УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ" ШТИП

ФАКУЛТЕТ ЗА ИНФОРМАТИКА

ШТИП

БИЛЈАНА ТЕОХАРЕВА-ФИЛИПОВА

ТЕХНИКИ НА ПОДАТОЧНО РУДАРЕЊЕ КАКО ПОДДРШКА НА БИЗНИС ОДЛУКИ

-МАГИСТЕРСКИ ТРУД-

Штип, мај 2012

Комисија за оценка и одбрана

Ментор: **Цвета Мартиновска**
Вон.проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Татјана Атанасова-Пачемска**
Вон.проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Александра Милева**
Доц. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Членови на комисија за оценка и одбрана

Претседател: **Татјана Атанасова-Пачемска**
Вон.проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Цвета Мартиновска-** ментор
Вон.проф. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Член: **Александра Милева**
Доц. д-р, Факултет за информатика
Универзитет „Гоце Делчев“ - Штип

Научно поле: **Информатика**

Научна област: **Информациони системи и технологии**

Дата на одбрана:_____

Дата на промоција:_____

ПОСВЕТА

Овој магистерски труд го посветувам на моите најмили**:**

Иван, Јордан и Павел

Тие се моето најголемо богатство и светлина која ме води низ патот по кој чекорам. За нив вреди да се чекори понатаму...

Публикувани трудови

Biljana Teohareva Filipova and Cveta Martinovska. (2012). Analysing Customer Profiles using Data Mining Techniques , 34th International Conference on INFORMATION TECHNOLOGY INTERFACES
ITI 2012
June 25-28, 2012, Cavtat / Dubrovnik, Croatia, **http://iti.srce.unizg.hr**

Билјана Теохарева- Филипова, Цвета Мартиновска. (2012). Техники и алатки на податочното рударење како поддршка на процесот на донесување бизнис одлуки, Годишен зборник на трудови на Економски факултет, Универзитет „Гоце Делчев“ - Штип

НАСЛОВ НА ТРУДОТ

ТЕХНИКИ НА ПОДАТОЧНО РУДАРЕЊЕ КАКО ПОДДРШКА НА БИЗНИС ОДЛУКИ

АПСТРАКТ

Податочно рударење (data mining) е процес на пронаоѓање на скриени законитости и врски меѓу податоците. Рударењето на податоци во базите може да резултира со зголемена точност на моменталните анализи. Добиените податоци во процесот на податочно рударење ја зголемуваат ефективноста и развојот на бизнис секторот и помагаат во донесување на правилни, проактивни одлуки од кои ќе зависи понатамошниот развој на некоја компанија.

Во овој магистерски труд се имплементираат техники на податочно рударење со цел да се користат како поддршка на одредени бизнис одлуки. Врз податоците од анкетни листови и од фискални сметки на купувачите во една македонска фирма применети се методи на дрва на одлучување, невронски мрежи, регресија, Баесови мрежи, асоцијативни правила, метод на потрошувачка кошница и кластерирање. Податоците се обработени со програмските алатки SPSS и Clementine. Со примена на техниките на податочно рударење се добиваат одговори на која категорија купувачи треба да се издаде карта на лојалност, кои фактори влијаат на купување производи со пониска цена, кои артикли се купуваат заедно, кои артикли најчесто се купуваат, кои артикли најчесто купувачите би сакале да бидат на акција и др.

По издавањето на карта на лојалност извршена е анализа на продажбата, која ќе покаже дека профитот на фирмата се зголемува и се утврдува колкав процент од тој профит припаѓа на лојалните купувачи. Врз основа на продажбата од претходниот период направено е предвидување на движењето на продажбата во следниот период.

Клучни зборови: бизнис интелигенција, складови на податоци, податочно рударење.

TITLE

DATA MINING TECHNIQUES AS A SUPPORT TO BUSINESS DECISION

ABSTRACT

Data mining is a process of finding hidden regularities and connections among data. Base data mining can result in an increased accuracy of the current analyses. Data received in the process of data mining increase the efficiency and development of the business sector and help in bringing right, pro-active decisions which will influence the further development of a particular company.

This Master's work implements data mining techniques in order to be used as a support to particular business decisions. Methods of decision trees, neuron networks, regression, Bayesian networks, associative rules and method of a spending basket were applied over data received from questionnaire leaves and fiscal accounts of customers in a Macedonian firm. These data were elaborated with the program tools SPSS and Clementine. Applying the data mining techniques brings answers to questions such as: what category of customers should be given loyalty cards, which factors influence on buying lower price products, which products are being bought together, which products are the most often being bought, which products would the customers like to be on an action etc.

After issuing the loyalty cards, analysis of the sale has been done which shows that the firm profit is increased and it has been found what percentage of this profit belongs to loyal customers. A projection of the selling in the next period has been made on the basis of the sale in the previous period.

.

Key words: business intelligence, data storage, data mining.

СОДРЖИНА

1. ВОВЕД

Во ерата на информации, мотивирани од развојот на подобро чување на податоци и техники за пронаоѓање на податоците, компаниите почнуваат да обезбедуваат повеќе информации за самите нив и нивните активности. Добиените информации се користат во поддршка на одлуки, истражувања и подобро разбирање на феноменот генерирање на податоци. Брзото и ефективно претворање на податоците во разбирливи информации дава добра можност за донесување правилни бизнис одлуки од кои понатаму ќе зависи успехот на некоја компанија. Кога потребните информации не се претставени разбирливо, или пак не се достапни, може да се донесат одлуки кои негативно ќе влијаат во понатамошниот развој на бизнис секторот.

Изразениот процес на глобализација со себе носи неизвесност, ризик и конкуренција. За да опстанат во современиот начин на водење на бизнис, компаниите мора секојдневно да се борат за оддржување на пазарот и остварување подобри бизнис резултати. Големата конкуренција на пазарот и современо развиените дистрибуирачки канали не дозволуваат да се остане „просечен“ туку мора да се тежнее кон самиот врв. За да го постигнат тоа менаџерите на компаниите секогаш мора да бидат чекор пред конкуренцијата, т.е. да ги предвидат потребите на своите клиенти.

Нов вид на технологија, чија цел е токму решавањето на проблемите со кои се соочуваат фирмите е ***Бизнис интелигенцијата (Business intelligence-BI).*** Бизнис интелигенцијата претставува широко множество од апликации и технологии за прибирање на податоците, лесен и брз пристап до истите, а се со цел да обезбеди адекватна поддршка во процесот на донесување одлуки. Поимот Бизнис интелигенција ги обединува методологиите, технологиите и платформите за *складирање на податоци (Data warehouse)*, *online аналитичка обработка на податоци (OLAP - Online Analytical Processing) и Податочно рударење (Data mining).* Податочното рударење е најважната компонента од фамилијата на бизнис интелигенцијата чија цел е пронаоѓање на скриените законитости и врски меѓу податоците и трансформација на истите во корисно знаење.

Работејќи со бази на податоци, менаџерите трошат значително време барајќи податоци од внатрешни и надворешни извори за да дојдат до бараните информации. Во минатото организациите дизајнирале и имплементирале

извршен информационен систем (EIS) и системи за поддршка на одлуки (DSS) за да обезбедат интегрирани презентации од разделени и неконзистенти податоци кои често се филтрираат од оперативните системи. Иако информациите понудени од овие системи доста често се значајни, интегрирани и корисни, менаџерите почнале да сфаќаат дека примената на техниките на податочно рударење врз поголеми количини податоци во базите би можело да резултира со повеќе точни информации и зголемена точност на моменталните анализи.

Во денешно време со интензивниот развој на информатичката инфраструктура на секоја фирма, а особено на оние поголемите, несомнено им се зголемува и потребата од чувањето на податоците со кои работат фирмите, нивните клиенти, добавувачи, остварениот приход и др. Најчесто се врши дневен влез на овие податоци и истите влегуваат во базата на податоци на самата фирма.

Информациите кои се добиени од различни извори може да бидат атрибутивни или нумерички и се однесуваат на факторите кои влијаат на работата на фирмата на потрошувачите, добавувачите, па дури и на конкуренцијата. Меѓутоа, ваквите податоци неадекватно структуирани и во различни формати, немаат некоја голема употреблива вредност. Неопходно е истите да се подготват, анализираат и врз основа на тоа да се дојде до информации кои на фирмата ќе можат да и обезбедат финансиски профит и успешност во работењето.

Со оглед на тоа што се работи за големи количини на податоци, невозможно е еден човек сам да ги извршува тие анализи. За тоа се направени различни софтерски пакети и алатки со чија помош се доаѓа на многу побрз и полесен начин до бараните одговори.

Целта на изработката на овој магистерски труд е да се нагласи и прикаже предноста од примената на модерните менаџерски алатки и техники, конкретно, предноста од користењето на техниките на податочното рударење во поддршката на бизнис одлуки. Затоа најпрво ќе биде даден еден општ теориски осврт на сè она што ќе се користи во овој труд, а потоа ќе се изврши избор на алатките за обработка на податоците кои се добиени и избор на соодветните техники на податочното рударење. Со помош на овие алатки и техники за обработка на податоците ќе се очекува да се донесат правилни

одлуки со кои би се подобрила работата на избраната фирма која се занимава со трговија на мало.

Со спроведувањето на анализите и добивањето на резултатите од истите, би требало да се потврди користа и предноста од имплементацијата на техниките на податочното рударење во работењето на некоја фирма, техники кои многу ретко се користат на македонскиот пазар, а доста често се употребуваат во САД и некои земји од Европа.

Во обработката на оваа тема методите кои ќе се користат се: квалитативна и квантитативна анализа, дескриптивна анализа, компаративен метод и др. Готовите и прибраните податоци ќе бидат прикажани табеларно и графички и со помош на бројни показатели.

1.1 Структура на магистерскиот труд

Во поглавјето 1 е даден воведен дел како опис на проблематиката која ќе се разработува.

Во поглавјето 2 е даден теоретски осврт на основниот концепт на бизнис интелигенцијата и нејзината примена во процесот на донесување бизнис одлуки.

Во поглавјето 3 е даден детаљен опис на податочните складови, од каде што се исцрпуваат податоците, опишан е ETL процесот и анализата на податоци.

Во поглавјето 4 се зборува за податочното рударење, неговите цели и задачи, значењето, типовите на атрибути кои се користат во процесот на податочното рударење и начинот на предпроцесирање на податоците.

Поглавјето 5 дава детаљен опис на техниките и методите на податочното рударство. Дел од овие техники ќе бидат и практично применети врз податоците добиени од анкетниот лист и фискалните сметки користејќи одредени алатки за податочното рударење

Поглавјето 6 дава опис на различните видови алатки кои се користат во податочното рударство, а со кои се интерпретираат резултатите од техниките на податочното рударство.

Поглавјето 7 го опишува експерименталниот дел од магистерскиот труд, односно ја опишува практичната имплементација на податоците добиени од анкетниот лист, базата на податоци и фискалните сметки, нивниот начин на

претставување и шифрирање, начинот на анализа на податоците и секако, добиените резултати од анализата, а сето тоа со примена на техниките на податочното рударење.

Поглавјето 8 ги опишува сите донесени заклучоци врз основа на извршените анализи, а врз основа на кои се очекува да се даде поддршка на одредени бизнис одлуки.

Во поглавјето 9 е даден заклучок произлезен од работата на овој магистерски труд.

2. Бизнис интелигенција

2.1 Основен концепт на бизнис интелигенцијата

Кога ќе помислиме на интелигенција, ние обично помислуваме на способноста на луѓето да го комбинираат наученото знаење со новите информации и на способноста да го променат однесувањето на начин со кој тие успешно ја завршуваат нивната задача или се адаптираат на новите ситуации. Поимот интелигенција се објаснува како ментална карактеристика која се состои од способност за учење од искуства, прилагодување на нови ситуации, разбирање и користење на апстрактни поими и користење на знаењето за снаоѓање во околината. Во таа смисла **бизнис интелигенцијата** е способност на некоја фирма да се прилагоди на новонастанатите услови на пазарот.

Бизнис интелигенцијата (Business Intelligence-BI) може да се дефинира како **процес на прибирање на расположливи интерни и релевантни екстерни податоци, и нивна конверзија во корисни информации со чија помош се донесуваат одлуки во бизнис секторот за да се постигне поголема профитабилност** [1].

Ако се користи строга дефиниција поимот *Бизнис интелигенција ги обединува методологиите, технологиите и платформите за складирање на податоци, online аналитичка обработка на податоци и податочно рударење кои на фирмите им овозможуваат креирање на корисни управувачки информации од податоците кои се наоѓаат на различни трансакциски системи и доаѓаат од различни екстерни и интерни извори* [29].

Позната уште и под името **Конкурентска интелигенција (competitive intelligence)** [17] претставува систематски и етички начин на добавување,

прибирање, анализирање и сортирање на јавно достапни информации за активностите на конкурентните фирми, врз основа на што може да се предвидат идните бизнис трендови, а со цел да се одржи и зацврсти сопствената компетентност на пазарот.

Бизнис интелигенцијата претставува еден од главните инструменти во процесот на донесување одлуки. Бизнис интелигенцијата интензивно почнува да се развива со автоматизацијата на фирмите, со која што доаѓа до т.н. „експлозија“ на податоци – податоците сè повеќе и повеќе се натрупувале, настанувале сè повеќе и повеќе нови бази на податоци до кои не можело да се дојде со лесен и едноставен начин, па затоа и не се користеле. Паралелно со тоа растела и свеста дека во таквите податоци лежи огромен потенцијал и вистинско богатство, но дека е потребно нешто со што тие податоци ќе се обработат, обединат и стават на располагање на менаџментот на фирмите. Гледано од техничка страна, бизнис интелигенцијата може да се опише како процес со кои сировите податоци се претвораат во информации, за потоа истите информации да се анализираат и користат во процесот на донесување одлуки.

Во реалноста бизнис интелигенцијата од една страна е начин на бизнис размислување кое овозможува да се донесат промислени и издржани бизнис одлуки врз основа на релевантни и ажурни информации, а не врз основа на претчувства и субјективно влијание. Но, од друга страна, од информатичка гледна точка, бизнис интелигенцијата е сложен информациски систем кој со помош на автоматизирани процедури ги прибира податоците од разни извори, ги обработува, трансформира и интегрира, и на тој начин им овозможува на корисниците пристап до квалитетни информации на интуитивен и лесно разбирлив начин.

Иако количината на податоци со кои се располага има важна функција, сепак тоа не е од пресудно значење. Концептот на бизнис интелигенцијата се заснова на следните темелни замисли [31] :

- Намерата на концептот на бизнис интелигенцијата не е создавање на голема количина на информации, туку генерирање на подобри, поквалитетни информации потребни при донесувањето на бизнис одлуките

- Бизнис интелигенцијата ги дава на располагање на корисниците само оние информации кои им се потребни, на начин на кој најмногу им одговара
- Со вистинска примена, концептот бизнис интелигенција ќе ја намали количината на информации, истовремено зголемувајќи го квалитетот на тие информации

Бизнис интелигенцијата функционира на следниов начин: Оперативните бази на податоци на компанијата ги следат трансакциите генерирани од водењето на бизнисот. Овие бази на податоци обезбедуваат податоци за магациниte (складовите) на податоци. Менаџерите ги користат алатките за бизнис интелигенцијата за да пронајдат скриени шаблони и значења во податоците. Менаџерите потоа дејствуваат според она што тие го научиле од анализирањето на податоците донесувајќи поинформирани и интелигентни бизнис одлуки.

Бизнис интелигенцијата поаѓа од тоа да секогаш треба да се тежнее кон потребите на клиентите и дава одговор на тие потреби, со што услугите кон истите ќе бидат поквалитетни, а менаџерите ќе можат да донесуваат правилни одлуки.

Зголемувањето на квалитетот на податоци треба да стане една од главните цели на менаџментот на една фирма. За да се создаде ова потребно е да се преземат сите можни мерки за преиспитување на квалитетот на податоците, кој може да бидат сместен во датотеки, бази на податоци, складови на податоци. За да може ова да профункционира потребни и од голема помош се бројни софтверски алати кои го анализираат интегритетот на податоците и спроведуваат статистички анализи на содржината на тие податоци, за на крај да генерираат извештаи за резултатите од обработката. До кое ниво да се навлезе во чистењето, односно претпроцесирањето на податоците и отстранувањето на грешките, треба да ни покаже анализата на трошоците, врз основа на која најчесто менаџментот на претпријатието се определува за вложување.

Клучно прашање во современото бизнис работење е: *Што е најважната претпоставка за преживување на едно претпријатие на денешните турбулентни пазари?* Одговорот е едноставен: **информацијата**, или

попрецизно, информацијата која овозможува преземање соодветна акција. Ова подеднакво важи за сите индустрии, почнувајќи од земјоделството, енергетиката и машинството, преку трговијата и банкарството, па сè до образованието и осигурувањето. Секоја фирма поседува знаење за своите клиенти, за да може да ги разбере нивните потреби, однесување и преференци. Знаењето е можно да се поседува, само доколку се поседуваат информации кои одржуваат некаква целина.

Информацијата е темел врз кој се гради знаењето, а истовремено информацијата се гради врз основа на податоците. Ако не се поседуваат доволно добри податоци, нема да се има добри информации, па секое знаење извлечено од тие информации ќе биде опасно за едно претпријатие.

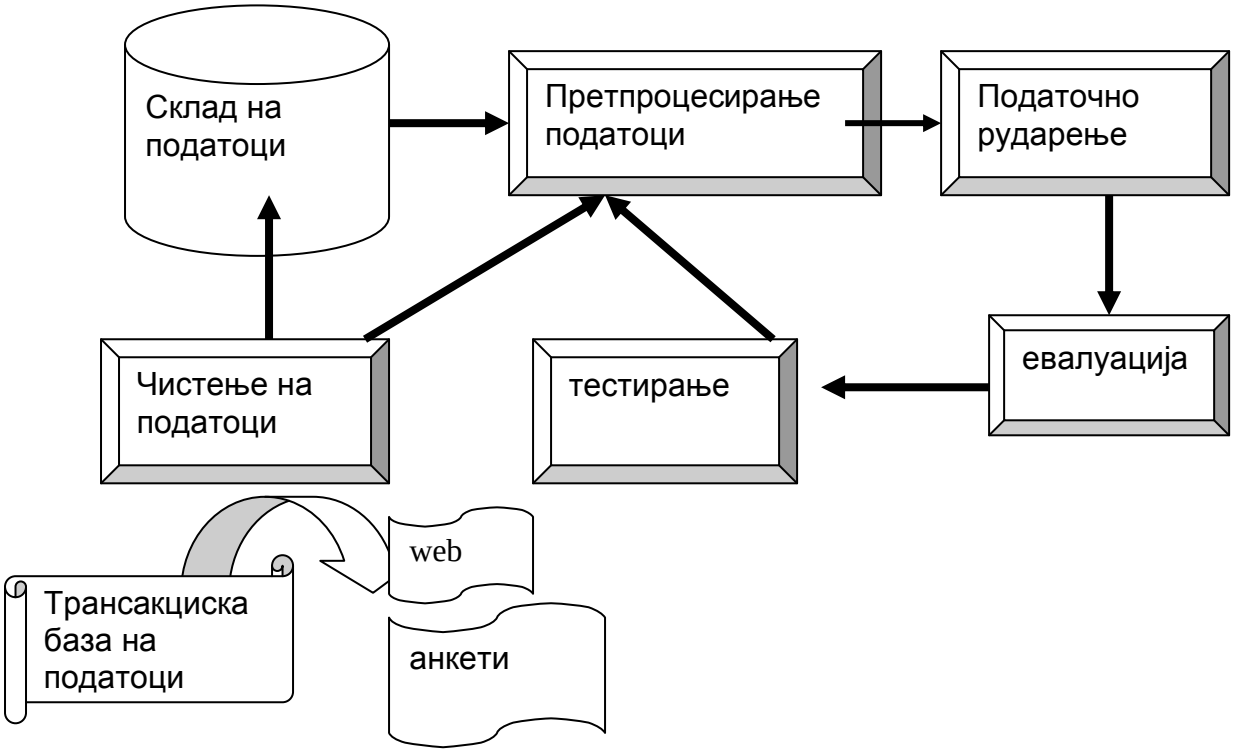
Сите информации во овој магистерски труд се добиени врз основа на базата на податоци на една македонска фирма која, располага со огромни информации, спроведената анкета на 120 купувачи и фискалните сметки на купувачите.

Брзото создавање на нови технологии придонесува во зголемување на ризикот од несвесна употреба на лоши податоци како основа за поддршка на бизнис одлуки, или пак неодговорно игнорирање на таквите ризици. Факт е дека во многу претпријатија постои свест за лошиот квалитет на одредени податоци, а и занемарување на преземање активност со кои би се исправиле таквите појави. Неквалитетни податоци на различен начин може да влијаат на работењето на едно претпријатие, а последиците од истото може да бидат катастрофални.

2.2 Примена на бизнис интелигенцијата

Секојдневно сме сведоци на разни примени на концептот на бизнис интелигенцијата во разни подрачја на делување. Многу големи и средни фирми во светот активно пристапуваат на поимот бизнис интелигенција, развиваат и имплементираат системи за подршка на истата, и ги користат во електронската трговија. Фирмите на овој начин успеваат да ги претворат информациите во бизнис интелигенција, бизнис интелигенцијата во организациско знаење, а колективното организациското знаење во зголемен профит.

Имајќи предвид, дека овој магистерски труд е ориентиран на податочното рударење, на слика 1 е даден шематски приказ на односот на системот за бизнис интелигенција и податочното рударење [31].



Слика 1. Модел на систем на бизнис интелигенција заснован на податочно рударење [31]

Figure 1. Model of business intelligence system based on data mining

Од сликата се забележува дека моделот на системот на бизнис интелигенцијата, заснован на податочно рударење, е врзан со трансакциска база на податоци и надворешни податоци прибрани од различни извори. Посредник меѓу пазарот и трансакциската база на податоци, па и методите кои ги генерираат правилата, е сегментот кој е задолжен за чистење и складирање на податоци. Потоа податоците влегуваат во дел од моделот кој е задолжен за генерирање на правила, а во себе ги има интегрирано споменатите методи кои тоа му го овозможуваат. Следен сегмент на моделот е модулот кој ги прилагодува правилата во формат кој е потребен за конкретниот експертен систем, така што податоците по записот во базата се подготвени за користење.

3.Складови на податоци

3.1 Основен концепт на складови на податоци

Во зависност од природата на бизнис процесите кои се поддржани од информационите технологии, постоечките информациони системи може да се поделат на две основни категории. Во првата категорија спаѓаат системите за извршување на секојдневните податочни трансакции *(On Line Transactional Processing – OLTP*) т.е. за управување со секојдневниот бизнис – за поддршка на дневните активности, како што се: управување со податоците за набавки, продажни и куповни трансакции, магацинско работење, финансии, човечки ресурси итн. Во втората категорија спаѓаат информационите системи кои се првенствено наменети да овозможат анализа, планирање и контрола на бизнис процесите т.е. наменети за извршување на аналитичките активности (*On Line Analytical Processing – OLAP)*. Овие информациони системи се фокусираат на трансформација на податоците од постоечките системи за обезбедување на навремени и релевантни информации за деловното работење на компанијата т.е. информации што ќе овозможат на менаџерските тимови да донесуваат брзи и прецизни одлуки за зголемување на ефикасноста и справување со конкуренцијата на пазарот. Фискалните сметки користени во изработката на овој труд се дел од системот за извршување секојдневни податочни трансакции.

Главната компонента на аналитичките системи се складовите на податоци (*Data Warehouse – DW*). Технологијата на складовите на податоци содржи архитектура, алгоритми, модели, алатки, организациони и управувачки модули за да овозможи доволно информации за донесување на вистински бизнис одлуки, најчесто реализирани како посебна база на податоци во која се чуваат интегрираните оперативни податоци. За разлика од операционите системи кои оперираат со детални, атомарни и тековни податоци, кои се користат од страна на трансакциските апликации, технологијата на складовите на податоци се стреми да обезбеди интегрирани, консолидирани и историски податоци за аналитичките апликации.

Меѓутоа, складовите на податоци се многу повеќе од архива на корпоративните податоци и многу повеќе од нов начин за пристап до корпоративните информации. Тие се основата на деловно интелигентните системи со кои се обезбедуваат потребните информации за извршните

менаџери. Намената на складовите на податоци не е само за извршување на комплексни прашалници, туку тие имаат поопшта намена- за добивање брзи, точни и прецизни, често интуитивни и „итри“ информации. Токму поради тоа т.е. поради нивниот неоспорен придонес во зголемувањето на ефективноста и ефикасноста на процесот на одлучување во деловниот и научниот домен, тие рапидно се раширија во индустријата во текот на последната декада. Тоа широко распространување е поддржано од значителни научно – истражувачки резултати и од брзиот развој на комерцијални алатки.

Системите на складови на податоци треба да обезбедат поддршка за аналитичките апликации со примена на идејата за чување на податоците во посебна, аналитичка база на податоци. Базата на податоци кај складовите на податоци најчесто е релациона база на податоци, што пред сè е наменета за извршување на прашалници и правење на анализи, а не за процесирање (извршување) на трансакции. Складовите на податоци најчесто содржат историски податоци добиени (извлечени) од трансакциските податоци, но може да вклучуваат и податоци од други извори. Покрај функционалностите на релационите бази на податоци, системите на складовите на податоци вклучуваат решение за извлекување (преземање), трансформирање и вчитување на податоците (*Data Extraction, Transformation and Loading - ETL*), обезбедуваат можности за *Online Analytical Processing (OLAP)* и податочно рударење (*Data Mining*), клиентски алатки за анализа и останати апликации што управуваат со процесот на преземање на податоците и нивно доставување до деловните корисници.

Често користен начин за запознавање со складовите на податоци е преку набројување на четирите основни карактеристики [15]:

- **Ориентираниот кон субјектот (*Subject Oriented*)**
- **Интегрираност (*Integrated*)**
- **Постојаност (*Nonvolatile*)**
- **Временска променливост (*Time Variant*)**

Спротивно на трансакциските (операционите) системи кои се обично имплементирани во контекст на процесите на компанијата, складовите на податоци се дизајнирани за да помогнат при анализа на податоците. Складот на податоци е **ориентиран кон „субјектит**е“ со која работи компанијата како што се клиентите, производите, бизнис трансакциите, активностите и полисите.

На пример, за да се дознае повеќе за податоците за продажба на некоја компанија, потребно е да се изгради склад на податоци наменет за продажбата.

Интеграцијата е поврзана со претходното својство. Складовите на податоци се полнат со податоци од различни хетерогени оперативни системи (како бази на податоци, датотеки и сл.) и дополнителни надворешни податочни извори (како демографски, статистички бази на податоци, *WWW* и сл.). Главната цел поради која се комбинираат податоци од различни извори е можноста за вкрстен и комбиниран пристап на податоци од различни апликации како што се продажба, маркетинг, финансии и производство. Податоците се складирани во различен формат, а во складот се чуваат во ист формат. Проблемите кои настануваат околу разликите во имињата и неконзистентноста на мерните податоци од различните извори, мора да бидат решени. Кога ќе се постигне оваа цел се добива интеграција на податоците во еден заеднички формат.

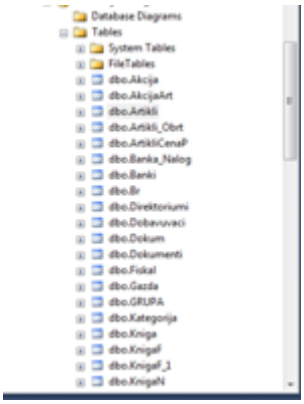
Складовите на податоци се статична колекција на податоци, а тоа значи дека пристапот до базата на податоци е претежно ориентиран кон читање на податоците. Промените во складовите на податоци се случуваат единствено кога промените на изворните податоци се распространуваат (рефлектираат) во складот на податоци. Тоа се случува релативно ретко т.е. кога податоците еднаш ќе бидат вчитани во складот на податоци, не би требало воопшто или многу ретко да се менуваат. Ова е логично бидејќи нивната намена е да се анализира она што е направено во минатото.

Една од најважните карактеристики за складовите на податоци е **временската променливост (*Time Variant*).** Потребата за пристап до „историски“, кои се воедно и временски променливи податоци кои претрпеле промена во одреден временски период, е една од примарните улоги на складовите на податоци. Голема количина на историски податоци се неопходни за анализа на трендовите во бизнисот кои може да се разберат преку најразлични погледи во податоците. Чувањето на историските податоци значи дека промените на некои податоци се запишуваат во складот без да се презапишат старите податоци. Обемот на историските податоци и периодот кои го тие опфаќаат е во релација со бенефитот и со расположливите ресурси.

Зборувајќи воопштено, постојат два извори на податоци за складовите на податоци и тоа: **внатрешни и надворешни податоци [37].**

Внатрешните податоци припаѓаат на фирмите и се генерирани по пат на трансакциски систем. Тие податоци ги опишуваат активностите што се случиле во фирмата. Надворешните податоци се прибираат надвор од фирмата, најчесто со посредство на специјализирани установи кои се занимаваат со прибирање и дистрибуција на информации. Надворешните податоци се од критична важност за стратешките одлуки, бидејќи со нивна помош претпријатието воочува поволни можности, но и закани. Различните видови на надворешни податоци може да се вбројат во конкурентни податоци (производи, услуги), економски податоци (движење на каматата, берзански податоци,..), економетриски, психометриски и др.

Како внатрешни податоци во овој магистерски труд се користат податоците од базата сместена на SQL Server 2008, фискалните сметки направени од страна на купувачите, а како надворешни спроведената анкета на 120 купувачи од продавницата. Дел од базата на податоци на фирмата во која се вршени анализите е прикажана на слика 2:



Слика 2. Приказ на базата на податоци користена во анализата

Figure 2. Review of the database used in the analyses

Една од најважните цели на складовите на податоци е интегрирање на внатрешните и надворешните податоци. Во овој магистерски труд се интегрирани сите обработени податоци од базата сместена локално на сервер, фискалните сметки и анкетниот лист.

За складовите на податоци најчесто се изработува димензиски модел, за разлика од трансакциските системи кои најчесто имаат модел на податоците заснован на објектите и нивните односи. Димензискиот модел дава подобри

можности за визуелизација на податоците. На луѓето им е природно да ги набљудуваат бизнис појавите низ призма на димензија. Ако појавата се следи во три димензии, станува збор за *коцка,* а во повеќе димензии за *хиперкоцка*. На секоја димензија од коцката и припаѓа по еден параметар од појавата и секоја точка во коцката има точно одредени вредности на секоја набљудувана димензија.

Моделите на податоци на трансакциските системи, пак, се изградени врз основа на други начела. Тие ги прикажуваат бизнис објектите и нивните меѓусебни односи кои произлегуваат од бизнис процесот. Табеларните модели на трансакциските системи се поприкладни за следење, односно управување со бизнис процесот, а димензиските модели на податочните складови за известување за ефектите на бизнис процесот. Но, и двата модела и димензискиот и објектниот се способни да го прифатат и опишат истото множество податоци, и од него да направат исто множество на извештаи или да извршат исти анализи.

3.2 *ETL* процеси

Како што е претходно кажано, податоците во складот на податоци влегуваат од најразлични извори, најчесто од трансакциските системи на фирмата. Најобемен процес во активностите со складирањето на податоците претставуваат процесите на интегрирање на податоците и организирање на нивната содржина. При тоа главна улога игра множеството процеси чија задача им е извлекувањето, трансформирањето и вчитувањето на податоците [30]. Се состои од сите активности поврзани со извлекување на податоците од изворните системи, прочистување и трансформирање на податоците во соодветен формат и на одредено ниво на детали, сместување на податоците во целниот податочен склад и подготовка за анализи.

Процесот на извлекување, трансформирање и вчитување на податоците т.е. *ETL* процесот (*Extraction, Transformation, Loading – ETL*) се користи во повеќе фази на преземање на податоците.

При самиот почеток на *ETL* процесот потребно е да се извршат низа подготвителни активности поврзани со *реформатирањето, усогласувањето и чистењето на податоците*. Изворните податоци добиени од различни датотеки и бази на податоци потребно е да се унифицираат, односно прикажат

во еден единствен формат. Во тој формат податоците ќе се користат во сите идни фази на обработка. Усогласувањето на податоците се користи со цел да се избегне редунданција на податоците. Освен тоа што во еден информационен систем сите податоци може да се појават на повеќе места, тие знаат да бидат и недоследни, односно нивните вредности не се секогаш исти на сите места во кои тие податоци се појавуваат. Поради тоа потребно е да се откријат и усогласат. Чистењето како подготвителна активност на *ETL* процесот има за задача да ги отстрани оние податоци кои не се комплетни, неточни, неконзистентни итн.

. Екстракцијата на податоците во овој магистерски труд е добиена од базата на македонска фирма сместена на сервер, фискалните сметки и анкетниот лист, при што се земени податоците за 2011 година, а се отфрлени претходните историски податоци.

3.3 Анализа на податоците

Анализа на податоци е процес на обработка на податоците од складот со цел да се извлече потребното и корисното знаење (заклучоци) од интегрираните податоци во него. Алатките за анализа на податоци нудат можности за поставување на прашалници (*Queries*) на складот на податоци, креирање на аналитички функции (прости агрегатни функции или статистички методи) и приказ на податоците на различни начини. Класичните алатки за анализа содржат алатки за генерирање на извештаи, табели, графици и симулации. Во овој контекст во последно време се развиени голем број на аналитички (*OLAP*) апликации како и алатки за податочно рударење или „копање” по податоците.

4.Податочно рударење

4.1 Поим, значење и архитектура на податочно рударење

Податочното рударење привлече големо внимание во областа на информатичката индустрија и во општеството како целина во последните години, што се должи на широката достапност на огромни количини на податоци и на непосредната потреба за претворање на таквите податоци во корисни информации и знаење. Стекнатите информации и знаење може да се користат за апликации почнувајќи од анализа на пазарот, откривање на измама

и задржување на клиенти, па сè до контрола на производството и научни истражувања.

Податочното рударење може да се гледа како резултат на природната еволуција на информатичката технологија. Брзиот раст, огромните количини на податоци, собирани и чувани во голем број на складови на податоци, ја има далеку надминато нашата човечка способност за разбирање. Како резултат на тоа, податоците собрани во големи податочни складови стануваат „податочни гробници"- архиви на податоци кои се ретко посетени. Следствено, важни одлуки често не се прават врз основа на податоци кои се богати со информации, туку со интуиција на човекот кој прави одлука, едноставно затоа што тој што ја носи одлуката нема доволно средства за да извлече вредни знаења вградени во огромните количини на податоци.

Податочно рударење (data mining) претставува истражување и анализа на големи количини на податоци со цел да се откријат правила и шаблони во истражуваните податоци кои имаат одредено значење. Една од дефинициите кои се користат за податочното рударење гласи: **Податочното рударење е составен, интерактивен и итеративен процес на извлекување и прикажување на корисното, имплицитно и иновативното знаење од податоците [24]**. Со оваа дефиниција важно е да се напомене и дека успешноста на примената на методите и алатите за оваа намена, првенствено зависи од стручноста и компетенцијата на оние кои ги толкуваат добиените резултати. Токму тие личности со своето знаење и искуство се способни да некој на изглед бесмислен примерок на смислен и коректен начин го претворат во вредна информација.

Постојат два типа на податочно рударење [27]:

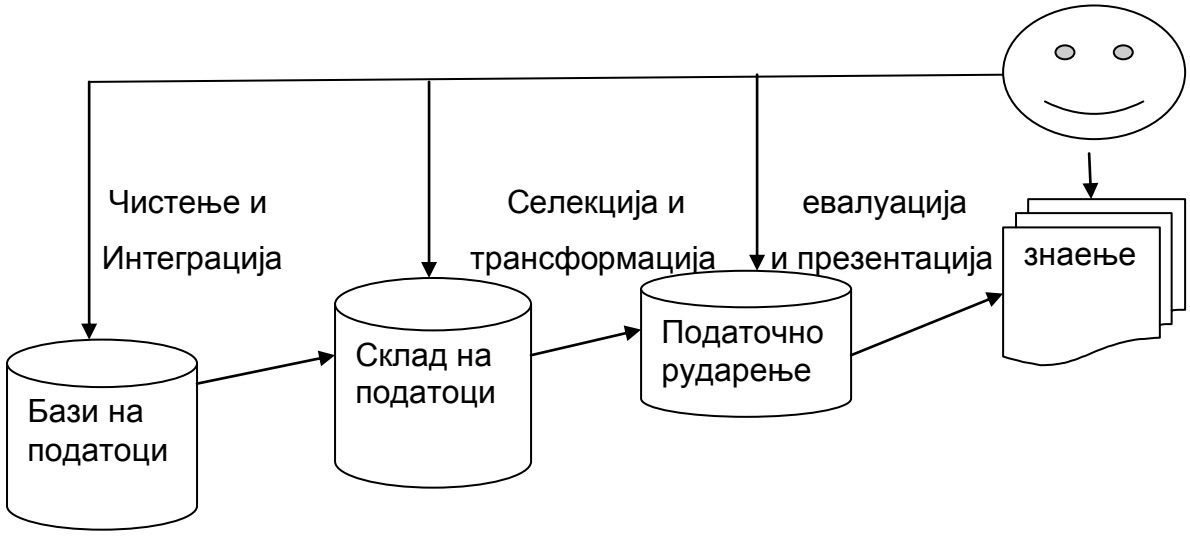
- *Верификација на хипотези*- целта е да се провери дали некоја идеја или впечаток за важноста на односите меѓу податоците е втемелен или не;
- *Откривање нови знаења*- меѓу некои појави може да постојат некои непознати, а важни статистички односи, кои човекот ниту со своето искуство, ниту со својот интелект може да ги долови;

Меѓу основните технологии за податочно рударење се: статистика, системи базирани на знаење, невронски мрежи, машинско учење, управување

со бази на податоци итн. Но сепак, основно јадро на сите тие процеси за откривање на знаење од податоците е аналитичката статистика. Од статистичка перспектива, податочното рударење се опишува како *компјутерски автоматизирана и истражувачка анализа на податоци од големи и сложени бази на податоци од различни платформи, локации, оперативни системи и софтвер*.

Бизнис процесите и научните инструменти лесно генерираат терабајт податоци. Јазот помеѓу можностите за прибирање на податоци и способноста за анализа на истите, како во бизнисот, така и во науката сè повеќе се шири. Податочното рударење е најкорисно таму кадешто постои постојана закана од голема поплава со информации.

Податочното рударење во принцип се занимава со градење на модели. Модел претставува алгоритам или множество на правила кои што ги поврзуваат одредени влезни променливи со за нив соодветни излезни променливи. Регресијата, невронските мрежи, одлучувачките дрва и повеќето останати техники на податочно рударење се занимаваат со градење на модели. Процесот на податочно рударење понекогаш се нарекува **процес на откривање на знаење во база на податоци (knowledge discovery in database)-KDD [11]** и истиот е прикажан на слика 3 подолу:



Слика 3. Процес на откривање на знаење во база на податоци

Figure 3. Process of knowledge discovery in database (KDD)

KDD како процес се состои од итеративна секвенца од следните чекори:

- **Чистење на податоци**
- **Интеграција на податоци**
- **Селекција на податоци**
- **Трансформација на податоци**
- **Податочно рударење**
- **Модел на евалуација**
- **Презентирање на знаење**

Гледано од оваа основа, архитектурата на еден типичен систем на податочно рударење ги содржи следниве компоненти [11], [12]:

- **База на податоци**;
- **Data warehouse сервер**- одговорен за земање релевантни податоци врз основа на барање на корисникот;
- **База на знаење**- водич за пребарување или евалуација на резултати;
- **Моторот на податочното рударење**- множество на функционални модули за задачи како асоцојација, анализа, класификација, кластерирање, анализа на оутлаери и др;
- **Модул за евалуација**- поврзан е со модулот рударење и целта му е да го ограничи пребарувањето на само интересни домени;
- **Кориснички интерфејс**- ја врши комуникацијата меѓу корисниците и системот за податочно рударење и овозможува пребарување на базата, оценува модели;

4.2 Цели и задачи на податочното рударење

Секој корисник што се занимава со податочно рударење ќе има соодветна задача која треба да ја изврши и цел што треба да ја постигне. Задачите може да бидат во форма на прашалник на податочно рударење кое би претставувало основна задача на податочното рударење. Основните задачи на корисникот му овозможуваат да комуницира со системот на податочно рударење, со цел да се испитаат добиените резултати од различни агли и длабочини. Задачите на податочното рударење го вклучуваат следново:

- **Множество на задача**- релевантни податоци кои треба да се обработуваат, вклучувајќи атрибути на база на податоци или релевантни атрибути;
- **Вид на знаење што треба да се обработи**- ги специфицира функциите на податочното рударење кои треба да се изведуваат;
- **Знаење во позадина** кое треба да се користи во процесот на откривање;
- **Мерки на заинтересираност и прагови на оценување**;
- **Очекувани репрезентации за визуелизација на откриените информации**- правила, табели, графикони, дрва на одлучување;

Целите на податочното рударство најчесто се:

- **Обработка на различни видови на знаења во базите на податоци**- широк спектар на анализа на податоци и задачи за откривање на знаење;
- **Основање на знаење во позадина**- водич во процесот на откривање и овозможување откриените модели да бидат изразени во точни термини;
- **Презентација и визуелизација на резултатите**- изразено со јазици на високо ниво и визуелни репрезентации;
- **Интерактивно рударење на знаења на повеќе нивоа-** се фокусира на потрага по модели и обезбедување на податочни побарувања врз основа на вратените резултати, односно му овозможува комуникација на корисникот со системот;
- **Паралелни и дистрибуирани алгоритми за податочно рударење-** делење на податоците на делови и паралелно извршување;
- **Ефикасност и приспособливост на алгоритмите за податочно рударење**- времето на извршување на даден алгоритам мора да биде предвидливо и прифатливо во големи бази на податоци;
- **Ракување со шум и нецелосни податоци**- методите на чистење и методите на анализа се справуваат со овие проблеми;

4.3.Типови атрибути

Постојат повеќе различни типови на атрибути кои се користат во обработката на податоци и тоа [6]:

- **Номинални**- ги класифицираат објектите според тип или карактеристика, при што категориите се заемно исклучиви и атрибутите немаат логичко подредување;
- **Ординални**- ги класифицираат објектите според тип или вид, но имаат логичко подредување и категориите се заемно исклучиви;
- **Интервални**- класифицираат според тип, имаат логичко подредување, но разликата меѓу секое од нивоата или категориите е еднаква и немаат нулта почетна точка;
- **Ratio**- тие се како интервални, но имаат нулта почетна точка;

Атрибутите може да бидат и:

- **Дискретни**- имаат конечно или бесконечно пребројиво множество на вредности и често се претставуваат како целобројни променливи;
- **Непрекинати-** имаат реални броеви како вредности на атрибутите;

4.4 Предпроцесирање и чистење на податоци

Податоците во реалниот свет не се чисти, па и самите податоци за рударење може да бидат:

- **Некомплетни**- недостигаат вредности на атрибути, недостигаат одредени атрибути кои ни се од интерес или содржат само агрегирани податоци;
- Да имаат **шум**- содржат грешки или оутлаери;
- **Неконзистентни**- содржат неконзистентни податоци во кодови или имиња;

Податочното чистење прави обид да се пополнат исчезнатите вредности, да се измазнат податоците додека се идентификуваат outlier-и и поправка на неконзистентност во истите. **Вредностите кои недостасуваат** (missing value) се пополнуваат со:

- Игнорирање;
- Рачно пополнување;
- Средна вредност на податоците;
- Некоја глобална константа;
- Користење на некоја најверојатна вредност;

Нечисти се и податоците кои содржат шум. ***Шум*** *е случајна грешка или отстапување од реалните вредности*. Техники за измазнување на податоците се:

- Binning;
- Регресија
- Кластерирање;

4.5 Мерки за избор на атрибути

Мерка за избор на атрибути претставува хевристика за избор на критериум за поделба којшто „најдобро" го дели дадено множество D кое се состои од подредени тренирачки вредности во индивидуални класи. Ако го делиме множеството D во помали партиции според критериумот за поделба, во идеален случај секоја партиција би била "чиста" т.е. сите подредени вредности во дадена партиција би и припаѓале на иста класа. Во принцип "најдобриот" критериум за поделба е оној кој најблиску резултира до такво сценарио. Мерките за избор на атрибути се исто така познати како правила за поделба, бидејќи тие одлучуваат на кој начин податоците во даден јазол да се поделат.

Една од мерките за избор на атрибути е намалување *на неизвесноста* за вредноста на влезниот податок доколку ја знаеме вредноста на излезниот податок. Нека во јазолот N се сместени подредените вредности од множеството D. Атрибутот со најголема сопствена информација се избира како атрибут според кој ќе се врши поделбата за јазолот N. Овој атрибут ја минимизира информацијата која што е потребна за да се класифицираат податоците во соодветни партиции и на тој начин се добиваат партиции со најмала "нечистотија". Информацијата потребна за класифицирање на податок од D е дадена со изразот:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

каде p_i е веројатноста дека одреден податок од множеството D припаѓа на класата. Логаритамската функција се пресметува со степен 2 бидејќи информацијата е енкодирана во битови. Info(D) е попозната како **ентропија** на D.

Втората мерка применува нормализација користејќи информациска поделеност, односно "split information", вредност дефинирана аналогно како и Info(D) на следниот начин:

$$SplitInfo(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2(\frac{|D_j|}{|D|})$$

Оваа вредност ја претставува потенцијалната информација генерирана со делење на множеството тренирачки вредности D на v партиции соодветни на v можни резултати при тестирање на атрибутот A. Во тој случај количникот на добивка се дефинира како:

$$\text{GainRatio(A)} = \frac{Gain(A)}{SplitInfo(A)}$$

Gini индексот е третата многу користена мерка за избор на атрибути кој најчесто се користи во алгоритмот CART. Тој всушност ја мери "нечистотијата" на множеството на подредени вредности D на следниот начин:

$$\text{Gini (D)} = 1 - \sum_{i=1}^{m} p_i^2$$

каде p_i е веројатноста дека подредената n-торка од D припаѓа на класата C_i.

Предложени се и многу други мерки за избор на атрибути. CHAID, алгоритам за креирање на дрва за одлучување кој што е популарен во областа на маркетингот користи мерка за избор на атрибути која што се базира на статистичкиот ChiSquare тест за независност.

5.Техники и методи на податочно рударење

Податочното рударење користи различни техники за да најде скриени поврзаности и односи во големи бази податоци, создава правила кои може да се користат за да се предвиди идно однесување и служи како водач за донесување одлуки.

Секоја анализа од податочното рударење подразбира делење на набљудуваната популација на примерок за учење и примерок за тестирање. На примерокот за учење се применуваат алгоритми кои врз основа на податоците ги распознаваат примероците, правилноста и вредноста на коефициентот на поставениот модел. Со примерокот за тестирање се врши проверка на

веродостојноста на добиеното решение. Доколку се појави значително отстапување добиено со примерокот за тестирање, постапката се повторува сè додека не се дојде до задоволителни резултати.

Податочното рударење вклучува техники од повеќе дисциплини како бази на податоци, складови на податоци, статистика, машинско учење, препознавање на шаблони, невронски мрежи, визуелизација на податоци, процесирање на слики и сигнали итн.

Многу проблеми кои се од интелектуален, економски и бизнис интерес може бидат решени со помош на следните 6 техники:

- **Класификација**
- **Проценка**
- **Предвидување**
- **Групирање по сродност**
- **Кластерирање**
- **Опис и профилирање**

Првите 3 техники се примери на насочено податочно рударење, каде што целта е да се најде вредноста на одредена целна променлива. Групирањето по сродност и кластерирање се ненасочени техики чија што цел е да откријат одредени структури во податоците не земајќи во предвид притоа ниедна целна променлива. Профилирањето е описна техника која може да биде насочена или ненасочена.

Со помош на методите на податочното рударење може прецизно да се изврши сегментација на пазарот, да се открие профилот на типичниот или лојалниот купувач, неговите афинитети за купување одредени групи артикли, да се открие сличноста меѓу одредени категории артикли итн. Методите за рударење на податоци се многубројни, од кои ќе бидат наведени само некои:

- **Метод на потрошувачка кошничка**
- **Невронски мрежи**
- **Дрва на одлучување**
- **Генетски алгоритми**
- **Генетско програмирање**
- **Мемориски засновано одлучување**

- **Асоцијативни правила**
- **Статистички методи и др.**

Методите за рударење на податоците треба да се користат во симбиоза со методите на предпроцесирање на податоците и методологиите за делење на податоците на примерок за учење и примерок за тестирање. По спроведените анализи на примероците аналитичарот добива информации за веродостојноста на моделот, и доколку се процени дека моделот не е веродостоен може да се менуваат параметрите на конкретните методи, или пак, тој метод се заменува со некој друг.

Знаењето кое е откриено со помош на податочното рударење може да биде употребено при донесување на разни одлуки, контрола на процеси, управување на информации и процесирање на пребарувања. Затоа се смета дека податочното рударење е една од најважните методологии применливи врз базите на податоци и информациските системи и една од најветувачките интердисиплинарни науки во информатичката технологија.

5.1 Податочно рударење и статистички методи

За статистичарите терминот податочно рударење долго време имал минорaтивно значење. Наместо наоѓање на корисни шаблони во големи количини на податоци, податочното рударење имало конотација на пребарување на податоци кои би се вклопиле во одредена статистичка дистрибуција на податоци. Во принцип двете дисциплини се многу слични и користат слични техники за решавање на слични проблеми, меѓутоа пристапот за решавање на проблемите кај податочното рударење се разликува од стандардниот статистички пристап во неколку области:

- Во податочното рударење се игнорираат грешките што настануваат при мерење на необработени податоци;
- Во податочното рударење се претпоставува дека располагаме со доволно податоци и процесирачка моќ за решавање на поставените проблеми;
- Во податочното рударење се претпоставува дека постои временска зависност секаде;
- Податоците се цензурирани и во скратена форма

5.1.1 Регресија

Убедливо најкористен пристап за нумерички предвидувања е статистичкиот метод *регресија*. Анализата со помош на регресија се користи за да се моделира врската помеѓу една или повеќе меѓусебе независни променливи (предиктори) и променлива која зависи од предикторите (која е од апсолутно непрекинат тип). Во контекст на податочно рударење предиктори се променливите кои се од интерес. Со процесот на регресија сакаме да ја предвидиме вредноста на зависната променлива. Методот на регресија многу често се користи за опис на врската меѓу променливите од примарен интерес (продажба, износ на фискална сметка) и т.н. предикторски променливи (месечен приход на купувачот, возраст, број на членови во семејството итн.), односно кога постоечките вредности се користат за предвидување на тоа какви ќе бидат останатите вредности.

Анализата со помош на регресија претставува добар избор кога сите променливи предиктори се од апсолутно непрекинат тип. Многу проблеми може да бидат решени со помош на линеарната регресија, додека за многу други проблеми оваа методологија претставува почетна точка за нивно решавање доколку се применат трансформации на променливите така што посочениот проблем од нелинеарен се претвора во линеарен.

Методот на регресија најчесто се применува за предвидување, на пример колку профит ќе генерира една категорија на потрошувачи или слично.

Кај линеарната регресија има една променлива предиктор x и една зависна променлива y. Таа претставува наједноставен облик на регресија каде што зависната променлива y претставува линеарна функција од независната променлива x, односно:

$$y=b_0+b_1x$$

каде b и b_0 се коефициенти на регресија. Коефициентите b_0 и b може да се сметаат и како тежини.

Вредноста на овие коефициенти може да се најде со методот на најмали квадрати. Нека D е множество кое се состои од тренажни вредности за предикторот x за одредена популација и соодветните вредности за зависната

променлива y. Множеството D се состои од подредени двојки од облик (x_i, y_i). Коефициентите на регресија може да се најдат од следните равенки:

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}$$

Полиномната регресија се користи често кога имаме само еден предиктор. Таа може да се моделира со додавање на полиномни членови на основниот линеарен модел. Во следниот пример полиномен модел на регресија се трансформира во линеарен модел:

$$Y=b_0+b_1x+b_2x^2+b_3x^3$$

Генерализираните модели на регресија ја претставуваат теориската основа со помош на која се применува линеарната регресија. Најчести типови на генерализирани линеарни модели се логистичката регресија и Поасоновата регресија. Логистичката регресија ја моделира веројатноста за случување на одреден настан како линеарна фукција од множество предиктори. Податоците пак кои се добиени од некакви броења на одреден настан се моделираат со Поасонова регресија.

Во случај кога не постои линеарна релација меѓу зависната и независните променливи, а зависната променлива обично е кодирана со 0 и 1, односно е бинарна, се користи бинарната логистичка регресија. При тоа се користи логистичка трансформација која ја дава линеарната релација меѓу веројатноста на набљудуваниот настан и вредноста на независната променлива X дадена со:

$$\log(odds) = \log\left(\frac{p}{1-p}\right) = b_0 + b_1x$$

каде p е веројатноста дека некој настан ќе се случи. Во случај кога постојат повеќе предикторски променливи имаме:

$$\mathbf{p} = \frac{\mathbf{e}^{\mathbf{b_0}+\mathbf{b_1}\mathbf{x_1}+\mathbf{b_2}\mathbf{x_2}+\ldots\ldots}}{\mathbf{1}-\mathbf{e}^{\mathbf{b_0}+\mathbf{b_1}\mathbf{x_1}+\mathbf{b_2}\mathbf{x_2}+\ldots\ldots}}$$

каде e=2,718. Коефициентот b_0 е неопходен за равенката и ја претставува вредноста log (odds) кога предикторот е 0, а b_1 е мерка за врската меѓу предикторот и log (odds) за појава на настанот кој нè интересира. Ако $b_1>0$

постои позитивна врска (корелација), ако $b_1<0$ корелацијата е негативна, а ако $b_1=0$ нема корелација.

5.1.2 Анализа на релевантноста на атрибутите

Анализата на податоци вклучува и интеграција на податоците, што ги комбинира податоците од повеќе извори во еден кохерентен склад на податоци. Овие извори може да вклучуваат повеќе бази на податоци, рамни датотеки или податочни коцки. Еден од проблемите на интеграцијата е тоа што еквивалентни реални ентитети од повеќе извори може да се совпаѓаат-проблем познат како идентификација на ентитети. Еден атрибут може да биде редундантен, ако тој може да се добие од друг атрибут или збир на атрибути. Некои редундантности може да се откријат со анализа на корелација, пресметувајќи го коефициентот на корелација. За нумерички атрибути се пресметува со:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\overline{A}\overline{B}}{N\sigma_A\sigma_B}$$

каде N-бројот на инстанци, a_i и b_i се тековните вредности на A и B за инстанцата i, $\bar{A}$ и $\overline{B}$ се средните вредности на A и B, а σ_A и σ_B се соодветни стандардни девијации за A и B. Ако $r_{A,B}>0$, тогаш A и B се позитивно корелирани. Ако $r_{A,B}=0$, тогаш A и B се независни и нема корелација меѓу нив. Ако $r_{A,B}<0$, тогаш се негативно корелирани, што значи дека како вредностите на еден атрибут растат, така на другиот опаѓаат. За категориски (дискретни) податоци корелациона анализа се прави со Ҳ-квадрат тест, каде:

$$\chi^2 = \sum \frac{(Observed - Expected)}{Expected}$$

5.2 Класификација

Класификацијата е една од најчесто употребуваните техники во податочното рударење. Класификацијата се состои во испитување на карактеристиките на даден објект и негово доделување на една од повеќе предефинирани класи. Објектите кои што се класифицираат најчесто се записи

во табели на бази на податоци или фајлови и чинот на класификација се состои во додавање на нова колона со код на класа од одреден вид.

Класификацијата се карактеризира со добро дефинирано множество на класи и множество за тренирање кои се состои од примери кои се класифицирани. Целта на класификацијата е да се изгради модел од одреден вид кој може да се примени на некласифицирани податоци со цел истите да се класифицираат.

Одлучувачките дрва и техниките на најблизок сосед се прикладни за извршување на класифицирачки операции. Во посебни околности се користат и невронски мрежи и анализа на врски.

5.2.1 Како работи класификацијата?

Класификацијата на податоци е процес во два чекора. Во првиот чекор класификаторот се гради, опишувајќи предодредено множество на податочни класи или концепти. Ова е **чекор на учење**, каде алгоритамот на класификација гради класификатор со анализа или учење од тренирачко множество составено од инстанци од базата на податоци и нивни класи на асоцијација. Инстанца X е претставена со n димензионален вектор на атрибути X=(x1,x2,.,xn). Секоја инстанца X се претпоставува дека припаѓа на претходно одредена класа. Индивидуалните инстанци кои го прават тренирачкото множество се наведени како тренирачки инстанци и се избираат од базата на податоци под анализа. Овој чекор е познат како надгледувано учење. Првиот чекор од процесот на класификација, исто така може да се гледа како на учење на мапирање или функција y=f(x), што може да ја предвиди поврзаната класа Y на дадена инстанца X.

Во вториот чекор моделот се користи за класификација. Најпрво се пресметува предвидената точност на класификаторот. Потоа се зема тест множество за чии инстанци не ги знаеме конечните класи во кои припаѓаат. Тие се независни од тренинг множеството, што значи не се користат при конструирање на класификаторот. Точноста на даден класификатор претставува процентот на точно класифицирани инстанци од тест множеството. Ако точноста на класификаторот е прифатлива, тогаш тој класификатор може да се употребува за класифицирање на идни инстанци со непознати класи приврзаности.

5.2.2 Точни и лажни позитивни и негативни инстанци

Во самиот процес на класификација може да се појават следниве инстанци:

- **TP**– точно позитивни: број на позитивни инстанци кои се класифицирани како позитивни;
- **FP**- лажно позитивни: број на негативни инстанци кои се класифицирани како позитивни;
- **FN**—лажно негативни: број на позитивни инстанци кои се класифицирани како негативни;
- **TN**- точно негативни: број на негативни инстанци кои се класифицирани како негативни;
- **N=FP+TN**-вкупен број на негативни инстанци

Табела 1. Однос на точни и лажни позитивни и негативни инстанци

Table 1. True and false positive and negative instances relation

True Positive Rate- Recall, Sensitivity	**TP/P**	**Однос на позитивни инстанци кои се точно класифицирани како позитивни**
FP Rate	FP/N	Однос на негативни инстанци кои се неточно класифицирани како позитивни
FN Rate	FN/P	Однос на позитивни инстанци кои се неточно класифицирани како негативни
TN Rate, Specificity	TN/N	Однос на негативни инстанци кои се точно класифицирани како негативни
Precision	TP/(TP + FP)	Однос на инстанци кои се позитивни и се класифицирани како позитивни
F1 Score	(2*Precision*Recall)/(Precision*Recall)	Мерка која ги комбинира Recall и Precision
Accuracy	(TP+TN)/(P+N)	Однос на инстанци кои се точно класифицирани
Error Rate	(FP+FN)/(P+N)	Однос на инстанци кои не се точно класифицирани

Лажно позитивни се појавуваат кога инстанците треба да се класифицирани како негативни, а се класифицираат како позитивни, додека лажно негативни се појавуваат кога инстанците треба да се класифицирани како позитивни, а се класифицираат како негативни;

5.2.3 Баесови мрежи

Баесовите класификатори претставуваат статистички класификатори. Тие можат да ја предвидат веројатноста за припадност на дадена подредена n-торка на податоци кон одредена класа. Овие класификатори се темелат на Баесовата теорема. Нивниот баесов класификатор претпоставува дека атрибутите се условно независни во поглед на вредностите на целните променливи. Оваа претпоставка може да биде премногу силна за средини, каде што постои зависност помеѓу променливите.

Баесовите мрежи (Bayesian belief networks - BBNs) се дизајнирани за да се овозможи претставување на условни независности меѓу парови на променливи. Баесовите мрежи имаат форма на насочен ацикличен граф (directed acyclic graph -DAG), каде што насочен значи дека лаците се само во еден правец, а ацикличен значи дека нема јамки во графот.

Секоја променлива во Баесовата мрежа е условно независна од нејзините следбеници во мрежата ако се познати нејзините родители. Така, имаме:

$$p(X_1 = x_1, X_2 = x_2, ..., X_m = x_m) = \prod_{i=1}^{m} p(X_i = x_i | parents(X_i))$$

Се забележува дека детето јазол зависи само од неговите родители. Баесовата мрежа претставува заедничка веројатносна дистрибуција за поредена група на променливи. Дистрибуцијата на овие набљудувања над заедничкиот простор се нарекува дистрибуција на заедничка веројатност (joint probability distribution). Баесовата мрежа претставува заедничка веројатносна дистрибуција преку обезбедување на:

1. Одреден сет на претпоставки во врска со условната независност на променливите
2. Табели на веројатност за секоја променлива.

Кога структурата на мрежата е позната и полињата со вредности се набљудувани, тогаш учењето на Баесовата мрежа е јасно. Локалните табели на веројатности се целосно определени, и сите други веројатности како joint, условени, prior, или posterior може да се пресметаат. Меѓутоа, кога некои од вредностите се скриени или непознати, ние треба да се свртиме кон други методи, за да ги пополниме сите записи на локални веројатности на дистрибуција во табелата. Непознатите записи во табелите со веројатности се сметаат како непознати тежини, и аналогно со невронските мрежи, тие можат да се применат за да го најде оптималното множество од тежини (веројатносни вредности) за дадени податоци.

За користење на Баесова мрежа како класификатор, единствено треба да се пресмета argmaxy P(y | x) со користење на дистрибуција P(U) претставена со Бајесовата мрежа:

$$P\ (y|x) = P(U) / P(x)$$

$$\alpha P(U) = \prod_{u \in U} p(u | pa(u))$$

и бидејќи сите променливи x се познати, не ни требаат дополнителни алгоритми, само пресметување на оваа формула за сите класни вредности.

5.3 Дрва на одлучување

Дрвата на одлучување се моќна и полуларна техника која се користи и за класификација и за предвидување. Атрактивноста на методите базирани на дрва на одлучување во голема мерка се должи на фактот што тие всушност претставуваат правила кои може многу лесно да се изразат на природен јазик за да може луѓето да ги разберат. Тие правила исто така може да се изразат во SQL изрази со цел да се извлечат податоци од одредена категорија..

Дрво на одлучување претставува структура која што може да се користи за да се подели голема колекција на податоци во повеќе мали множества од податоци со користење на едноставни правила за одлучување. Со секоја следна поделба членовите на добиените множества стануваат послични помеѓу себе.

Моделот на дрво на одлучување се состои од множество на правила за поделба на хетерогена популација во помали хомогени групи во однос на некоја целна променлива. Дрво на одлучување може да се конструира

автоматски со примена на еден од неколкуте алгоритми за дрва на одлучување врз множеството на моделот кое се состои од прекласифицирани податоци. Целната променлива обично е од категориски тип. Моделот на дрвото на одлучување се користи за да се пресмета веројатноста дека даден запис припаѓа на секоја од категориите или да се класифицира одреден запис со негово доделување на класата на која тој најверојатно припаѓа.

Дефиниција на дрво на одлучување. Нека е дадена база на податоци $D= \{t1,t2,\ldots,tn\}$ каде $t_i=\{t_{i1},t_{i2},\ldots,t_{in}\}$ и шемата на базата на податоци ги содржи следниве атрибути $\{ A_1,A_2,..A_n\}$. Нека е дадено и множеството на класи $C= \{C_1 ,..,Cm\}$. **Дрво на одлучување** (Decision Tree – DT) или **класификациско дрво** е дрво поврзано со D, што ги има следниве својства [39]:

- Секој внатрешен јазел се означува со атрибут A_i
- Секоја гранка се означува со предикат што може да се примени врз атрибутот поврзан со родителот;
- Секој лист се означува со класа C_j .

Податокот влегува во дрвото преку коренот на дрвото. Коренот на дрвото применува тест за да одреди во која насока да го насочи податокот односно кое дете- јазол треба понатаму да го обработува податокот.

Постојат различни алгоритми за избор на почетниот тест, но целта е секогаш иста: Да се избере тестот кој најдобро ги разликува целните класи. Овој процес се повторува додека податокот не пристигне до јазол кој претставува лист на дрвото. Сите податоци кои пристигнуваат на одреден лист на дрвото се класифицирани на истиот начин. Постои единствен пат од коренот на дрвото до секој лист. Тој пат го претставува правилото кое се користело за да се класифицира податокот на начинот на кој е класифициран. Различни листови може да ја прават истата класификација, меѓутоа до секој лист се стигнува по различен пат. .

Една од негативните работи кај дрвата за одлучување е тоа што најчесто работат со атрибути именки, а повеќето множества содржат нумерички атрибути, а и проблем се бројот на вредностите кои недостасуваат во множествата при нивната обработка, бидејќи во тој случај не може да се одреди точната гранка по која би се движеле по дрвото.

Најважните фактори кои влијаат за градење на дрвата се големината на множеството што се третира и начинот како се определува најдобриот атрибут

за поделба. Мерките за избор на атрибути исто така се познати како правила за поделба, бидејќи тие одлучуваат на кој начин да се поделат податоците во даден јазол.

Алгоритмите за индукција на дрвата за одлучување можат да го изградат дрвото и потоа да извршат негово поткастрување за класификацијата да биде поефикасна. Со техниките на поткастрување, делови од дрвото можат да се отстранат или да се комбинираат за да се редуцира вкупната големина на дрвото. Поткаструвањето може да се изведува додека се креира дрвото, со тоа да се избегне дрвото да стане премногу големо, или вториот пристап е да се поткастри дрвото по градењето. Временската и просторна комплексност на алгоритмите за индукција на дрва за одлучување зависи од големината на податоците за обучување q, бројот на атрибути h и обликот на креираното дрво. Во најлош случај, дрвото може да биде многу длабоко и да не е многу разгрането. Додека се гради дрвото, за секој од јазлите, секој атрибут ќе биде испитан за да се определи дали тој е најдобриот атрибут за поделба. Тоа доведува временската комплексност на дрвото на дрвото да биде $O(hq \log q)$. Времето што е потребно да се класифицира база со големина n се базира на висината на дрвото.

Постојат повеќе алгоритми за конструкција на дрва на одлучување меѓу кои спаѓаат ID3, C4.5, CART, CHAID и др. Повеќето алгоритми за креирање на дрва на одлучување го користат пристапот на top-down дизајн кој започнува со множество на тренирачки подредени листи на вредности и соодветните класи на кои тие им припаѓаат.Тоа множество на тренирачки вредности рекурзивно се дели на помали множества со текот на градењето на дрвото

5.3.1 ID3 Алгоритам

Алгоритмот ID3 врши испитување на сите кандидати атрибути и го избира оној атрибут A што ја максимизира информациската добивка $G(A)$, го конструира дрвото, и потоа го користи истиот процес рекурзивно за да конструира дрва за одлучување за останатите подмножества $\{ D_1,..,D_N\}$. За секое подмножество D_I ако сите примероци od D_i се позитивни, се креира ,,yes” лист и се запира; ако сите примероци се негативни се креира ,,no” лист и се запира; во секој друг случај се избира друг атрибут на истиот начин како и опишаниот и се прави нова поделба.

Една од главните предности на ID3 е тоа што не е потребно некакво претходно знаење за проблемот. Тоа значи дека алгоритмот може да се примени на секое синтаксички добро формирано множество за обучување.

5.3.2 C4.5 Алгоритам

C4.5 е проширување на ID3 кој овозможува работа со вредности за атрибути кои недостасуваат, континуирани домени, поткастрување на дрвото на одлучување, изведување правила и др.

Проширувањето на алгоритмот се состои од воведување на бинарна поделба за континуалните атрибути. Прагот за поделба обично се поставува на средина меѓу вредностите кои ги даваат границите на концептот. Кога дрвото за одлучување тестира еден дискретен атрибут имаме по една гранка за секоја вредност на атрибутот. Бидејќи поделбата кај континуалните атрибути е само бинарна, постои разлика меѓу континуалните и нумеричките атрибути: откако ќе се изврши гранење на еден номинален атрибут се искористува неговата целокупна информација, а кај континуалните атрибути при секоја поделба имаме потреба од нова информација. Според тоа, еден дискретен атрибут може да се тестира само еднаш во патеката од коренот до еден лист, додека нумерички атрибут може да се појави повеќепати.

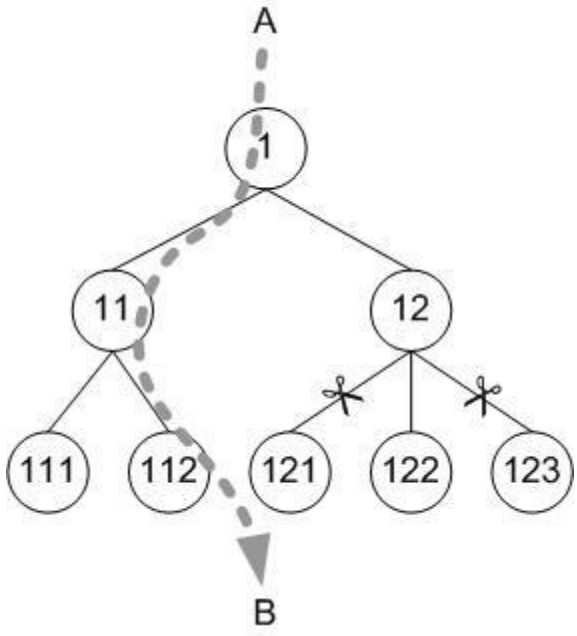
Во случај кога го применуваме дрвото за класификација на инстанца во која недостасуваат вредности за одредени атрибути ја изведуваме следната постапка: се дели инстанцата на делови со користење на нумеричка тежинска шема и испраќаме дел од инстанцата по секоја од гранките пропорционално на бројот на инстанци за обучување во соодветната гранка; различни делови од инстанцата ќе пристигнат до листовите и одлуките потоа мораат да се рекомбинираат со користење на истите тежини за да се добие резултантната класа.

Кај C4.5 постојат две стратегии за поткастрување и тоа: **пост-поткастрување (postpruning**) и **пред-поткастрување (prepruning) [39] .** Ако вршиме индукција на комплетното дрво и потоа вршиме негово поткастрување тогаш ја користиме стратегијата на пост-поткастрување.

Пред-поткаструвањето се користи кога сакаме да запреме со индукцијата на поддрва во текот на градењето на дрвото. Кај алгоритмите за индукција на дрва за одлучување најчесто се користи стратегијата на пост-поткастрување.

За пост-поткаструване постојат две значително различни операции и тоа: замена на поддрво со лист (subtree replacement) и подигнување на поддрво (subtree raising). За секој јазол алгоритмот за учење мора да одлучи дали ќе изврши замена на поддрво со лист, ќе изврши подигнување на поддрвото или ќе го остави дрвото непоткастрено (непроменето).

Идејата на замената на поддрво со лист е да селектира некои од поддрвата и да ги замени со листови. Ваквата операција ќе доведе до намалување на точноста врз множеството за обучување ама може да доведе до зголемување на точноста ако дрвото се примени врз независно избрано множество за тестирање. Кај ваквото поткаструване се движиме од листовите кон коренот. Подигнувањето на поддрвото е покомплексна операција и не секогаш јасно дали се исплати да се користи. Во овој случај се заменува поддрво со неговото најчесто користено поддрво. Овде се подигнува поддрвото од тековната локација на јазол кој се наоѓа погоре во дрвото. Пример на изглед на дрва со поткаструване е даден подолу:



Слика 4. Изглед на дрво со поткаструване (преземено од http://antognini.ch/2008/12/what-are-hints/)

Figure 4. Tree with pruning review (taken from http://antognini.ch/2008/12/what-are-hints/)

5.3.3 CHAID алгоритам

CHAID алгоритамот е акроним за Chi-Squared Automatic Interaction Detection, кој го користи Ҳ квадрат тестот за наоѓање на следната најдобра

поделба на јазол. Овој алгоритам се користи за да се изгради предиктивен модел, врз основа на класификација, и е еден од најчесто користените алгоритми кои користат дрва за одлучување. Овој алгоритам врши детекција на интеракција, односно корелација меѓу променливите во множеството податоци. Постојат два вида на Ҳ квадрат тест врз основа на кој работи овој алгоритам, и двата се за категориски променливи. Х квадрат тестот за испитување на квалитетот на совпаѓање (goodness of fit) ја истражува пропорцијата на случаите кои спаѓаат во разни категории на една променлива и ги споредува со хипотетички вредности на тие проорции. Вториот, Ҳ квадрат тест за тестирање на независност одредува дали се поврзани две или повеќе категориски променливи. При тоа во алгоритамот се користат техниките за разделување и спојување на атрибутите кои се користат во анализата. За стопирање на разгранувањето на дрвото овој алгоритам користи статистички методи. Дозволува повеќекратно разгранување на дрвото и дава добар визуелен преглед на екран. При секое делење на дрвото, алгоритамот ја разгледува предикторската променлива која ако се подели понатаму дали подобро ќе ја опишува категоријата на целната променлива. Со цел да одлучи каде да изврши поделба врз основа на предикторската променлива, CHAID алгоритамот ја тестира хипотезата чија цел е зависноста меѓу поделената променлива и категориите од таа променлива. Ако тестот покаже дека целната и предикторската променлива се независни, алгоритамот го прекинува разгранувањето на дрвото. Во спротивно, поделбата се креира и се бара понатамошна најдобра поделба. Поради тоа, овој алгоритам е доста погоден за анализи каде целта е да се опише или разбере врската меѓу целната променлива и множеството променливи од кои таа зависи.

CHAID техниката бара влезните променливи да имаат дискретни вредности. Поделбата на атрибутите се врши со помош на р- вредности на Ҳ-квадрат дистрибуција. Р вредноста се корегира за да се овозможи споредба на повеќе различни поделби. Изборот на најдобрата поделба оди чекор по чекор. Прво на секоја вредност на влезната променлива и се доделува посебна гранка од дрвото. Потоа гранките се спојуваат и повторно делат во зависност од р вредноста. Оригиналниот Kass- ов CHAID алгоритам запира кога никакво делење или спојување не ја дава оригиналната р вредност. Техниката на спојување на гранки продолжува се додека не се постигне поделба на две

гранки. Тогаш, помеѓу сите поделби се избира онаа со најголема p вредност. Откако за секоја влезна променлива ќе се избере најдобра поделба, се врши корекција на p вредноста и тогаш од сите поделби се избира онаа со најголема p вредност. До поделба доаѓа само ако p вредноста е помала од зададената. Конструкцијата на дрвото завршува кога сите корегирани p вредности кои се разгледуваат за поделба се поголеми од дозволените, зададените.

5.4 Невронски мрежи

Невронските мрежи се моќна општонаменска алатка која се користи за предвидување, класификација и кластерирање. Најмоќните невронски мрежи се биолошките. Човечкиот мозок им овозможува на луѓето да генерализираат од своите искуства. Компјутерите, од друга страна, настојуваат да следат низи од експлицитни инструкции. Невронските мрежи го надополнуваат овој недостаток со моделирање на невронските конекции од човечкиот мозок на компјутер.

Со тренирање на невронската мрежа се гради модел кој што потоа може да биде искористен за да се пресмета резултатот кај случаи каде излезниот резултат не е познат. Процесот на тренирање на невронските мрежи е всушност процес на подесување на тежините со цел да се постигне најдобра комбинација на тежини за пресметување на посакуваните резултати. Мрежата прво се креира со произволни тежини па почетните пресметки кои поприлично се разликуваат од точните резултати. Меѓутоа тренирачките вредности се репроцесираат со цел намалување на излезната грешка, па така невронската мрежа постепено дава сè поточни пресметки на излезните променливи од тренажното множество. Тренажниот процес завршува кога пресметките повеќе не напредуваат во својата точност.

Невронските мрежи се состојат од јазли или единици кои директно се поврзани. Секоја единица која одговара на еден неврон прво ја пресметува тежинската сума на своите влезови, а потоа применува т.н активациона функција која треба да задоволи две основни барања [34]:

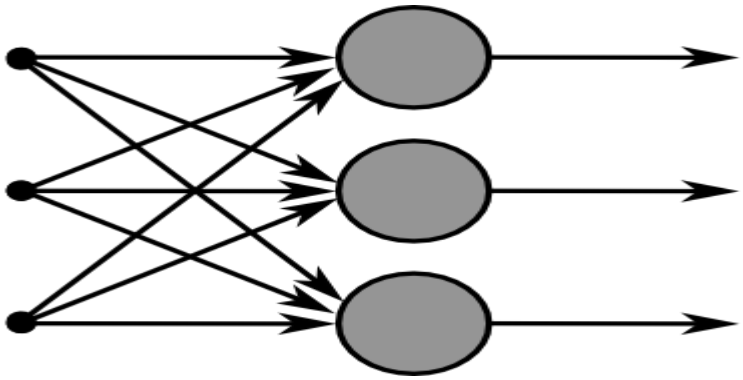
- Единицата- неврон треба да биде активна кога се присутни одредени пожелни влезови, како и да не биде активна кога се присутни одредени непожелни влезови

- Активационата функција треба да биде нелинеарна

Постојат четири видови на невронски мрежи и тоа [34]:

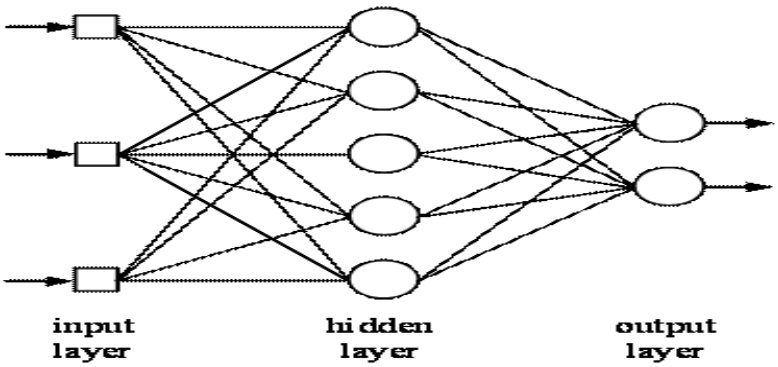
- Еднослојни мрежи без повратни врски (single-layer feedforward networks);
- Повеќеслојни мрежи без повратни врски (multi-layer feedforward networks);
- Мрежи со повратни врски (recurrent networks);
- Скалести мрежи (lattice structures);

Подолу следи графички приказ на видовите невронски мрежи:



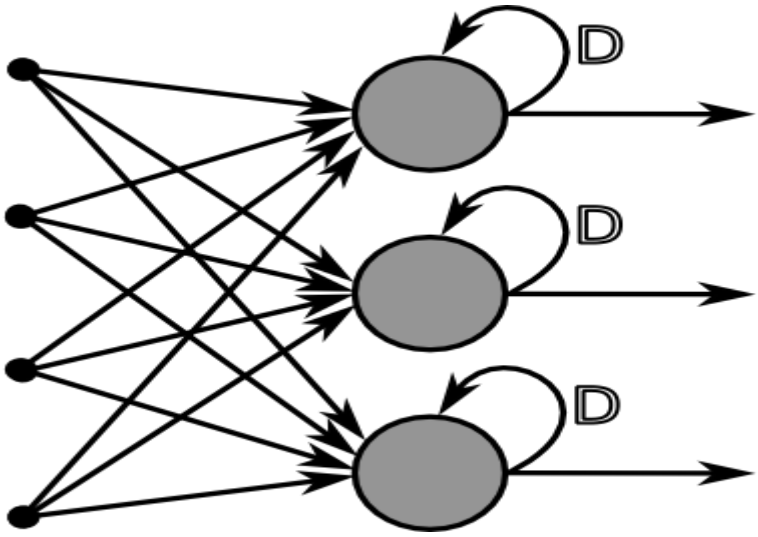
Слика 5. Еднослојна мрежа без повратни врски (преземено од http://en.wikibooks.org/wiki/Artificial_Neural_Networks/Feed-Forward_Networks)

Figure 5. Single-layer feedforward networks (taken from http://en.wikibooks.org/wiki/Artificial_Neural_Networks/Feed-Forward_Networks)



Слика 6. Повеќеслојна мрежа без повратни врски (преземено од http://ljs.academicdirect.org/A15/053_070.htm)

Figure 6. Multi-layer feedforward networks (taken from http://ljs.academicdirect.org/A15/053_070.htm)



Слика 7. Мрежа со повратни врски (преземено од http://en.wikibooks.org/wiki/Artificial_Neural_Networks/Recurrent_Networks)

Figure 7. Recurrent network (taken from http://en.wikibooks.org/wiki/Artificial_Neural_Networks/Recurrent_Networks)

Невронските мрежи наоѓаат голема примена за прогноза на идни трендови. Тие даваат одговор на прашања како на пример: Ако цената на артикалот X се намали за одреден процент, за колку ќе се зголеми побарувачката за тој артикал и слични прашања.

5.5 Метод на потрошувачка кошница

Методот на потрошувачка кошница во своето основно значење се сведува на откривање на асоцијативни правила кои ги покажуваат паровите на артикли и нивото на веројатноста дека ќе бидат купени заедно. Ова асоцијативно правило математички може да се опише како X>=Y, каде x∩y=0. Нивото на поддршка на множеството артикли X се изразува како однос на трансакциите кои содржат множество трансакции во однос на вкупниот број трансакции.

Сигурноста може да се дефинира како процент од трансакциите, кои ако го содржат артикалот X, го содржат и арикалот Y, во однос на сите трансакции кои го содржат X, што се претставува на следниов начин:

Сигурност(X>=Y)=поддршка (XUY)/поддршка (X).

Користејќи ја условната веројатност, сигурноста може да се изрази како:

Сигурност(X>=Y)=P(Y/X)=P(X∩ Y)/P(X)

Методот на анализа на потрошувачката кошница ги открива скриените правила во многубројните трансакции кои се однесуваат на продажбата на робата. Овој метод предупредува на веројатноста да, ако купувачот купи производ X, ќе купи и производ Y. Многубројните информации кои се користат за оваа анализа се во базите на податоци, каде секоја сметка претставува еден купувач, а малопродажната сметка покажува на множеството производи кои овој купувач ги купил во малопродажба. Ваквите анализи не се ограничени само на производите, туку предмет на анализите може да бидат и добавувачите и производителите, или комбинација на производите со добавувачите или производителите.

Овој метод дава можност за избор на одредено ниво на слобода при утврдувањето на правила со оглед на веројатноста која се појавува во трансакциите, а кои се поврзани со појавата на парови производи или други множества во трансакциите. При тоа, потребно е да се утврди долната граница на веројатност која влијае на формирањето на правила. Овој критериум е искористен во процесот на анализа заедно со примена на *a priori алгоритамот*, кој најчесто се користи во процесот на анализа на потрошувачката кошница. Основната претпоставка на овој метод е генерирање на правила во форма:

ако a тогаш b

ако c и d тогаш e,

каде променливите a, b ,c, d и e претставуваат производи. Секоја од овие променливи може да се појавува со одреден степен на сигурност. Во случај кога го имаме правилото: ако a тогаш b, степенот на сигурност покажува колкава е веројатноста на појавување на променливата b во трансакциите, ако во трансакциите учествува променливата a. Најсигурно е она правило кое има највисок степен на веројатност. A priori алгоритамот ги елиминира вредностите кои не се појавуваат во доволен број на трансакции, со што се намалува големината на примерокот за обработка.

Gri алгоритамот е, исто така, алгоритам кој се користи кај методот на потрошувачка кошница и применува асоцијативни правила како и apriori алгоритамот. Асоцијативните правила кај овој алгоритам се појавуваат во облик: Ако претходен, тогаш последица на следен, односно претходната променлива е последица на следната. Овој алгоритам врши екстракција на

множество од правила од податоците, извлекувајќи ги правилата со најголема вредност на информацијата, односно со највисока содржинска информација. Содржината на информацијата се мери со помош на индекс кој како правила ги зема поддршката и точноста на асоцијативните правила. За да се креираат *Gri* асоцијативни правила потребно е да се има една или повеќе влезни променливи, и една или повеќе излезни променливи. Правилата извлечени со овој алгоритам е мошне лесно да се интерпретираат. Некои од записите во алгоритамот може да предизвикаат повеќе правила. Овој алгоритам добро се справува со излез на повеќе променливи и за разлика од apriori, *Gri* се справува со нумерички и атрибути именки.

Покрај *apriori и GRI алгоритамот* во процесот на анализа на потрошувачката кошница се користи и *методот на дрва на фрекфентни примероци* за решавање на проблеми со комбинаторна експлозија, причинети од растот на бројот на можни комбинации. Со оваа метода со поминување низ трансакциската база се бележат фреквенциите на појавување на артиклите во базата и се врши сортирање врз основа на фрекфенцијата на појавување, а се занемаруваат нефрекфентните артикли. После сортирањето се пристапува кон градење на дрва на примероци.

Методот на потрошувачката кошница е наменет, пред сè за откривање на законитостите во купувањето на групи артикли во малопродажба со цел зголемување на продажбата, заедно со маркетингшки потези, како што е давање попуст за продажба на производот Y, под услов за купување на производот X, при што продавачот пронаоѓа интерес во зголемување на коефициентот на обрт, а со остварување на помалку провизија. Овој метод ја користи веројатноста за предвидување на однесувањето на потрошувачот при купувањето, па затоа овој метод може да се искористи за препознавање на моделот на однесување на потрошувачот при купувањето, или препознавање на интересите на корисникот на картичка, преку сегментација на потрошувачот по одредени критериуми и следење на нивното однесување во текот на времето. Со овој метод исто така се утврдува дека артиклите кои со анализа се покажало дека се купуваат заедно, односно во парови е добро да бидат поставени заедно на полиците, сè заради зголемување на коефициентот на профит.

Асоцијативните алгоритми, освен во класичен облик на анализа на фрекфентноста на појавување во парови, може да бидат користени и во негациска смисла, односно форма од типот: Ако артикал А, тогаш НЕ артикал В.

5.5.1 Групирање по сродност или асоцијативни правила

Задачата на групирањето е да одреди кои нешта одат заедно, односно укажуваат на тоа колку често некои случувања се појавуваат заедно. Најчест пример кој се дава во овој контекст е да се одреди кои артикли припаѓаат заедно при рекламирање на промоции на производи во супермаркет. Афинити групирањето претставува едноставен пристап за генерирање на правила од податоците што ги имаме на располагање.

5.6 Проценка

Во практика, проценката често се користи за да се изврши задачата на класифицирање. Пристапот на проценка има голема предност во тоа што записите кои се проценети може да се рангираат во зависност од проценката која им е доделена.

Примери на задачи за проценка се:

- Проценка на бројот на деца во одредена фамилија
- Проценка на вкупниот приход на дадено домаќинство
- Проценка на веројатноста дека одреден клиент ќе одговори на рекламна кампања
- Проценка за тоа кој производ најмногу се бара итн

Моделите на регресија и невронските мрежи се прикладни техники за извршување на вакви задачи. Истите ќе бидат прикажани со визуелни резултати во овој магистерски труд.

5.7 Кластерирање

Кластерирањето има за задача да сегментира одредена хетерогена популација во повеќе хомогени подгрупи или кластери. Кластерирањето се разликува од класификацијата по тоа што не зависи од предефинирани класи за да ја изврши својата задача. Кај класификацијата на секој запис му е доделена предефинирана класа врз основа на модел кој е развиен со тренирање на прекласифицирани примери. Кај кластерирањето записите се

групираат заедно врз основа на сличноста помеѓу нив. Крајниот корисник треба да одлучи кое значење треба да го додели на кластерите кои се добиле врз основа на извршената анализа. Кластерирањето често се извршува како предзадача на некоја друга форма на data mining или моделирање.

Проблемот кој најчесто се појавува кај кластерирањето е тоа што најчесто се добиваат поголем број одговори на зададениот проблем и тешко е да се утврди точниот број на кластери кој одговара на проблемот. Друг проблем е тоа што и откога кластерирањето ќе заврши со формирање на множество кластери, точното значење на секој кластер не мора да биде очигледно и затоа во кластерирањето се користат повеќе техники и алгоритми.

Најчесто користен алгоритам при кластерирањето е K-means алгоритамот кој работи на следниов начин: Секоја точка чиј центар е најблизок се доделува на групата. Центарот (центроид) е точка која е добиена со аритметичка средина за секоја димензија од точки поодделно.

Пример: множеството податоци има три димензии, а групата има две точки : $X = (x_1, x_2, x_3)$ i $Y = (y_1, y_2, y_3)$. **Тогаш центарот Z е центар:** $Z = (z_1, z_2, z_3)$, каде $z_1 = (x_1 + y_1)/2$ and $z_2 = (x_2 + y_2)/2$ and $z_3 = (x_3 + y_3)/2$.

Потоа пак се пресметува центарот-средината на новите групи и се повторува се додека некое доделување на некоја точка не е променето.

5.8 Предвидување

Во принцип, предвидувањето е исто како и класификацијата или проценувањето со таа разлика што записите се класифицираат според одредено однесување во иднината или според одредена вредност која ја карактеризира некоја идна ситуација. Кај задачите на предвидување единствениот начин да се провери точноста на класификацијата е да се чека и да се набљудува. Главната причина за третирањето на предвидувањето како засебна задача од класификацијата и проценката е тоа што во моделите на предвидување постојат додатни проблеми кои ги засегаат привременитe врски помеѓу влезните променливи или предикторите и излезните променливи. Било која од техниките која се користи за класификација или проценка може да се прилагоди за извршување на задачи за предвидување со користење на примероци за тренинг каде што вредноста на променливата која што треба да се предвиди е веќе позната.

Изборот на техники кои што ќе се употребат за извршување на задачи од ваков тип зависи од природата на влезните податоци, типот на вредноста која треба да биде предвидена и важноста која му се придава на објаснувањето кое произлегува од предвидената вредност.

5.9 Профилирање

Понекогаш целта на податочното рударење е едноставно да опише што се случува во комплицирана база на податоци на начин што го зголемува нашето разбирање на процесите што ги произвеле податоците во базата на податоци. Доволно добар опис на однесувањето често сугерира објаснување за проблемот што се разгледува. Во најлош случај, добар опис ќе ни даде идеја каде да започнеме со барање на објаснување.

Одлучувачките дрва се моќна алатка за профилирање. Асоцијациските правила и кластерирањето исто така може да се употребат за профилирање.

6. Алатки за податочно рударење

Програмите за податочното рударење се појавуваат во различни облици, и тоа како **самостојни програми** кои може да подржуваат само еден метод, како **дел од некој програмски систем** за развој на апликации, како **самостојни алатки за рударење податоци**, како **модул од други програмски системи,** како систем за управување со базите на податоци, или како **дел од статистички пакети**, или пак како **готови решенија за поединечни проблемски подрачја.**

Одредени програмски пакети имаат опција за читање формат на податоци на туѓо софтверско решение, но ако таа опција не постои, податоците секогаш може да бидат разменети во класичен текстуален формат. Пред податоците да бидат подготвувани за вчитување во друг софтверски пакет, многу е важно истите претходно да бидат организирани на начин кој ќе биде адекватен за прифаќање.

Алатките за податочно рударење извршуваат анализа на податоци и може да откријат важни модели на податоци, кои придонесуваат во голема мера за бизнис стратегии, бази на знаење, и научни и медицински истражувања. Проширувањето на јазолот меѓу податоците и информациите

повикува на систематски развој на алатките за податочно рударење кое ќе ги претвори „податочните гробници“ во „златни грутки" на знаење.

Постојат голем број на програмски алатки за рударење на податоците присутни на светскиот пазар, произведени од реномирани водечки фирми. Поделени се според областа каде се применуваат и тоа:

- *Алатки како дел од статистички програмски пакети*: Enterprise Miner (SAS), Clementine (SPSS), Statistika и др.
- Специјализирани алатки за општа или бизнис употреба: Intelligent Miner (IBM), Data Miner (SAS), SPSS Statistics и др.
- *Алатки вклучени во OLAP процесите:* Hiperion, Pentaco, IBM Cognos 8 Business Intelligence (BI) и др.
- *Алатки вклучени во системот за управување со податоци*: Microsoft SQL Server Business Intelligence, Darwin (Oracle) и др.
- *Математички програмски пакети*: MathLab, Matematica и др.

Покрај погоре наведените алатки постојат и многу други на пазарот, како: Advanced Miner, Affinium Model, DataDetective, DataLab, Kalidara Advisor, XLMiner и др. Водечките алатки за рударење на податоците како: Enterprise Miner (SAS), SPSS Statistics, Clementine (SPSS), Modeler, Darwin (Oracle), и др. се доста скапи за одредени фирми и поединци и чинат неколку десетина илјади долари. Фирмите кои не можат или не сакаат да вложат во софтвер за ваква намена обично користат слободен софтвер. Познати алатки кои се слободен софтвер за податочно рударење се: WEKA, Orange, R, Tanagra, Rapid Miner, KEEL, KNIME, MiningMart, MLC++ и др.

Според истражувањата направени од страна на KDD Nudgets и RexerAnalytics во мај 2010 година луѓето кои се вклучени во процесот на податочно рударење на прашањето: Која алатка за податочно рударење ја користат во последните 12 месеци? се изјасниле вака:

Табела 2. (Преземена од http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html)

Table 2. Taken from http://www.kdnuggets.com/polls/2010/data-mining- analytics-tools.html)

RapidMiner (345)	37.8%
R (272)	29.8%
Excel (222)	24.3%
KNIME (175)	19.2%
Your own code (168)	18.4%
Pentaho/Weka (131)	14.3%
SAS (110)	12.0%
MATLAB (84)	9.2%
IBM SPSS Statistics (72)	7.9%
Other free tools (67)	7.3%
IBM SPSS Modeler (former Clementine) (67)	7.3%
Microsoft SQL Server (63)	6.9%
Statsoft Statistica (57)	6.2%
Other commercial tools (56)	6.1%
SAS Enterprise Miner (50)	5.5%
Zementis (34)	3.7%
Orange (25)	2.7%
Oracle DM (19)	2.1%
KXEN (19)	2.1%
Salford CART Mars other (15)	1.6%
VisuaLinks (12)	1.3%
Viscovery (10)	1.1%
Angoss (8)	0.9%
TIBCO Insightful Miner (7)	0.8%
Miner3D (7)	0.8%
REvolution Computing (4)	0.4%
Megaputer Polyanalyst/TextAnalyst (3)	0.3%
Portrait Software (2)	0.2%
Data Applied (2)	0.2%
Centrifuge (2)	0.2%
PRSD Studio (1)	0.1%
Clario Analytics (1)	0.1%
Bayesia (1)	0.1%

Првите топ 5 пак на листата на RexerAnalytics биле:

Табела 3. Топ листа на алатки за податочно рударење според RexerAnalytic во 2010 година (преземена од http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html

Table 3. Tools charts of data mining according to Rexer Analytic in 2010 (taken from http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html)

1	**R**
2.	SAS Enterprise Miner
3.	IBM SPSS Statistics
4.	IBM SPSS Modeller
5.	WEKA

Ако се направи една споредба со истражувањата на овие фирми од 2007 година ќе се види дека постои разлика во првите топ 10 програмски пакети за податочно рударење, односно во мај 2007 година топ листата била следна:

Табела 4. Топ листа на алатки за податочно рударење според RexerAnalytic во 2007 година (преземена од http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html)

Table 4. Tools charts of data mining according to Rexer Analytic in 2007 (taken from http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html)

1.	**SPSS Clementine**
2.	Salford System
3.	Yale (cera RapidMiner)
4.	SAS Enterprise miner
5.	Angos Knowledge Studio
6.	KXEN

7.	WEKA
8.	R
9.	Microsoft SQL Server
10.	MatLab

Rapid Miner претходно познат како YALE е слободен софтвер кој има широк ранг на техники кои се користат во податочното рударење. Изграден е врз основа на WEKA алатки и дава добри методи на визуелизација на резултатите.

SAS Enterprise Miner е аналитички софтвер кој користи клиент-сервер архитектура врз основа на JAVA платформа која дозволува паралелни процесирања. Тој поддржува широк ранг на техники за податочно рударење како: дрва на одлучување, невронски мрежи, анализа на потрошувачка кошница, регресија и др.

R е слободен софтвер кој спаѓа во статистичките пакети користени во рударењето на податоци, кој најчесто се користи во биоинформатиката и социјалните науки.

IBM SPSS Statistics SPSS (Statistical Package for the Social Sciences) е една од најраспространетите програми за статистичка анализа пред сè во социјалните науки. Се користи од страна на голем број маркетинг организации, државни институции и компании за истражување на пазарот, образованието, здравјето и др. SPSS е програма која исто така се користи во податочното рударење, анализа на текст со цел да се добијат релевантни податоци од извршените анализи.

IBM SPSS Modeller порано познат како **Clementine** е статистички пакет во кој се поддржани голем број техники на податочното рударење. Се користи како модел за предвидување во многу области како трговија, продажба користејќи одредени линк анализи.

WEKA е пакет за рударење на податоци, кој е слободен софтвер и вклучува широк ранг на техники. Лесниот кориснички интерфејс овозможува флексибилност во анализите. Во него се вклучени голем број алгоритми за машинско учење и тоа го прави погоден за многу академски истражувања.

Алатките за податочно рударење е потребно да се избираат во зависност од проблемот кој треба да се реши, бидејќи секоја од алатките има предности и мани во решавањето на одредени проблеми. Пример, Clementine како алатка за податочно рударење во праксата се покажал како една од најдобрите алатки со голем број на функции кои се од големо значење за податочното рударење. Weka, RapidMiner, R и др. иако се меѓу водечките алатки за податочно рударење, сепак во споредба со Clementine заостануваат во одредени карактеристики. Но, не секогаш и секоја алатка е применлива за одредени множества податоци, и затоа корисникот ја бира алатката која ќе може да се примени врз дадениот случај.

Водечки компании во производството на алатки за податочно рударење се SAS, SPSS, IBM, Microsoft и др.

7. Експериментален дел

7.1 Користени податоци во магистерскиот труд

Податоците користени во изработката на овој магистерски труд се добиени од „Компанија Моневи“, македонска фирма која се занимава со трговија на мало и големо. Во текот на изработката на овој труд беа посетени и други фирми, но при разгледувањето на базите на податоци кај истите, утврдено е дека „Компанија Моневи“ има многу добар имплементиран софтвер, односно база во која се чуваат огромни количини на податоци, и истите ми беа дадени на располагање. Дел од податоците се од доверлив карактер, така да истите се обработени без нарушување на доверливоста.

„Компанија Моневи“ во својата база на податоци располага со финансиски податоци од правни и физички клиенти. Правните клиенти се од територијата на Република Македонија. Претпријатието „Компанија Моневи“ располага со Дисконт за сувомеснати производи на големо, продавница за колонијални производи и мини ресторан. Има 32 вработени, од кои 4 го сочинуваат менаџерскиот тим.

Од базата на податоци утврдено е дека „Компанија Моневи“ работи со правни лица од територијата на Р.Македонија. Во дисконтот на сувомеснати производи има 100 различни артикли, додека во продавницата за колонијални производи околу 6000 артикли. Сите податоци кои се користени во анализата се добиени од базата на податоци во фирмата, но за потребите на анализите е

изготвен и анкетен лист, а обработени се и фискални сметки од купувачите. Во фирмата постои многу добро изграден склад на податоци, и со разговорот на менаџерот на фирмата е утврдено дека се вложени доста пари во софтверот, кој секако има пресудна улога за работата на фирмата. Софтверот користен во фирмата не е софтвер за податочно рударење, но сепак исцрпените податоци од него, а и од другите анализи, ќе бидат доволни за да се применат техниките на податочното рударење користејќи некои од алатите за податочно рударење.

При разгледувањето на податоците од фирмата утврдено е дека истата бележи раст во последните три години, но при разговорот со менаџерскиот тим на истата утврдено е дека конкуренцијата на пазарот е сè поголема, односно во средината која работи фирмата се појавиле многу големи конкурентски фирми и потребно е донесување на правилни и подржани бизнис одлуки во борбата со конкуренцијата.

Бидејќи фирмата располага со магацин на големо, продавница на мало и мини ресторан, при разгледувањето на софтверот се гледа дека базата на податоци е сместена на сервер, но може да биде и разгледувана локално. Сите податоци кои се наоѓаат во базата може да бидат експортирани во Excel табели и истите се многу добро документирани. Кои податоци се важни за анализа, а кои не е утврдено при разговор со менаџерскиот тим на фирмата. При анализата на податоците утврдив дека во базата на податоци не се чуваат значајни податоци за физички лица, туку само за правни лица, а бидејќи продавницата работи исклучиво со физички лица, анализите се насочени токму на физичките лица, односно на купувачите во продавницата на мало. За таа цел спроведена е анкета на купувачите во продавницата, и анализирани се фискални сметки од истата.

Анкетата во продавницата се спроведуваше речиси 30 дена со цел да бидат опфатени различни профили на купувачи. За истите купувачи веднаш се нумерираа и нивните фискални сметки кои исто така ќе бидат анализирани. Анкетата се спроведе во периодот мај-јуни 2011 година. По спроведената анкета се издадоа карти на лојалност на целната категорија купувачи која се доби преку техниките на податочно рударење. Потоа се користеа сите податоци од базата преку SQL Server 2008 во периодот од 1.01.2011 до 31.12.2011. На истите се применија одредени техники на податочно рударење со цел да се направи споредба што се случува со продажбата пред и по

моментот на издавање на карта на лојалност на купувачите. Картата на лојалност послужи како модел за вреднување на купувачите.

7.2 Практична имплементација на податоците од анкетниот лист

За потребите на анализите за клиентот купувач е изготвен и анкетен лист, кој анонимно е пополнет од страна на 120 купувачи во подружницата продавница со која располага фирмата. Со податоците добиени од анкетниот лист ќе бидат направени анализи за купувачите во продавницата, со што се очекува да биде подобрена работата на истата. Анкетниот лист е прикажан во Прилог 1 од овој магистерски труд.

Во бизнисот податочното рударење најчесто се користи во подрачјето на маркетингот, кој се повеќе и повеќе е насочен кон поединечниот купувач и управување на односите со купувачите (Customer Relationship Managment-CRM), односно создавање, одржување и подобрување на односите со купувачите. Целта е придобивање на нови купувачи и задржување на веќе постоечките преку согледување на желбите и потребите на купувачот, сегментација на профилот на купувачот, разбирање на неговото однесување и предвидување на идното однесување..

Знаењето кое на фирмата може да и обезбеди опстанок на пазарот се однесува токму на разбирањето на купувачите, односно разбирање на нивните потреби, а со тоа може да се обезбеди и зголемена лојалност на купувачите. Со помош на извршената анкета ќе се дојде до целната категорија купувачи на кои треба да и биде издадено карта на лојалност, а која пак карта на лојалност ќе биде употребена во следни анализи.

Успешно спроведување на анкетниот лист е еден од пресудните фактори од кои понатаму ќе зависи и успехот на проектот со кој е потребно да се изврши анализа на податоците. Една од најважните работи во составувањето на анкетниот лист е тоа што е потребно да се постават прашања со кои ќе се искристализираат клучните индикатори и категории важни за анкетирањето.

Во овој анкетен лист се обидуваме да добиеме одговор на прашањата од типот: Каков е профилот на купувачот во продавницата, што преферира да купува, кому е потребно да се издаде карта на лојалност, кои артикли би требало да бидат на акција и др.

Најголем дел од прашањата во анкетните листови се поделени на два дела: затворени и отворени. Затворено прашање е кога на анкетираните им се понудуваат повеќе оданпред дефинирани одговори. Од нив се бара да се заокружи еден од понудените одговори. Понудените одговори обично се од видот Да/Не, машко/женско и.т.н. Во нашиот анкетен лист од вкупно 14 прашања 11 се од затворен вид. Но, понекогаш не можеме да претпоставиме кои одговори би ги дале анкетираните личности, па затоа се задаваат и отворени прашања. Во анкетниот лист ние имаме 3 отворени прашања.

Екстракцијата на податоци во овој дел значи е добиена од одговорите на прашањата од анкетниот лист. Анкетниот лист се спроведуваше повеќе денови во просториите на самата продавница за колонијални производи. Иако беше анонимен, сепак голем дел од купувачите едноставно одбиваа да го пополнат поради недостиг на време. Можеби најдобар начин за спроведување на анкетата е испраќање на анкетниот лист во електронска форма до купувачите, но бидејќи во базата на податоци не постоеја податоци за физичките лица, како на пример нивна e-mail адреса, тоа беше едноставно неизводливо.

Како што е претходно спомнато во овој труд најголем дел од времето се потроши за трансформација на податоците од анкетниот лист, затоа што податоците требаше да се трансформираат во формат погоден за обработка во програмскиот пакет **SPSS Statistics,** во кој ќе се работи анализата на податоците од анкетниот лист и ќе се применат техниките за податочно рударење. Дел од податоците ќе бидат обработени и во програмскиот пакет **Clementine.**

Затворените прашања се претворија во нумерички формат кој е потребен за работа во SPSS Statistics. То значи дека сите понудени одговори се шифрираа со броеви, додека за отворените прашања беа разгледани сите анкетни листови со цел да може да се извлечат неколку категории, кои би можеле да се шифрираат со броеви.

Во врска со отворените прашања по разгледувањето на анкетните листови се забележа дека најголем број од анкетираните не ги одвоиле производите во група туку ги ставиле заедно и затоа направив 3 категории и тоа: прва категорија ќе бидат прехранбени артикли во кои ќе бидат ставени и лебот, млекото, зејтинот, месото, сувомеснатите и млечните производи, втора категорија ќе бидат хигиенските производи како прашок, омекнувач, средства

за чистење, трета категорија пијалоци во кои ќе припаѓаат безалхохолни и алхохолни пијалоци и четврта категорија останато во која ќе припаднат производите кои не се наведени во првите три категории. На ист начин ќе бидат и рангирани производителите на овие артикли.

Пред да почнеме со работа во SPSS Statistics, најпрво ги дефинираме имињата на променливите, а потоа и атрибутите за истите. Во нашиот случај променливи се:

1. **Redenbr**- променлива која ги дефинира редните броеви на анализираните случаи
2. **Vozrast**- се движи во интервал од 20 до над 50 години. Таа е поделена во четири категории кои се шифрирани со броевите од 1-4 и тоа:

Табела 5. Шифрирање на променливата vozrast

Table 5. Encoding the variable age

vozrast	шифрирање
20-30 години	1
30-40 години	2
40-50 години	3
Над 50 години	4

3. **M.prihod**- се движи во интервал од 0 до над 25000 денари. Исто така е поделена во четири категории кои се шифрирани со броевите од 1-4 и тоа:

Табела 6. Шифрирање на променливата м.prihod

Table 6. Encoding the variable monthly income

M.prihod	шифрирање
0-5000 денари	1
5000-15000 денари	2
15000-25000 денари	3
Над 25000 денари	4

4. **Bracna sostojba**- прима две вредности: самец шифрирано со 1 и оженет шифрирано со 2.

5. **Br.deca**- од 1-над 3 деца. Поделена е во 5 категории шифрирани со броевите од 1-5 и тоа:

Табела 7. Шифрирање на променливата br.deca

Table 7. Encoding the variable number of children

Br.deca	шифрирање
1 дете	1
2 деца	2
3 деца	3
Повеќе од 3 деца	4
Немам деца	5

6. **Smetka-** фискална сметка на анкетираниот купувач во категорија од 0-над 3000 денари. Поделена е во пет категории кои се шифрирани со броевите од 1-5 и тоа:

Табела 8. Шифрирање на променливата smetka

Table 8. Encoding the variable account

Smetka	шифрирање
0-500 денари	1
500-1000 денари	2
1000-2000 денари	3
2000-3000 денари	4
Над 3000 денари	5

7. Променливи p1 до p10 кои ги дефинираат одговорите на поставените прашања во анкетниот лист. Во зависност од дадените одговори, тие имаат различни вредности:

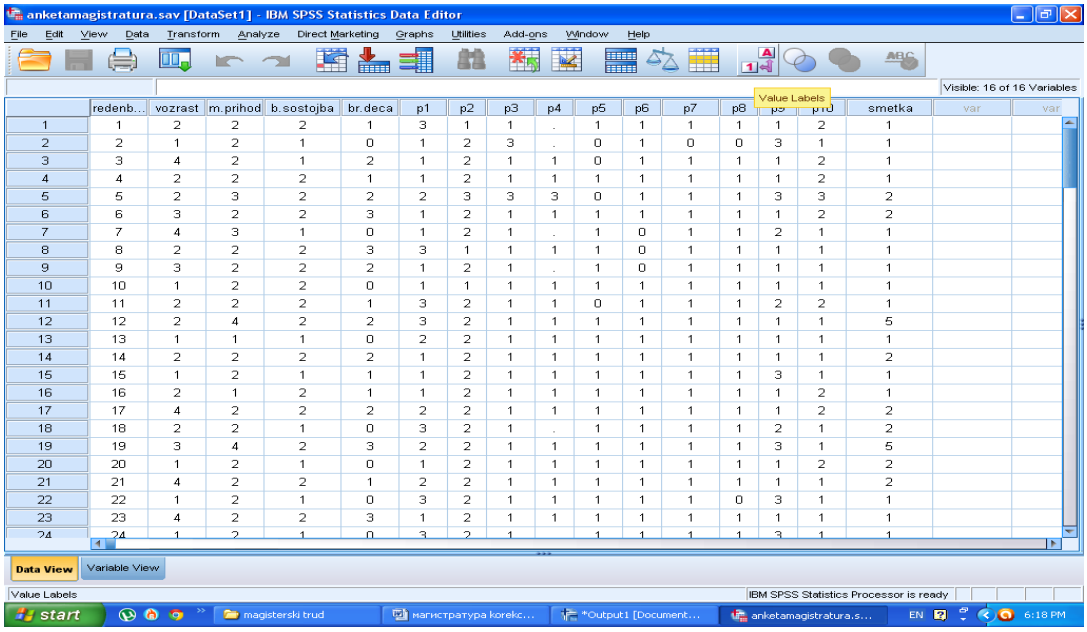
Табела 9. Променливи и нивно шифрирање

Table 9. Variables and their encoding

Одговор на прашање	Име на променлива	Шифирање
1. Колку често купувате производи (артикли) од	P1	Често-1 Многу често- 2

нашата продавница?		Ретко- 3
2.Цените на артіклите во нашата продавница во споредба со останатите се:	P2	Високи-1 Исти- 2 Ниски- 3
3.Наведете најмалку 5 артикли кои најчесто ги купувате кај нас	P3	Прехранбени производи- 1 Хигиенски производи- 2 Освежителни пијалоци- 3 Останато- 4
4.Наведете го производителот на 5-те погоре наведени артикли:	P4	Прехранбени производи- 1 Хигиенски производи- 2 Освежителни пијалоци- 3 Останато-4
5.Дали поniската цена е одлучувачки фактор за Вас да купите еден артикал?	P5	0-не 1-да
6.Дали сметате дека квалитетот на производот е поврзан со цената?	P6	0-не 1-да
7.Дали сте задоволни од услугите на нашите купувачи?	P7	0-не 1-да
8.Дали би сакале да имате карта на лојалност како лојален купувач?	P8	0-не 1-да
9.Дали практикувате да купувате производи на акција?	P9	0-не 1-да
10.Наведете најмалку три артикли кои би сакале да бидат на акција !	P10	Прехранбени производи- 1 Хигиенски производи- 2 Освежителни пијалоци- 3 Останато

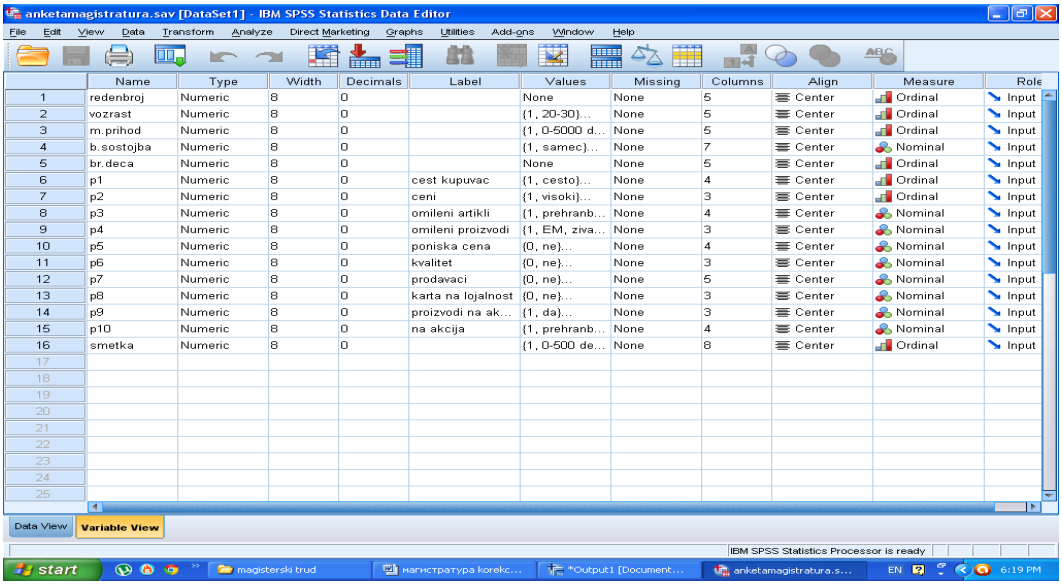
Сите податоци во нашиот случај се нумерички, а нивните атрибути се од различен вид, односно постојат номинални и интервални атрибути. Од сите анкетирани случаи се формира датотека во формат погоден за работа во SPSS Statistics. Променливите се гледаат во приказот Variable View на слика 5 прикажана подолу:



Слика 8. Приказ на променливите во SPSS Statistics

Figure 8. Variable view in SPSS Statistics

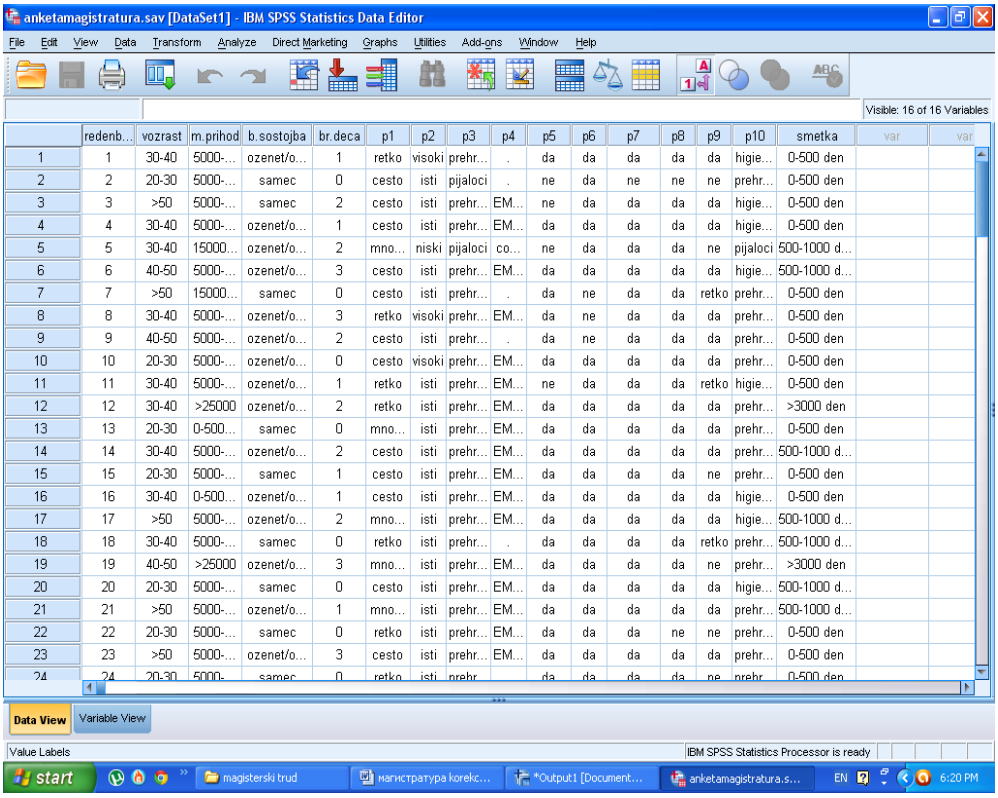
а, шифрирани во Data View изгледа вака:



Слика 9. Приказ на податоците во SPSS Statistics

Figure 9. Data View in SPSS Statistics

Нивните вредности се:



Слика 10. Приказ на променливите по категории

Figure 10. Variable view in categories

Предпроцесирањето на податоците од анкетниот лист покажа дека вредности кои недостасуваат има само кај променливите p4 и p10, односно:

Табела 10. Приказ на вредности кои недостасуваат во p4

Table 10. Review of the missing values in p4

p4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	EM, ziva, promes, Radovo, IMB, Milki, Floriol,...	106	88.3	93.0	93.0
	DUEL, SPIN, P&G<	3	2.5	2.6	95.6
	coca cola, sinalko, bitolska pivara, prilepska,..	5	4.2	4.4	100.0
	Total	114	95.0	100.0	
Missing	System	6	5.0		
Total		120	100.0		

Од табелата 10 се гледа дека кај променливата p4 недостасуваат само 6 вредности, и истите ќе се игнорираат затоа што нема да влијаат на резултатите од обработката.

Табела 11. Приказ на вредности кои недостасуваат во p10

Table 11. Review of the missing values in p10

p10

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	prehranbeni artikli	73	60.8	69.5	69.5
	higienski proizvodi	25	20.8	23.8	93.3
	pijaloci	4	3.3	3.8	97.1
	ostanato	3	2.5	2.9	100.0
	Total	105	87.5	100.0	
Missing	System	15	12.5		
Total		120	100.0		

Од табелата 11 се гледа дека кај втората променлива недостасуваат 15 вредности, но тие се поврзани со одговорот од прашањето 9, односно се гледа дека само оние купувачи кои не купуваат производи на акција или ретко ги купуваат не ги навеле артиклите што сакаат да бидат на акција, и истите не се важни за понатамошните анализи. Поради тоа се активира опцијата Exclude cases pairwise во SPSS Statistics, која е еден од начините за справување со вредностите кои недостасуваа, односно во понатамошните анализи ќе бидат исклучени овие случаеви само за оние анализи за кои им недостасуваат некои од вредностите, односно и таквите случаеви ќе бидат анализирани секогаш кога тоа е можно. Останатите податоци се чисти и не содржат вредности кои недостасуваат или пак отстапуваат од вредноста на променливата.

7.3 Фактори кои влијаат на купување производи со пониска цена

Првата техника која ќе ја употребиме за да процениме некои работи во продавницата е регресијата. Во нашиот случај ќе користиме бинарна логистичка регресија која е најпогодна за обработка на податоци од анкетен лист. Таа овозможува испитување на моделот за предвидување на категориски променливи со две или повеќе категории. Првото прашање на кое сакаме да добиеме одговор е да процениме дали пониската цена како одлучувачки фактор за купување на некој производ зависи од месечниот приход, брачната состојба и бројот на деца. За нашата анализа ни треба:

- Една категориска променлива (пониска цена), променливата p5 која прима две категории да/не шифрирани со 1 и 0
- Три категориски променливи: месечен приход, брачна состојба и број на деца кои имаат повеќе категории

Што ќе постигнеме: Логистичката регресија служи за оцена на тоа колку добро едно множество од предикторски променливи ја предвидува или објаснува категориската зависна променлива, во нашиот случај пониска цена. При тоа ќе направиме анализа на релевантноста на атрибутите кои се користат при обработката. Анализата е направена во SPSS Statistics. Категориската зависна променлива p5 (пониска цена) се префрла во полето Dependent, а останатите предикторски променливи (m.prihod, b.sostojba, br.deca) во полето Covariates. Добиените резултати се прикажани во табелите 12-20.

Табела 12. Приказ на анализираните случаи
Table 12. Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	120	100.0
	Missing Cases	0	.0
	Total	120	100.0
Unselected Cases		0	.0
Total		120	100.0

Од табелата 12 се гледа дека нема вредности кои недостасуваат, односно се анализирани 120 купувачи. Бројот на анализирани случаи е еднаков со бројот на анкетирани купувачи.

Табела 13. Табела на класификација
Table 13. Classification Table

Observed			Predicted		
			p5		
			da	ne	Percentage Correct
Step 0	p5	Da	77	0	100.0
		Ne	43	0	.0
	Overall Percentage				64.2

Табела 14. Приказ на мерка на значајност
Table 14. Signifance measure review

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.583	.190	9.365	1	.002	.558

Во табелата 13 стои дека дека правилно класифицирани случаи имаме 64.2%. Мерката на значајност Sig во табела 14 е 0,002, односно е помала од 0,05 што значи имаме добар предвидувачки модел.

Табела 15. Приказ на вредност и мерка на значајност на променливите користени во анализата
Table 15. Review of the value and significance measure of the variables used in the analyses

			Score	df	Sig.
Step 0	Variables	m.prihod	4.643	1	.031
		b.sostojba	.370	1	.543
		br.deca	.005	1	.941
	Overall Statistics		4.946	3	.176

Ако ја погледнеме табелата 15 во предикторските променливи ќе забележиме дека вредноста на Sig<0,05 само кај месечниот приход, што значи дека само променливата месечен приход предвидува дека купувачот ќе одговори со да на прашањето дека пониската цена е одлучувачки фактор за купување на некој производ.

Hosmer and Lemeshow-от тест е статистички тест кој оценува дали избраниот модел на логистичка регресија е добро калибриран, така што веројатноста на предвидувањата на моделот се рефлектира врз појавата на настани во податоците. Со овој тест податоците се групирани според нивниот процент на предвидената веројатност на постоење на случај во зависност од моделот. Тестот всушност оценува дали или не набљудуваниот настан одговара на очекуваниот настан. Овој тест врши идентификување на подгрупи. Се пресметува со:

$$H = \sum_{g=1}^{n} \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)}$$

Og-набљудуван настан, Eg- очекуван настан

Ng-број на групи, π_g- предвиден ризик

Табела 16. Табела на Hosmer and Lemeshow Test
Table 16. Hosmer and Lemeshow Test table

Step	Chi-square	df	Sig.
	4.185	6	.652

Табела 17. Приказ на набљудувани и очекувани вредности за Hosmer and Lemeshow Test
Table 17. Review of observed and expected values of Homser and Lemeshow test

		p5 = da		p5 = ne		
		Observed	Expected	Observed	Expected	Total
Step 1	1	12	10.055	1	2.945	13
	2	5	5.037	2	1.963	7
	3	11	11.149	5	4.851	16
	4	19	19.341	9	8.659	28
	5	8	9.244	6	4.756	14
	6	5	5.985	5	4.015	10
	7	7	8.352	8	6.648	15
	8	10	7.837	7	9.163	17

Бидејќи Hosmer and Lemeshow Test е еден од најсигурните тестови за предвидување на моделот, применет е и тој, но кога тој е во прашање се смета дека имаме добро предвидување само кога значајноста е поголема од 0, 05. Во нашиот случај тоа е 0.652, што ни дава за право дека имаме модел кој врши добро предвидување. Од табела 16 на овој тест ја гледаме и вредноста на Ҳ-квадрат тестот кој како што е спомнато погоре се користи за корелациона анализа кај релевантноста на атрибутите. Неговата вредност е 4.185 со ниво на значајност 0.652, што е повеќе од 0,05, а тоа значи дека моделот е подржан.

Табела 18. Табела на класификација за Hosmer and Lemeshow
Table 18. Classification table of Homser and Lemeshow

	Predicted		
	p5		
Observed	da	ne	Percentage Correct

Step 1	P5	Da	71	6	92.2
		Ne	39	4	9.3
	Overall Percentage				62.5

Во табелата 18 кај Hosmer and Lemeshow Test забележуваме дека 92.2% точно се класифицирани случаите кои одговориле со да, а вкупниот процент на точно предвидени случаи е 62.2%, што е слично со првото класифицирање. Ако погледнеме во класификацискатa табела исто така ќе забележиме дека само месечниот приход има ниво на значајност <0,05 што повторно потврдува дека само оваа предикторска променлива, а не брачната состојба и бројот на деца влијаат на тоа да пониската цена е одлучувачки фактор за купување на некој производ.

Табела 19. Приказ на мерка на значајност и коефициенти на корелација

Table 19. Significance measure and correlation coefficient review

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	m.prihod	-.575	.280	4.210	1	.040	.562
	b.sostojba	-.306	.630	.236	1	.627	.736
	Br.deca	.139	.271	.262	1	.609	1.149
	Constant	2.289	1.078	4.507	1	.034	9.870

a. Variable(s) entered on step 1: m.prihod, b.sostojba, br.deca.

Коефициентите B во табелата 19 се коефициенти кои покажуваат на насоката на врската (кои фактори ја зголемуваат веројатноста на одговорот да, а кои ја намалуваат) и тие може да бидат позитивни или негативни. Негативните коефициенти покажуваат дека зголемувањето на вредноста на независната променлива ја намалува веројатноста дека таа ќе одговори со да во зависната променлива (во нашиот случај пониската цена како одлучувачки фактор). Во нашиот случај негативен коефициенти имаме повторно и кај месечниот приход, што значи дека колку повеќе се зголемува приходот на купувачот, толку повеќе тој нема да ја зема пониската цена како одлучувачки фактор за купување производ. S.E претставува стандардна грешка на коефициентите. EXP(B)=e^B, односно:

- $e^{b1}=e^{(-0,575)}=0,562$
- $e^{b2}=e^{(-0.306)}=0,736$

- $e^{b3}=e^{(0.139)}=1.149$

Wald χ^2 се користи за да се провери значајноста на коефициентите во моделот и се пресметува со: $Wald\ \chi^2 = \left(\frac{коефициент}{SE}\right)^2$

Предикторските променливи може да бидат анализирани и посебно по категории, за да се види и точно која категорија од променливата има најголема мерка на значајност врз зависната променлива. При тоа имаме:

Табела 20. Приказ на мерка на значајност и коефициенти на корелација за сите категории на предикторските променливи

Table 20. Significance measure and correlation coefficient review for all categories of predictable variables

			Score	df	Sig.
Step 0	Variables	m.prihod	9,581	3	,022
		m.prihod(1)	4,787	1	,029
		m.prihod(2)	,971	1	,024
		m.prihod(3)	6,220	1	,013
		b.sostojba(1)	,370	1	,543
		br.deca	1,564	4	,815
		br.deca(1)	,370	1	,543
		br.deca(2)	,784	1	,376
		br.deca(3)	,073	1	,787
		br.deca(4)	,082	1	,774
	Overall Statistics		10,753	8	,216

Од табелата 20 се гледа дека навистина месечниот приход како променлива најмногу има влијание врз тоа да купувачите купуваат производи со пониска цена, односно и во ниту една подкатегорија на променливите br.deca и br.sostojba немаме ниво на значајност <0,05.

7.4 Значајни статистички информации извлечени од анкетниот лист

Други информации кои се добиваат од анкетниот лист се дали купувачите се задоволни од продавачите- (променлива p7 шифрирана со 1-да, 0-не), дали купуваат производи на акција- (променлива p9 шифрирана со 1-да, 0-не), кои производи најчесто купувачите сакаат да бидат на акција-

(променлива p10 шифрирана во четири категории), кои артикли им се омилени, кои производители им се омилени итн. Овие информации се од големо значење за задржување на купувачите, за избор на артикли и производители во следниот период и.т.н

За опис на овие променливи едноставно само ќе употребиме статистичка анализа, со цел на менаџерот на компанијата да му ги презентираме резултатите од анализата на анкетниот лист, за тој да донесе правилни одлуки за понатамошното работење на продавницата на мало.

Подолу следуваат табеларните прикази од извршените анализи и добиените резултати:

Табела 21. Одговор на прашањето: Дали сте задоволни од продавачите?

Table 21. Answer to the question: Are you satisfied of the sellers?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ne	2	1.7	1.7	1.7
	da	118	98.3	98.3	100.0
	Total	120	100.0	100.0	

98.3% од анкетираните се задоволни од услугата на продавачите- процент кој е навистина голем, што значи дека во продавницата не е потребно да се размислува за промена на истите.

Табела 22. Одговор на прашањето: Колку често се купуваат производи на акција?

Table 22. Answer to the question: How often are the products being bought on action?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	da	78	65.0	65.0	65.0
	retko	27	22.5	22.5	87.5
	ne	15	12.5	12.5	100.0
	Total	120	100.0	100.0	

65% практикуваат да купуваат производи на акција- добар процент за идно планирање на акциите врз одредени артикли.

Табела 23. Одговор на прашањето: Која категорија производи купувачите сакаат да бидат на акција?

Table 23. Answer to the question: What category of products do the customers want to be on action?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	prehranbeni artikli	73	60.8	69.5	69.5
	higienski proizvodi	25	20.8	23.8	93.3
	pijaloci	4	3.3	3.8	97.1
	ostanato	3	2.5	2.9	100.0
	Total	105	87.5	100.0	
Missing	System	15	12.5		
Total		120	100.0		

60.8% од купувачите на акција сакаат да бидат прехранбените производи, 20.8% хигиенски производи- откриена е целната категорија на производи кои треба да се најдат на акција.

Табела 24. Фрекфенција на омилени производители

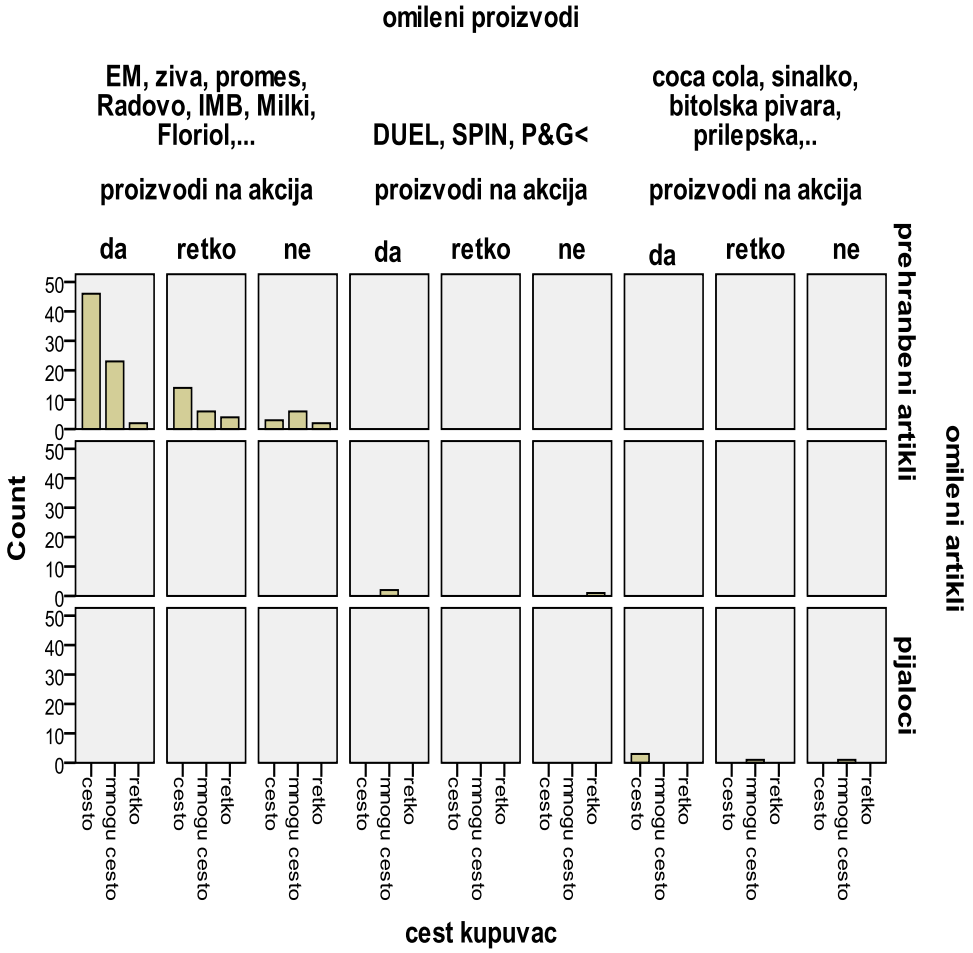
Table 24. Favorite products frequency

	Frequency	Percent	Valid Percent	Cumulative Percent
EM, ziva, promes, Radovo, IMB, Milki, Floriol,...	106	88,3	93,0	93,0
DUEL, SPIN, P&G<	3	2,5	2,6	95,6
coca cola, sinalko, bitolska pivara, prilepska,..	5	4,2	4,4	100,0
Total	114	95,0	100,0	
System	6	5,0		
Total	120	100,0		

Омилени производители за прехранбените производи се: од сувомесните производи „Екстра Меин“, „Жива“, „Промес“, од млечните производи- „Здравје Радово“, „ИМБ“, „Милки“. Омилени производители на хигиенски производи се: Дуел, Спин, П&Г, а најбарани производители на

освежителни пијалоци- „Синалко“, „Битолска пивара“, „Прилепска пивара“. Зголемената соработка токму со овие производители би била поддржана бизнис одлука на менаџментот на фирмата со која би се очекувал зголемен профит во работењето на истата.

Потврда дека честите купувачи на акција најчесто би сакале да бидат прехранбените производи од производителите „ЕМ“, „Жива“, „Промес“, „ИМБ“ и др. е прикажана и преку слика 11 подолу:

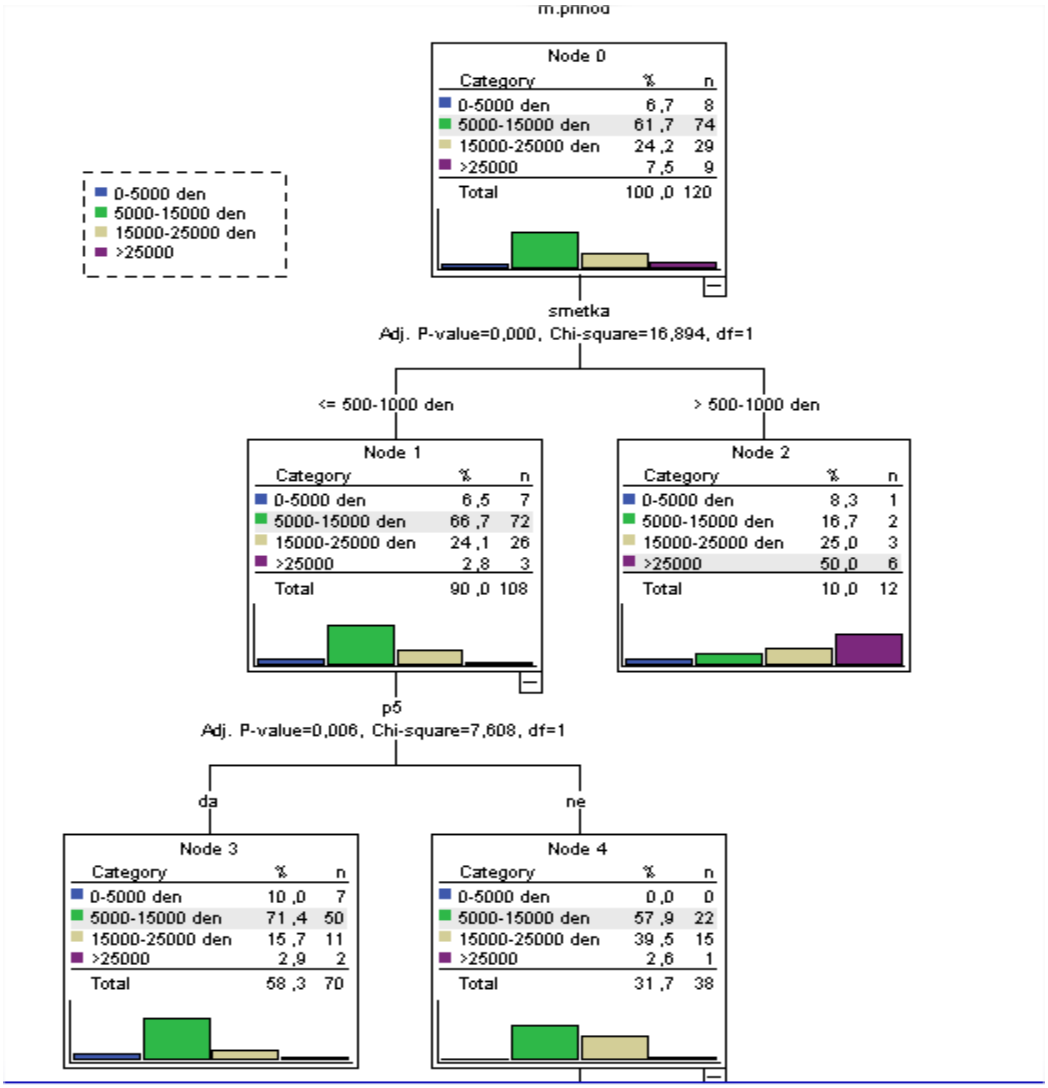


Слика 11. Графички приказ на омилените производи и производители на акција од лојалните купувачи

Figure 11. Graphic review of the favorite products and the products on action according to loyal customers

7.5 Примена на класификација за утврдување на фактори кои влијаат на висината на сметката

Податоците кои се анализираат потекнуваат од анкетни листови и фискални сметки на купувачите. Во овој дел се користат техники на класификација за да се добие одговор дали месечниот приход на купувачите влијае на тоа колкава сметка прават во продавницата, и дали истите поради тоа како одлучувачки фактор ја земаат пониската цена. Пониската цена е дадена со променливата p5. Со примена на дрвата на одлучување и CHAID алгоритамот добиваме:



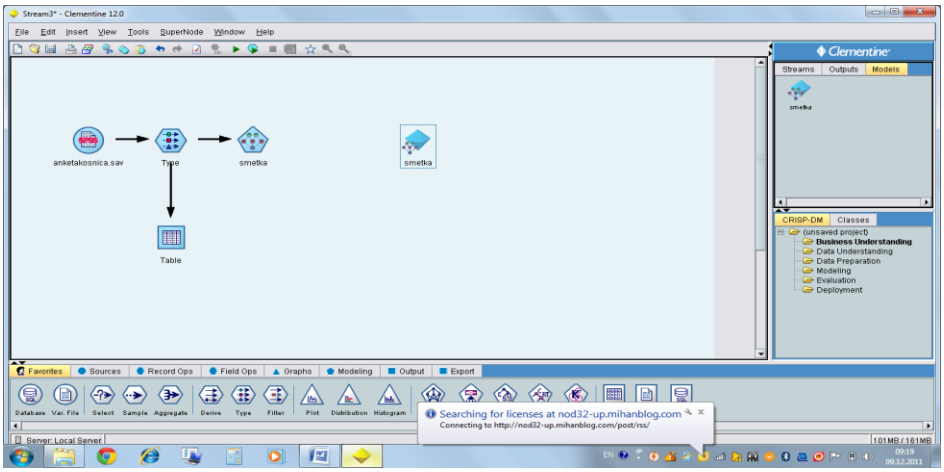
Слика 12. Дрво на одлучување за влијание на месечниот приход врз сметката направена од купувачите

Figure 12. Decision tree of the monthly income influence on the account made by the customers

Од слика 12 се гледа дека месечниот приход влијае на сметката која што ја прават купувачите, но најчесто на категоријата купувачи со приход од 5000-15000 денари. Истата категорија купувачи (која сочинува 66,7% од вкупниот број купувачи) прави сметка ≤500-1000 денари, додека сметка >500-1000 денари прават категоријата на купувачи со над 25000 денари месечен приход. За категоријата купувачи кои прават сметка ≤500-1000 денари пониската цена во 71,4% од случаите е одлучувачки фактор за купување производ со пониска цена. Од податоците јасно е дека во продавницата најчесто купуваат купувачи со приход од 5000-15000 денари, кои најчесто прават сметка ≤500-1000 денари. Тоа би значело доколку цените на производите во продавницата се пониски, најчестата категорија купувачи најверојатно ќе прави и поголеми сметки.

Како модел на класификација може да се употребат и Баесовите мрежи. Баесовите мрежи се употребени со цел да се испита од кои фактори зависи висината на сметката што ја прават купувачите. Како фактори се разгледуваат месечниот приход, брачната состојба и бројот на деца.

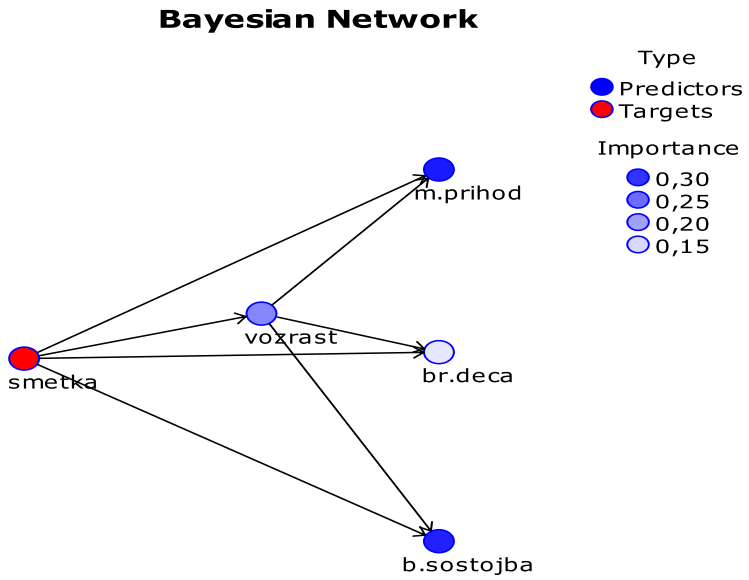
Направена е и споредба на резултатите добиени со Баесовата мрежа имплементирана во Clementine со резултатите од дрвото на одлучување опишано погоре. За Баесовата мрежа како целна променлива се зема сметката што ја прават купувачите, а останатите како предикторски променливи. Пред да се примени Clementine за добивање на Баесовата мрежа најпрво се трансформираат податоците од фискалните сметки на купувачите во формат погоден за работа во Clementine, и се врши поврзување . При тоа се добива:



Слика 13. Clementine приказ на поврзување на датотеките

Figure 13. Clementine view of the data files connecting

Баесовите мрежи се всушност графички модели кои ја прикажуваат веројатностната релација која се темели на условните веројатности меѓу множеството променливи. Се состојат од нециклични графови и табели на условни веројатности.



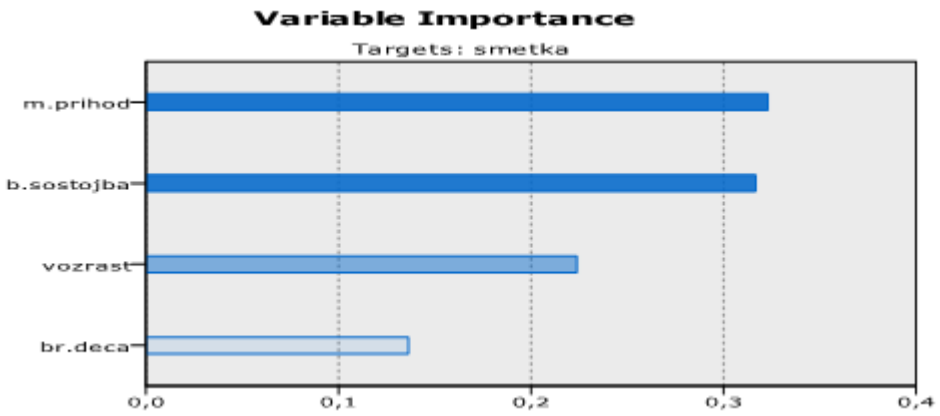
Слика 14. Баесова мрежа за фактори од кои зависи висината на сметката

Figure 14. Bayesian networks of the factors which influence the amount of the account

Од Баесовата мрежа се гледа дека најголемо влијание на сметката што ја прават купувачите во продавницата има месечниот приход, брачната состојба, возраста, па бројот на деца. Дека месечниот приход влијае на сметката што ја прават купувачите се покажа и со дрвата на одлучување.

Исто така, од Баесовата мрежа се гледа дека возраста на купувачот е поврзана само со сметката на купувачот, месечниот приход со сметката и возраста, бројот на деца со сметката и возраста, брачната состојба со сметката и возраста. Очигледно е дека возраста на купувачот значително влијае на останатите променливи и затоа од голема важност ќе биде и да ја предвидиме возраста на лојалните купувачи во понатамошните анализи.

Графичкиот приказ на важноста на променливите во однос на целната променлива е прикажан на слика:



Слика 15. Графички приказ на важноста на предикторските променливи за променливата Smetka

Figure 15. Graphic review of the importance of the predictable variables to the variable account

Табелите на условни веројатности кои се составен дел на Баесовата мрежа содржат информации за условните веројатности меѓу променливите. Тие табели се креираат врз основа на емпириските податоци, во нашиот случај обработените податоци од анкетниот лист и фискалните сметки на купувачите. За таа цел сите континуирани променливи се трансформираат во категориски вредности со што се олеснува процесот на автоматизирано машинско учење кое се спроведува преку специјализирани алгоритми, а кои се составен дел од софтверот за Баесови мрежи, во начиот случај Clementine.

Табела 25. Условната веројатност на променливата smetka

Table 25. Conditional probability of the variable account

Probability				
1	2	3	4	5
0.6	0.3	0.058	0.016	0.025

Табелата на условната веројатност на сите категории од целната променлива сметка покажува дека условната веројатност е најголема кај купувачите со сметка од категорија 1, односно 0-500 денари и изнесува 0.6, а најмала кај категоријата купувачи со сметка од 2000-3000 денари и изнесува 0.025.

Локалните табели на веројатности за секоја променлива и нејзината зависност се:

Табела 26. Условна веројатност на променливата м.prihod

Table 26. Conditional probability of the monthly income

Parents		Probability			
vozrast	smetka	1	2	3	4
1	1	0.086	0.782	0.130	0
1	2	0.142	0.714	0.142	0
1	3	0	1	0	0
1	4	0	0	0.5	0.5
2	1	0.062	0.625	0.312	0
2	2	0.058	0.588	0.294	0.058
2	3	1	0	0	0
2	5	0	0	0	1
3	1	0	0.583	0.416	0
3	2	0	0.25	0.5	0.25
3	3	0	0.333	0.333	0.333
3	5	0	0	0	1
4	1	0.047	0.761	0.190	0
4	2	0.125	0.625	0.125	0.125
4	3	0	0	0.5	0.5

Променливата vozrast како што е и претходно објаснето е шифрирана со четири категории: 1- 20-30 години, 2- 30-40 години, 3- 40-50 години, 4- над 50 години. И променливата smetka се појавува во пет категории: 1- 0-500 денари, 2- 500-1000 денари, 3- 1000-2000 денари, 4- 2000-3000 денари и 5- над 3000 денари.

Од табелата 26 е јасно дека најголема е веројатноста да купувачите кои имаат месечен приход од 5000-15000 (категорија 2) денари се на возраст од 20-30 години и прават сметка од 1000-2000 денари. Исто така купувачите на возраст од 30-40 години прават сметка и над 3000 денари, но имаат приход над 25 000 денари.

Најмала, пак, веројатност постои на пример кај купувачите на возраст од 20-30 години да направат сметка од 0-500 денари, а да имаат приход над 25 000 денари. Или, пак, купувачите на возраст од 30-40 години да направат сметка над 3000 денари, а да имаат приход 0-5000 денари.

Вакви и уште многу други законисти произлезени од условните веројатности на променливите во анализата јасно се гледаат од табелата.

Табела 27. Условна веројатност на променливата b.sostojba

Table 27. Conditional probability of the marital status

Parents		Probability	
vozrast	smetka	1	2
1	1	0.652	0.347
1	2	0.428	0.571
1	3	0	1
1	4	0.5	0.5
2	1	0.125	0.875
2	2	0.058	0.941
2	3	0	1
2	5	0	1
3	1	0	1
3	2	0	1
3	3	0	1
3	5	0	1
4	1	0.142	0.857
4	2	0.125	0.875
4	3	0	1

Променливата b.sostojba која се јавува во две категори: 1- оженет/омажена и 2-самец е прикажана со нејзината условна веројатност во табела 27. Најголема е веројатноста да купувачите на возраст од 20-30 години кои прават сметка од 1000-2000 денари се самци. Иста е пак веројатноста кај купувачите на возраст од 20-30 години кои прават сметка од 2000-3000 денари да бидат самци или оженети.

Табела 28. Условна веројатност на променливата vozrast

Table 28. Conditional probability of the age

Parents	Probability			
smetka	1	2	3	4
1	0.319	0.222	0.166	0.291
2	0.194	0.472	0.111	0.222
3	0.142	0.142	0.428	0.285
4	1	0	0	0
5	0	0.666	0.333	0

Од табелата 28 се гледа дека купувачите кои прават сметка над 3000 денари сигурно не се на возраст од 20-30 и над 50 години.

Табела 29. Условна веројатност на променливата br.deca

Table 29. Conditional probability of the number of children

Parents		Probability				
zrast	smetka	0	1	2	3	4
1	1	0.739	0.173	0.043	0.043	0
1	2	0.428	0.428	0.142	0	0
1	3	0	0	1	0	0
1	4	0.5	0.5	0	0	0
2	1	0	0.5	0.312	0.187	0
2	2	0.058	0.117	0.764	0.058	0
2	3	0	0	1	0	0
2	5	0	0	1	0	0
3	1	0.166	0.25	0.583	0	0
3	2	0	0.5	0.25	0.25	0
3	3	0	0.333	0.666	0	0
3	5	0	0	0	1	0
4	1	0.095	0.047	0.666	0.142	0.04
4	2	0	0.25	0.75	0	0
4	3	0	0.5	0.5	0	0

Од табелата 29 на условна веројатност на променливата br.deca јасно се гледа дека најголема веројатност се појавува во категоријата 3 деца, односно купувачите со 3 и повеќе деца очигледно е дека прават повисоки сметки во продавницата.

7.6 Издавање карта на лојалност - модел за вреднување на купувачите

Во овој дел е опишано креирање на невронска мрежа за да се најде категорија на купувачи за кои е најдобро да се издаде карта на лојалност, која понатаму пак ќе претставува инструмент за следење на купувањето. Во анкетните листови голем број од купувачите одговориле дека би сакале да имаат карта на лојалност, а тоа секако би го зголемило бројот на купувачи во продавницата, а со тоа и профитот на истата. Се користи невронската мрежа во која зависна е променливата карта на лојалност(променливата p8 која е шифрирана со 0 за не за карта за лојалност, и со 1 за да за карта на лојалност). Променливите од кои претпоставуваме дека зависи се: возраста, бројот на деца и брачната состојба. Крајните јазли на невронската мрежа се категоријата купувачи кои одговориле дека сакаат да им се издаде карта на лојалност. Од невронската мрежа се гледа само на која возраст се, колку деца имаат и дали

се оженети/омажени или самци. Користејќи го повторно SPSS Statistics добиваме:

Табела 30. Приказ на анализирани случаи
Table 30. Case Processing Summary

		N	Percent
Sample	Training	89	74,2%
	Testing	31	25,8%
Valid		120	100,0%
Excluded		0	
Total		120	

Во табелата 32 се забележува дека валидни се 100% од податоците, и дека 74,2% се искористени за тренирање, а 25,8% за тестирање.

Табела 31. Мрежна информација
Table 31. Network Information

Input Layer	Factors	1	br.deca
		2	vozrast
		3	b.sostojba
	Number of Units[a]		11
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		8
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	karta na lojalnost
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy

Во табела 31 има информации за зависната променлива и факторите од кои зависи, број на скриените и излезните јазли.

Табела 32. Приказ на збирниот модел
Table 32. Model Summary review

Training	Cross Entropy Error	16,259
	Percent Incorrect Predictions	5,6%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error
	Training Time	00:00:00,203
	Percent Incorrect Predictions	3,2%

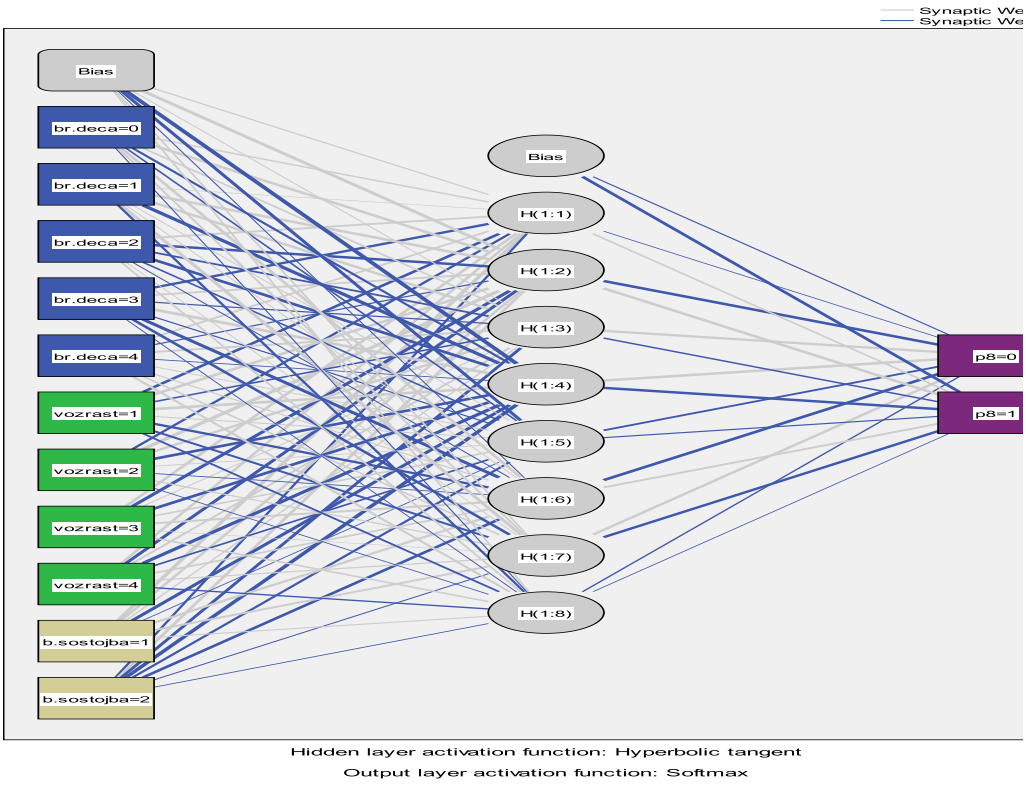
Во табела 32 се гледа дека само 5.6% од тренирачките вредности се неточно проценети, што е многу мал процент, односно имаме добар предвидувачки модел.

Табела 33. Класификација со помош на невронска мрежа

Table 33. Classification with neural network

Sample	Observed	Predicted ne	Predicted da	Percent Correct
Training	ne	0	5	.0%
	da	0	85	100.0%
	Overall Percent	.0%	100.0%	94.4%
Testing	ne	0	2	.0%
	da	0	28	100.0%
	Overall Percent	.0%	100.0%	93.3%

Од табелата 33 се гледа дека 94.4% од тренирачкото, и 96.8% од тестирачкото множество се вистински класифицирани



Слика 16.Приказ на невронска мрежа за издвојување профил на купувач за издавање карта на лојалност

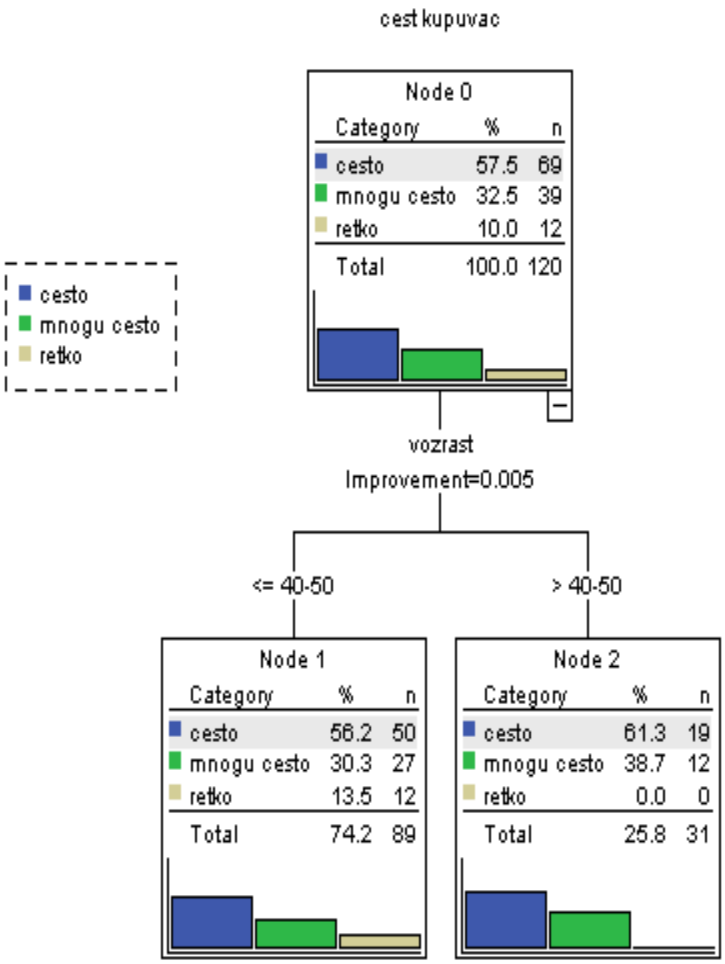
Figure 16. Neural network review of distinguishing a profile of a customer for issuing a loyalty card

Од слика 16 на невронската мрежа се гледа дека невроните на влезниот слој се променливите од кои зависи предикторската променлива, карта на лојалност, која пак е претставена со излезниот неврон и тоа со двете категории. Излезниот неврон со вредност 0 е за оние купувачи кои не се заинтересирани за издавање карта на лојалност, додека вредноста 1 е за купувачите кои би сакале да им се издаде карта на лојалност. Бидејќи 94% од анкетираните купувачи одговориле со да за издавање карта на лојалност, тоа значи дека многу е важно да се најде профилот на најлојалните купувачи. Купувачите со издадена карта на лојалност во понатамошните пресметки ќе бидат модел за вреднување на купувањето. Зголемување на бројот на лојални купувачи и придобивање на нови, секако дека е една од најстратешките цели за било која фирма. Цел на анализа ќе бидат и нивните трансакции во продавницата, нивните преференци и.т.н.

Од слика 16 исто така јасно се гледа дека целна група за издавање на карта на лојалност се купувачи на возраст од 30-40 години и оние кои имаат 2 или 3 деца. Многу слаба врска постои со брачната состојба, но сепак се гледа дека категоријата 2 (оженет/омажен) има влијание. Од визуелниот приказ исто така се заклучува дека картата на лојалност би немала влијание врз купувачите со возраст над 50 години.

Издавањето карта на лојалност се очекува дека ќе го зголеми степенот на лојалност на постоечките купувачи, и ќе активира нови, со тенденција на нивно задржување. При потполнувањето на картата на лојалност би требало да се зачуваат и дополнителни податоци за купувачите, како нивната адреса, телефонски број и електронска пошта, како истите би биле полесно контактирани за сите идни случувања во продавницата.

За да се спроведе попрецизна сегментација ќе употребиме и дрво за одлучување за да видиме кои се најчести купувачи и потоа ги споредиме. При тоа добиваме:

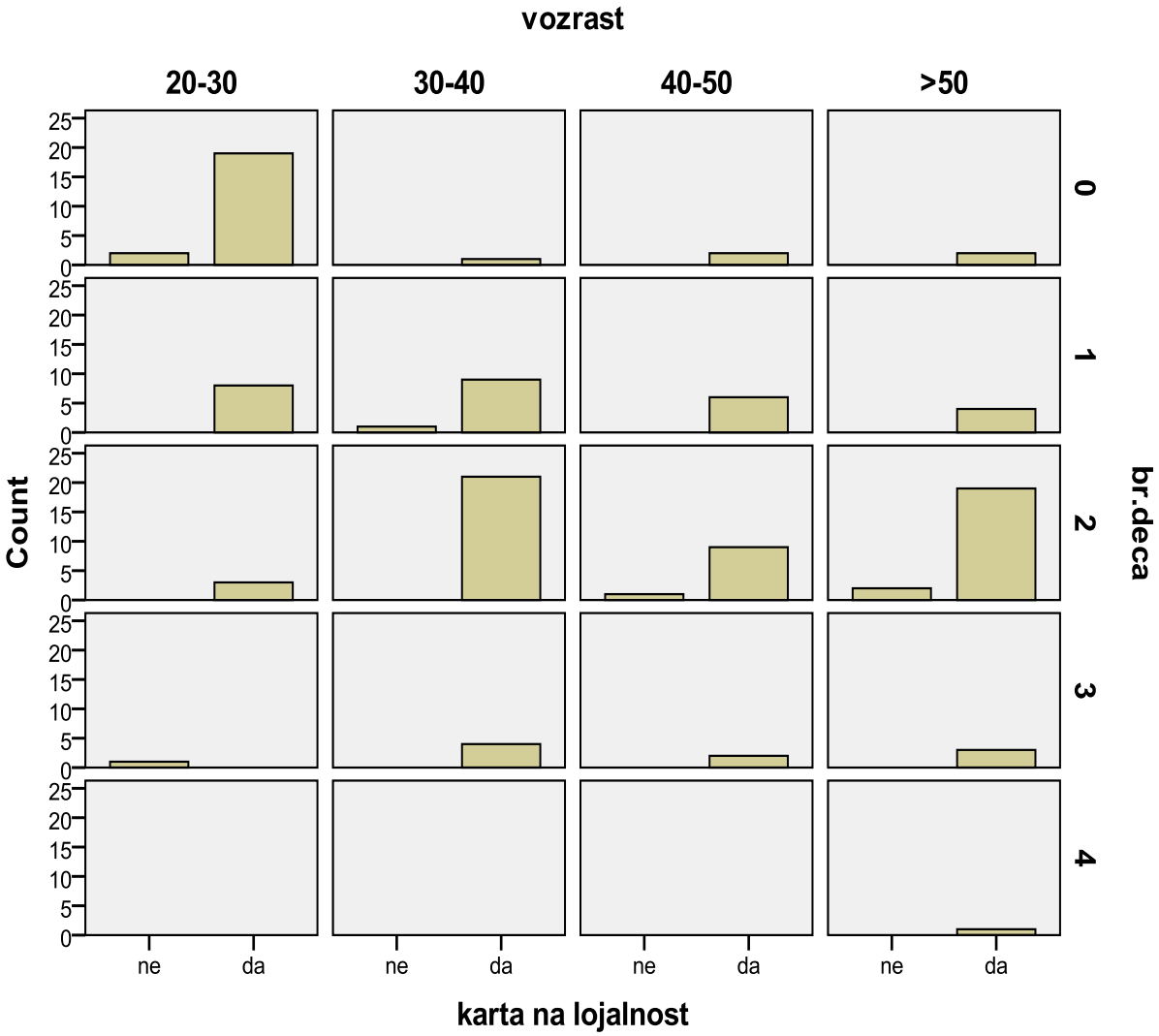


Слика 17. Дрво на одлучување за најчести купувачи

Figure 17. Decision tree of the most often customers

Од самото дрво се гледа дека честиот купувач се класифицира според возраста, дека најголем број купувачи кои често и многу често купуваат во продоавницата се на возраст се предвидени со 74.2% и се на возраст ≤ 40-50 . Тоа го потврдува фактот дека целна категорија за издавање на карта на лојалност во оваа компанија ќе биде оваа возраст, односно возраста 30-40 години.

Ако употребиме и графички приказ на меѓусебната зависност на овие променливи ќе добиваме:

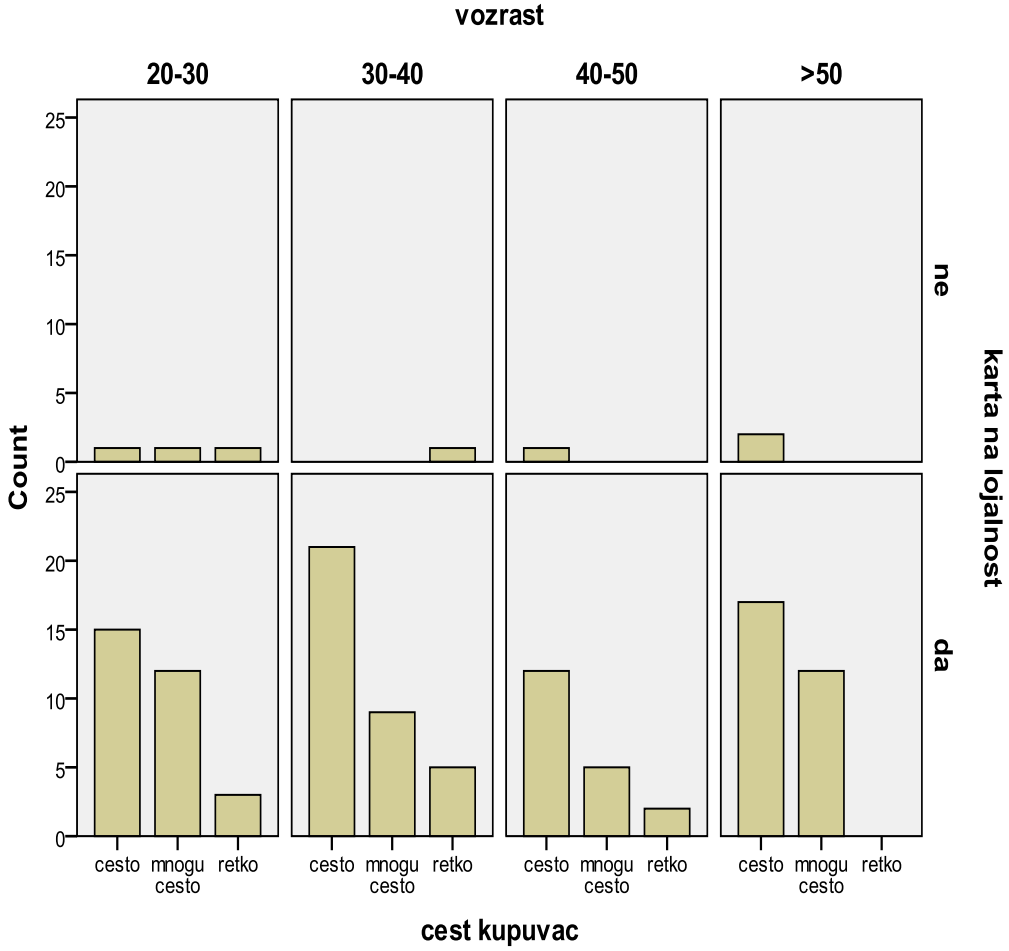


Слика 18. Графички приказ на карта на лојалност во зависност од број на деца и возраст на купувач

Figure 18. Graphic review of a loyalty card depending on the number of children and the age of the customer

Од слика 18 јасно се гледа дека најчиста ситуација за издавање карта на лојалност имаме кај категоријата купувачи на возраст од 30 до 40 години и со две деца.

Графичкиот приказ на тоа дали оваа категорија купувачи земена како целна за издавање карта на лојалност се чести купувачи е прикажан на слика 19:



Слика 19. Графички приказ на зависност на карта на лојалност од возраст и лојален купувач

Figure 19. Graphic review of the loyalty cards dependence on the age and loyal customer

Од слика 19 се гледа дека од категоријата чести и многу чести купувачи кои сакаат да им се издаде карта на лојалност најмногубројни се оние на возраст од 30 до 40 години. Многу честа категорија лојални купувачи се и категоријата купувачи над 50 години, но од претходните анализи се гледа дека тоа не е целната група за издавање карта на лојалност.

7.7 Најбарани артикли и производители

Невронските мрежи ќе ги употребиме и за да видиме што најчесто купуваат редовните купувачи и од кој производител. Значи зависна променлива е променливата p1, но кај неа посебно не интересираат вредностите 1 и 2 кои шифираат одговор на прашањето често и многу често, а како фактори кои

влијаат ќе ги земеме променливите p3 и p4 кои се од категориски тип и имаат 4 категории. Со примена на невронските мрежи добиваме:

Табела 34. Приказ на анализирани случаи
Table 34. Case Processing Summary

		N	Percent
Sample	Training	75	67.6%
	Testing	36	32.4%
Valid		111	100.0%
Excluded		9	
Total		120	

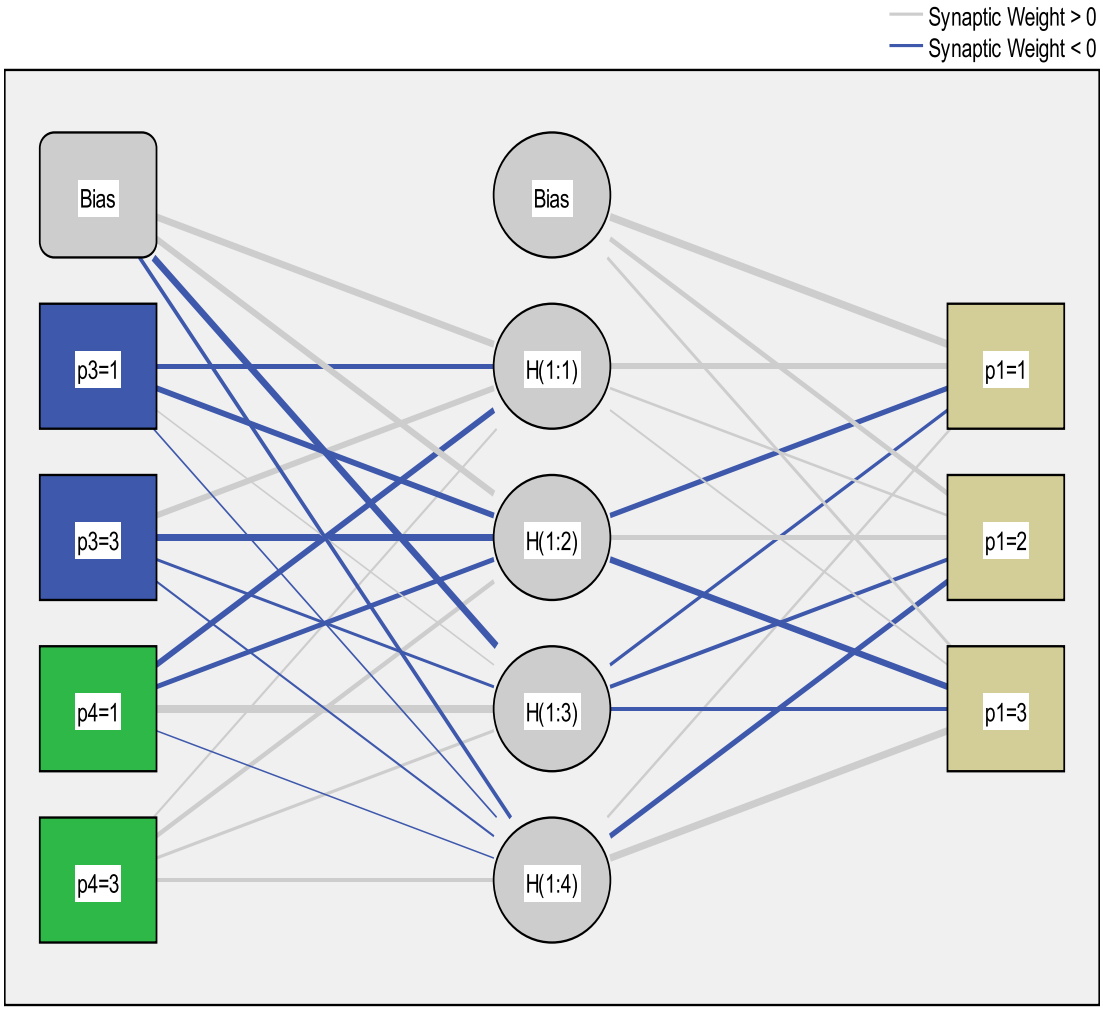
Табела 35. приказ на класификација по категории
Table 35. Classification in categories review

		Predicted			
Sample	Observed	cesto	mnogu cesto	retko	Percent Correct
Training	cesto	48	0	0	100.0%
	mnogu cesto	21	0	0	.0%
	Retko	6	0	0	.0%
	Overall Percent	100.0%	.0%	.0%	64.0%
Testing	cesto	18	0	0	100.0%
	mnogu cesto	16	0	0	.0%
	Retko	2	0	0	.0%
	Overall Percent	100.0%	.0%	.0%	50.0%

Она што се забележува на прв поглед од табелите 34 и 35 е дека класификацијата дава само 64% точност на тренирачкото множество и 50% на тестирачкото, но бидејќи се забележува дека категоријата често е 100% точно предвидена, пресметките се насочени токму на категоријата чести купувачи.

Од архитектурата на невронската мрежа дадена на слика 20 очигледно е тоа дека променливата p1 со вредност 1 (често) предвидува дека најчесто се бараат категоријата прехранбени производи, и производителите на прехранбени производи. Исто така, јасно се гледа дека освен категоријата прехранбени артикли и освежителни пијалоци, не се појавуваат категориите 2 и

4, што значи дека истите не се јавуваат како артикли на категоријата чести купувачи.



Слика 20. Приказ на невронска мрежа

P3=1 прехранбени артикли p3=освежителни пијалоци p4=1 производители на прехранбени артикли p4=3 производители на освежителни пијалоци

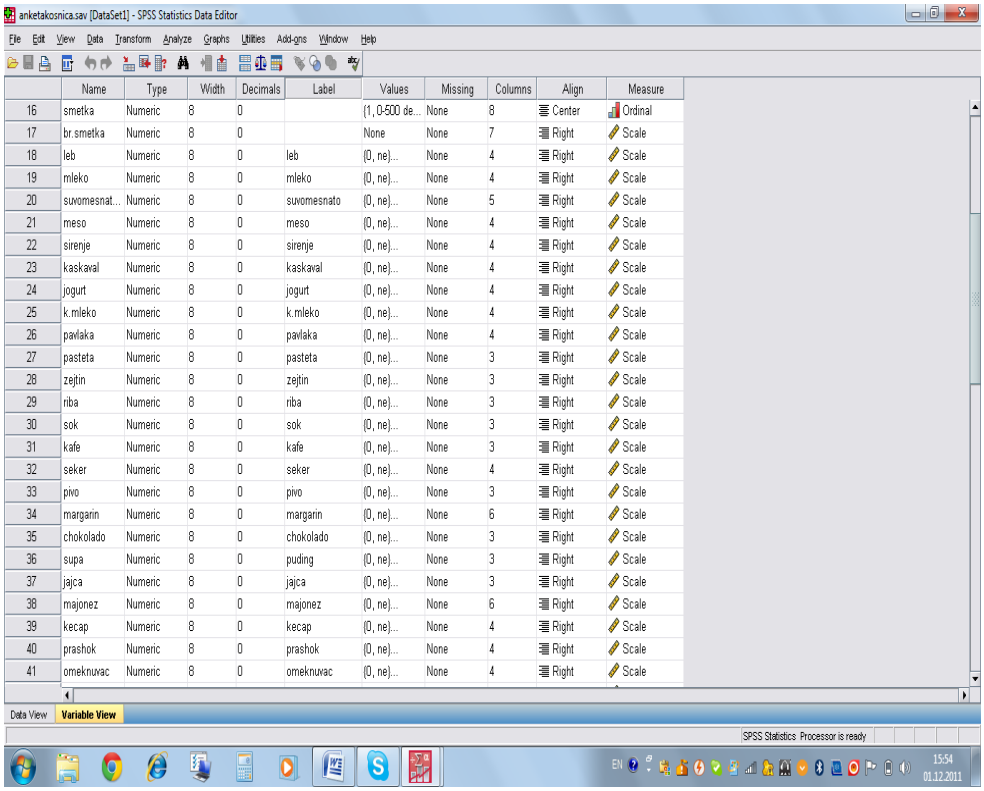
Figure 20. Neural network review

p3=1 consumables p3=refreshing drinks p4=1 consumables producers p4=3 refreshing drinks producers

7.8 Групирање на артиклите по сродност

За понатамошните анализи во продавницата се користеа фискалните сметки на купувачите кои беа анкетирани. Фискалните сметки беа детално прегледани, и од нив се издвоија одредени артикли кои најчесто беа присутни

во тие фискални сметки, но и анализирани беа фискалните сметки од подолг период во продавницата. За таа цел сите фискални сметки беа внесени и обработени во програмскиот пакет SPSS Statistics, со шифра и производи кои се купуваат заедно. Целта е да се примени Методот на потрошувачка кошница со откривање на асоцијативни правила кои ги покажуваат паровите на артикли и нивото на веројтноста дека ќе бидат купени заедно. При тоа како променливи се дефинираа Br.smetka- бројот на фискалната сметка и 28 артикли како: леб, млеко, сувомесни производи, месо, сирење, кашкавал, јогурт, кисело млеко, павлака, паштета, зејтин, риба, сок, кафе, шеќер, пиво, маргарин, чоколадо, супа, јајца, мајонез, кечап, прашок, омекнувач, средство за садови, сунѓер и шампон. Купувањето на овие артикли во SPSS Statistics се шифрира со 0-не и 1-да. Истите беа внесени како променливи заедно со променливата број на фискална сметка, при што се доби нова изворна датотека. Со користење на изворната датотека се добива следниот прозорец:

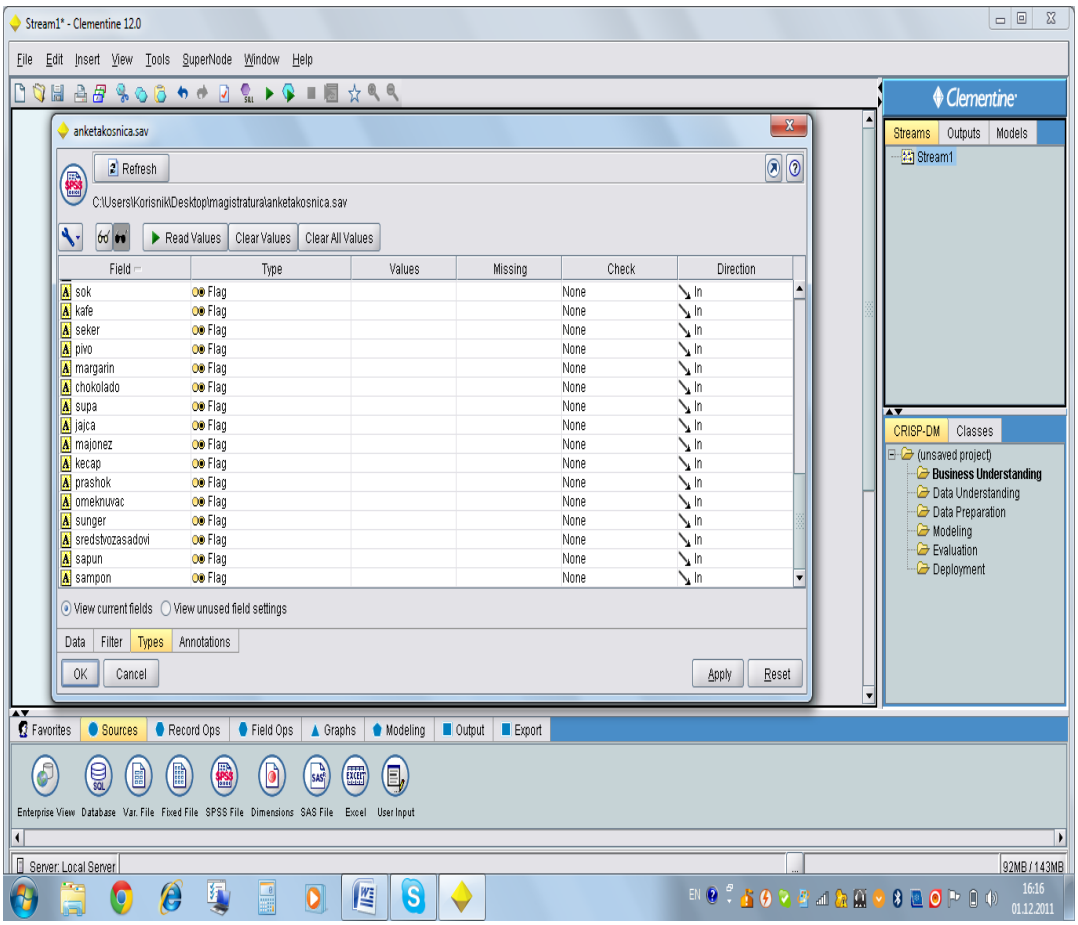


Слика 21. Приказ на променливите за анализа на потрошувачка кошница во SPSS Statistics

Figure 21. Variables review of selling basket analysis in SPSS Statistics

За да се направи анализата со помош на овој метод на податочното рударење, податоците се обработија во програмскиот пакет Clementine, пакет кој што го поддржува овој метод и дава добар визуелен приказ. Екстрахираните податоци од фискалните сметки беа трансформирани во формат погоден за работа во SPSS Statistics, а сега пак истите ќе бидат трансформирани во програмскиот пакет Clementine. Clementine како пакет дозволува вчитување на податоците од SPSS Statistics формат.

При вчитување на датотеката во Clementine потребно е да се измени типот на променливите, односно да се трансформира во тип дефиниран како Flag, а погоден за шифрирање на податоци со Да и Не, односно 1 и 0. Овој тип на променливи (атрибути) го нема во SPSS Statistics. По трансформацијата добиваме:

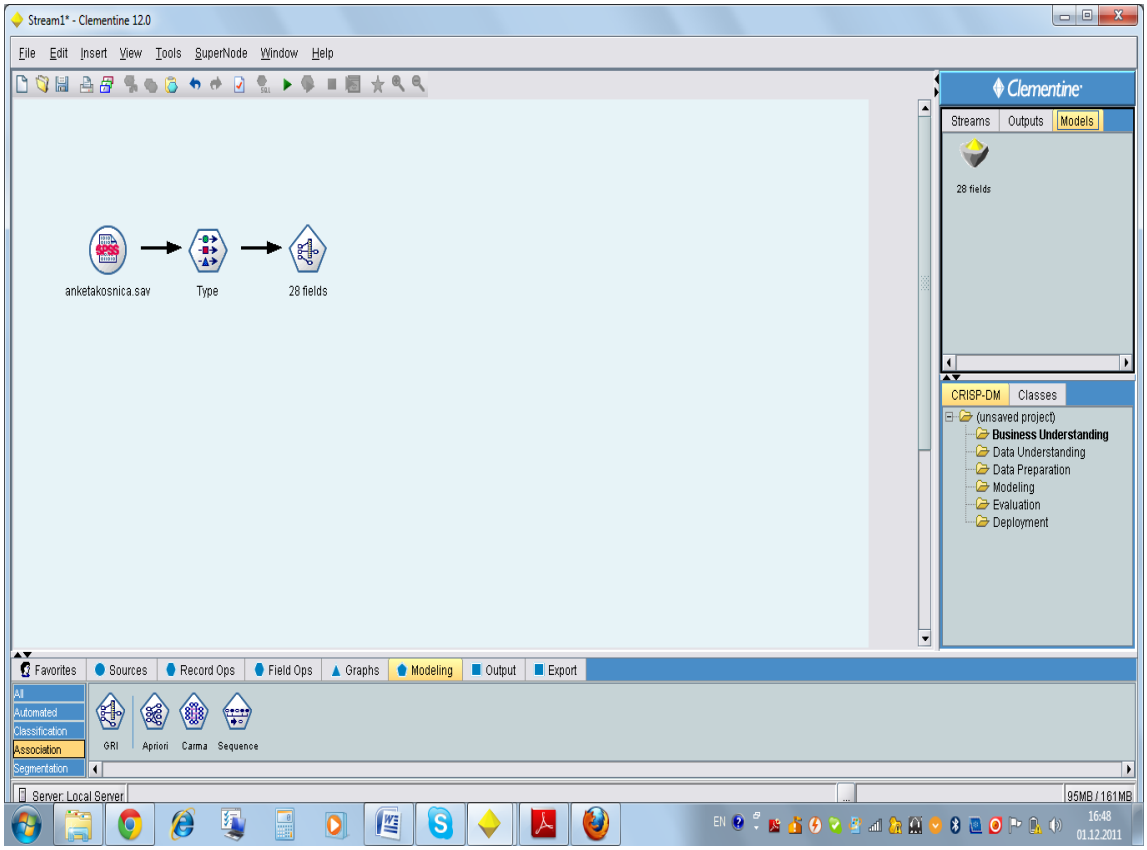


Слика 22. Приказ на типот на променливите во Clementine

Figure 22. Review of variables type in Clementine

Преку овој пример ќе се прикаже моделирање преку асоцијативни правила и моделирање на приказ на артиклите купени заедно, односно врските меѓу артиклите купени заедно.

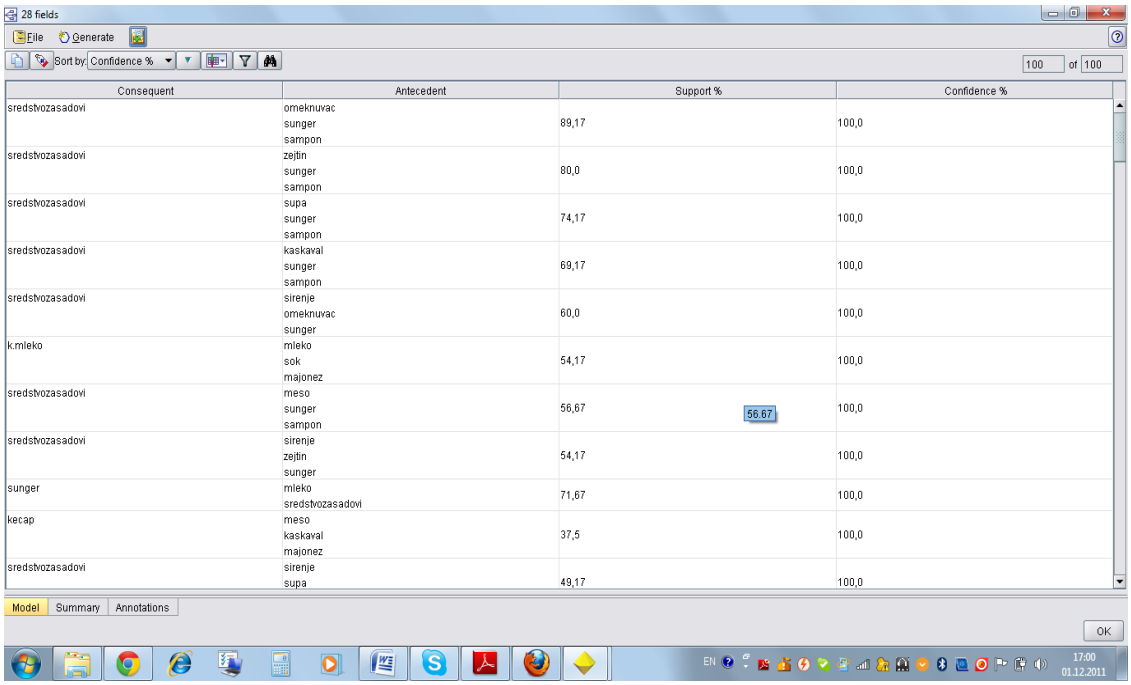
Најпрво се конектира Type Node со изворната датотека и се поврзува со Table Node. Потоа се користи Gri алгоритамот за моделирање на асоцијативни правила при што ги екстрахира правилата со највисок степен на информациска содржина врз основа на индекс што ги зема подршката и точноста на правилата како основа за моделирање. При тоа сите променливи кои сакаме да ги анализираме се користат како влез и излез, односно се означуваат со Both, а сите останати кои не влегуваат во анализата со None. При тоа во Clementine тоа изгледа вака:



Слика 23. Излез од GRI алгоритамот

Figure 23. GRI algorithm exit

Со извршување на GRI алгоритамот ја добиваме и табелата на асоцијативни правила со различен степен на подршка и сигурност. Истата подолу е прикажана во само еден дел, поради нејзиниот обемен излез.



Слика 24. Приказ на асоцијативни правила со подршка и сигурност

Figure 24. Review of associative rules with support and security

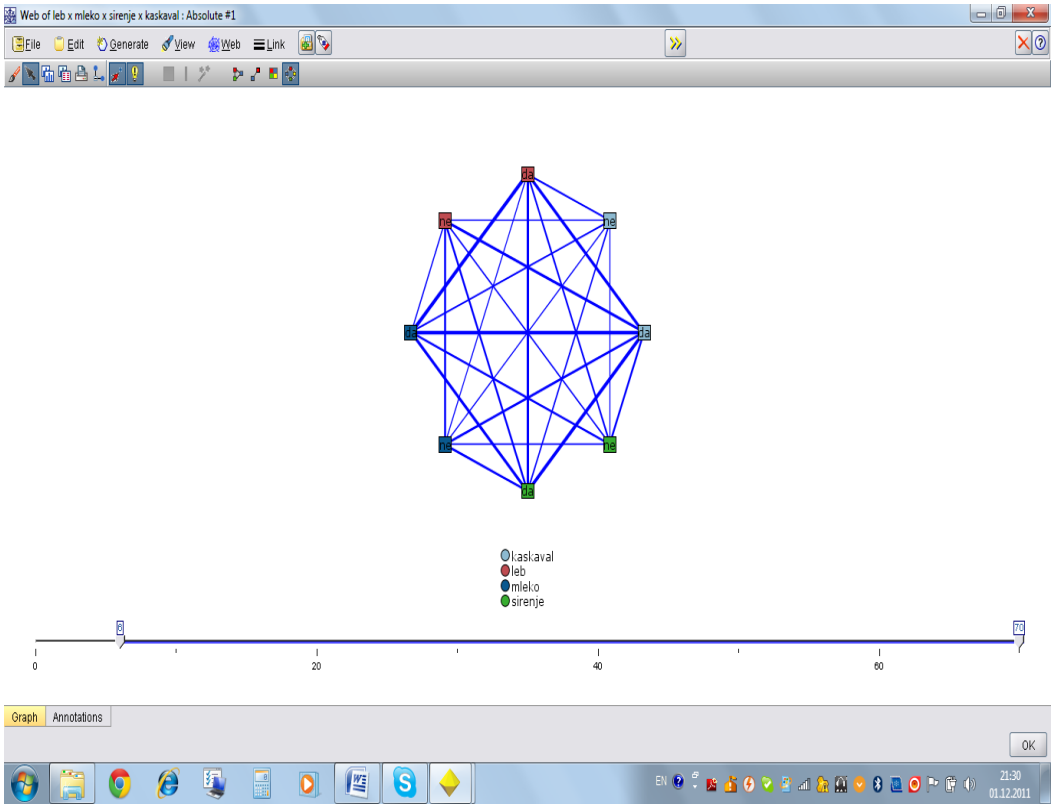
Од табелата ќе ги издвоиме оние артикли кои имаат најголем степен на поддршка и сигурност, односно ги добиваме следниве асоцијативни правила:

- Ако купува средство за садови, тогаш купува и сунѓер, прашок и омекнувач со 100% сигурност и 89,17% поддршка
- Ако купува средство за садови, тогаш купува и сунѓер, шампон и зејтин со 100% сигурност и 80% поддршка
- Ако купува средство за садови, тогаш купува и сунѓер и шампон со 99,8% сигурност и 90,08% поддршка
- Ако купува средство за садови, тогаш купува и прашок, сунѓер и шампон со 99,07% сигурност и 90% поддршка
- Ако купува средство за садови, тогаш купува и сунѓер, омекнувач и шампон со 99,07% сигурност и 90,07% поддршка
- Ако купува средство за садови, тогаш купува и прашок, омекнувач и сунѓер со 98,18% сигурност и 91,67% поддршка итн
- Ако купува млеко, тогаш купува и леб, кисело млеко и шеќер со 93,3% сигурност и 50% поддршка
- Ако купува млеко, тогаш купува и леб, јогурт и кисело млеко со 92,42% сигурност и 55,05% поддршка

- Ако купува јајца, тогаш купува и чоколадо со 99,01% сигурност и 84,17% поддршка
- Ако купува јајца, тогаш купува шекер и чоколадо со 98,98% сигурност и поддршка од 81,67%
- Ако купува млеко, тогаш купува и леб со 89,04% сигурност и 60,83% поддршка

Вакви асоцијативни правила може да се издвојат уште многу, и тоа се однесуваат за сите 28 артикли кои се купуваат заедно. Бројот на артиклите кои се купуваат заедно може да се подеси на 2, 3 и повеќе. Бидејќи графот за овие 28 артикла е доста голем и непрегледен, артиклите ќе ги групирам во група по сродност на артиклите за да можеме да добиеме добар визуелен преглед.

При тоа добиваме:



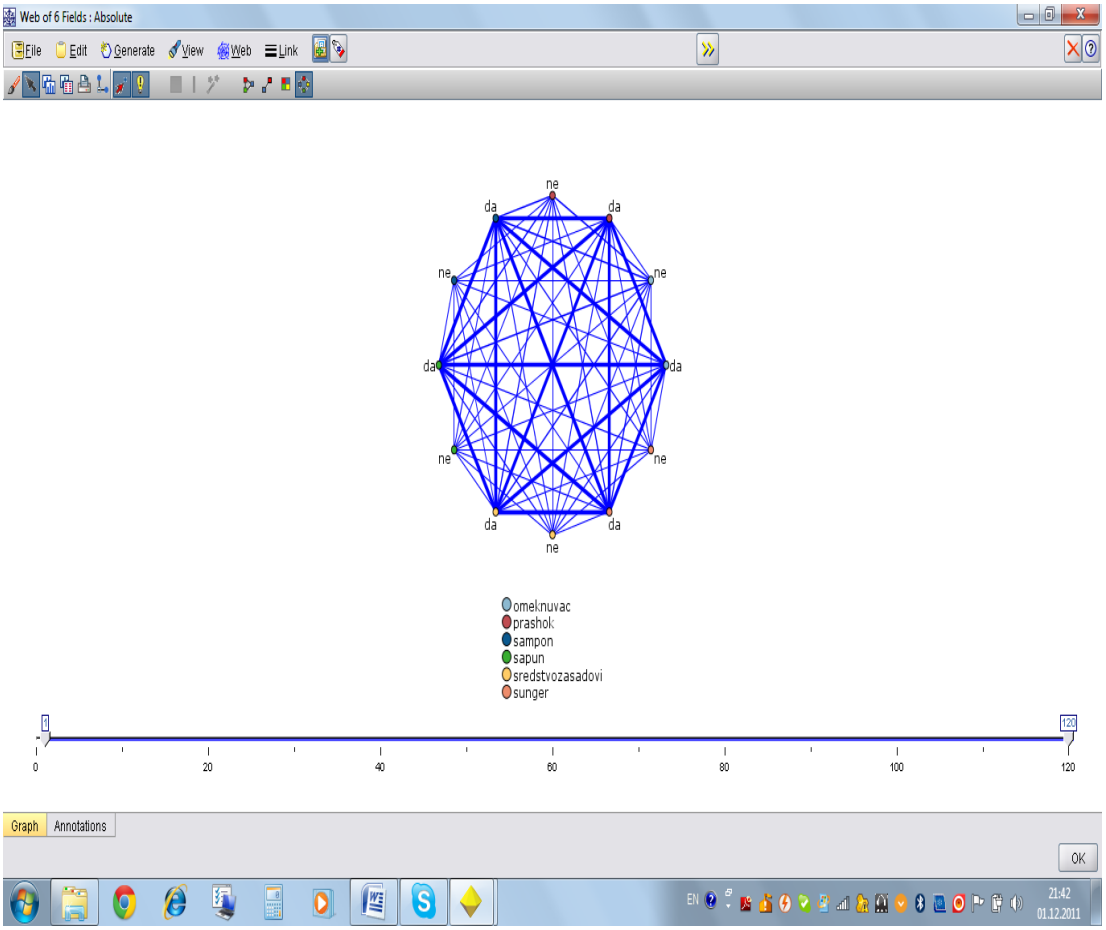
Слика 25. Граф на прехранбени артикли кои се купуваат заедно

Figure 25. Graphics of the consumables which are being bought together

Задебелените линии во графот ја покажуваат посилната врска, а потенките слабата врска. Од графот е очигледно дека добиваме групи на купувачи кои заедно купуваат:

- сирење, кашкавал и млеко
- млеко, леб и кашкавал
- леб, сирење и кашкавал

На следната слика се прикажани групата на хигиенски производи кои најчесто се купуваат заедно:



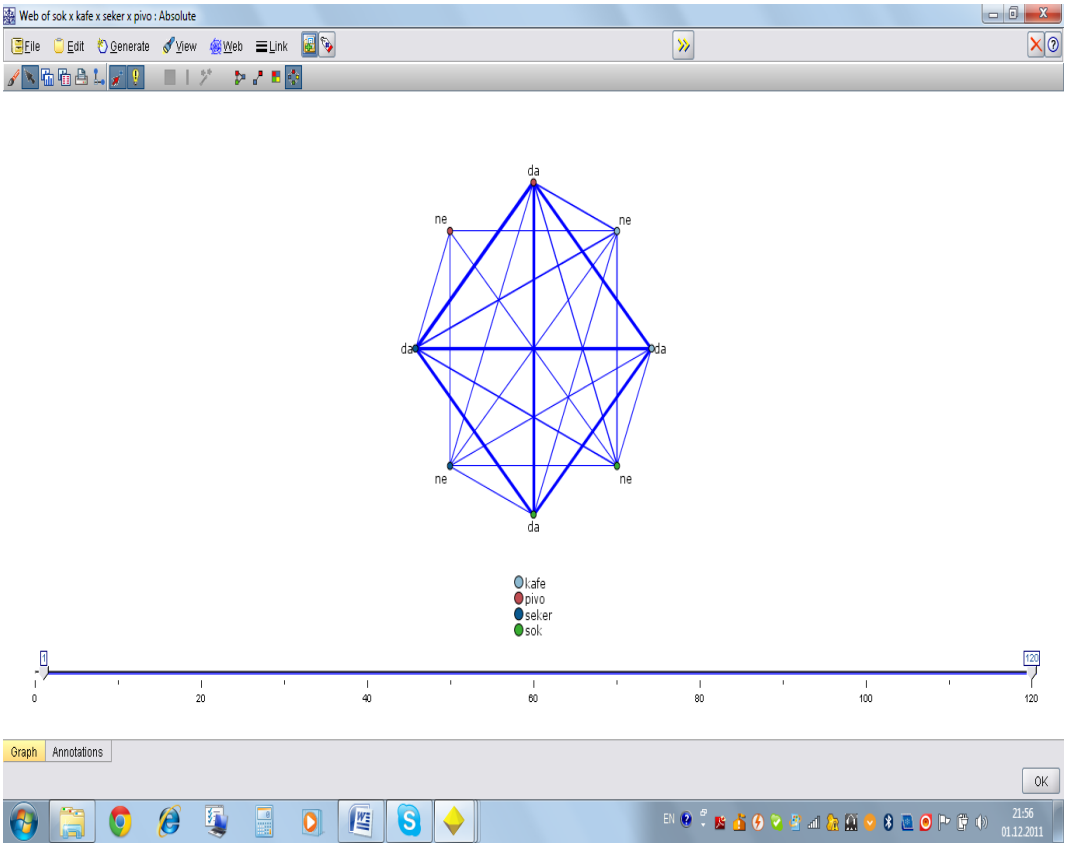
Слика 26. Група на хигиенски производи кои се купуваат заедно

Figure 26. Group of hygiene products which are being bought together

Од задебелените линии на графот кои ја покажуваат силната врска ќе ги издвоиме купувачите кои заедно купуваат:

- омекнувач, сапун и сунѓер
- шампон, средство за садови и сунгер
- сунѓер, шампон и прашок
- прашок и омекнувач
- омекнувач, шампон и средство за садови

Сега заедно ке ги групираме некои од останатите артикли:



Слика 27. Приказ на граф на кафе, пиво, сок и шекер

Figure 27. Review of graphics of coffee, beer, juice and sugar

Од графот се гледа дека заедно се купуваат:

- сок, кафе и пиво
- кафе, шеќер и пиво
- кафе, шеќер и сок

На ист начин добиваме дека заедно се купуваат

- маргарин, мајонез и јајца
- кечап, јајца и мајонез
- јајца, кечап и супа
- јајца, мајонез и чоколадо
- маргарин, супа и чоколадо
- чоколадо, леб и маргарин
- сирење, кашкавал и месо
- сувомеснати производи, кашкавал и сирење

Она што е исто значајно е и тоа што со овој метод се издвојуваат и производите кои не се купуваат заедно. Тоа се гледа и од графот и од

табеларниот приказ на ниво на поддршка. Анализирајќи ги графовите се издвојува следново:

- Купувачот кој купува мајонез, не купува чоколадо
- Купувачот кој купува чоколадо, не купува кечап
- Купувачот кој купува средство за садови, не купува прашок
- Купувачот кој купува јогурт, не купува сувомесни производи
- Купувачот кој купува јогурт, не купува млеко
- Купувачот кој купува млеко, не купува кисело млеко

Се појавуваат и случаеви кои одат во парови, да ако купувачот не купува еден производ, тогаш не купува и друг.

На овој начин може да добиеме повеќе групи на артикли кои купувачот ги купува заедно. Тоа би значело дека артиклите кои се купуваат заедно во продавницата пожелно е да стојат близу еден до друг.

7.9 Анализа на период по издавање на карта на лојалност

Откако се спроведе анкетата во продавницата на мало во „Компанија Моневи“ во периодот јуни-јули беа издадени 100 карти на лојалните купувачи, а како целна категорија се зедоа оние кои се добија при анализите со техниките за податочно рударење. Картата на лојалност ги содржи основните податоци:

- Број на карта
- Име и презиме
- Адреса и место на живеење
- Година на раѓање
- E-mail адреса
- Телефонски број
- Брачна состојба
- Број на деца
- Месечен приход

Податоците од пополнетите карти на лојалност со кои купувачите ќе добиваат одреден попуст при купувањето и ќе собираат поени за попуст на одредени производи беа внесени во базата на податоците во фирмата, и категоријата

купувачи на кои им беше издадена карта на лојалност се внесе како категорија Lojaliti во базата преку бројот на издадената карта на лојалност.

Од базата на податоци во „Компанија Моневи„ која функционира на SQL Server 2008, се гледа секоја трансакција, секоја продажба за одредено множество артикли.

Како што е кажано и претходно издадената карта на лојалност ќе претставува модел за вреднување на купувањето/продажбата. Во зависност од анализата на издадените карти на лојалност креирани се клучните индикатори за лојалност на купувачите прикажани во табела.

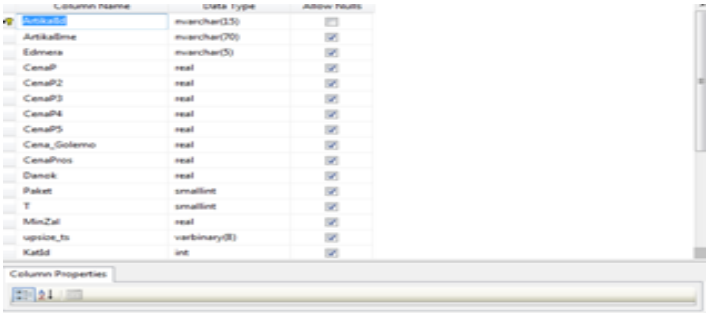
Табела 36. Клучни индикатори за лојалност

Table 36. Loyalty key indicators

Бр. бодови на картичка	одзив на кампања	Учество во купувањето	лојалност
мал	мала	мала	мала
мал	средна	мала	Мала
мал	мала	висока	Средна
среден	Средна	Висока	Висока
висок	Висока	висока	висока
.....			

Врз основа на клучните индикатори за лојалност на купувачите, се креира и релациска табела во зависност од бројот на картата на лојалност, степенот на лојалност, профитабилноста и перспективноста на купувачите.

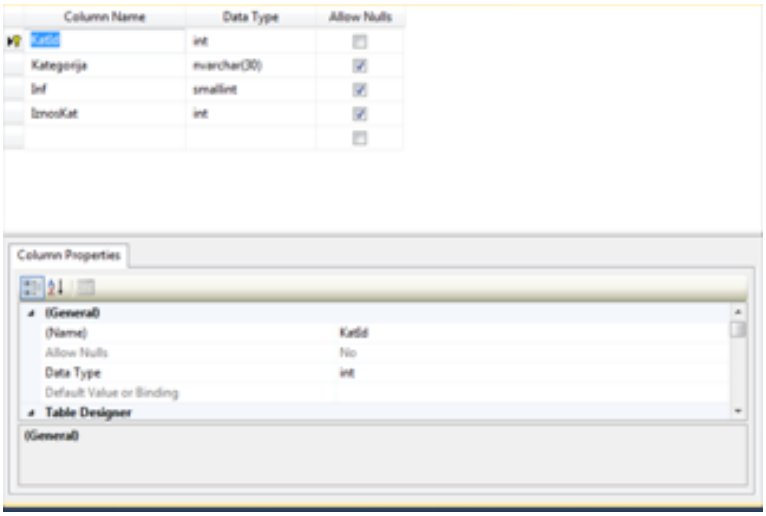
За нашата анализа ќе бидат употребени неколку полиња од базата на податоци. Полето Artikli кое во моментот на анализата располага со 18.123 податоци. Основното поле со име ArtikalId е прикажано на слика 28:



Слика 28. Приказ на поле ArtikalId

Figure 28. Articles field review

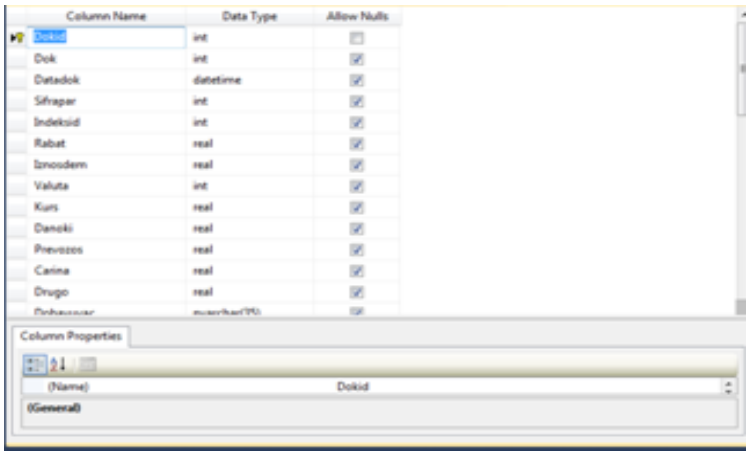
Како што се забележува ова поле содржи повеќе податоци како име, цена, данок и др. Бидејќи во ова поле се сместени сите артикли, за полесна манипулација се разгледува полето Kategorija кое претставува категорија на артикли од 318 категории на артикли. Истото е прикажано на слика 29:



Слика 29. Приказ на полето Kategorija

Figure 29. Category field review

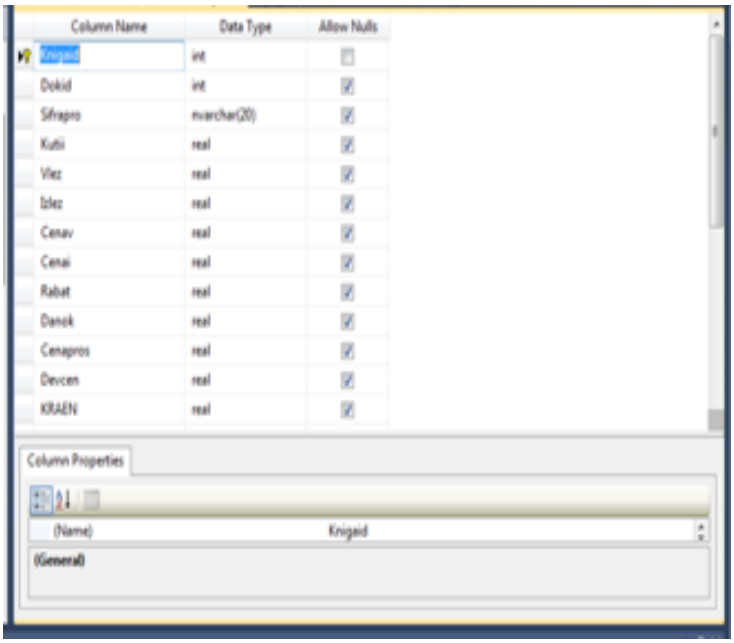
Полето Dokumenti ги содржи сите издадени документи, односно во нашиов случај издадените фискални сметки. Ова поле содржи 389.133 податоци и е прикажано на слика 30:



Слика 30. Приказ на полето Dokumenti

Figure 30. Documents field review

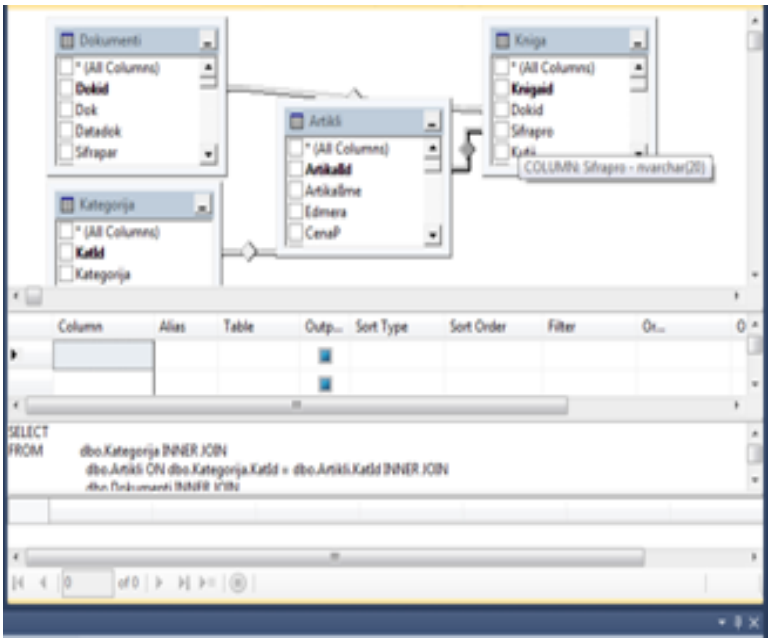
Полето Kniga содржи 1.498.979 податоци и во него се сместени податоците од artikal*kolicina*cena. Истото е прикажано на слика 31:



Слика 31. Приказ на полето Kniga

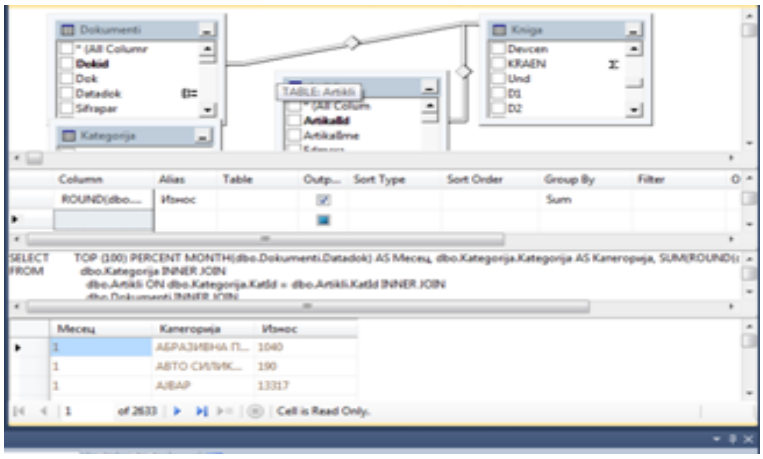
Figure 31. Book field review

За анализа на тоа која група на купувачи колку промет прави во периодот по издавањето на картата на лојалност се креира Query/View од базата и притоа се земаат полињата Artikli, Dokumenti, Kniga и Kategorija. Креирањето се гледа на слика 32 и 33 подолу:



Слика 32. Query/View приказ

Figure 32. Query / View



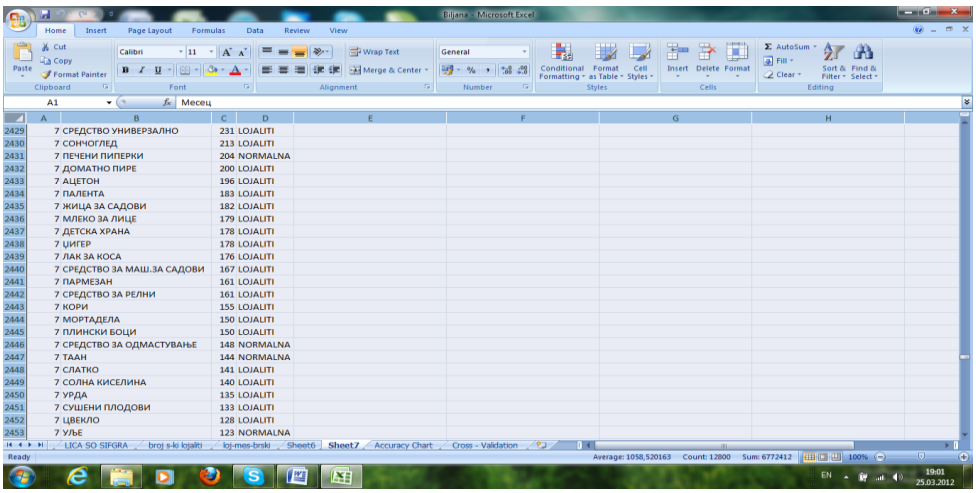
Слика 33. Query/View приказ

Figure 33. Query / View

Групирањето на податоците е направено врз основа на категорија. Со цел да видиме колкав процент на продажбата отпаѓа на купувачите на кои им е издадена карта на лојалност сите податоци ги експортираме во Excel. При тоа бидејќи секоја карта на лојалност има свој број, а во периодот јуни-јули 2011 беа издадени 100 пробни карти на лојалност се гледа дека во периодот од јули 2011 до март 2012, 98 од нив се користени во продажба.

Во периодот јули – октомври 2011 се добива дека купувачите на кои им е издадена карта на лојалност направиле 2561 сметки од вкупно 3200 сметки, а тоа е 80% од продажбата во продавницата.

На слика 34 е прикажан дел од направените сметки на групата купувачи без карта на лојалност, означена како normalna, и lojaliti- групата на купувачи со лојална карта:

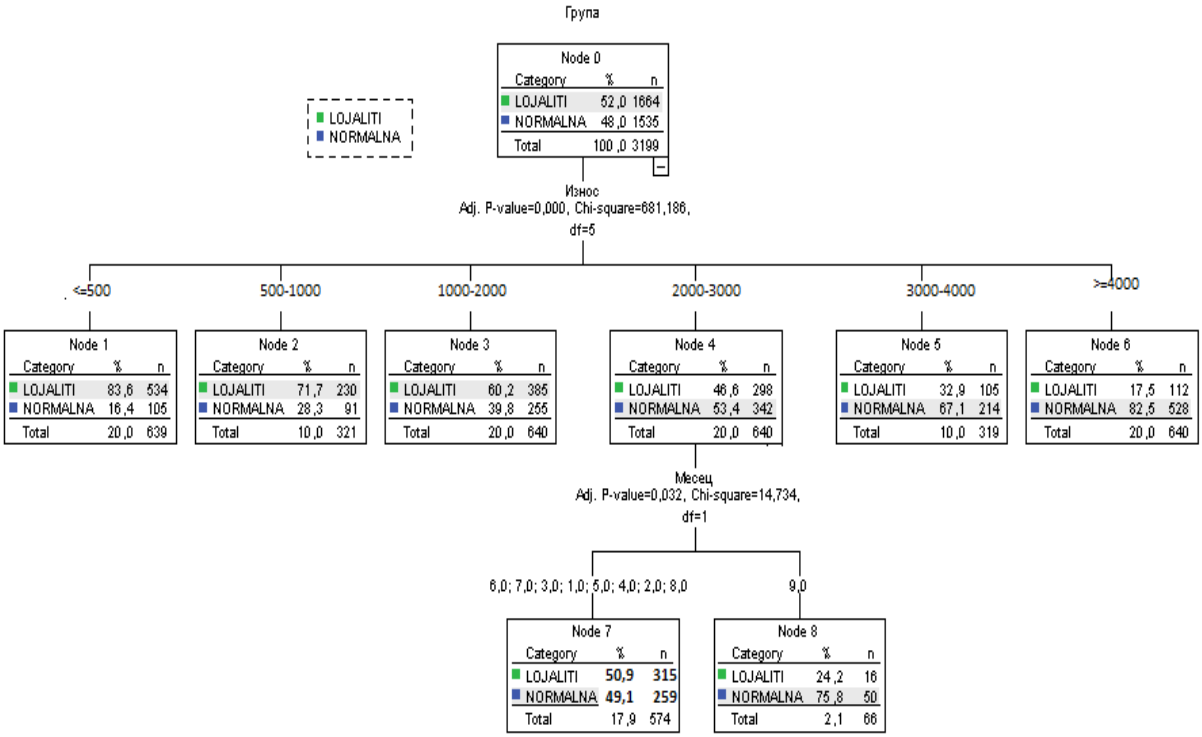


2429	7 СРЕДСТВО УНИВЕРЗАЛНО	231	LOJALITI
2430	7 СОНЧОГЛЕД	213	LOJALITI
2431	7 ПЕЧЕНИ ПИПЕРКИ	204	NORMALNA
2432	7 ДОМАТНО ПИРЕ	200	LOJALITI
2433	7 АЦЕТОН	196	LOJALITI
2434	7 ПАЛЕНТА	183	LOJALITI
2435	7 ЖИЦА ЗА САДОВИ	182	LOJALITI
2436	7 МЛЕКО ЗА ЛИЦЕ	179	LOJALITI
2437	7 ДЕТСКА ХРАНА	178	LOJALITI
2438	7 ЦИГЕР	178	LOJALITI
2439	7 ЛАК ЗА КОСА	176	LOJALITI
2440	7 СРЕДСТВО ЗА МАШ.ЗА САДОВИ	167	LOJALITI
2441	7 ПАРМЕЗАН	161	LOJALITI
2442	7 СРЕДСТВО ЗА РЕЛНИ	161	LOJALITI
2443	7 КОРИ	155	LOJALITI
2444	7 МОРТАДЕЛА	150	LOJALITI
2445	7 ПЛИНСКИ БОЦИ	150	LOJALITI
2446	7 СРЕДСТВО ЗА ОДМАСТУВАЊЕ	148	NORMALNA
2447	7 ТААН	144	NORMALNA
2448	7 СЛАТКО	141	LOJALITI
2449	7 СОЛНА КИСЕЛИНА	140	LOJALITI
2450	7 УРДА	135	LOJALITI
2451	7 СУШЕНИ ПЛОДОВИ	133	LOJALITI
2452	7 ЦВЕКЛО	128	LOJALITI
2453	7 УЉЕ	123	NORMALNA

Слика 34. Направени сметки на купувачи со и без карта на лојалност

Figure 34. Accounts made by the customers with and without loyalty card

Ако ги погледнеме направените сметки на лојалните купувачи по месеци и по број на сметки ќе видиме дека голем дел од продажбата во месеците по издавање на картата на лојалност отпаѓа на купувачите со издадена карта на лојалност. Ќе употребиме повторно дрва на одлучување користејќи ја како зависна променлива категоријата лојални и нормални купувачи, а како предиктивни променливата iznos- износ на сметката што е направена и mesec- месецот во кој е направена сметката за да видиме категоријата лојални купувачи кој износ на сметки ги прави и во кои месеци најчесто. Податоците превземени од базата најпрво се пренесени во Excel, а потоа во пакетот SPSS Clementine, при што се добива следново дрво на одлучување:



Слика 35. Дрво на одлучување за категорија лојални купувачи

Figure 35. Decision tree of the loyal customers category

Од дрвото на одлучување се забележува дека групата купувачи со и без издадена карта на лојалност прави различен износ на сметки и тоа дека сметка со помал износ најчесто прават купувачите со издадена карта на лојалност, а со поголем износ купувачите без карта на лојалност. Ако се направи споредба со извршените анализи при издавањето карта на лојалност се гледа дека по

издавањето карта на лојалност, лојалните купувачи сега прават повисоки сметки од порано и тоа и сметки од 1000-2000 денари. Анализата на фискалните сметки покажува и дека најголем број од сметките се со вредност до 2000 денари, а многу поретки се оние над 3000 денари, што значи дека на лојалните купувачи отпаѓа и најголем дел од продажбата.

Од месецот е зависна само категоријата купувачи кои прават сметка од 2000-3000 денари, и тоа оваа сметка најчесто ја прават купувачите без карта на лојалност во месец септември, додека во останатите месеци ваква сметка лојалните купувачи прават 50,9%, а останатите 49,1%. Ако ја погледнеме табелата на класификација прикажана подолу:

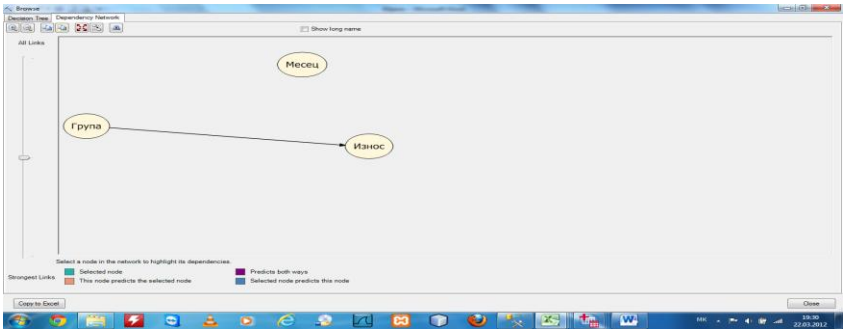
Табела бр 37. Класификација на група купувачи

Table 37. Classification of the group of customers

Observed	Predicted		
	LOJALITI	NORMALNA	Percent Correct
LOJALITI	1149	515	69,1%
NORMALNA	451	1084	70,6%
Overall Percentage	50,0%	50,0%	69,8%

Од табелата на класификација јасно се гледа дека групата купувачи со издадена карта на лојалност е предвидена со точност од 69,1%, а останатите со 70.6%.

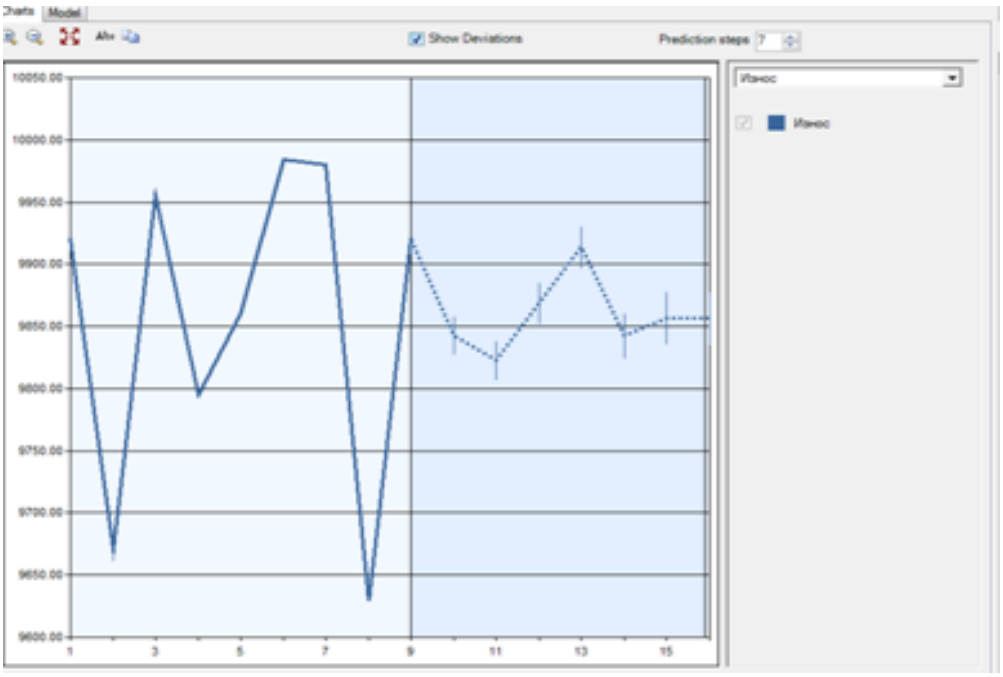
Дека групата купувачи со и без издадена карта на лојалност е зависна од износот на сметка што се прави се гледа и од самата база на податоци, односно јасно се гледа дека постои врска прикажана на слика 35:



Слика 35. Јачина на врска меѓу групата купувачи и износот на сметка

Figure 35. The strength of the connection between group of customers and the account amount

Анализирајќи ја продажбата во претходниот период, и тоа од месец јануари 2011 до месец септември 2011 година, употребувајќи го методот на предвидување и тоа директно во Excel земајќи ги податоците од базата може да се предвиди и продажбата за следниот период. Графичкиот приказ е даден на слика 36:



Слика 36. Предвидување на продажба

Figure 36. Selling products

Од слика 36 јасно се гледа во периодот по издавање на картата на лојалност, односно од месец јули, па натаму се зголемува профитот на фирмата, а се предвидува и истиот за следниот период. Ваквото предвидување е направено врз основа на издадените 100 карти на лојалност, а во следниот период се очекува да се издадат многу повеќе.

Во периодот по издавањето на картата на лојалност може да се примени и кластерирањето со цел да се издвојат основните сегменти на купувачи, односно да се утврдат категориите на производи кои се купуваат од лојалните купувачи. За таа цел кластерирањето ќе се изврши директно во Excel 2010 со инсталирани алатки за податочно рударење во него. За спроведување на кластерирањето кое ќе се изврши над издвоените категории со и без издадена

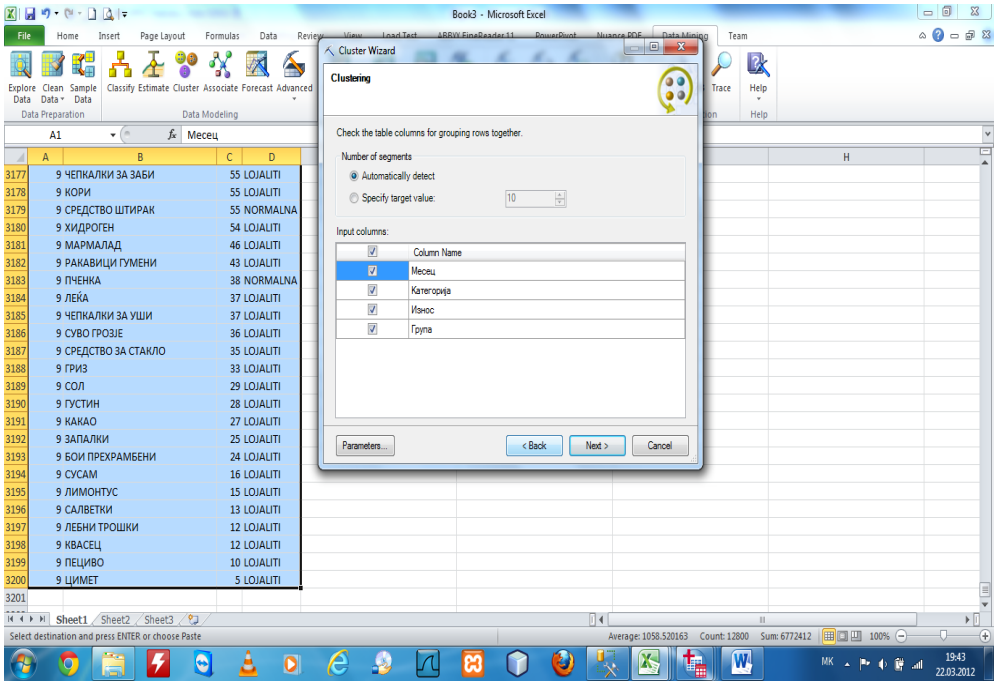
карта на лојалност, податоците се превземени од базата со помош на SQL на следниов начин:

SELECT <atributi>FROM tabela WHERE <uslov> AND KATEGORIJA.

Јасно е дека еден купувач може истовремено да купува и повеќе категории производи, но тоа нема влијание на текот на анализите при кластерирањето. Во нашиов случај во процесот на кластерирање се користени следниве променливи:

- Месец- месецот на продажба
- Категорија- категорија на артикли
- Износ- износот на сметката што ја прават купувачите
- Група- купувачи со карта на лојалност- лојални и купувачи без издадена карта на лојалност

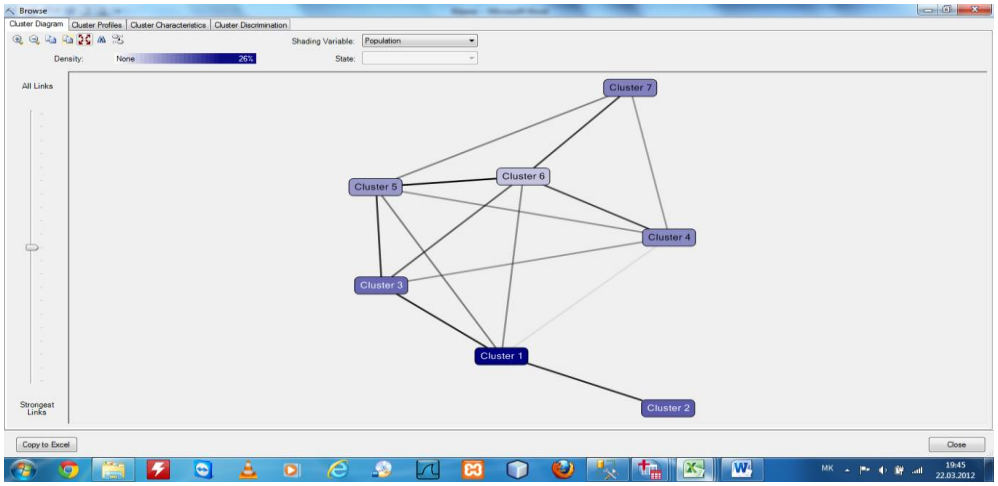
Слика 37 го прикажува издвојувањето на променливите од категориите:



Слика 37. Приказ на променливите издвоени за кластерирање

Figure 37. Review of the variables selected for clustering

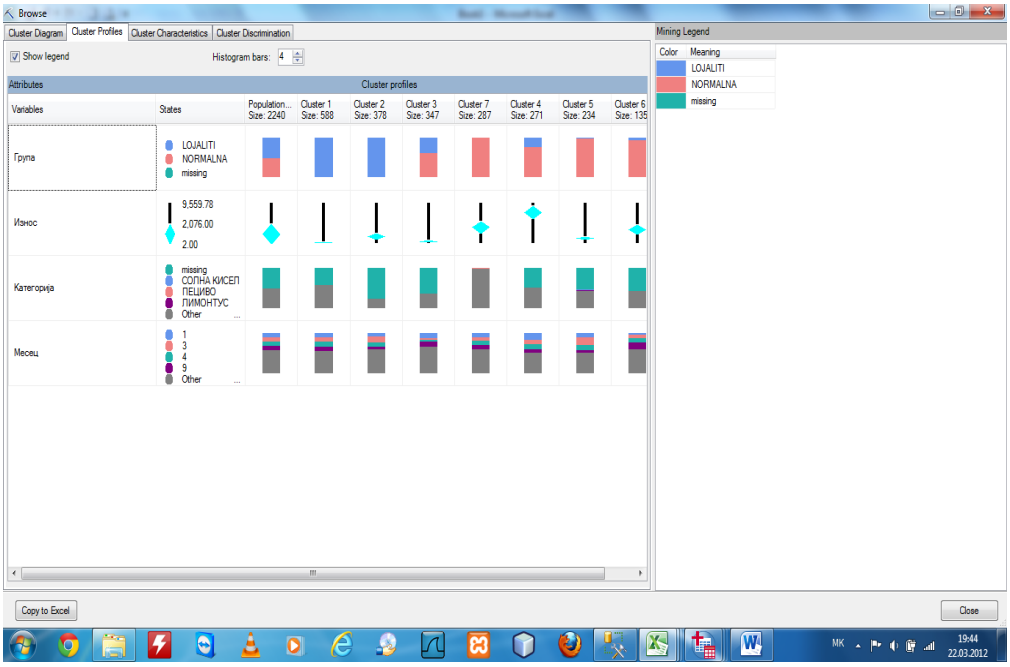
По извршеното кластерирање се добиваат 6 кластери кои се прикажани на слика 38:



Слика 38. Број на кластери во процесот на кластерирање

Figure 38. Number of clusters in the process of clustering

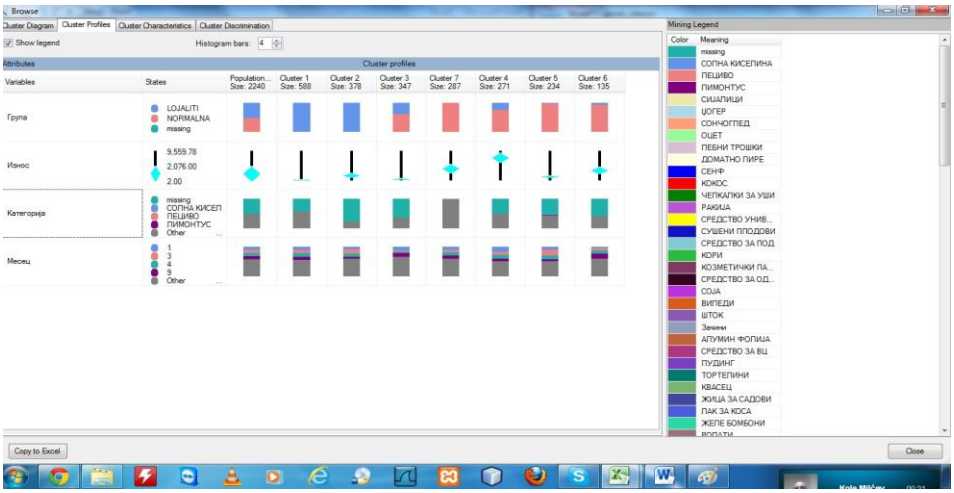
Ако ги погледнеме кластерите поединечно го добиваме следново:



Слика 39. Приказ на кластерите добиени во процес на кластерирање

Figure 39. Review of clusters received in the process of clustering

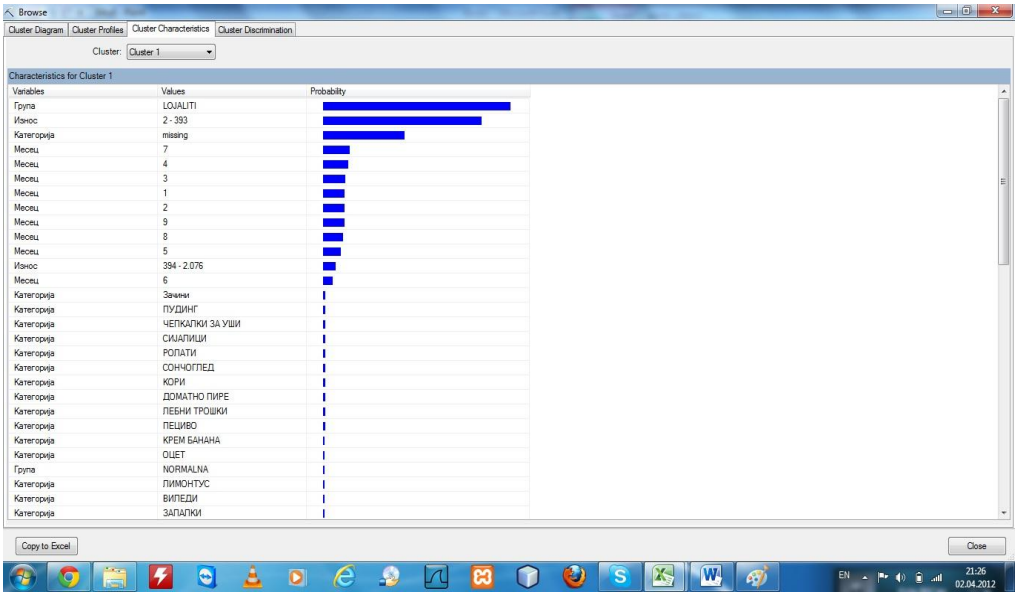
Од слика 39 јасно се гледа дека групата купувачи со издадена карта на лојалност се издвојува во процесот на кластерирање во 4 од 6 добиени кластери, но кластерите 1 и кластерот 2 целосно отпаѓаат на категоријата лојални купувачи. Која категорија на производи се издвојува е дадено на слика 40:



Слика 40. Издвоена категорија производи

Figure 40. Selected category of products

Карактеристиките на издвоениот кластер 1 се гледаат на слика 41:



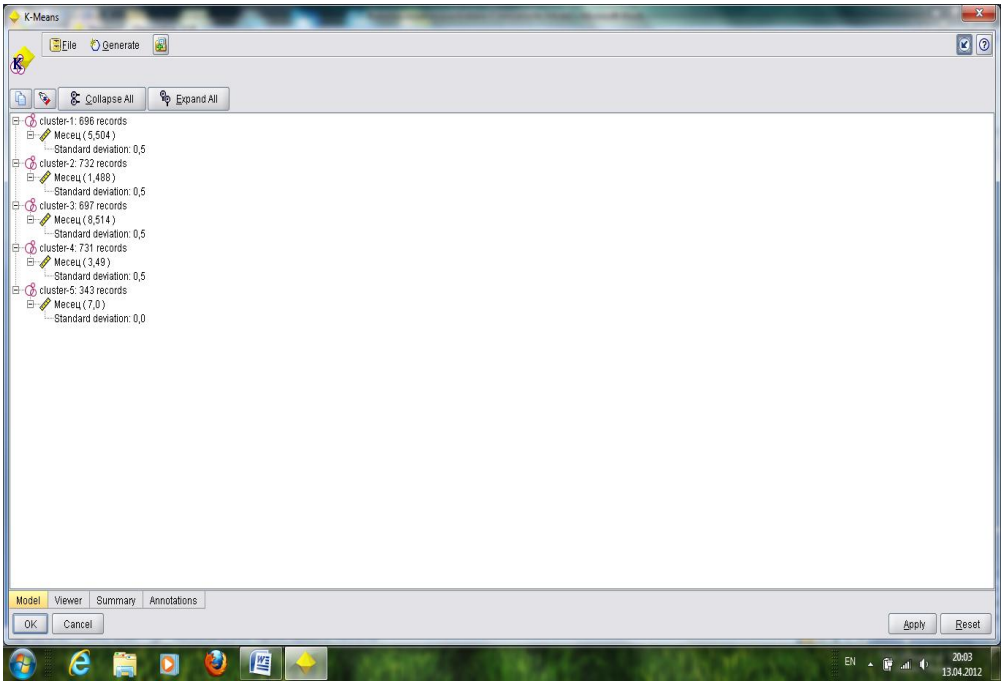
Слика 41. Карактеристики на издвоен кластер

Figure 41. Characteristics of the selected cluster

Бидејќи во базата податоци производите не се групирани по категории како прехранбени, хигиенски, освежителни и останати, туку производите се со нивните имиња, од неа и од кластерирањето се издвојуваат два сегмента на лојални купувачи и тоа прв- оние кои најмногу купуваат прехранбени артикли, како тортелини, ролати, кори, кечап, мајонез, млеко и др. Истите поретко купуваат сувомесни производи. Како втор сегмент на лојални купувачи се оние кои купуваат хигиенски производи, како средство за садови, плочки, средства за чистење, а многу поретко купуваат прехранбени артикли. Ако направиме

споредба со добиените резултати од статистичката анализа кај податоците во анкетниот лист, категоријата хигиенски производи не заземаше значајно место во омилените артикли на анкетираните. Но, сега лојалните купувачи со издадена лојална карта се издвојуваат како големи потрошувачи на хигиенските производи. Бидејќи, претходно е направена и анализа на групирање по сродност, истите хигиенски производи може да се продаваат во парови или по групи. Ова важи и за прехранбените производи.

Со кластерирање ги издвојуваме и месеците во кои најчесто овие два сегмента на купувачи купуваат. Добиени се следниве кластери:



Слика 42. Месеци во кои најчесто купуваат лојалните купувачи

Figure 42. Months in which loyal customers mostly buy

Од слика 42 се гледа дека двата сегмента лојални купувачи најчесто купуваат во месеците 1, 3, 5, 7, 8. Тоа значи дека во овие месеци потребно е да се направат промотивни активности на категоријата прехранбени производи и хигиенски производи. Дури би требало и да се планира да се комбинираат двете категории на производи на следниов начин- за сегментот купувачи кои купуваат прехранбени производи, заедно со нив би требало на акција да се стават и дел од хигиенските производи, и обратно, за сегментот купувачи кои купуваат најчесто хигиенски производи, во акција со нив би требало да одат и некои прехранбени производи, со цел да купувачите на прехранбени производи

да почнат да трошат повеќе и хигиенски производи, и обратно. Тоа секако ќе им ги зголемува и поените на картата на лојалност, а со тоа ќе се зголеми и продажбата во продавницата.

8. Поддршка на одредени бизнис одлуки

Врз основа на анализата направена со купувачите од анкетниот лист и фискалните сметки на купувачите и базата на податоци на менаџерот на компанијата му се соопштени сите резултати, врз основа на кој тој треба да донесе правилни бизнис одлуки. Некои од нив се:

1.Купувањето производ со пониска цена е единствено зависно од месечниот приход на купувачот, а овде самиот менаџер не може да влијае, затоа што се покажува дека поголем број од анкетираните купувачи имаат помал приход. *Намалувањето на цената на некои производи, посебно прехранбените би го зголемил бројот на купувачи со помал приход.*

2*. Веднаш е потребно да изготви карта на лојалност за купувачите, а целна група треба да му бидат купувачите на возраст од 30 до 40 години и со две или три деца.* Картата на лојалност секако дека значи дека продавницата ќе го зголеми бројот на лојални купувачи, а со тоа и профитот на истата. Издавањето карта на лојалност ќе допринесе и за полесно спроведување на следни маркетиншки кампањи од страна на фирмата, затоа што лојалните купувачи многу почесто ќе бидат известувани за сите акции кои ги спроведува фирмата, а и за прилагодувањето на продавањето на артиклите според нивните потреби.

3. *Најчесто барани артикли во продавницата му се прехранбените артикли*, а тоа значи дека тие и ќе му донесат најголем профит.

4.*Утврдени се групите артикли кои се купуваат заедно, така да пожелно е во продавницата истите да бидат поставени поблизу еден до друг.*

5.Утврдени се и артиклите кои не се купуваат заедно

6. Најчесто барани производители му се фирмите „Екстра Меин“, „Жива, Промес“, „Здравје Радово“, „Жито Лукс“, „Жито Балкан“, „Идеал Шипка“ и со истите би требало да ја одржува соработката

7.*Не е потребно да врши промена на продавачите во неговата продавница*

8.Потребно е почесто да прави акции, посебно на прехранбените и хигиенските производи, бидејќи најголем дел од купувачите го бараат тоа.

9. Издавањето карта на лојалност на целната категорија купувачи се покажа како правилна бизнис одлука, бидејќи само со издадени 100 карти на лојалност 98 од нив активно се користат во продажбата, а во одреден период на нив отпаѓа и 80% од вкупната продажба. Менаџерот треба да размислува на издавање нови карти на лојалност, па дури и испраќање на одредени понуди и каталози по електронска пошта на лојалните купувачи.

10. Предвидена е продажбата во следниот период, односно се предвидува приходот кој би можел да биде остварен во текот на следниот период, па во моментот кога е предвидена дека таа се намалува, менаџерот може истата да ја зголеми со ударни акции во тој период.

11. Промотивните активности на прехранбените и хигиенските производи е најдобро да бидат во месеците 1, 3, 5, 7 и 8 и др.

12. За сегментот купувачи кои купуваат прехранбени производи, заедно со нив би требало на акција да се стават и дел од хигиенските производи, и обратно, за сегментот купувачи кои купуваат најчесто хигиенски производи, во акција со нив би требало да одат и некои прехранбени производи, со цел да купувачите на прехранбени производи да почнат да трошат повеќе и хигиенски производи, и обратно.

Трговските претпријатија кои настојуваат да ја зголемат продажбата во главно се одлучуваат да применат две основни стратегии. Првата стратегија се однесува на активирање на нови клиенти (купувачи), додека втората се фокусира на веќе постоечките купувачи. Со донесување на подржани одлуки врз основа на обработените податоци во овој магистерски труд би можело да се применат и двете стратегии. Во множеството производи потребно е да се изберат оние производи или пак производ, кои купувачот кој преферира купување на одреден производ или множество производи најверојатно ќе прифати и додатно купување во нашиот случај, во продавницата мотивиран со понудата, посебно со оние производи кои се на акција и со собирањето поени со картата на лојалност. За да веројатноста на прифаќање на понудата биде поголема, таквите понуди во главно се формулирани на следниов начин: ако купувате производ X, остварувате право на купување на производот У и Z со одреден процент на попуст. Тоа значи дека во продавницата може да се зголеми профитот и доколку на лојалните купувачи кои најчесто се одлучиле за

категоријата прехранбени производи *им се понуди да купат производи за кои тие не биле толку заинтересирани, но сега истите се на попуст и се продаваат заедно со прехранбен производ кој купувачите го преферираат.*

Вакви и уште многу други одлуки ќе му помогнат на менаџерот за понатамошни стратешки активности на пазарот.

9. Заклучок

Анализата на пазарот и спроведувањето на анкета имаат свои предности во откривањето на пазарните трендови кои не може да се анализираат само со помош на базите на податоци. Идеално е да се комбинираат и двете методологии за да се добие што е можно појасна слика за продажбата.

Во овој магистерски труд се изврши комбинација на анализи од спроведената анкета, фискалните сметки и базата на податоци. Истите беа обработени со алатките за податочно рударење и анализирани со помош на техниките на податочното рударење и се претвори во корисни информации кои дадоа поддршка на одредени бизнис одлуки во компанијата каде се спроведе истражувањето и анализите. Беа користени оние техники на податочно рударење кои одговараа на анализираните податоци и се покажа дека квантитетот на податоците не игра пресудна улога во ерата на информации, туку квалитетот на истите. Одредени техники на податочно рударење како пристапот до најблизок сосед, мемориски засновано одлучување и генетските алгоритми не беа применливи во овој магистерски труд поради природата на податоците, но тоа не значи дека истите не можат да се применат на податоци екстрахирани од други извори и на друг начин, или пак да бидат вклучени во останатите техники на податочно рударење.

Сите анализи спроведени во овој труд се однесуваат на една продавница со различни категории на артикли. Резултатите кои произлегоа докажуваат дека техниките на податочно рударење на многу ефикасен начин помагаат да се донесат одредени бизнис одлуки, од кои понатаму ќе зависи и иднината на една фирма. Исти вакви анализи секако дека е можно да бидат направени и во други гранки, не само во продажбата, туку и во производството, индустријата, банкарството, здравството, образованието и др.

Во нашата земја податочното рударење, односно техниките на податочно рударење сè уште не се разбираат сериозно и не се доволни имплементирани во системите за поддршка на одлуки. Кај нас сè уште се користат статистички анализи за да се дојде до некои информации, или пак иако добро функционалните софтвери се добри складови на податоци, истите едноставно се обработуваат на традиционален начин. Овој магистерски труд е доказ дека техниките на податочно рударење се многу поефективни од

традиционалните статистички обработки на податоците, дека е можно со помал квантитет на податоци да се дојде до многу важни информации.

Податочното рударење како составен дел од Бизнис интелигенцијата е новата сила која им е потребна на компаниите за да донесат правилни бизнис одлуки. Секако, овде не треба да се занемари фактот дека човекот е главниот чинител во процесот на донесување одлуки, а информацискатa технологија е само средство со чија помош се доаѓа до квалитетни информации кои може да влијаат на процесот и видовите на донесување на одлуки.

Користена литература

[1] Badami V. (2003) Payback on Business Intelligence

[2] Berry M., Linoff G.(1997). Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley & Sons

[3] Berry M., Linoff G. (2000). Mastering Data Mining, John Wiley & Sons

[4] Bigus J.P. (1996). Data Mining with Neural Networks, McGraw-Hill

[5] Bramer Max. (2007). Principles of Data Mining, Springer-Verlag

[6] Capena P. (1997). Discovering Data Mining: From Concept to Implementation. Prentice Hall, Englewood Cliffs, New York

[7] Clement R T. (1996). Making Hard decision- An introduction to Decision Analysis, second edition. Dexburry Press, Pacific Grove

[8] Ciric Bojan. (2006). Poslovna inteligencija, Data status

[9] Edgar E. Peters. (1999). Patterns in the dark: Understanding risk and financial crisisi with complexity theory, John Wiley & Sons, Inc

[10] Han J., Kamber M. (2000). Data Mining: Concepts and Techniques, Morgan Kaufmann

[11] Han J., Kamber M. (2001). Data mining: concepts and techniques, Morgan Kaufmann, San Francisco

[12] Immon W.H. (1994). Using the Data Warehouse, John Wiley & Sons, New York

[13] Immon W.H. (1996). Building the Data Warehouse, John Wiley & Sons, New York

[14] Immon W.H., Welch J.D., Glassey K.L.(1997). Managing the Data Warehouse, John Wiley & Sons, New York

[15] Immon W.H., Rudin K., Buss C.K., Sousa R. (1999). Data Warehouse Performance, John Wiley & Sons, New York

[16] Institut Ruđer Boškovid, Otkrivanje znanja dubinskom analizom podataka, Priručnik za istraživače i studente, http://lis.irb.hr/Prirucnik/prirucnik-otkrivanje-znanja.pdf

[17] Javorović, B., Bilandžić, M. (2007). Poslovne informacije i business intelligence, Golden marketing – Tehnička knjiga

[18] JORION PHILIPPE. (2007). Financial Risk Manager Handbook, Fourth Edition, John Wiley & Sons, Inc

[19] Kienan, B. (2000) Small business solutions: E-commerce, Microsoft Press Redmond, Washington
[20] Klepac Goran I Leo Mrsic. (2006). Poslovna inteligencija kroz poslovne slucajeve, Lider, Tim
[21] Лаудон Кенет., Џејн.П.Лаудон. (2010). Менаџмент на информациски системи, Аламина
[22] Liautaud B. (2001). E-Business Intelligence: Turning information into Knolwledge into Profit, Mc-Graw Hill, New York
[23] Ljubetić Višnja. (2005). UPRAVLJANJE ZNANJEM PRIMJENOM ALATA POSLOVNE INTELIGENCIJE, MAGISTARSKIRAD, http://www.skladistenje.com/download/Visnja_Ljubetic.pdf
[24] Mattison R. (1996). Data Warehousing, Strategies Technologies and Techniques, McGraw-Hill
[25] Middlebrooks A, Graig T. (1999). Market Leadership Strategies for Service Companies: Creating Growth, Profits and Customer Loyality, NZC Publishing Group, New York
[26] Mercer David. (2005). Building Online Stores with osCommerce, Professional Edition
[27] MRŠIĆ Leo. (2004). Primjena metoda rudarenja podataka u trgovini tekstilnim I srodnim proizvodima http://www.skladistenje.com/download/MrsicLeo_MagistarskiRad.pdf.
[28] Њуболд Пол, Вилијам Л.Карлсон, Бети Торн. (2010). Статистика за бизнис и економија, Магор
[29] Oreščanin D.(2003). BI – hit ili mit, Banka poseban prilog
[30] Pallant Julie. (2009). SPSS prirucnik za prezivljavanje, Mikro knjiga
[31] Panian Ž., Klepac G. (2003). Poslovna inteligencija, Masmedia
[32] Panian Zeljko i suradnici. (2007). POSLOVNA INTELIGENCIJA Studije slučajeva iz hrvatske prakse, Narodne novine
[33] Richeldi, M., P. Lanzi. (1996). Performing effective feature selection by investigating the deep structure of the data, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp.379-383.AAAI Press
[34] Taylor, J.G. (1996). Neural networks and their applications, John Wiley&sons Inc

[35] Van Greuning Hennie, Sonja Brajovic Bratanovic.(2009). Analyzing Banking Risk *a Framework for Assessing Corporate Governance and Risk Management* 3rd Edition, THE WORLD BANK

[36] Vujovic Slavko.(2005). Elektronsko poslovanje i poslovna inteligencija, Cugura Print

[37] Wang John. (2005). Encyclopedia of Data Warehousing and Mining Second Edition, Idea Group Publishing

[38] Westphal C., Blaxton T., (1998). Data Mining Solutions, Methods and Tools for Solving Real-World Problems, J. Wiley & Sons

[39] Witten I.H., Frank E. , (2001). Data Mining, Morgan Kaufmann Publishers

[40] Young, Peter C.; Tippins, Steven C. (2001). Managing Business Risk : An Organization-wide Approach to Risk Management, Amacom

ПРИЛОГ БР.1

Компанија Моневи

АНКЕТЕН ЛИСТ

Почитуван купувач!

Пред тебе се поставени неколку прашања. Ве замолуваме да во врска со истите добро размислите и искрено одговорите!

Возраст:

а)20-30 години

б)30-40 години

в)40-50 години

г)над 50 години

Месечен приход

а) 0-5000 денари

б) 5000-15000 денари

в) 15000-25000 денари

г) над 25000 денари

Брачна состојба

а)самец

б)оженет/омажен

Број на деца

а)1 дете в)3 деца

б)2 деца г)повеќе од 3 деца

д)немам деца

1.Колку често купувате производи(артикли) од нашата продавница?

а)често

б)многу често

в)ретко

2.Цените на артиклите во нашата продавница во споредба со другите се:

а)високи б)исти в)ниски

3.Наведете најмалку 5 артикли кои најчесто ги купувате во нашата продавница!

_____ _____ _____

_____ _____ _____

4.Напишете го производителот на 5-те погоре наведени артикли!

_____ _____ _____ _____ _____

5.Дали поништата цена е одлучувачки фактор за Вас да купите еден артикал?

а)да б)не

6.Дали сметате дека квалитетот на производот е поврзан со цената на истиот?

а)да б)не

7. Дали сте задоволни со услугите од нашите продавачи?

а)да б)не

8.Дали би сакале да имате Карта на лојалност од нас, со која би добивале одредени бенифиции од нашата фирма како лојален купувач?

а)да б)не

9.Дали практикувате да купувате производи на акција?

а)да б)ретко в)не

10. Наведете најмалку три артикли кои би сакале да бидат на акција во текот на следниот период?

_____

Ви благодариме на соработката!