ryu, K. (2005). *DINESCAPE, emotions and behavioral intentions in upscale restaurants*. PhD dissertation. Kansas: Kansas State University.

Salomon, D. & Motta, G. (2010). *Handbook of Data Compression*, 5th edition. NY: Springer.

Sant' Anna, A. & Wickström, N. (2011). Symbolization of time-series: An evaluation of SAX, Persist, and ACA. In: *Proceedings of the IEEE 4th International Congress on Image and Signal Processing*: 2223–2228.

Sapra, S. (2014). A Useful Role for Data Mining in Economics. *Business and Economics Journal*, 5(3): editorial.

Scarpa, B. (2011). Data Mining. In: Lovrić, M. (Ed) *International Encyclopedia of Statistical Science*: 336-338. Berlin: Springer-Verlag.

Senić, R. (2000). *Marketing menadžement*, III izmenjeno i dopunjeno izdanje. Kragujevac (Srbija): Ekonomski fakultet Univerziteta u Kragujevcu.

Senić, R. & Senić, V. (2008). *Menadžement i marketing usluga*. Kragujevac: sopstveno izdanje.

Sharma, S. (1996). *Applied Multivariate Techniques*. NewYork: John Wiley & Sons, Inc.

Shmueli, G., Patel, N.R. & Bruce, P.C. (2005). *Data Mining In Excel: Lecture Notes and Cases*. USA: Resampling Stats, Inc.

Shmueli, G., Patel, N.R., & Bruce, P.C. (2010). *Data Mining For Business Intelligence Concepts, Techniques, and Applications in Microsoft Office Excel with Xlminer*. Hoboken (New Jersey): John Wiley and Sons, Inc.

Shmueli, G. (2010). To explain or to predict?. *Statistical Science*, 25(3): 289-310.

Sim, J. (2014). Consolidation of Success Factors in Data Mining Projects. *GSTF Journal on Computing* (*JoC*), 4(1): 66-73.

Славковић, М. (2013). *Стратегијско управљање људским ресурсима у економији заснованој на знању*. Докторска дисертација. Крагујевац: Економски факултет Универзитета у Крагујевцу.

Smyth, P. (2001). Data Mining at the Interface of Computer Science and Statistics. In: Grossman, R. L., et al. (Eds.), *Data Mining for Scientific and Engineering Applications*: 35-61. *KIllwer Academic Publishers.*
Доступно на: *https://link.springer.com/chapter/10.1007/978-1-4615-1733-7_3*

Soldić-Aleksić, J. (2004). Statistika i data mining: sličnosti i razlike. *Statistička revija*, LIII (1-4): 40-51.

Soldić-Aleksić, J. (2009). Prediktivni model segmentacije tržišta: primena modela logističke regresije i CHAID procedure. *Marketing*, 40(3): 129-138

Soldić-Aleksić, J. (2013). Problem nedostajućih podataka u istraživanju podataka. У: *Zbornik radova XL simpozijuma o operacionim istraživanjima* (*SYM-OP-IS* 2013): 474-479. Zlatibor: Fakultet organizacionih nauka Univerziteta u Beogradu.

Soldić-Aleksić, J. (2015). *Primenjena analiza podataka*. Beograd: Ekonomski fakultet Univerziteta u Beogradu

Soldić-Aleksić, J. & Chroneos Krasavac, B. (2009). *Kvantitativne tehnike u istraživanju tržišta: Primena SPSS računarskog paketa*. Beograd: Ekonomski fakultet Univerziteta u Beogradu.

Солдић-Алексић, J. & Chroneos Красавац, Б. (2016). Савремени аспекти дигиталне економије: утицај феномена *big data*. У: *Зборник радова XLIII Симпозијума о операционим истраживањима* (*SYM-OP-IS* 2016): 237-241. Тара: Министарство одбране и Војска Србије.

Soon, L.-K. & Lee, S.H. (2007). An Empirical Study of Similarity Search in Stock Data. In: *Proceedings of the 2nd International workshop on Integrating Artificial Intelligence and Data Mining – AIDM'07*: 31-38. Australia: Australian Computer Society.

Stamenković, M. & Milanović, M. (2014). Outlier detection in function of quality improvement of business decisions. In: *Proceedings of the International scientific conference–Enterprises in hardship: economics, managerial and juridical perspectives*: 173-184. Messina: Faculty of Economics University of Messina.

Stamenković, M., Milanović, M. & Mimović, P. (2012). Simbolički prikaz vremenskih serija: *SAX* pristup. U: *Zbornik radova XXXIX Simpozijuma o operacionim istraživanjima SYM-OP-IS 2012*: 19-22. Tara: Visoka građevinsko-geodetska škola, Beograd.

Stevens P., Knutson, B. & Patton, M. (1995). Dineserv: a tool for measuring service quality in restaurants. *Cornell Hotel & Restaurant Administration Quarterly*, 36(2): 56-60.

Straf, M.L. (2003). Statistics: The Next Generation. *Journal of the American Statistical Association*, 98(461): 1-6.

Tan, Q., Oriade, A. & Fallon, P. (2014). Service quality and customer satisfaction in chinese fast food sector: a proposal for CFFRSERV. *Advances in Hospitality and Tourism Research* (*AHTR*), 2(1): 30-53.

Thearling, K. (2003). An Introduction to Data Mining. *White Paper*.
Доступно на: *http://akira.ruc.dk/~bulskov/undervisning/E2003/data_mining.pdf*

The Economist (2010). *Data, data everywhere: A special report on managing information*. USA: The Economist Newspaper Ltd.
Доступно на: *https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf*

Tufféry, S. (2011). *Data Mining and Statistics for Decision Making*. Chichester (UK): John Wiley & Sons Ltd.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading (MA): Addison-Wesley.

Uriarte, Jr. F.A. (2008). *Introduction to Knowledge Management: A Brief Introduction to the Basic Elements of Knowledge Management for Non-practitioners Interested in Understanding the Subject*, Jakarta (Indonesia): ASEAN Foundation.

Vercellis, C. (2009). *Business intelligence: Data Mining and Optimization for Decision Making.* Chichester (UK): John Wiley & Sons Ltd.

Vlachos, M., Gunopulos, D. & Das, G. (2004). Indexing Time Series Under Conditions of Noise. In: Last, M., Kandel, A. & Bunke, H. (Eds), *Data Mining in Time Series Databases*: 67-100. Singapore: World Scientific Publishing Co. Ltd.

Wang, J. (2003). *Data Mining Challenges and Opportunities*, London: IRM Press.

Wang, R. & Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4): 5-34.

Wei, L., Keogh, E., Xi, X. & Yoder, M. (2008). Efficiently finding unusual shapes in large image databases. *Data Mining & Knowledge Discovery*, 17(3): 343-376.

Wendler, T. & Gröttrup, S. (2016). *Data Mining with SPSS Modeler, Theory, Exercises and Solutions*. Switzerland: Springer International Publishing.

Wilson, A., Zeithaml, V.A., Bitner, M.J. & Gremler, D.D. (2008). *Services Marketing*. NewYork: McGraw-Hill Education

Witten, I.H., Frank, E. & Hall, M.A. (2011). *Data Mining*, 3rd edition. Boston: Morgan Kaufmann Publishers (Elsevier, Inc.)

Wu, K.-P., Wu, Y.-P. & Lee, H.-M. (2014). Stock Trend Prediction by Using *K*-Means and AprioriAll Algorithm for Sequential Chart Pattern Mining. *Journal of Information Science and Engineering*, 30(3): 653-667.

Yang, Y., Webb, G.I. & Wu, X. (2010). Discretization Methods. In: Maimon, O. & Rokach, L. (Eds), *Data mining and knowledge discovery handbook*, 2nd edition: 101-116. New York: Springer Science+Business Media, LLC.

Yeoh, W. & Koronios, A. (2010). Critical Success Factors for Business Intelligence Systems. *Journal of Computer Information Systems*, 50(3): 23-32

Yi, B.K. & Faloutsos, C. (2000). Fast Time Sequence Indexing for Arbitrary Lp Norms. In: *Proceedings of the 26st International Conference on Very Large Databases*: 385–394.

Yu C.H. (2010). Exploratory Data Analysis in the Context of Data Mining and Resampling. *International Journal of Psychological Research*, 3(1): 9-22.

Zekić-Sušac, M., Frajman-Jakšić, A. & Drvenkar, N. (2009). Neuronske mreže i stabla odlučivanja za predviđanje uspješnosti studiranja. *Ekonomski vjesnik: Review of Contemporary Entrepreneurship, Business, and Economic Issues*, XXII(2): 314-327.

Zhang, S., Zhang, C. & Yang, Q. (2003). Data Preparation for Data Mining. *Applied Artificial Intelligence*, 17(5/6): 375-381.

Zhou, Z.-H. (2003). Three perspectives of data mining. *Artificial Intelligence*, 143(1): 139-146.

Zhu, X., Wu, X. & Chen, Q. (2006). Bridging Local and Global Data Cleansing: Identifying Class Noise in Large, Distributed Data Datasets. *Data Mining and Knowledge Discovery*, 12(2-3): 275-308.

Интернет извори:

http://www.gartner.com/it-glossary/data-mining
http://www.kdnuggets.com
http://www.kdnuggets.com/faq/classification-vs-prediction.html
https://unstats.un.org/unsd/wsd/News3.aspx
http://www.itl.nist.gov/div898/handbook/eda/eda.htm
http://data.worldbank.org/
http://web.worldbank.org/archive/website01503/WEB/0__CO-10.HTM
http://www.theglobaleconomy.com
http://zse.hr/
http://www.belex.rs/

http://www.mnse.me/
http://www.sase.ba/
https://www.blberza.com/
http://www.mse.mk/
http://www.ljse.si/
http://www.bvb.ro/
http://www.bse-sofia.bg/
http://www.bsse.sk/
http://www.bse.hu/
http://www.gpw.pl/
http://www.pse.cz/
http://www.bse.hu/

П Р Е Г Л Е Д С Л И К А

П Р Е Г Л Е Д Т А Б Е Л А

БИОГРАФИЈА АУТОРА

Марина (Будимир) Милановић је рођена 08.06.1967. године у Аранђеловцу, где је завршила основну и средњу школу. Дипломирала је на Економском факултету Универзитета у Крагујевцу са просечном оценом 9,29. Награђена је од стране Универзитета за постигнути успех у току студија.

Последипломске студије, магистарска група Управљање предузећем–принципи и пракса, смер Квантитативни методи, на Економском факултету у Крагујевцу, завршила је са просечном оценом 9,88. Новембра 1999. године је стекла звање Магистар економских наука. Докторске академске студије уписала је на Економском факултету Универзитета у Нишу.

Након завршетка основних студија радила је 18 месеци у привреди, а затим, као професор предметне наставе у средњој економској школи у Аранђеловцу. Радни однос на Економском факултету Универзитета у Крагујевцу, као асистент-приправник на предмету Статистика у економији и менаџменту, засновала је 1995. године. У звање асистент на наставном предмету Основи статистике, изабрана је у мају 2000. године. Одлуком Наставно-научног већа, изводила је вежбе и на наставном предмету Управљање квалитетом. На наставном предмету Основи статистике била је ангажована до јула 2013. године. Током наставно-педагошког рада, у Анкетама студената оцењивана је највишим оценама у свим облицима наставе. Од јула 2013. године, на Економском факултету Универзитета у Крагујевцу обавља послове у оквиру Библиотечке службе и Службе за информационо комуникациону подршку.

Као истраживач, била је ангажована на Интерним пројектима Факултета. У периоду од 2002. до 2005. године учествовала је у реализацији пројекта *Развој корпоративног управљања у условима транзиције*, који је финансиран од стране Министарства за науку, технологије и развој Владе Републике Србије.

Објавила је већи број радова, као аутор или коаутор, у међународним и националним часописима и зборницима радова са научних домаћих и међународних конференција. У свом раду користи стандардне рачунарске алате (*Microsoft Office*), као и специјализоване пакете за обраду и презентацију статистичких података (*EduStat, IBM SPSS, STATA*).

Универзитет у Нишу
Економски факултет

ИЗЈАВА О АУТОРСТВУ

Изјављујем да је докторска дисертација, под насловом ИЗВОЂЕЊЕ ЗАКОНИТОСТИ ИЗ ЕКОНОМСКИХ ПОДАТАКА ПРИМЕНОМ *DATA MINING* ПРИСТУПА, која је одбрањена на Економском факултету Универзитета у Нишу:

- резултат сопственог истраживачког рада;
- да ову дисертацију, ни у целини, нити у деловима, нисам пријављивала на другим факултетима, нити универзитетима;
- да нисам повредила ауторска права, нити злоупотребила интелектуалну својину других лица.

Дозвољавам да се објаве моји лични подаци, који су у вези са ауторством и добијањем академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада, и то у каталогу Библиотеке, Дигиталном репозиторијуму Универзитета у Нишу, као и у публикацијама Универзитета у Нишу.

У Нишу, 06. септембра 2018. године

Аутор дисертације: Мр Марина Милановић

Потпис аутора дисертације _____

Универзитет у Нишу
Економски факултет

ИЗЈАВА О ИСТОВЕТНОСТИ ШТАМПАНОГ И ЕЛЕКТРОНСКОГ ОБЛИКА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Име и презиме аутора: Марина Милановић

Наслов дисертације: ИЗВОЂЕЊЕ ЗАКОНИТОСТИ ИЗ ЕКОНОМСКИХ ПОДАТАКА ПРИМЕНОМ *DATA MINING* ПРИСТУПА

Ментор: Проф. др Винко Лепојевић

Изјављујем да је штампани облик моје докторске дисертације истоветан електронском облику, који сам предала за уношење у Дигитални репозиторијум Универзитета у Нишу.

У Нишу, 06. септембра 2018. године

Потпис аутора дисертације _____

Универзитет у Нишу
Економски факултет

ИЗЈАВА О КОРИШЋЕЊУ

Овлашћујем Универзитетску библиотеку „Никола Тесла“ да, у Дигитални репозиторијум Универзитета у Нишу, унесе моју докторску дисертацију, под насловом: ИЗВОЂЕЊЕ ЗАКОНИТОСТИ ИЗ ЕКОНОМСКИХ ПОДАТАКА ПРИМЕНОМ *DATA MINING* ПРИСТУПА.

Дисертацију са свим прилозима предала сам у електронском облику, погодном за трајно архивирање.

Моју докторску дисертацију, унету у Дигитални репозиторијум Универзитета у Нишу, могу користити сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons), за коју сам се одлучила.

1. Ауторство **(CC BY)**
2. Ауторство – некомерцијално **(CC BY-NC)**
3. Ауторство – некомерцијално – без прераде **(CC BY-NC-ND)**
4. Ауторство – некомерцијално – делити под истим условима **(CC BY-NC-SA)**
5. Ауторство – без прераде **(CC BY-ND)**
6. Ауторство – делити под истим условима **(CC BY-SA)**

У Нишу, 06. септембра 2018. године

Аутор дисертације: Мр Марина Милановић

Потпис аутора дисертације