Универзитет "Св. Климент Охридски"- Битола

Економски факултет - Прилеп

Дејан Здравески, дипл. ек.

ОТКРИВАЊЕТО НА ЗНАЕЊЕ ВО АНАЛИТИЧКИТЕ БАЗИ НА ПОДАТОЦИ (DATA MINING) ВО ФУНКЦИЈА НА ДЕЛОВНОТО РАБОТЕЊЕ НА КОМПАНИИТЕ

- МАГИСТЕРСКИ ТРУД -

Ментор:

Проф. д-р Коста Сотироски

Прилеп, 2008 година

С о д р ж и н а

I DEL

VOVEDNI NAPOMENI, CELI I METODOLOGIJA

NA ISTRA@UVAWETO

II DEL

INFORMACISKATA TEHNOLOGIJA I NEJZINATA PRIMENA

VO DELOVNITE SISTEMI

1. Informaciite vo delovnoto odlu~uvawe
2. Globalen pazar
3. Pazarot kako inicijator za razvoj na informaciskite tehnologii
4. Aspekti na primena na informaciskite tehnologii
5. Menaxersko izvestuvawe
6. Sistemi za poddr{ka na odlu~uvaweto
7. Efekti od primenata na informaciskata tehnologija vo delovniot svet

III DEL

DATA MINING I IZGRADBA NA MODEL ZA DATA MINING

1. Definicija za Data Mining
2. Arhitektura i funkcionirawe na Data Mining
3. Skladi{te na podatoci (Data Warehouse)
   - 3.1 Celi, struktura i komponenti na Data Warehouse
   - 3.2 Arhitektura na Data Warehouse
   - 3.3 Logi~ki dizajn na Data Warehouse
   - 3.4 Primer za izgradba na Data Warehouse vo opredelen deloven sistem
4. Izgradba na model za Data Mining
   - 4.1 Definirawe na celi
   - 4.2 Pronajduvawe i izbor na podatoci
   - 4.3 Podgotovka na podatocite

4.4 Izgradba na model za Data Mining

4.5 Vrednuvawe na modelot za Data Mining

4.6 Implementacija na modelot za Data Mining

IV DEL

OSNOVNI METODI I ALGORITMI ZA DATA MINING

1. Metodi i algoritmi {to se koristat za Data Mining
   1.1 Klasterirawe
   1.2 Klasificirawe
   1.3 Procenka i predviduvawe
   1.4 Analiza na verojatnosta
   1.5 Regresiona analiza
   1.6 Steblo na odlu~uvawe
   1.7 Nevronski mre`i
   1.8 Geneti~ki algoritam
   1.9 Nejasna logika
2. Klasifikacija na metodite za Data Mining
3. Izbor na metod
4. Analiti~ka obrabotka na podatocite – OLAP

V DEL

PRIMENA NA DATA MINING VO DELOVNIOT SVET

1. Oblasti na primena na Data Mining
   1.1 Primena na Data Mining vo marketingot
   1.2 Primena na Data Mining vo bankarstvoto i finansii (osobeno vo oblasta na investiciite)
   1.3 Primena na Data Mining vo proizvodstvoto
   1.4 Primena na Data Mining vo otkrivaweto na gre{ki
   1.5 Primena na Data Mining vo telekumunikacii
   1.6 Ostanati oblasti na primena na Data Mining
2. Specifi~ni oblasti na primena na Data Mining

2.1 Data mining na vremenski serii

2.2 Text Data Mining

2.3 Web Data Mining

VI DEL

ASPEKTI NA PRIMENA NA METODITE ZA DATA MINING VO DELOVNOTO RABOTEWE NA KOMPANIITE

ZAKLU^NI SOGLEDUVAWA

KORISTENA LITERATURA

VOVEDNI NAPOMENI, CELI I METODOLOGIJA

NA ISTRA@UVAWETO

Компаниите денес работат во окружување што е проследено од висок степен на несигурност и мали можности за предвидување на иднината. Во такви услови на динамично и турбулентно окружување, каде што на едната страна е конкуренцијата, а на другата страна се се поголемите очекувања и барања на потрошувачите, растат и потребите за квалитетно управување. Ваквите радикални промени во деловното окружување ја наметнуваат потребата за користење такви управувачки алатки и технологии што овозможуваат сеопфатно, брзо и ефикасно користење на сите достапни податоци и информации, важни за управување како внатре во компаниите, така и надвор од нив.

Кон крајот на деведесеттите години од минатиот век, а особено во дваесет и првиот век и во Р.Македонија се повеќе се чувствува трендот на глобализацијата и силните надворешни влијанија и притисоци на нестабилното окружување. Компаниите се соочени со конкуренција на сите подрачја од деловното работење, па поради тоа тие се принудени што подоследно да ги користат сите свои расположливи потенцијали.

Од постојните потенцијали што им стојат на располагање, секако најдрагоцено во современиот деловен свет е знаењето, т.е. интелектуалниот капитал. Знаења можат да се мерат и вкалкулираат во вкупната вредност на компанијата. На тоа укажува и студијата изготвена од страна на Brokings institute што покажала дека во текот на осумдесеттите години книговодствената вредност на материјалниот имот изнесувала 62% од пазарната вредност на компанијата. Десет години подоцна тој процент опаднал на 38%, а најновите истражувања покажуваат дека таа вредност сега изнесува околу 10-15% од пазарната вредност на компанијата. Според тоа, може да се каже дека со развојот на технологијата, пред се на информацискатa технологија економијата на една држава и нејзините развојни потенцијали се помалку зависат од машините, опремата и капиталот, а се повеќе од луѓето и нивниот креативен потенцијал. Технолошкиот развој, новите технолошки и научни откритија и нивната непосредна примена условуваат перманентните промени и тенденцијата кон иновативност да стануваат единствен можен начин за егзистенција како на компаниите, така на општеството во целина. Колку технологијата е поразвиена, толку тоа бара се поголеми вештини и

способност на вработените, односно на разновидни профили на високообразовани вработени. Тежиштето на работата се префрла од физичката кон интелектуалната и психичката енергија на луѓето. Во македонски услови таквите знаења и вештини често стагнираат или се губат, главно заради неодговорниот третман во смисла на постојаното учење и зголемувањето на компетенциите. Откривањето и управувањето со знаењето е неизоставен елемент на стратегијата на успешните глобални компании. Така, стратегиските потреби на организацијата за креирање и рекреирање ново знаење достигнуваат синергија по пат на процесирање на информациите со помош на напредни информациско-комуникациски технологии и користење иновативни и креативни алатки за откривање на знаењето. Познавањето на современите трендови и менаџерските алати, како и нивната примена е неопходно за развој на секоја економија, односно за стекнување конкурентска предност и обезбедување опстанок и напредок на деловните субјекти на пазарот.

Еден од современите концепти што овозможува можност за сеопфатно и продуктивно користење на податоците и информациите, е концептот за складирање податоци (Data Warehouse) и откривање знаење од аналитичките бази на податоци (Data Mining). Според досегашните информации, само 20% од компаниите користат повеќе од 50% од податоците што им стојат на располагање за зголемување на својата конкурентска предност. Ваквите концепти овозможуваат искористување и на преостанатите податоци и нивно претворање во употребливи информации. Поради растечките потреби за информации компаниите секојдневно креираат голем број податоци за нивното работење и безброј факти за нивните потрошувачи, производи, операции, вработени итн. Најголем дел од овие податоци буквално се "заклучени" во илјадниците компјутерски системи. Само мал дел од податоците што се земаат, процесираат и складираат се употребуваат од страна на компаниите за донесување одлуки. Новите информациски технологии, со силни перформанси, овозможуваат складирање на големи количини на податоци, нивна обработка и искористување за добивање информации што ќе служат како поддршка на одлучувањето и ќе отвараат големи можности за унапредување на деловните процеси, т.е. ќе обезбедуваат анализа и следење на ефикасноста на деловниот систем. Data Warehouse и Data Mining се едни од најновите дисциплини

и тие се уште се во фаза на интензивен развој. Основата за нивниот развој главно, лежи во интензивниот развој на информациските технологии и новите сознанија за информациските потреби на компаниите. Концептот на Data Warehouse е дел од одговорот, преку информацискатa технологија, да се излезе во пресрет токму на овие идентификувани потреби. Тоа е едноставно концепт што може многу да придонесе за успехот и стабилноста на компанијата, односно одржливоста на глобалниот пазар. Data Warehouse овозможува олеснување на интегрирањето на податоците, генерирани во еден свет на неинтегрирани информациски системи. Ваквиот константен циклус на генерирање податоци, нивно складирање и комплексноста при истражувањето и анализата резултираат со развој на нови и моќни алатки што би помогнале да надмине овој предизвик. Data Warehouse и Data Mining можат да бидат оружје кое би се користело за откривање на скриените поволности што се содржани во огромниот магацин на податоци и тие стануваат стандард за обезбедување компетитивна предност.

Во овој труд ќе бидат презентирани некои од Data Mining техниките што се резултат на долг процес на истражување и производ на еден нивен континуиран развој. Имено еволуцијата започнува кога деловните податоци на компаниите беа складирани во компјутерите, при што континуирано се подобруваше пристапот до нив, а во последно време се развиени технологии што им овозможуваат на корисниците да управуваат со собраните податоци во реално време.

Во последниот дел од трудот ќе бидат изнесени заклучни согледувања за ефектите од примената на Data Warehouse и Data Mining, што претставуваат две моќни организациски алатки. Исто така, ќе биде согледано колку нивното познавање може да придонесе за зголемување на конкурентската предност на компаниите, што денес претставува еден од клучните фактори за успех во бизнис работењето.

Примената на овие концепти, како и секоја друга новина е предизвик, но и неизвесност, а понекогаш предизвикува и отпор кај одделни интересни групи како во компаниите, така и во определени традиционално инертни структури во

државните институции. Таквиот отпор, а особено неразбирањето на корисноста од воведувањето на ваквите нови концепти и методологии честопати ги минимизира веќе постигнатите резултати што се остварени со големо залагање и напори пред се, од поединците или мали групи стручњаци. Со овој труд се нагласува и се потенцира предноста што ја носи примената на современите менаџерски алатки коишто ги проучува и развива современата економска и информациска наука. Конкретно, во овој случај, ќе бидат прикажани предностите што ги носи прифаќањето и практичната примена на концептот на Data Warehouse и Data Mining . Прифаќањето на ваквите концепти од страна на компаниите е еден од начините што може да ја приближи Р. Македонија кон кругот на успешните и просперитетни земји. Токму таквото "паметно" искористување на интелектуалниот капитал и дава шанса на нашата земја во што пократок период, ако не ги достигне, барем да се доближи до развиените држави и успешно да се вклучи во глобалниот пазарен натпревар, со нагласок на она окружување кон коешто природно тежнее, во рамките на Европската унија.

Предмет на истражување во овој труд претставува примената на методите за Data Mining во насока на подобрување на одлучувањето во компаниите. Најпрвин ќе се направи теоретска елаборација на основните поими и разгледување на основните карактеристики на Data Mining методите за потоа да биде прикажана и нивната практична употреба во рамките на одделна компанија и бенефициите што би се добиле од нивната примена. Во овој труд се претставени основните и најчесто користени методи за Data Mining, при што акцентот ќе биде ставен на оние методи што се најприфатливи за употреба, во зависност од количината, достапноста и видот на податоците. Исто така, намерата на овој труд ќе биде и да се даде преглед на постојната состојба во Република Македонија во оваа област, и тоа преку анализа на постојните состојби. Натаму, пак, ќе бидат разгледани карактеристиките на расположливите и користените алатки за Data Mining и која од тие алатки, во какви услови, може да најде најсоодветна примена.

Основната цел на овој труд е да се овозможи подлабоко осознавање на концептот на Data Warehouse и Data Mining како од теоретски, така и од практичен аспект и да се прошират знаењата за оваа информациска технологија, како главен

двигател во остварувањето конкурентска предност на компаниите. Целта се состои во нагласување на неопходноста и корисноста од примената на ваквите нови информациски технологии, што можат да имаат практична вредност во подобрувањето на процесот на деловното одлучување во компаниите.

Во натамошното разгледување на складирањето на податоците и откривањето на знаењето од аналитичките бази на податоци, како **основни задачи на истражувањето,** се следниве:

- разгледување на најприсутните информациски термини и детектирање на значењето на појавата на новите информациски технологии во се поизразените процеси на глобализација,
- поконкретно објаснување за основните поими складиште на податоци (Data Warehouse) и откривање на знаење од аналитичките бази на податоци (Data Mining)
- осврт на процесот за изградба на модел за Data Mining. Притоа ваквиот процес е разработен според фазите на неговото одвивање
- утврдување на применливоста на одделни методи и техники за Data Mining и потенцирање дека таквите методи можат да придонесат за подобрување на деловното одлучување.
- потврдување на претпоставката дека довербата во моделот за Data Mining зависи од квалитетот, достапноста и количината на расположливите податоци и колку некои од методи повеќе или помалку се соодветни за определена ситуација
- истражување на практичната примена на ваквите методи и техники, и тоа преку користење на неколку алатки за Data Mining на едноставни модели за Data Mining.

Со елаборирањето на овие задачи и нивната разработка се остварува и целта на самото истражување за на крајот да се изведе заклучок во врска со темата. Или, поточно, се дава одговор на прашањата поврзани со примената на алатките за Data Mining, т.е. колку се тие соодветни и прифатливи, колкави се нивните можности, предности и недостатоци, како можат да придонесат во подобрување на процесот на одлучување, итн.

Методологијата на истражувањето на теоретските основи на концептот за Data Warehouse и Data Mining и практичната примена на овие информациски технологии е определена од самите карактеристики на овој концепт, што е заснован на повеќе дисциплини како статистика, системи на бази на податоци, вештачка интелегенција итн. Токму поради тоа, најпрвин е образложен теоретскиот дел при што се објаснети термините, моделите, методите и техниките што најчесто се користат при примената на овој концепт, а потоа е прикажана и можноста за употреба на соодветните методи и техники, за на крајот да биде прикажана и практичната примена на некои Data mining алатки.

Сложеноста на ваквата проблематика налага користење различни графички презентации, дијаграми, табели, соодветни блок – дијаграми и модели како и користење на аналитички и математичко статистички методи. На крајот, секако, ова подразбира и примена на компјутерски, т.е. Data Mining апликации за реализација на моделите.

Со помош на овие методи и модели се дава можност да се запознаат основните особености на концептот на Data Warehouse и Data Mining, да се разгледаат квалитативните и квантитативните својства на резултатите добиени со примената на овој концепт, како и различните решенија на одделни деловни проблеми што можат да се добијат со примената на ваквите современи менаџерски алатки. Овој труд ги спојува новите научни достигнувања на економската теорија и информацискатa технологија, којашто служи како алат за унапредување на самиот деловен систем.

I ДЕЛ
ИНФОРМАЦИСКАТА ТЕХНОЛОГИЈА И НЕЈЗИНАТА ПРИМЕНА ВО ДЕЛОВНИТЕ СИСТЕМИ

1.Информациите во деловното одлучување

Информациите и знаењето се клучни ресурси за успешно деловно работење. Менаџерите постојано се соочуваат со растечкиот број интерни и екстерни информации. Од нив се бара брзо донесување определени одлуки, па пристапот до целосни, навремени и точни информации се повеќе станува неминовност, а информациите добиваат стратегиска важност. Одлуките, по правило, се донесуваат врз основа на информациите, а со реализација на донесените одлуки доаѓа до одделни промени во компанијата или во окружувањето. “ **Информацијата е таков податок што во текот на деловната активност обезбдува нови сознанија коишто врз основа на постојните знаења и искуства кажуваат повеќе за промената на некоја состојба внатре во организацијата или окружувањетоа што претходно проверена и анализирана претставува соодветна основа за донесување одлука за влијанието врз понатамошниот тек на деловната активност”.**[1]

Доколку информацијата се набљудува како **ресурс,** тогаш може да се забележи дека нејзините карактеристики се исклучителни. Но може да се каже дека и доколку информацијата се дели или се пренесува, таа не губи од својата вредност. Информацијата не се покорува на законот за зачувување на материјата и енергијата. Така, денес, сите останати ресурси почнуваат да зависат од информациските ресурси, бидејќи без поседување информации за состојбата, местото и начинот на нивното користење, тие стануваат бескорисни.

Составот на информациите мора да биде прилагоден на вистинските информациски потреби на менаџментот и насочен кон донесување целисходни деловни одлуки, во зависност од специфичноста на одделните дејности. Токму затоа, најчесто се поставува барање интерните информации да го прикажуваат работењето на компаниите синтетички (за целата компанија) и аналитички (на пример, разложувајќи ги извештаите по организациони единици), па мораат да искажуваат определени состојби, трендови или отстапувања. Екстерните

[1] www.corporateinformation.com

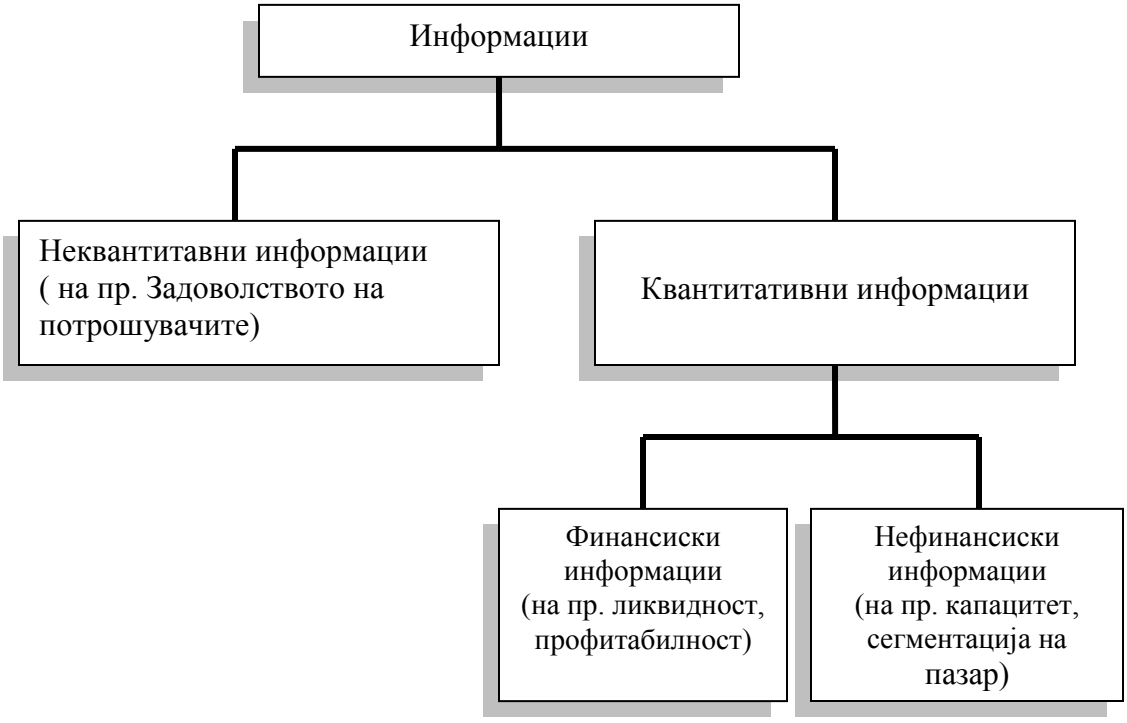
информации, пак, треба да ги искажуваат економските трендови и учинокот на конкуренцијата.

Значењето на информацијата можно е да се разгледува од три аспекта и тоа:

- **синтактичко**. Информацијата содржи порака што е вистинита и потполна,
- **семантичко**. Содржината на информацијата мора да биде јасна и разбирлива за корисниците,
- **прагматичко**. Информацијата треба да ги задоволува потребите на корисникот и да ја редуцира низвесноста при донесувањето на деловни одлуки

Во процесот на обработка и/или анализа податоците се претвораат во информации што се презентираат во најприфатлив облик за корисникот. Информациите го зголемуваат знаењето за фактите, обележјата или настаните.Честопати информациите за тактичкото ниво на управување се податоци што треба да се обработат за да се добијат специфични информации за стратегиското ниво на управување. На слика 1 е даден процесот на развој од податоци-до мудрост, а на слика 2 е дадена вообичаената класификација на информациите за потребите на управувањето.

Слика 1. Од податоци до мудрост[2]



Слика 2. Класификација на информациите за потребите на управувањето

[2] www.attar.com

Важно е да се напомене дека битен услов за ефикасна обработка и користење на информациите е нивната дигитализација. Информациите, во дигитален облик, лесно се пренесуваат низ компјутерските мрежи и се користат во базите и складиштата на податоци, односно во информациските системи.

2.Глобален пазар

Не постои единствена и општоприфатена дефиниција за **глобализацијата.** Сепак доколку е по,требно да се даде некаква дефиниција, тогаш таа би била во согласност со ставот на најголемиот број теоритичари кои се согласуваат со фактот дека глобализацијата е процес што води кон интеграција на целиот свет во сите цивилизациски аспекти. Според тоа, глобализацијата претставува квантитативно и квалитативно проширување и надминување на границите во движењето на активностите и идеите. Некои аргументи укажуваат дека глобализацијата претставува квалитативен чекор напред, во однос на претходните трендови како, на пример "интернационализацијата"[3].

Појавата на новите информациски технологии, пред се, напредокот на комуникациските технологии ги приближи сите учесници на пазарот, додека локалната економија, прерасна во глобална економија. Просторната оддалеченост меѓу партнерите во деловното работење веќе не претставува никаква пречка, а единствено што се јавува како проблем можат да се наведат малите разлики во законодавствата на одделни земји. Така, расцепканите мали локални пазари се поврзуваат во еден голем глобален пазар.

Нормалното функционирање на компаниите на ваквиот глобален пазар бара подобрување на продуктивноста и зголемување на нивната конкурентност. Во таа смисла, користењето на информациските технологии, а особено експлозивниот развој на интернетот претставуваат основа за остварување на претходно наведените цели. Ова, осебено поради тоа што глобализацијата и информациските технологии, како **двигатели на информатичката револуција** во светот имаат поголемо и

[3] Dale Dauten, Economists miss the boat in ignoring globalization's impact, Article june 10, 2007

поквалитативно влијание од кој било друг феномен досега[4]. Сите технологии што денеска се користат брзо напредуваат, но информациските технологии, без сомнение, имаат најголемо влијание врз глобалната економија, правејќи го собирањето, обработката и пренесувањето на информациите многу побрзо и со помали трошоци, а со самото тоа ја зголемуваат продуктивноста и го подобруваат квалитетот на производите и услугите во сите индустрии и услужни дејности.

Ова, секако, е поттикнато од брзиот развој на информациската технологија. Денес, се повеќе се споменува поимот дигитално општество, односно електронски и виртуелни групи на соработници на глобалниот пазар, кои се поврзани со напредни комуникациски технологии. Наместо индустриското општество, предност добива дигиталното општество, особено со својот услужен сектор.

3.Пазарот како иницијатор за развој на информациските технологии

Основна премиса во деловното работење на компаниите во современите услови е постигнувањето задоволство кај клиентите, и тоа преку задоволување на многубројните и разновидни потреби. Тоа подразбира дека најголемо внимание се обрнува на клиентите(потрошувачите), нивното задоволство, преференции, барања и желби. Конкурентската предност на производителите на производи и давателите на услуги е можно да се постигне само преку градење позитивен имиџ, уникатност на производите и услугите, различни нови погодности, бонуси, награди и намлување на цените. На овој начин може да се зголеми лојалноста на клиентите, односно да се равие нивната доверба.

Информациските тенологии стануваат еден од најважните аспекти од работењето на компаниите, бидејќи не само што овозможуваат автоматизација на деловните процеси, туку можат да процесираат огромна количина податоци со голема брзина, да го забрзуваат изведувањето на вообичаените операции и да овозможуваат следење и управување со некои процеси. Тоа со вообичаениот

[4] Fred W. Riggs, Henry Teune, Andjelko Milardovic (ur)- Mali Leksikon Globalizacije, Pan Liber, 2001, str 56

"хартиен" начин беше преопширно и прекомплицирано, така што современите технолошки решенија ги оптимизираат тековите на материјали и информации, ги намалуваат трошоците и влијаат врз промената на начинот на работење на компаниите.

За константно унапредување на производите и услугите компаниите денес се повеќе вложуваат во развојот на информациските технологии. Иновациите на подрачјето на информациските и комуникациските технологии недвосмислено овозможија понуда на такви производи и услуги какви што до пред неколку децении не можеа ниту да се замислат. Новите технологии овозможија зголемување на уделот на информациите и знаењето во конечниот производ и услуга. Поради тоа, се почесто се зборува за **економија на знаење** (knowledge economy), бидејќи знаењето станува главен чинител во пазарниот натпревар меѓу компаниите[5].

Во овој труд натаму ќе стане збор токму за новата технологија за откривање на знаењето од аналитичките бази на податоци (Data Mining), односно во поширок контекст business intelligence.

Информациските технологии можеби се развиваа и побрзо, отколку што пазарот можеше да ја спознае нивната примена. Меѓутоа, пазарот, со појавата на ваквите напредни технологии, бара брз развој само на определени гранки од информациските технологии. Типична потврда на ова е примерот на интернетот што требаше да служи како информациска инфраструктура само за одделни научни кругови, но набрзо беа препознаени неговите можности, главно, од аспект на понудата на производи и услуги со помош на електронската трговија (e-commerce).

Но, не секогаш воведувањето на новите информациски тенологии обезбедува бенифиции. Новите придобивки од воведувањето овие информациски технологии едноставно можат да бидат приграбени само од определен богат и моќен слој на населението. Така, на пример, **Зелената револуција**[6] во Индија, што беше резултат на имплементацијата на новите информациски технологии само го

[5] Singh, S., Using Information Technology Effectively, Information & Management 24, 1999
[6] Ernest J. Wilson, Globalization, Information Technology and Conflict in the Second and Third Worlds, Rockefeller Brothers Fund, Inc., 1999, str. 58

зголеми јазот меѓу богатите и сиромашните во земјата. Оттука за да се редуцираат ваквите ефекти што водат кон зголемување на нееднаквостите е потребно внимателно развивање и креирање на IT проектите. Поради тоа, работната група на UNCSTD[7] (UN Commission on Science and Technology for Development) укажува на фактот дека ваквата лоша ситуација и јазот меѓу развиените и неразвиените може да доведе до креирање и на социјално непремостлив јаз.

4.Аспекти на примена на информациските технологии

Подрачјето на примена на информациските технологии може да се подели во зависност од затворениот круг елементи на деловните активности. Притоа можат да се разликуваат информациски технологии што овозможуваат поддршка на производителите на производи и давателите на услуги; информациски технологии што овозможуваат поддршка на оние кои ги добиваат услугите или производите; и информациски технологии што ја овозможуваат инфраструктурата којашто е потребна за да се овозможи размената на услугите и производите.

Давателот на услугата или производителот на производите користи информациска технологија за да го унапреди постојниот збир услуги или производи, но и за да продуцира нови. Користените технологии се трансакциски и димензионални бази на податоци, алати на системите за поддршка на одлучувањето и алати за пронаоѓање нови начини на примена на производите и услугите, пред се, со помош на алатите за Data Mining.

Оние пак кои ги користат производите и услугите употребуваат информациски технологии кои овозможуваат директен пристап до бараните производи и услуги.

На крајот, инфраструктурата за размена на производите и услугите претставува основа за каналите на дистрибуција на купените производи и услуги. Тоа се комуникациски технологии во облик на мрежни ресурси, како, на пример, интернетот.

[7] www.uncstd.org

Набљудувајќи го влијанието на информациските технологии од **макро** аспект може да каже дека тие одиграа клучна улога во создавањето брз развој на некои нови стопански гранки, притоа насочувајќи и концентрирајќи ги активностите на луѓето кон дејностите што на различни начини се поврзани со информациите. На **микро** план, пак, во рамките на компаниите, со интензивна примена на информациските технологии се доаѓа до структурни промени што најмногу се одразуваат врз деловните процеси и активностите поврзани со третманот на информациите, во смисла на нивната автоматизација и информатизација.

5.Менаџерско известување

Околината во којашто компаниите дејствуваат и мора да се прилагодуваат можеби е најбитниот фактор за успешното деловно работење, бидејќи и нивниот опстанок всушност, зависи од прилагодливоста и информираноста на одделните организации за потесната и поширокатa средина. Од менаџерите се бара секојдневно донесување деловни одлуки што често се од голема важност и што може да имаат несогледливи последици врз деловното работење, при што се трошат огромни средства и се преземаат различни ризици. Одлуките што се донесуваат мораат да бидат конкретни и брзи, односно да се засноваат на веродостојни информации што се црпат од посебно изготвените менаџерски извештаи. Според тоа, основна цел на менаџерските извештаи е давање информации за донесените на деловни одлуки. Токму затоа на менаџерот мораат да му се овозможат точни показатели, ажурни информации и прегледни извештаи во стандардизирана форма. Преку менаџерското известување се документираат деловните настани, се поттикнуваат и надгледуваат деловните процеси и се проектираат идните активности. Според тоа, ваквото известување ја извршува една од најважните функции, а тоа е информирањето. Со оглед на важноста на извештаите во процесот на одлучување и управување потребно е тие да бидат **јасни, веродостојни, точни и објективно составени,** а со самото тоа и доверливи.

При составувањето на извештаите неопходно е да се почитуваат определените начела[8]:

- **транспарентност**. Извештаиите треба да бидат јасни, едноставни и прегледни.
- **флексибилност**. Системот за известување треба да биде прилагодлив спрема динамиката на работењето и потребите на корисниците,
- **компарабилност**. Сродните извешати или истите типови извештаи што се прават за различни временски периоди мора да даваат можност за меѓусебна споредба,
- **оптималност.** Современите информациски системи можат да обработуваат големи количини податоци, а на крајот е потребно да се добијат едноставни и употребливи информации, па затоа се јавува проблем при изборот на податоците за решавање на одделни специфични ситуации.

Менаџерите континуирано носат важни одлуки, а одлуките донесени врз основа на достапните информации не можат да се вбројат во **добри или лоши одлуки**, туку се третираат како одлуки што се донесени врз основа на добри или лоши информации[9]. Еден извор на информации се трансакциските бази на податоци, а извештаите врз основа на тие информации, најчесто, ги користат менаџерите на оперативно ниво. Втор извор на информации се димензионалните бази на податоци (складиштата). Овие информации се користат како за известување на средното ниво на менаџментот така и на стратегиското ниво на менаџментот. Третиот извор информации е изведен со помош на **откривањето на знаењето од аналитичките бази на податоци** (Data Mining), и тоа од тековните и минатите информации, а претставува збогатување на постојните информации заради донесување подобри деловни одлуки. Притоа, напредокот на ваквото известување е забележлив на сите менаџерски нивоа.

[8] Saban, Josip, Menadzerska Izvjesca, HGR 6/98, Zagreb, 1998, str.22
[9] www.mindtools.com

6.Системи за поддршка на одлучувањето

Системите за поддршка на одлучувањето се специфична класа **компјутеризирани информациски системи** што ги поддржуваат активностите коишто се поврзани со донесувањето одлуки во самата организација. Информациите што можат да бидат собрани и презентирани со помош на апликациите за поддршка на одлучувањето се[10]:

- сите тековни информации, вклучувајќи ги историските и релационите податочни извори, складиштата на податоци (data warehouses) и маркетите на податоци (data marts);
- релативните обеми на продажба меѓу еден и друг период;
- проектираниот приход, врз основа на претпоставките за продажба на нов производ;
- консеквенциите од прифаќањето на различните алтернативи на одлучување, со што се овозможува стекнување искуство во контекст на начинот на којшто се опишани

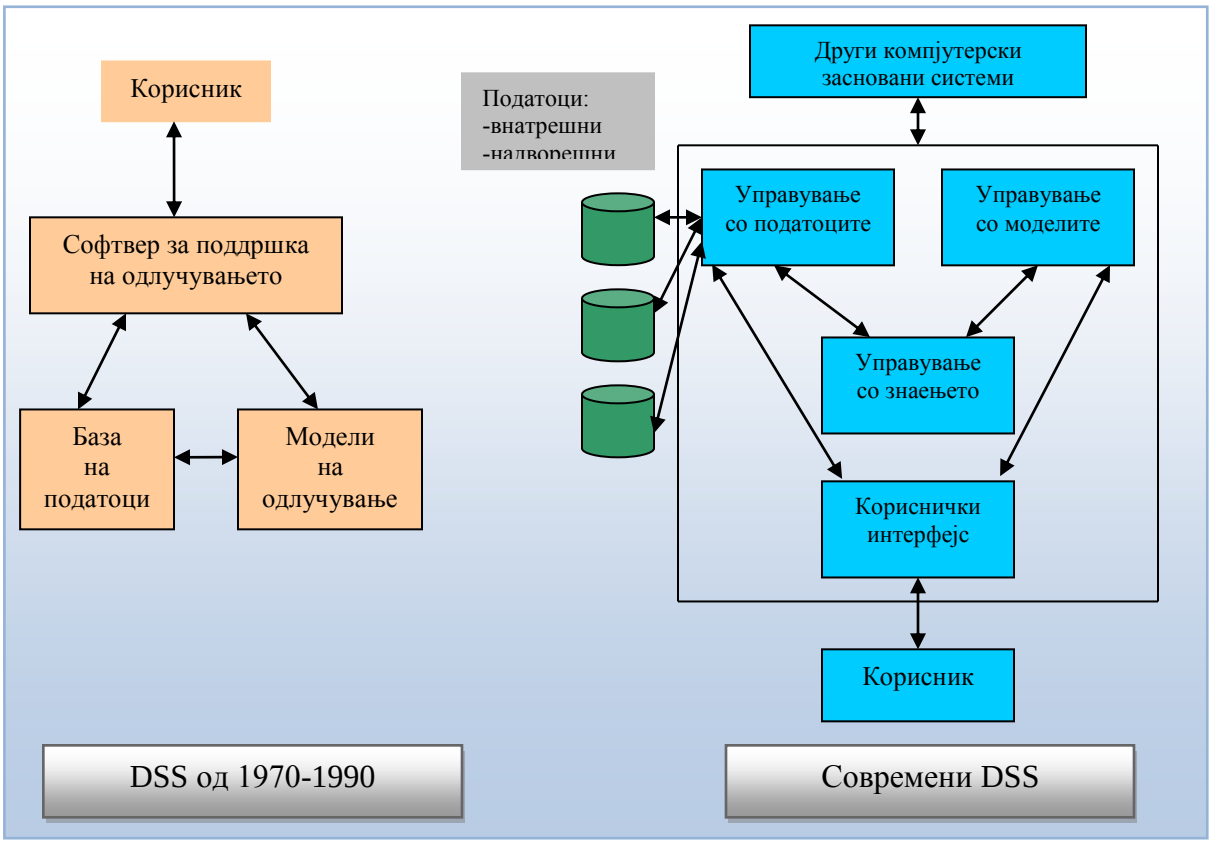
Главната цел на системите за поддршка на одлучувањето (Decision Support System-DSS) е да им помогнат на оние кои донесуваат одлуки при решавањето на проблемите и носењето на самите одлуки. Од дефиниција што ги дефинира системите за поддршка на одлучувањето произлегува дека тие се системи што им помагаат на луѓето да ја изберат најдобрата од мноштвото алтернативи што им стојат на располагање. DSS ги оценува сите алтернативи, во согласност со функцијата за **оценка/корисност** на системот.

Според тоа, системите за поддршка на одлучувањето се автоматизирани, така што овозможуваат подобрување на процесот на донесување одлуки. Ваквото подобрување е најмногу видливо на највисоките нивоа на менаџерско одлучување. Нивни основни карактеристики се: вклучуваат податоци и модели; дизајнирани се да овозможат помош на менаџерите при решавањето на **полуструктуирани и неструктуирани** проблеми; и имаат за цел повеќе да ја подобрат ефективноста на одлучувањето (точноста, ажурноста , квалитетот), отколку ефикасноста во

[10] Carlson C., Turban E., Decision Support System, Elsevier, 2002, str.45

одлучувањето (ниски трошоци за обработка на податоците и донесувањето на одлуките). Тоа укажува на фактот дека со воведувањето на системите за поддршка на одлучувањето се отстранува еден од главните недостатоци на управувачките информациски системи, а тоа е поддршката на одлучувањето само при решавање структуирани проблеми. Системите за поддршка на одлучувањето се компјутерски засновани системи и на менаџерите им помагаат да креираат квалитетни одлуки, особено при решавањето на **неструктуираните проблеми,** низ директна интеракција меѓу корисникот на податоците и моделите.

Современите системи за поддршка на одлучувањето, како резултат на се посложените услови на стопанисување и развојот на компјутерската технологија, се надополнети со низа карактеристики во однос на претходната генерација. Разликите можат да се согледаат од слика 3[11]



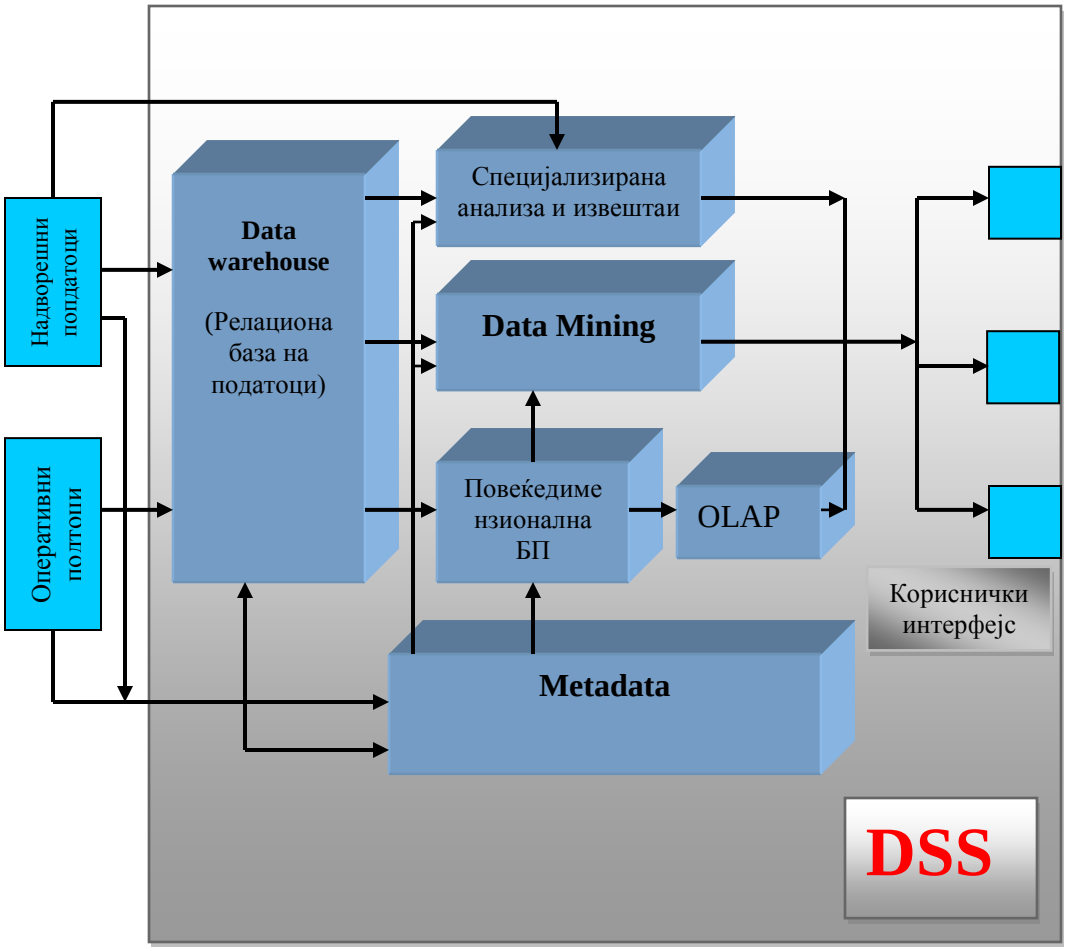
Слика 3. Традиционални и современи DSS

[11] Маргарита Јанеска, Менаџмент Информациски Системи, Универзитет "Св. Климент Охридски", Економски Факултет, Прилеп, 2004, стр 215

Системите за поддршка на одлучувањето можат да бидат **податочно или модел-ориентирани** системи. Податочно ориентираните DSS алатки овозможуваат добро разбирање на податоците, главно, преку **сегментација**, **делење**, **раситнување**, **drilling-down** (пребарување по податоците од горе-надолу), **rolling-up** (пребарување по податоците од долу-нагоре) и други операции. Од друга страна, модел ориентираните DSS алатки го поддржуваат развојот на моделите на одлучување во форма на дрво на одлучување (различен формат од дрвата на одлучување кај data mining), дијаграми на влијание и мултиатрибутни модели.

Алатите на системите за поддршка на одлучувањето се поддржани од димензионалните бази на податоци, алатите за откривање скриени шаблони меѓу податоците и апликациите што овозможуваат примена на нови знаења. Таквата поддршка е прикажана на слика 4.

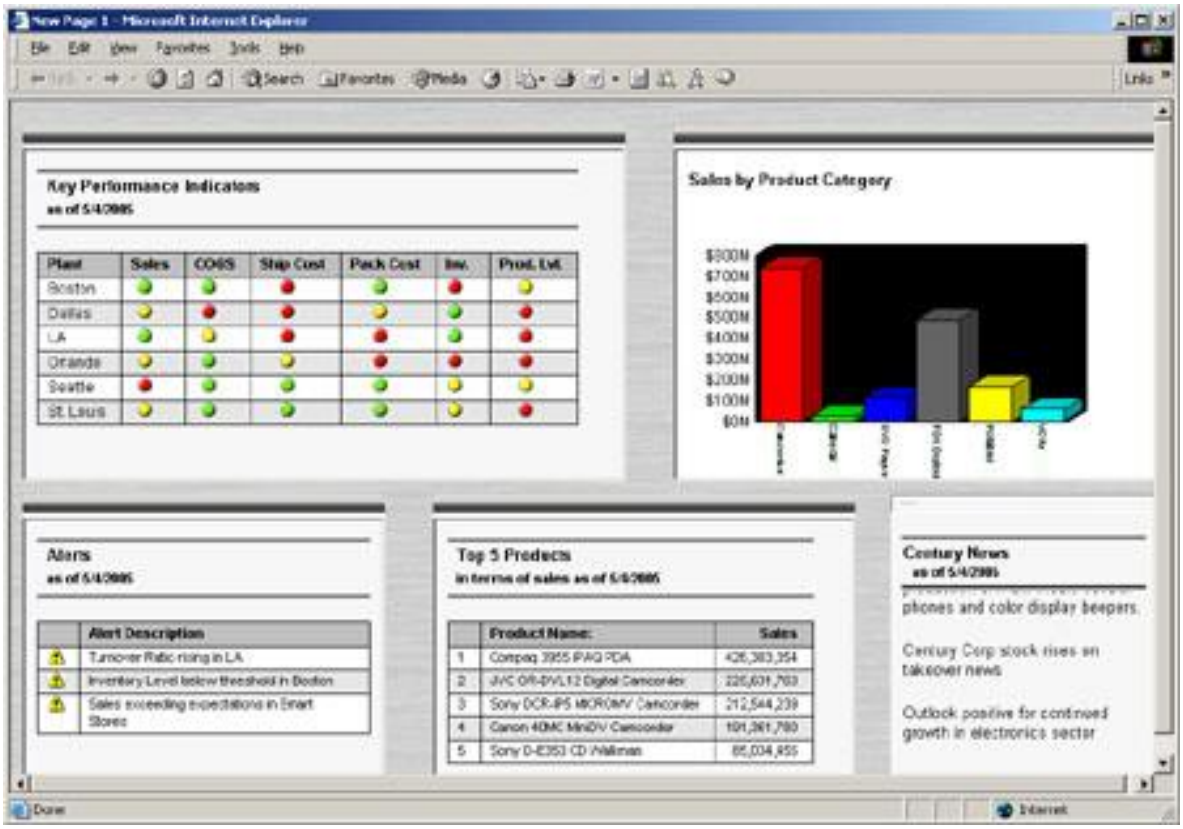
Системите за поддршка на одлучувањето им обезбедуваат на менаџерите кратки и прецизни информации што не е можно да се добијат на друг начин. Овие системи се ефикасни поради можноста за автоматизирано донесување деловни одлуки. Истоа така во рамките на компанијата системите за поддршка на одлучувањето можат да обезбедуваат поддршка за одлучување на поголема група менаџери кои меѓусебно се поврзани (мрежно), клиент – серевер околина или како Desktop DSS за поединечен корисник на компјутерот во канцеларијата на менаџерот.



Слика 4. DSS[12]

Најдобрите системи за поддршка на одлучувањето вклучуваат високо ниво на збирни извештаи или графикони и му овомножуваат на корисникот користење алатки за детализирање на информациите. Таков пример на апликација за поддршка на одлучувањето што дава можност за преглед на информациите, е претставен на слика 5.

[12] E.G.Mallach, Decision Support and Data Warehouse Systems, McGraw-Hill, New York,2004, str 169



Слика 5. Апликација на DSS[13]

Важноста на системите за поддршка на одлучувањето во целокупниот информациски систем на компанијата добива на значење поради потребата за брзо донесување одлуки. Проблемите што се јавуваат постојано се менуваат, било поради влијанието на новите фактори што предизвикуваат развој на еднократни и неповторливи ситуации, било поради промените во самата структура на проблемот што се манифестира како резултат на зголеменото искуство на доносителот на одлука. Системите за поддршка на одлучувањето се **пофлексибилни**[14] од конвенционалните управувачки информациски системи (MIS). Овие системи имаат вградени механизми на флексибилност што овозможуваат наменско прилагодување, пред се, поради промената на содржината на самата проблематика.

[13] www.sas.com
[14] Маргарита Јанеска, Менаџмент Информациски Системи, Универзитет "Св. Климент Охридски", Економски Факултет, Прилеп, 2004, стр 208

Според сите наведени карактеристики на системите за поддршка на одлучувањето би можело да се извлече заклучок дека со нивното користење на менаџерите од тактичко и стратегиско ниво им се овозможува, да донесуваат подобри, поконзистентни и поефективни одлуки. Потребата за нивна примена, исто така, ја налага и самата несигурност во деловното одлучување, како и исклучителната вредност на информацијата во денешниот бизнис амбиент.

7.Ефекти од примената на информацискатa технологија во деловниот свет

Информациските и комуникациските технологии (ICT) имаат долгорочно влијание врз потенцијалот за економски раст и стануваат еден од клучните ресурси за стекнување конкурентска предност и зголемување на приходите. Поради тоа, во март 2000 година во Лисабон, Европската унија поставувајќи си за цел да стане "**најконкурентна и најдинамичка светска економија, заснована на знаење**" во следните 10 години, дефинираше дека ваквата цел може да се оствари само со интезивно и најдобро можно користење на информациските технологии. Како резултат на ваквата Лисабонска стратегија подоцна беше донесен **eEurope 2002 Action Plan,** како водечка основа за модернизирање на европската економија и трансформирање во економија заснована на знаење[15].

Генерално , на макроекономско ниво, интензивираната употреба на информациските технологии ја зголемува продуктивноста на компаниите, а оттука произлегува и конкурентноста во севкупната национална или светска економија. Зголемената продуктивност, предзвикана од развојот на IT секторот, претставува извор за нови работни места во одделни сектори, и тоа за сметка на оние работни места што исчезнуваат. Крајниот резултат од употребата на информациските технологии води до отворање работни места во оние сектори каде што доминантно влијание има токму информацискатa технологија.

[15]www.e-europestandards.org

На микроекономско ниво, пак, информациските технологии претставуваат водечка сила во имплементирањето на промените на организациските пазарни структури. Брзите технолошки движења имаат главно влијание врз структурата и животните циклуси на компаниите. Информацискатa технологија најпрво го редуцира економското влијание на меѓусебната оддалеченост на компаниите и ги намалува трошоците за пристап до информациите, а со тоа се зголемува делокругот на компетитивност внатре во самите пазари. Исто така, информациските технологии овозможуваат намалување на трошоците за основање мали и средни компании со што потенцијално, т.е. дополнително ја зголемуваат конкурентноста во определен сектор. На крајот, ваквите технологии даваат можност за креирање нова корпорациска средина за производство и испорака на услуги онаму каде што може да се стане лидер, преку обезбедување квалитет и преку истовремено намалување на трошоците. Од самото тоа можеби и произлегува една од најважните улоги на информациските технологии, а тоа е развојот на нови производи и видоизменети услуги.

Ефективното користење на информациските технологии денес станува се покомплексно. Брзите промени во стандардите и развојот на новите алатки, како и нивната интероперабилност бараат постојано следење и должно внимание. Во таа смисла, за да се овозможи ефективно користење на предностите што ги нуди информацискатa технологија, компаниите треба да си обезбедат пристап до компетентни, профитабилни и целно ориентирани сервиси што ќе бидат креирани во согласност со нивните потреби. Од оваа причина потребно е земјите да имаат кохерентни и усогласени IT политики и правила.

Користењето на **интернетот** се повеќе навлегува во секојдневното живеење на луѓето и станува интегрален дел од нивниот живот. Денес просечната светска интернет популација ја сочинуваат 16% од вкупниот број население. Притоа, распонот е од околу 2% во Африка до речиси 70% во Северна Америка. Оттука може да се констатира дека интернетот навистина го измени светот и го направи виртуелен. Во табелата 1 е дадено користењето на интернетот во светски рамкиза периодот 2000-2007 год.

КОРИСТЕЊЕ НА ИНТЕРНЕТ ВО СВЕТОТ						
Региони во светот	**Население**		**Корисници на Интернет**		**Учество на корисниците (%)**	**Раст во светот (2000-2007)**
	Проценка (2007)	%	Број	%		
Африка	915.210.928	14	23.649.000	2.6	2.3	424%
Азија	3.667.774.066	56	380.400.713	10.4	36.5	233%
Европа	807.289.020	12	294.101.844	36.4	28.2	180%
Среден исток	190.084.161	3	18.203.500	9.6	1.7	454%
Северна Америка	331.473.276	5	227.470.713	68.6	21.8	110%
Латинска Америка	553.908.632	9	79.962.809	14.7	7.8	351%
Австралија/Океанија	33.956.977	1	17.872.707	52.6	1.7	135%
Вкупно	6.499.667.060	100	1.043.104.886	16	100	189%

Табела 1. Користење на интернет во светот[16]

Од табелата може да се забележи дека Азија со над 380 милиони интернет корисници прераснува во 1/3 од вкупната популација интернет корисници. Покрај Северна Америка, и Европа е едно од високоразвиените подрачја за користење на интернетот. Кога ќе се анализираат поединечните земји може да се забележи дека користењето на интернетот многу варира од земја до земја. Покрај САД особено висок степен на користење интернет имаат и голем број европски земји (над 50%). Пенетрацијата на интернетот во одделни европски земји е прикажана на слика 6.

Набљудувајќи ги овие показатели јасно е дека улогата и влијанието на интернетот во денешни услови речиси е несогледливо. "**Мрежата над мрежите**" е преполна со Web страници што го менуваат начинот на живеење. Web претставува многу силна алатка за секојдневно комуницирање, информирање, деловна соработка и забава. Според тоа, заостанувањето во денешниот електронско-информатички свет може да биде многу опасно, а покрај некои специфични знаења од корисникот се бара уште и постојано учење.

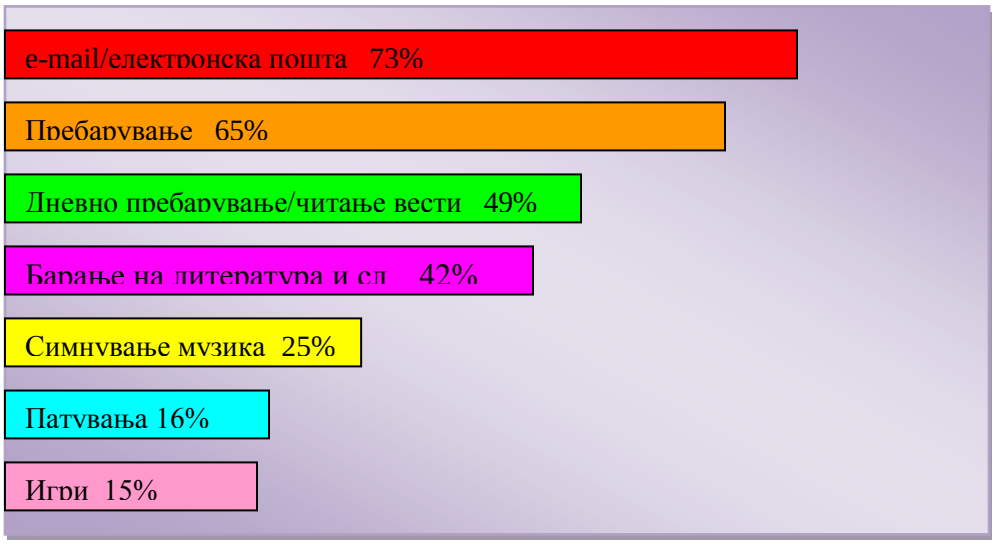
[16] www.worldstats.org

Слика 6. Пенетрација на Интернет во Европа[17]

Од овој преглед за нас особено е важно да се иницира испитувањето улогата и учеството на интернетот во **Република Македонија**. Имено набљудувано во однос на вкупната популација на земјата, користењето на интернетот е застапено кај приближно 392.677 жители или тоа е 19,2 % од вкупниот број население во Р. Македонија. Но, македонската информатичка реалност покажува дека е потребно уште многу работа за да се подобри ситуацијата во оваа област. За успешен развој на **информатичко општество** државата треба уште повеќе да ги поттикнува и поддржува сите напредни електронски услуги. Покрај зголемувањето на нивото на стандардот на граѓаните, исто така, особено важна улога игра и нивото на образование, пред се, општата информатичка писменост, како и познавањето на

[17] www.worldstats.org

странските јазици(посебно англискиот јазик). На слика 7 е дадена целта на пребарувањето преку интернет, односно за што најчесто се користи интернетот.



Слика 7. Најчеста цел на користење на Интернет[18]

[18] www.worldstats.org

II ДЕЛ
DATA MINING И ИЗГРАДБА НА МОДЕЛ ЗА DATA MINING

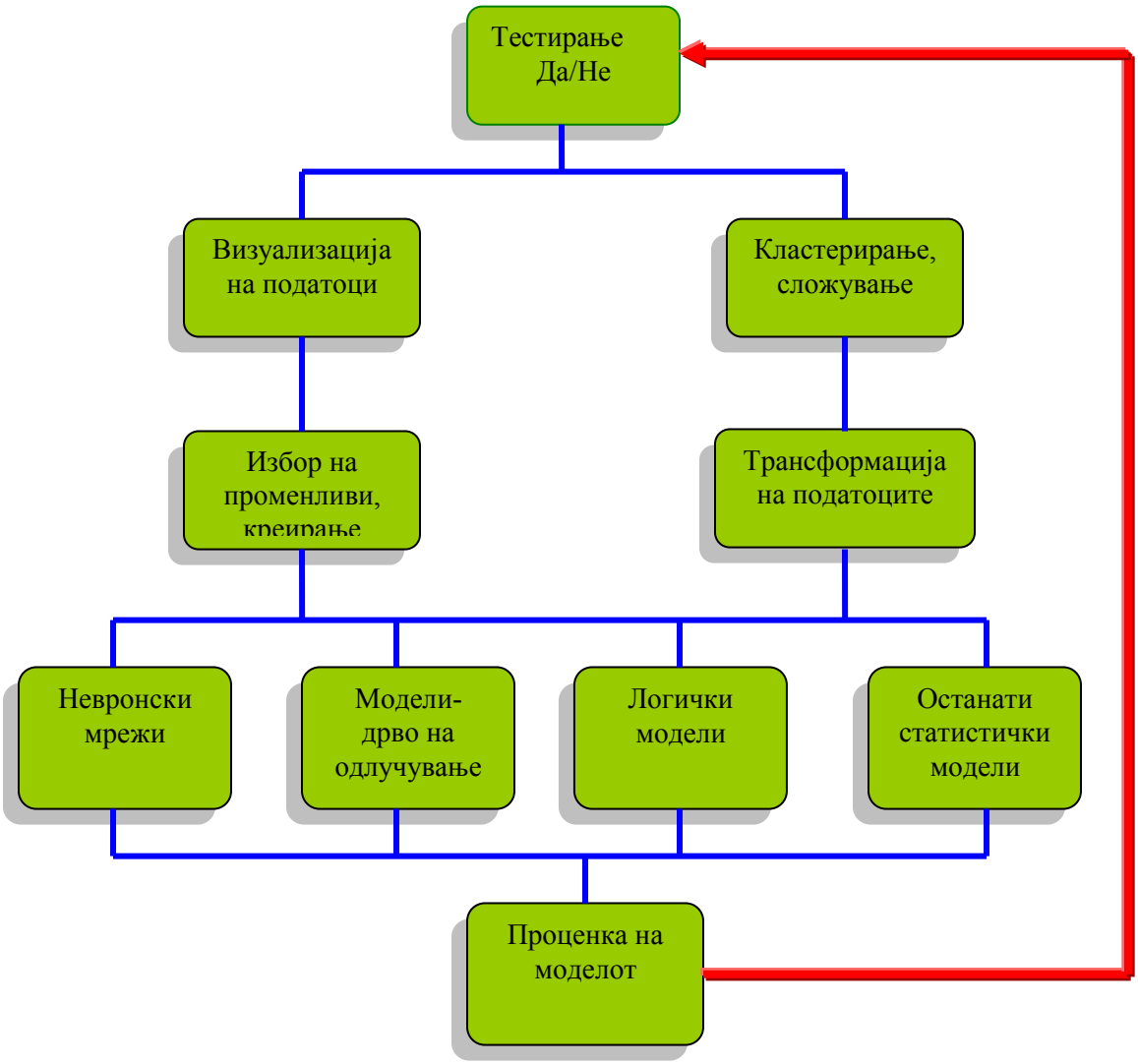
1.Дефиниција за Data Mining

Data Mining е идеја заснована на едноставна аналогија. Зголемувањето на димензиите на базите на податоци (складиштата на податоци) резултираше во креирање огромни количини податоци. Базата на податоци претставува централно место на информацискиот систем и се дефинира како збир на меѓусебно поврзани податоци за на оптимален начин да послужат во различна примена.[19] Големите количини податоци претставуваат вреден ресурс за секоја компанија. Единствено што останува е тие податоци рационално и навреме да се искористат. Оттука, вадењето на скриените предиктивни модели од таквите големи бази на податоци и фокусирањето на компаниите на најважните податоци од нивните огромни бази може да се направи со помош на една нова моќна технологија со голем потенцијал, а тоа е Data Mining.

Всушност, Data Mininig претставува сет на активности што се користи за изнаоѓање нови, скриени и неочекувани шаблони или модели во податоците. Според **SAS Institute, Data Mining е дефиниран како процес на селектирање, истражување и моделирање на големите количини податоци, преку откривање на претходно непознати шаблони од податоците за стекнување знаење, а со тоа и конкурентска предност[20].** Во овој случај процесот на Data Mining вклучува пет чекори: тестирање, истражување, модифицирање, моделирање и проценка (sample, explore, modify, model and assess). Од овие чекори произлегува препознатливиот акроним на SAS Institute преку којшто се објаснува концептот на Data Mining, а тоа е познатата **SEMMA** што е прикажан на слика 8. Оттука не може веднаш да се забележи разликата меѓу **SEMMA** и традиционалната методологија што се користи за статистичка анализа, туку треба да се потенцира користењето на

[19]Zoran Skocir, Ivan Matasic, Boris Vrdoljak, Organizacija obrade podataka, Sveuciliste u Zagrebu, Fakultet elektrotehnike i Racunarstva, 2000, str 24
[20] SAS Institute:www.sas.com/feature/4qData Mining/whatisData Mining.html

напредните Data Mining алатки, за разлика од традиционалната статистичка анализа.



Слика 8. SAS Institut - SEMMA[21]

Дефинициите што се предлагаат за Data Mining тесно корелираат со концептите коишто се нудат за откривање на знаење од аналитичките бази на податоци, иако не може да се направи конкретна дистинкција меѓу овие три пристапа. Сепак, како најточно може да се прифати мислењето според коешто

[21] SAS Institute:www.sas.com/feature/4qData Mining/whatisData Mining.html

поддржувачите на Data Mining веруваат дека откривањето на знаење од аналитичките бази на податоци (Knowledge Discovery in Database-KDD) претставува процес што користи компјутерски заснована анализа на податоците и, како резултат на тој процес, се добива научно засновано знаење. **Според тоа, вистинското откривање знаење од базите на податоци претставува процес на автоматизиран Data Mining, без примена на човечка интервенција или иницијатива.**[22] Поради тоа процесот на Data Mining често се изедначува со процесот на откривање знаење од базите на податоци, иако се смета дека Data Mining е само еден чекор од процесот на откривање знаење. Сепак, и кога Data Mining се зема како дел од KDD, тоа претставува **најзначајниот и најобемниот дел** од откривањето на знаењето од аналитичките бази на податоци. Дури и оние теоретичари[23], кои го застапуваат мислењето дека Data Mining е дел од KDD, се согласуваат дека процесот на откривање знаење од базите на податоци, во суштина, се сведува на откривањето скриени шаблони меѓу податоците. Според нив, Data Mining претставува:

[22] Joseph M. Firestone, Data Mining and Knowledge Discovery in Database, White Paper No. Two, 1999
[23] U. Fayyard, G. Piatetsky, Shapiro, P. Smyth & R. Uthurusamy, Advances in Knowledge Discovery & Data Mining, Cambridge MA (The AAAI Press/The MIT Press) 2000

По откривањето на тие шаблони меѓу податоците многу полесно се извршува интерпретацијата на добиените резултати. Уште повеќе, што во последно време со развојот на нови алатки за Data Mining, самите алатки содржат механизам за интерпретација на резултатите, односно претставување на знаењето во лесно разбирлива форма за корисниците. Исто така, современите алатки, што се нудат на пазарот на Data Mining алатките во определена мера даваат можност за извршување на останатите чекори од процесот на откривање знаење од аналитичките бази на податоци, како што се претпроцесирањето или трансформацијата на податоците. Поради тоа, Data Mining често се користи и како синоним за откривањето знаење од аналитичките бази на податоци. Исто така, за Data Mining се употребуваат и термините екстракција на знаење, податочна анализа, податочно рударење, копање по податоците, интелегентна анализа на податоците, истражувачка анализа на податоците, анализа на податочни шаблони итн. Од самата **SEMMA** може да се забележи дека Data Mining се искористи и да се опишат традиционалните процеси на анализа или, со други зборови кажано, Data Mining претставува традиционално моделирање и анализа поддржано со новите алатки и техники и инкорпорирано во комерцијалната рамка на Data Warehouse.

Користејќи ги информаците што се содржани внатре во аналитичките бази на податоци, Data Mining може да понуди одговори на одделни прашања што се можеби превидени од страна на оние кои донесуваат одлуки во компанијата како, на пример:

- кои производи можат да бидат понудени на определен потрошувач;
- која е веројатноста дека определен потрошувач ќе одговори на планираната промоција;
- колкава е веројатноста за тоа кој потрошувач на каков начин ќе плати;
- колкава е веројатноста некој потрошувач да биде преземен од конкуренцијата;
- кој е најдобар канал на дистрибуција за производот итн.

Овие типови прашања можат да бидат одговорени изненадувачки брзо ако информациите, скриени во огромната количина податоци, бидат соодветно

лоцирани и искористени. Data Mining технологијата може да генерира нови бизнис можности со користењето на следниве способности:

- **автоматско предвидување на трендовите и однесувањето**. Data Mining ги автоматизира процесите за пронаоѓање предиктивни информации од огромните бази на податоци. Прашањата што традиционално да се одговараат бараа многу време и ресурси, но сега можат да се одговорат директно со користење на податоците и тоа многу брзо. Типичен пример за решавање на предиктивен проблем е **целниот маркетинг**. Преку користење на историски податоци, со Data Mining, можат да се идентификуваат целите што ќе овозможат најбрзо враќање на вложените инвестиции;
- **автоматско откривање на претходно непознати модели**. Со помош на Data Mining алатките се откриваат претходно непознатите и скриени шаблони на податоци во базата на податоци, и тоа во еден чекор. Еден ваков пример е анализата на продажбата за да се идентификуваат наизглед наповрзаните производи што заеднички можат да понудат поголема вредност. Друг пример на откривање вакви форми на податоци е идентификувањето на проблемите во врска со трансакциите со кредитните картички и откривањето на анимозитетни податоци што можат да предизвикаат голем број потешкотии при внесувањето на шифрите на кредитните картички.

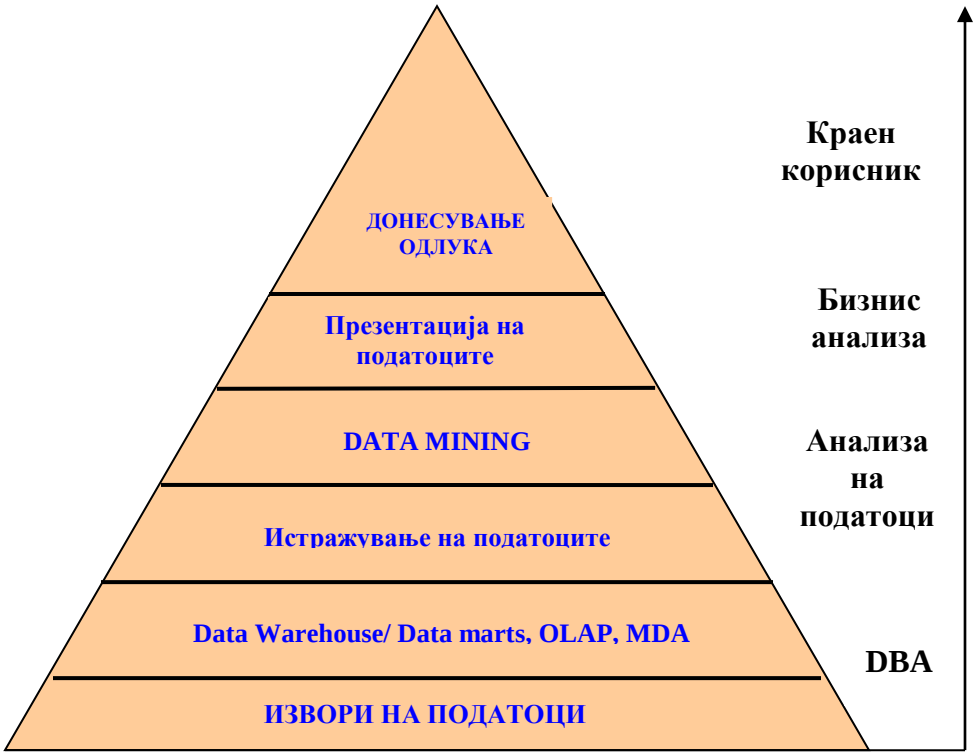
Data Mining може да се набљудува и како **збир на аналитички техники**[24] со коишто се анализираат големите количини податоци што се собрани од различни извори. Притоа, овие техники и алатки не се користат засебно, туку меѓусебе се комбинираат и надополнуваат за остварување што подобар преглед врз збирот на наизглед неповрзаните податоци. Врз основа на тоа се добиваат корисни информации потребни за деловно и стратегиско одлучување од најниското ниво па се до топ менаџментот. Со помош на алатките на Data Mining се предвидуваат

[24] J. Han and M. Kamber, Data Mining:Concepts and Techniqeues, Morgan Kaufmann, San Francisko, 2004, str. 105

идните трендови и однесувања, а тоа ги прави бизнисите проактивни. **Проспективната** анализа, што ја овозможува Data Mining, се движи меѓу минатите случувања, а е поддржана со алатките на системите за поддршка на одлучувањето. Притоа Data Mining има **две главни групи цели**. Првата група цели укажува дека основен таргет на Data Mining е автоматизирањето на процесот за донесување одлуки низ креирање модел за предвидување, сегментирање на набљудуваните објекти во одделни групи и проценка на вредноста на набљудуваните објекти. Втората група цели укажува на тоа дека Data Mining овозможува подобро разбирање што се се случува со набљудуваните податоци и нивната меѓузависност со други елементи.

Голем број компании веќе имаат огромен квантитет на податоци. Од аспект на компанијата, Data Mining може да се вклучи и во дефинирањето на еден релативно нов поим, а тоа е "**деловна интелегенција**"(business intelligence). Деловната интелегенција може да се дефинира како откривање на прикриено деловно знаење од рутински собраните податоци, што се однесуваат на деловните трансакции[25]. Според самата дефиниција, деловната интелегенција ги обединува методологиите, технологиите и платформите за Data Warehouse, OLAP процесирањето на податоци (On-line Analyitical Processing) и Data Mining (во последно време и порталите), што им овозможуваат на компаниите креирање корисни управувачки информации од податоците за работењето, а што се наоѓаат дисперзирани во различните трансакциски системи и доаѓаат од различни интерни и екстерни извори. Тоа е прикажано на слика 9.

[25] Panian Z., Klepac G., Poslovna Intelegencija, Masmedia, Zagreb, 2003



Слика 9. Data Mining и Business intelligence

Business intelligence се користи за интерпретација на историските податоци, потоа за предвидување на идните трендови како и за определување и мерење на показателите од деловното работење. Друга причина за појавата на ваквите business intelligence е инхерентаната особина на трансакциските системи за генерирање извештаи во ограничен број, а извештаите можат да бидат генерирани само од страна на искусни програмери. Исто така, генерираните извештаи имаат големи ограничувања, пред се, од аспект на нивните аналитички можности.

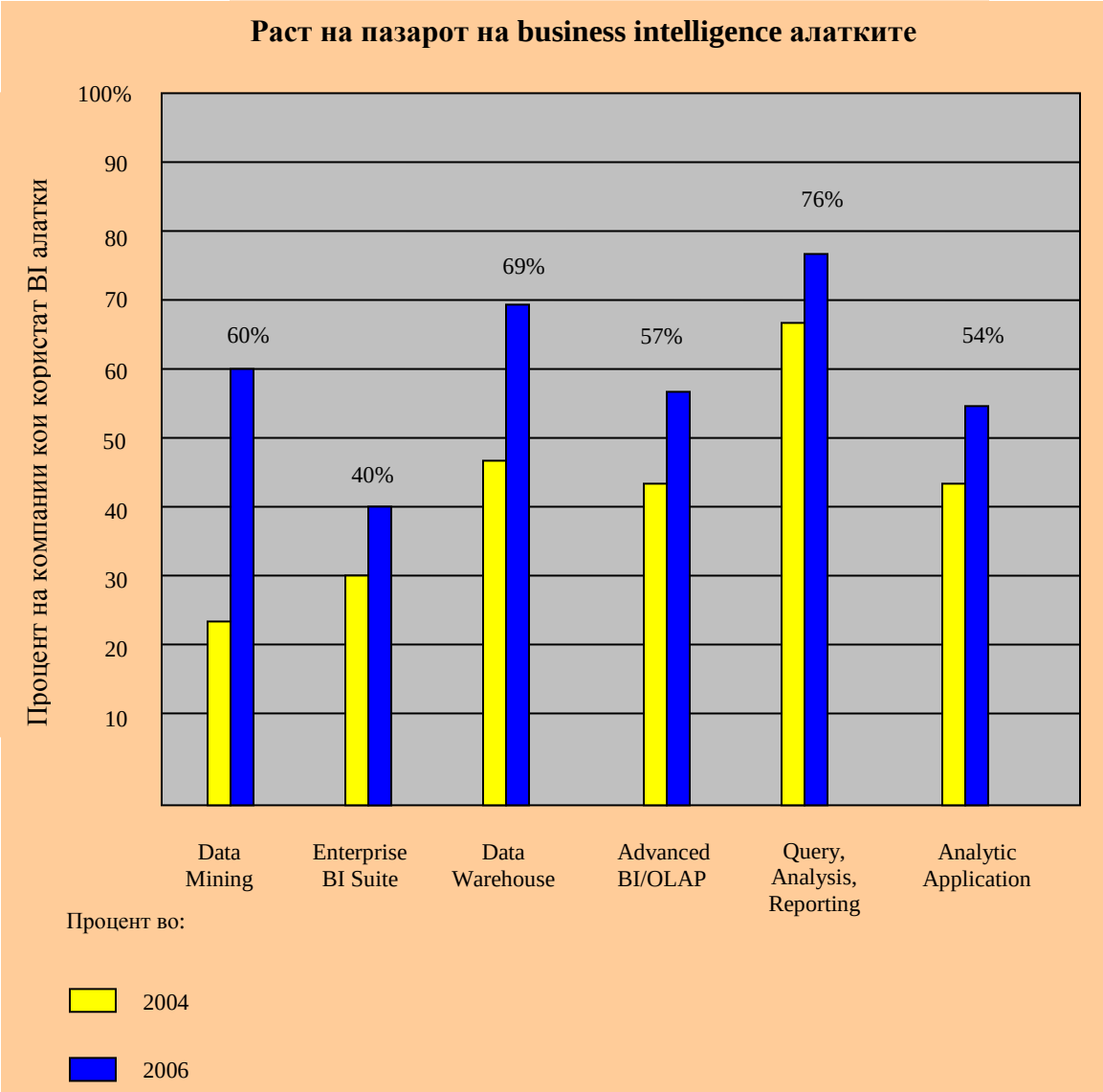
Data Mining техниките можат брзо да бидат имплементирани на постојниот софтвер и хардвер и, на тој начин, да ја зголемат вредноста на информациските ресурси. Кога се имплементирани на клиент сервер информациските системи се со повисоки перформанси Data mining алатките можат да анализираат гломазни бази на податоци и да ги дадат потребните одговори.

Овде може да се каже уште дека Data Mining води потекло, односно се развива врз основа на повеќе дисплини, како што се: статистиката, вештачката интелегенција, машинското учење, системите на бази на податоци итн. Самиот факт дека во Data Mining методите и техниките спаѓаат некои статистички методи, потоа методите на вештачка интелегенција или методите на машинско учење како и одделните нивни комбинации, укажува на потеклото на оваа дисциплина, а тоа поконкретно може да се види и од слика 10.

Слика 9. Потекло на Data Mining

Растечките барања на компаниите за навремени и точни информации ја наметнуваат потребата за се поголемо користење алатки што ќе им овозможуваат собирање, складирање и обработка на информациите. На слика 10 е прикажано зголемувањето на користењето на ваквите алатки на светскиот пазар. Имено 80% од анкетираните 2000 водечки светски компании сметаат дека Data Mining ќе претставува критичен фактор за идниот успех на пазарот. Од самото тоа е условен

и големиот раст на Data Mining софтвер пазарот, што веќе се мери во вредност од околу 780 милиони долари и на тој пазар учествуваат околу 70 компании.[26]

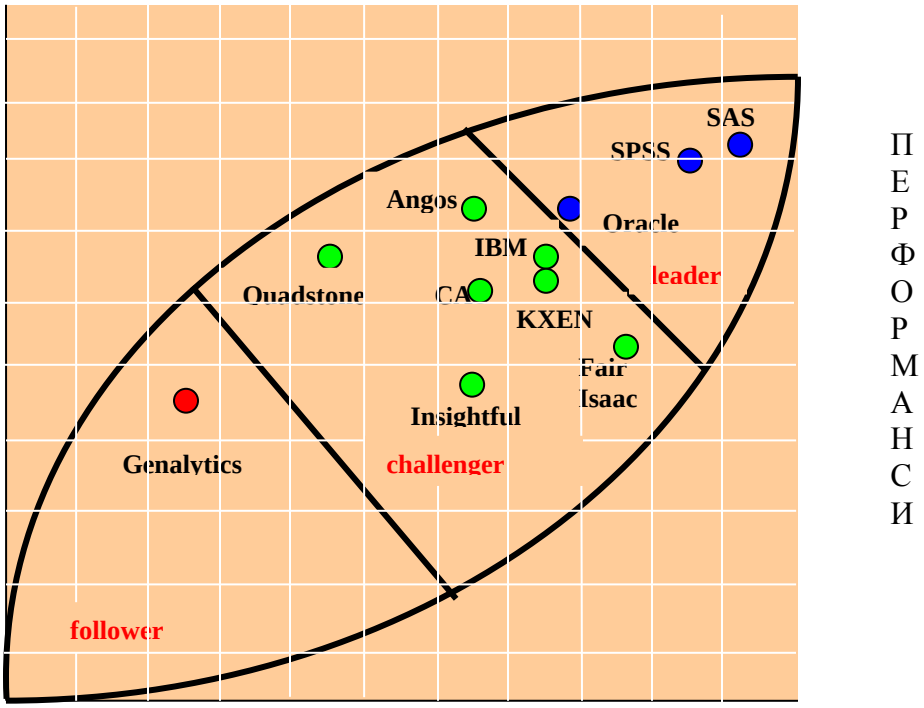


Слика 10. Состојба на пазарот на business intelligence алатките[27]

Од прегледот, што е даден на претходната слика, може да се забележи дека има видливо зголемување на користењето на сите алатки за обезбедување, складирање и обработка на информациите. Она што е впечатливо на овој графикон е зголемувањето на користењето на Data Warehouse, а особено на Data Mining

[26] Meta Group Inc., METAspectrum Market Summary 2007
[27] G. Gardarin, Data PRO users Survey, 2006

алатките. Оттука произлегува заклучокот за сознанието на компаниите дека Data Warehouse и Data Mining претставуваат две **моќни организациски алатки,** што можат многу да придонесат во зголемувањето на конкурентската предност.Тоа за компаниите денес претставува еден од клучните фактори во унапредувањето на бизнис работењето. А, тоа, пак, особено доаѓа до израз доколку постои можност компанијата да ги набави **водечките производи** од оваа технологија. На следниот графикон се прикажани водечките Data Mining алатки што моментално се нудат на пазарот.



Слика 11. Состојба на пазарот на Data Mining алатки[28]

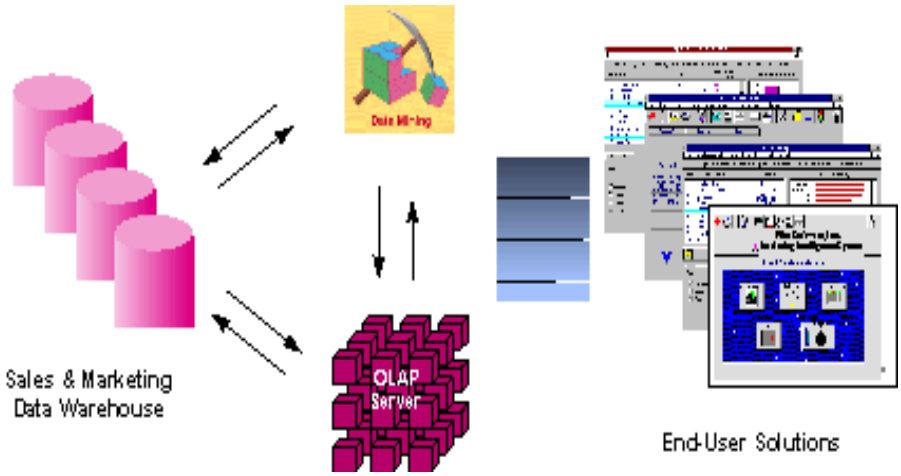
На графиконот е дадена состојбата на Data Mining пазарот каде што се јавуваат определен број на големи компании што се **водечки понудувачи** на пазарот на Data Mining алатки, но постојат и некои други памали продавачи.

[28] Meta Group Inc., METAspectrum Market Summary 2007

Според Meta Group Inc. и ова нивно истражување, се очекува пазарот на Data Mining алатки да расте за 10%-20% годишно. Особено се очекува голем пробив во иднина на оние компании што продаваат **бизнис ориентирани апликации**. Најчесто користени алатки за Data Mining се алатките на SPSS и SAS. Иако SPSS може да најде голема примена и во бизнис окружувањето сепак неговата применае сведена повеќе на академско ниво. За разлика од SPSS, SAS производите се повеќе бизнис ориентирани, благодарение на интеракцијата на стручњаците од компанијата и крајните корисници (најчесто од бизнис окружувањето).

2.Архитектура и функционирање на Data Mining

За да можат да се применат Data Mining техниките е потребно целосно нивно интегрирање со базата на податоци, како интерактивни флексибилни бизнис алатки за анализа. Многу Data Mining алатки оперираат надвор од складиштата на податоци, барајќи дополнителни чекори за екстрактирање, импортирање или анализирање на податоците. Аналитичките бази на податоци можат да се применат за подобрување на бизнис процесите во компанијата, и тоа во областите како што се: маркетинг промоцијата, детекцијата на грешки, промоција(advertising) на нов производ. На слика 12 е илустрирана архитектура за напредна анализа на една огромна база на податоци.



Слика 12. Data Mining архитектура[29]

[29] www.thearling .com

Како што може да се забележи, иделна почетна точка е содржината на базата на податоци, како комбинација од внатрешните податоци во спрега со надворешните пазарни податоци. Повратната информација претставува одлична основа за натамошни предвидувања. Ваквата база на податоци може да биде имплементирана на различни системи на релациони бази на податоци како што се:Sybase, Oracle, Redbrick а може да се оптимизира и за флексибилен, и брз пристап до податоците. OLAP серверот овозможува пософистициран end-user бизнис модел што може да биде применет при управувањето со базата на податоци. **Повеќедимензионалната структура** му дозволува на корисникот да ги анализира податоците онака како што му се потребни за неговата работа – сумирани по производи, региони или според некој друг клучен параметар. Data Mining серверот мора да биде интегриран со базата на податоци и со OLAP серверот. Натаму, главниот шаблон на метадата ги дефинира Data Mining целите за специфични бизнис ситуации како, на пример, менаџмент кампања, предвидување, оптимизација на промоцијата итн. Ваквата интеграција овозможува оперативните одлуки да бидат директно имплементирани. Со растот на базата на податоци со нови одлуки и резултати компанијата може континуирано да ги одделува најдобрите искуства и да ги применува во иднина за донесување на што поквалитетни одлуки. Овој дизајн претставува фундаментално поместување од конвенционалните системи за поддршка на одлучувањето. За разлика од едноставната испорака на податоци до крајните корисници низ прашалници и извештаи, Advanced Analysis Server корисничките бизнис модели[30] ги применува директно на базата на податоци, и како резултат на тоа, се добива проактивна анализа врз основа на релевантните информации. На крајот, известувањето и визуелизацијата, како и другите аналитички алатки, можат да бидат применети за идните планови и реализација на идните активности.

[30] www.thearling .com

Овде се поставува прашањето за **функционирањето на Data Mining**[31] односно како со помош на Data Mining можат да се откријат работи што дотогаш не биле познати или да се разбере што ќе се случува во иднина?

Техниката што се користи во Data Mining за оваа работа се нарекува **моделирање.** Моделирањето, всушност, претставува градење модел за некоја позната ситуација и придружување на моделот кон ситуацијата што не е позната. Денес, со користењето на компјутерите, во компјутерот се внесуваат голем број податоци за различни ситуации за коишто е дефиниран одговорот и потоа Data Mining софтверот треба да ги пребара податоците во базата на податоци, односно да ги одвои карактеристичните податоци што се вклопуваат во дадениот модел. Кога моделот еднаш ќе биде изграден, тој може да се користи за различни ситуации за коишто не е познат одговорот. Но, тука се поставува прашањето: доколку се разбере дека веќе има изградено модел што може да помогне во одделна ситуација, дали тој модел навистина е добар, т.е. дали може да ги задоволи потребите на конкретната ситуација? Прво што мора да се направи е да се тестира моделот на постојната база на податоци за онаа ситуација за којашто веќе се знаат одговорите. Тестирањето може да се изврши со елиминирање или изолирање на некои од податоците во текот на Mining процесот. По комплетирањето на Mining процесот, резултатите што ќе се добијат можат да се споредат со оние податоци што биле изолирани за да се утврди валидноста на моделот. Ако моделот функционира, тогаш таквите опсервации можат да се задржат и за изолираните податоци.

3.Складиште на податоци (Data Warehouse)

3.1 Цели, структура и компоненти на Data Warehouse

Складиште на податоци (Data Warehouse) е место каде што вработените во организацијата континуирано можат да пристапуваат до потребните податоци и да ги користат за процесот на одлучување. Сепак, при проучувањето на која било област е потребно да се даде соодветна дефиниција за појавата што се проучува. Па

[31]George M. Marakas, Modern Data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str 71

проучувањето на Data Warehouse, може да се започне од самата дефиниција што гласи:

Data Warehouse е копија на трансакциски податоци, специфично структуирани за да овозможат пребарување, анализа и известување.[32]

Тргнувајќи од оваа дефиниција, може да се добие претстава што, всушност, претставува Data Warehouse. Најпрвин, складиштето претставува база на податоци што всушност е **копија на податоците за трансакциите**. Ова упатува на фактот дека организацискиот Data Warehouse и бројните трансакциски системи и базите на податоци во организацијата претставуваат два одделни ентитети. Понатаму бидејќи Data Warehouse содржи копија на трансакциони податоци, а не вистинските податоци што се генерирани во оригиналната трансакција, податоците не можат да бидат ажурирани или променети штом еднаш се доделени на Data Warehouse. Податоците во складиштата се статички и еднаш внесени подразбираат непроменливост, без разлика на тоа во која форма или облик се наоѓаат.

Од дефиницијата произлегува дека податоците **специфично се структуирани.** Ова упатува на фактот дека податоците во складиштата што истовремено се резултат на трансакциските податоци, на некој начин, се трансформирани во определена специфична структура. Така, трансакциските податоци од различните извори можат да се наоѓаат во Data Warehouse, користејќи ја оваа специфична структура. Комплексноста на ваквата трансформација може да се претстави кога бројот на трансакциските податоци ќе се спореди со начините на коишто секој податок ќе биде претставен во одделните бази на податоци.

Конечно, дефиницијата укажува уште на целта за којашто е формиран Data Warehouse, а тоа е **пребарувањето, анализата и известувањето**. Data Warehouse претставува нешто како централно складиште на организациските податоци што ќе овозможува истражување на новите релации, трендови и скриени вредности. Исто така, можат да се наведат и некои општи цели што треба да ги задоволи ваквото складиште на податоци а тие се[33]:

[32]George M. Marakas, Modern Data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str 5

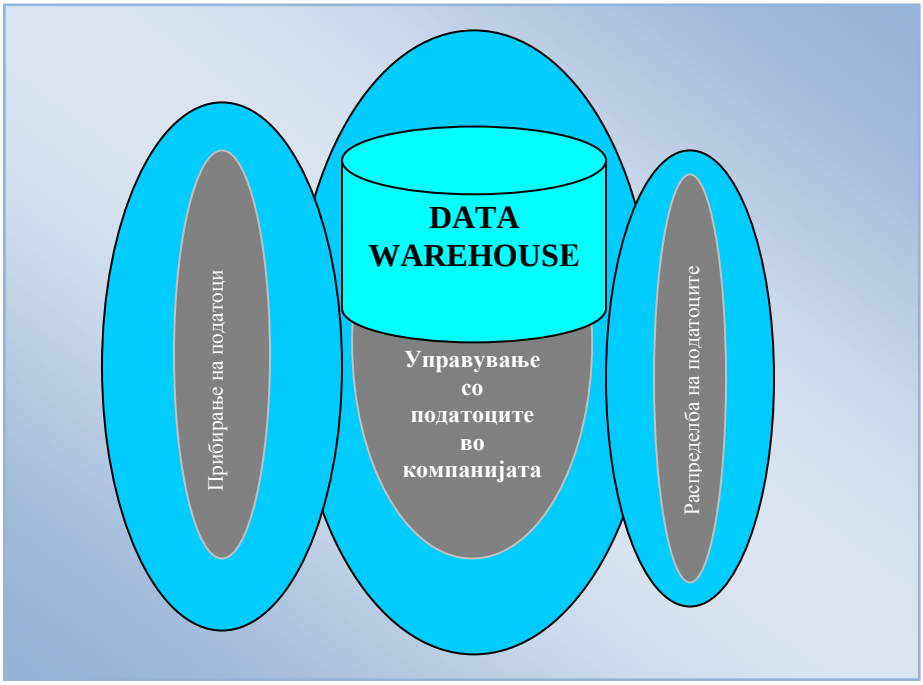
[33] Inmon, W., H., Building the Data warehouse, Second Edition, John Viley & Sons, New York, 1999.

- складиштето на податоци овозможува пристап до корпорациските податоци,
- податоците во складиштето на мораат да бидат конзистентни,
- податоците во складиштето можат да се сепарираат и да се комбинираат преку која било димензија
- складиштето на податоци не ги претставува само податоците, туку тоа е и збир на алатки за пребарување (прашалници), анализа и презентација на податоците,
- податоците од складиштето овозможуваат квалитетен редизајн на деловниот систем

За да може да функционира складиштето на податоци мора да биде во постојан допир со околината, бидејќи информациите што доаѓаат однадвор се важни за пронаоѓање, анализа и вреднување на податоците што се стават во процесот на одлучување. Имено за да има целосен увид во одвивањето на деловниот процес важни се податоци што можат да укажат на промените. Поради тоа, многу е важна временската димензија на складиштето на податоци, главно, заради проценка на идното однесување врз основа на определените анализи (трендови, временски серии и др.). Притоа потребни се т.н. историски податоци што се пренесуваат од трансакциските бази на податоци во складиштето на податоци што пак, претставува повеќедимензионална база на податоци. Покрај ваквите историски податоци од окружувањето доаѓаат и некои надворешни податоци, при што во складиштето податоците се собираат и се агрегираат, а потоа со помош на аналитички и статистички методи се обработуваат и се презентираат во облик на извештаи или се применуваат во автоматизираните системи на одлучување. Податоците, во форма на извештаи, се користат во процесот на одлучување од страна на највисокото раководство. Кога станува збор за складиштето на податоци се мисли на огромната количина податоци. Примерот на Data Warehouse на Wal*Mart може да покаже за што станува збор. Всушност тоа е Data Warehouse што содржи 24 TB податоци (на пример: 1 TB податоци "собира"

доволно текст за околоу два милиони книги), организирани во 5 милијарди записи, од изминатите 5 квартали, што го користатат илјадници корисници и што треба да генерира над 60 000 извештаи неделно[34].

Според тоа, може да се заклучи дека Data Warehouse се наоѓа меѓу двете крајности: прибирање на податоците и распределба на податоците. Тоа е прикажано на слика 13.



Слика 13. Местото на Data Warehouse во организацијата[35]

Ваквата речиси генерализирана перспектива на Data Warehouse дозволува да се категоризира неговата улога во следниве примарни функции:

- Data Warehouse е директна рефлексија на различните бизнис правила во компанијата,

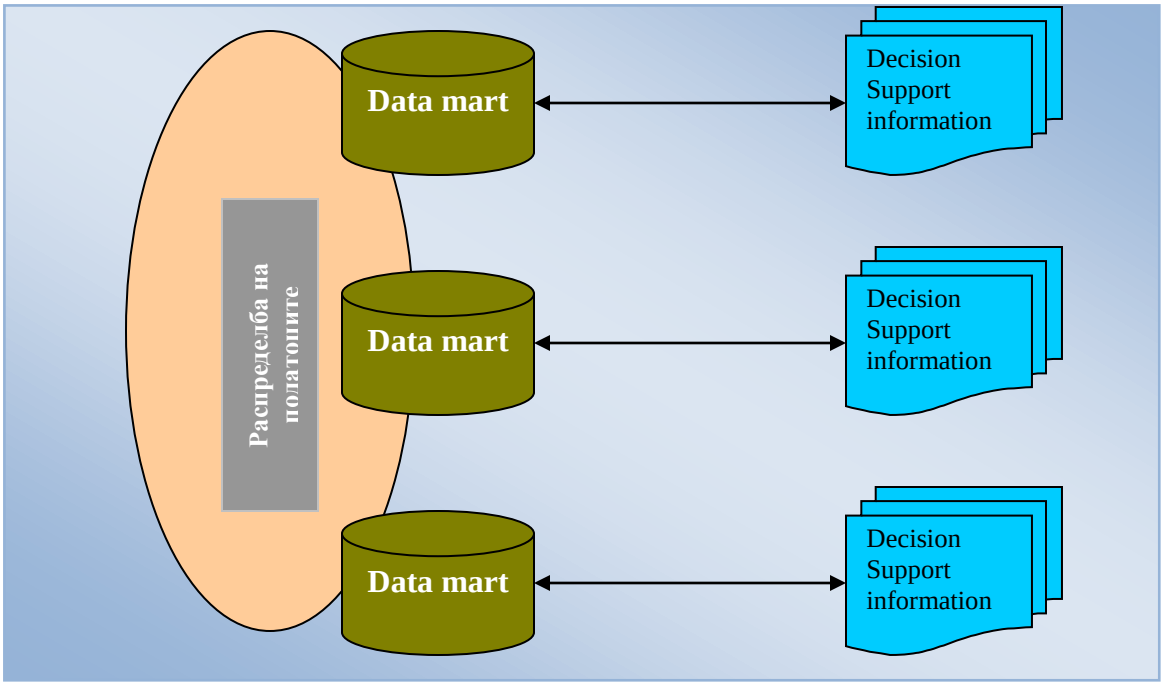
[34] www.teradata.com
[35]George M. Marakas, Modern Data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str 7

- Data Warehouse претставува колекција на интегрирани целно ориентирани стратегиски информации,
- Data Warehouse е историски магацин на стратегиски информации за претходната поврзаност меѓу кои било податоци или нивните врски, или и двете истовремено,
- тој е извор на информации што дополнително се доставуваат до Data Mart (маркетите на податоци)

Притоа Data Mart не е ништо друго туку, **помала, пофокусирана Data Warehouse**. Во многу случаи организациите наоѓаат за корисно да креираат Data Mart за специфични бизнис единици што имаат потреба за анализа на специфични податоци. Иако големото складиште на податоци може да ги задоволи овие потреби, енормната количина податоци во него може да ја редуцира ефикасноста при анализата на определени специфични подрачја. Како што е прикажано на слика 14, Data Mart се наоѓа меѓу распределбата на податоците и директниот извор на податоци, пристапен од крајните корисници, а придружен од актуелниот Data Mart.

Data Mart ги задоволува следниве примарни функции:

- тој е одраз на бизнис правилата за спефичните функции или бизнис единици,
- Data Mart ги добива своите податоци од релативно стабилен, чист и интегриран извор на податоци-организацискиот Data Warehouse,
- Data Mart е сет на табели, дизајнирани за директен пристап на корисниците кои имаат потреба за анализа на податоците, наспроти сетот од предефинирани параметри,
- тоа е сет на табели што е дизајниран за агрегација, и
- конечно, Data Mart не претставува типичен извор на податоци за традиционална статистичка анализа или Data Mining.



Слика 14. Позицијата на Data Mart во организацијата

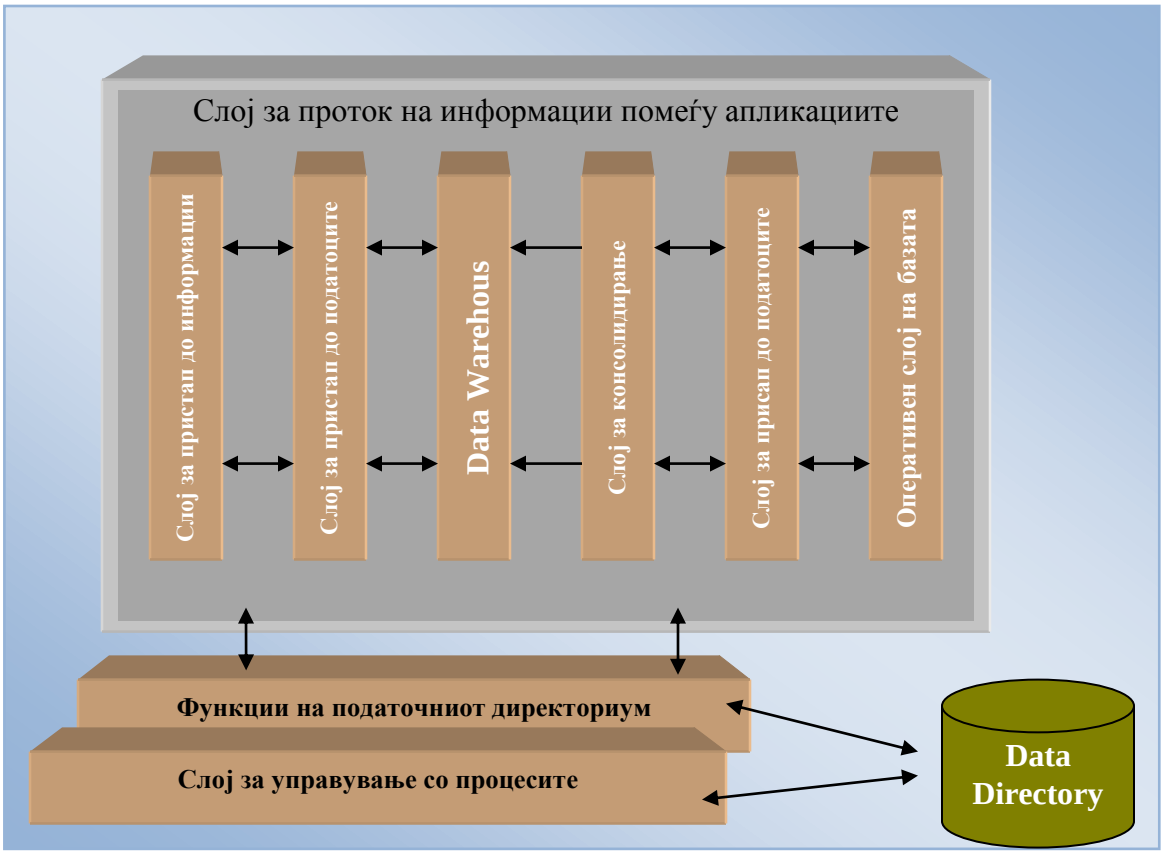
Податочниот дел од складиштето на податоци, како компоненти, ги содржи **основните податоци, агрегираните податоци и мета податоците[36].** Основните податоци доаѓаат од трансакциските бази на податоци и тие се историски, сурови, необработени податоци. Агрегираните податоци се на едно повисоко ниво и, како што самото име кажува, тие се добиваат со агрегирање на основните податоци врз основа на методите за пронајдување податоци со коишто се изнајдуваат различни типови податоци. Мета податоците, пак, се податоци што укажуваат на другите податоци. Без мета податоците не би било можно корисникот да комуницира со податоците од складиштето, бидејќи нема да знае колку тие се стари, како табелите се структуирани, која е точната дефиниција на определен податок или каде податоците се јавиле.

3.2 Архитектура на Data Warehouse

Data Warehouse архитектурата е метод преку којшто можат да бидат презентирани целокупната **структура податоци, комуникацијата,**

[36] Singh, S., Interactive Data Warehousing, Prentice Hall, 2000

процесирањета и презентацијата на крајните корисници внатре во компанијата[37]. На слика 15 се претставени различните меѓусебно поврзани елементи што ја сочинуваат структурата на Data Warehouse.



Слика 15. Структура на Data Warehouse

Оперативен и надворешен слој на базата на податоци

Овој слој го претставува изворот на податоци за Data Warehouse. Поточно овој слој ги опфаќа најнапред системите за процесирање на оперативните трансакции и надворешните секундарни бази на податоци. Целта на Data Warehouse е да ги ослободи информациите заклучени во оперативните бази на податоци и да ги искомбинира со информациите од другите надворешни извори. Како дополнителна цел на Data Warehouse е да има минимално влијание врз перформансите и операциите на системите што се наоѓаат во овој слој. Големите

[37]George M. Marakas, Modern Data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str. 51

организации дополнително добиваат информации од надворешните бази на податоци. Лесната пристапност до web страниците и интернетот го прават лесно и економично инкорпорирањето на ваквите податоци во Data Warehous на компанијата.

Слој за пристап до информациите

Крајните корисници имаат контакт директно со слојот за пристап до информациите. Вообичаено, овој слој е претставен со алатките што крајните корисници секојдневно ги користат за издвојување и анализа на податоците коишто се наоѓаат внатре во Data Warehouse. Овој слој се заснова на хардверот и софтверот што е вклучен за печатење на табелите, извештаите и графиконите за анализа и презентација. На овој слој DSS ги користат податоците од Data Warehouse за поддршка на различните типови и категории при донесувањето на одлуките внатре во организацијата.

Слој за пристап до податоците

Како што може да се забележи од сликата, слојот за пристап до податоците служи како некој вид интерфејс или интермедијатор помеѓу оперативниот и информацискиот слој и самиот Data Warehouse. Овој слој се протега над сите различни бази на податоци што се содржани внатре во Data Warehouse и го олеснува заедничкиот пристап од страна на крајните корисници. Така, овој слој не ги раздвојува само различните бази на податоци од системските фајлови, туку ги одделува и мноштвото работни од мрежните протоколи. Успешно дизајнираниот Data Warehouse им обезбедува на крајните корисници универзален пристап до податоците , па теоретски, крајните корисници треба да имаат можност да пристапат до некои или до сите податоци на компанијата што им се потребни за извршување на нивната работа, без оглед на локацијата или алатките за пристап до бараните податоците. Улогата на овој слој е да го овозможи таквиот пристап.

Metadata слој

За да се обезбеди универзален пристап до податоците апсолутно е потребно да се формира некој вид податочен директориум или складиште на Metadata информации. Metadata, всушност, претставуваат податоци за тоа како и каде се сместени податоците внатре во Data Warehouse. Податоците во складиштето можат да бидат различно организирани, во зависност од методологијата што се користи, но во секој случај секоја табела со податоци во определен временски период од одделна локација се вчитува, доживува некаква трансформација, се комбинира со некои други табели за, на крајот, да се добие оптимизирана табела за изработка на мултидимензионалната коцка и за добивање извештаи. Сите податоци за тој процес, како и трансформациите, односно податоците што се резултат на тој процес се нарекуваат метаподатоци и ја дефинираат структурата и содржината на Data Warehouse. Метаподатоците ги сочинуваат документирани дефиниции за изворите на податоци, релациите меѓу OLAP податоците, OLAP структурите (како што се: димензии, хиерархии, методи за калкулирање со податоците) предефинираните OLAP апликациски шаблони. Исто така, во метаподатоци спаѓаат и дефинициите за влезните табели од различните трансакциски системи. Таквите дефиниции честопати можат да бидат многу корисни, бидејќи овозможуваат лесно снаоѓање во големата количина табели, и на тој начин, добивање квалитетни резултати. Примерот за Metadata вклучува директориум за тоа каде податоците се складирани, правилата за сумирање и пречистување на податоците или можните записи за оперативните извори на податоци.

Слој за управување со процесите

Слојот за управување со процесите е фокусиран на распоредување на различните задачи што мораат да бидат извршени за да се изгради и управува со Data Warehouse. Овој слој, исто така, може да служи како контрола на работите од поголема важност што мора да бидат извршени за да се обезбеди ажурноста на Data Warehouse. Работите, како на пример, периодично download-ирање(симнување) од дефинираните оперативни извори на податоци, распоредувањето на сумираните и оперативни податоци, пристапот и download- ирањето од надворешните извори на

податоци, односно ажурирањето на Metadata се типични за овој слој на Data Warehouse.

Слој за протек на информации помеѓу апликациите

Овој слој, пак, има задача да ги трансферира информациите низ компјутерската мрежа. Имено овој слој уште е познат како midleware. Ваквиот слој, на пример, може да биде користен за да се одделат апликациите (операциски или информациски) од точниот формат на податоците од двата краеви. Овој слој, исто така, може да биде користен за прибирање на трансакциите или пораките и нивна испорака до определена локација во определено време. Во таа смисла, слојот за протек на информациите меѓу апликациите може да биде разгледуван како транспортен систем што го поддржува Data Warehouse.

Физички слој на Data Warehouse

Физички слој е оној слој каде што актуелните податоци се користат за поддршка во одлучувањето во целата организација. Во некои случаи се мисли дека Data Warehouse е логички и виртуелен поглед на податоците бидејќи понекогаш Data Warehouse всушност, и не ги складира податоците до коишто се пристапува преку него.

Слој за консолидирање на податоците

Овој слој претставува крајна компонента на Data Warehouse. Впрочем овој слој ги вклучува сите процеси што се потребни за селектирање, едитирање, сумирање, комбинирање и вчитување на Data Warehouse и податоците за пристап на информациите од операциските и надворешните бази на податоци. Ваквото консолодирање вклучува комплексно програмирање, но се почесто на пазарот се нудат алатки што имаат за цел да ја редуцираат сложеноста што ги придружува овие процеси.

Како основни архитектури на Data Warehouse можат да се набројат следниве[38]:

- **виртуелна Data Warehouse-**архитектура што им дозволува на крајните корисници пристап до оперативните складишта на податоци, користејќи алатки коишто го овозможуваат нивото на пристап до податоците. Овој пристап треба да понуди идеална комбинација со голема флексибилност и минимална редундантност на податоците. Виртуелниот Data Warehouse често нуди можност за ниска почетна цена за организациите коишто сакаат да проценат кој тип податоци им е потребен на крајните корисници;
- **централизиран Data Warehouse**- претставува единечна физичка база на податоци што ги содржи сите податоци за специфичните функционални подрачја, оддели, дивизиони или компанија. Овој пристап, најчесто, се применува во ситации кога корисниците имаат потреба од вообичаени податоци и најголем број од крајните корисници веќе се приклучени на главниот компјутер или мрежата на компанијата. Ваквата архитектура, најчесто, е користена и претставува еден вид стандард при имплементирањето на Data Warehouse;
- **дистрибуираниот Data Warehouse**-укажува дека компонентите на Data Warehouse се дистрибуирани на бројни различни физички бази на податоци. Колку што донесувањето на одлуките во рамките на организацијата е на пониско ниво, толку и податоците потребни за донесување на таквите одлуки се спуштаат на ниво на локалните клиентски станици. Постарите Data Warehouse, најчесто, го користат овој пристап, бидејќи е полесно да се креираат неколку помали бази на податоци, одколку една голема сложена база на податоци.

[38]R. Kimball, The Data Warehouse Toolkit:Practical Techniques for Building dimansional Data Warehoses, John Viley & Sons, New York, 1999

3.3 Логички дизајн на Data Warehouse

Логичкиот дизајн повеќе е концуптален и апстрактен, отколку што е физичкиот дизајн. Во логичкиот дизајн се бараат **логичките врски и односи** меѓу објектите. Во физичкиот дизајн се бара најдобриот пат за ажурирање и пребарувањето низ податоците. Оттука потребно е ориентираноста да биде повеќе кон оние корисници кои бараат агрегирани податоци, наместо кон оние кои бараат индивидуални податоци. Понекогаш крајните корисници не знаат да дефинираат што им е потребно, туку тоа го сознаваат дури откако ќе го видат решението. Логичкиот дизајн се фокусира на информациските потреби, без вложување некои посебни напори околу непотребните детали.

Процесот на логички дизајн вклучува распоредување на податоците во серија на логички врски, наречени **ентитети и атрибути**[39]. Во димензионалното моделирање, паралелно со барањето ентитети и врски меѓу нив, се настојува да се идентификува која информација припаѓа на **вредносната табела,** а која информација припаѓа на придружените **димензионални табели.** Притоа се дефинираат работните целини или полињата на податоци, т.е. се дефинираат врските меѓу нив и имињата на атрибутите за секоја работна целина.

Елементите што помагаат да се определи шемата на складиштето на податоци се модел за изворот на податоци и потребите на крајните корисници. Освен тоа, изворниот модел може да се добие од постојните податочни модели и реинженерингот на деловните процеси за на крајот да се добие логичкиот податочен модел на складиштето на податоци.

Во процесот на дизајнирање на Data Warehouse најважен е разговорот со крајните корисници. Всушност, постојат две причини за тој разговор: првата е да се добие претстава за потребите и очекувањата на корисниците, а интервјуто е важен дел за проектниот тим што ќе го дизајнира и развива Data Warehouse; втората е кај корисниците да се подигне нивото на сознание за областа за којашто се развива таквиот Data Warehouse и евентуално да се корегираат нивните очекувања. По разговорот со крајните корисници е потребно проектантите на Data Warehouse да

[39] Zoran Skocir, Ivan Matasic, Boris Vrdoljak, Organizacija obrade podataka, Sveuciliste u Zagrebu, Fakultet elektrotehnike I Racunarstva, Zagreb, 2000, str. 55

стартуваат да го развиваат логичкиот дизајн. Меѓотоа, понекогаш се потребни и дополнителни разговори со корисниците. Имено потребно е да се согледаат сите значајни факти од аспект на корисникот и да се идентификуваат процесите што би се одвивале во текот на работата.

3.4 Primer za izgradba na Data Warehouse vo opredelen deloven sistem

При изградба на моделот за Data Mining е потребно де се најде доверлив и ажурен извор на податоци. Голем дел од успешната изградба на моделот зависи токму од тоа. Оттука, во овој дел од трудот ќе биде прикажан еден поедноставен модел на Data Warehouse за потребите на определен деловен систем. Овој модел ќе претставува само основа за да се добие подобар увид во оваа проблематика.

Трансакцискиот систем мора да биде развиен така што ќе може да се користи во повеќекорисничка околина којашто поставува високи барања во однос на пропустливоста на системот[40]. Тој мора да биде сигурен и стабилен при изведувањето на апликациите што опслужуваат голем број корисници кои, истовремено, користат определени потсистеми и бази на податоци. Корисникот пристапува на апликацијата преку клиентски терминали на серверот на којшто се наоѓа апликацијата и базата на податоци. Поточно, корисникот пристапува преку своето корисничко име и лозинка со што пристапува до определен сервер и може да ги користи неговите ресурси во согласност со правата што се определени од соодветниот администратор. Локалната клиентска станица не треба да има јаки хардверски перформанси, додека кај компјутер-серверот неговите технички карактеристики можат да бидат решавачки во поглед на брзината и пропустливоста, па и стабилноста на целокупниот систем.

За користење определени апликации е потребна дополнителна **авторизација** што се определува посебно за секоја група апликации, а внатре во нив авторизација и за одделните делови на апликацијата, т.е. дозвола за покренување на определени процедури.

[40] Zoran Skocir, Ivan Matasic, Boris Vrdoljak, Organizacija obrade podataka, Sveuciliste u Zagrebu, Fakultet elektrotehnike i Racunarstva, 2000

Складиштето на податоци може да се развива во различни насоки, во зависност од барањата на корисникот за аналитички извештаи. За таа цел базата на податоци може да ги содржи сите потребни елементи од релациски табели, па се поединечни референтни атрибут-табели. Проектирањето и изградбата на ваквото складиште на податоци може да се врши според определена методологија. Една од методологиите за изградба на Data Warehouse во еден деловен систем предлага процес што може да се одвива во неколку чекори и тоа[41]:

- **проектирање на општа структура на системот.**Сите проектанти и крајни корисници се соочуваат со исти проблеми. Првата група проблеми е утврдување што треба да се направи за да се задоволат барањата на крајните корисници. Постапката на проектирање е од итеративна, природа што значи кога ќе се заврши еден од чекорите, потребно е да се разгледаат сите претходни за да се провери дали се променило нешто во основниот концепт;
- **проектирање извештаи.** Постапката на проектирање мора да се расчлени до најситни детали што дотогаш се познати. Така секој нов чекор се започнуваме со повторно согледување на основните цели на проектот. Притоа, се поставува прашањето: дали најпрвин да се направи скица на извештајот или да се дефинира кои податоци или текст да се појават на него? Всушност, треба да се прават двете работи истовремено;
- **проектирање на структурата на податоци.** Откако ќе се дефинира кои се излезните податоци, во зависност од нив се организираат податоците внатре во системот. Значи, треба да се направи список на полињата со податоци што ќе бидат потребни за да се обезбедат бараните излезни податоци. Еден од најдобрите начини за тоа е да се направи список на сите полиња што се јавуваааат во поединечните извештаи. Притоа треба да се внимава на оние полиња што се повторуваат во двата извештаи, односно што се однесуваат на ист податок. Таквите полиња не треба да

[41] Cary N. Prague and Michael R. Irvin, Access 2002 Bible, Hungry Minds, 2002

се повторуваат на списокот што ќе се однесува на сите полиња. Вториот чекор во оваа постапка би бил утврдување некаква логичка поврзаност меѓу полињата, што подоцна ќе се групираат во логичката структура на табелите. Потешкиот дел, при проектирањето на структурата на податоци, е комбинирањето на податоците, зашто е потребно да се утврдат полињата во табелите врз основа на коишто ќе се направат извештаите. Бидејќи од извештаите се исцрпени сите податоци, потребно е тие да се разделат според функциите.Меѓусебното поврзување на податоците што се раздвоени според функциите ќе се врши со помош на полињата што содржат единствени вредности. Со додавање на таквите полиња се овозможува на секоја група податоци меѓусебно да се поврзат. Тоа се полињата на примарен клуч;

- **проектирање на врски.** Податоците меѓусебно се поврзувааат за оние од една табела да бидат достапни и во друга табела. Тоа значи дака не мора истите податоци да се чуваат на две различни места. Врските меѓу табелите се воспоставуваат така што повеќе табели делат иста вредност. Притоа, не е неопходно полињата во тие табели да бидат еднакви, туку е потребно вредностите во полињата да се поклопуваат. Вредноста на полето во една табела овозможува со нејзина помош да се пребаруваат податоци во друга табела. Битен е начинот на којшто ќе се земат предвид податоците што ќе се јават повеќепати.

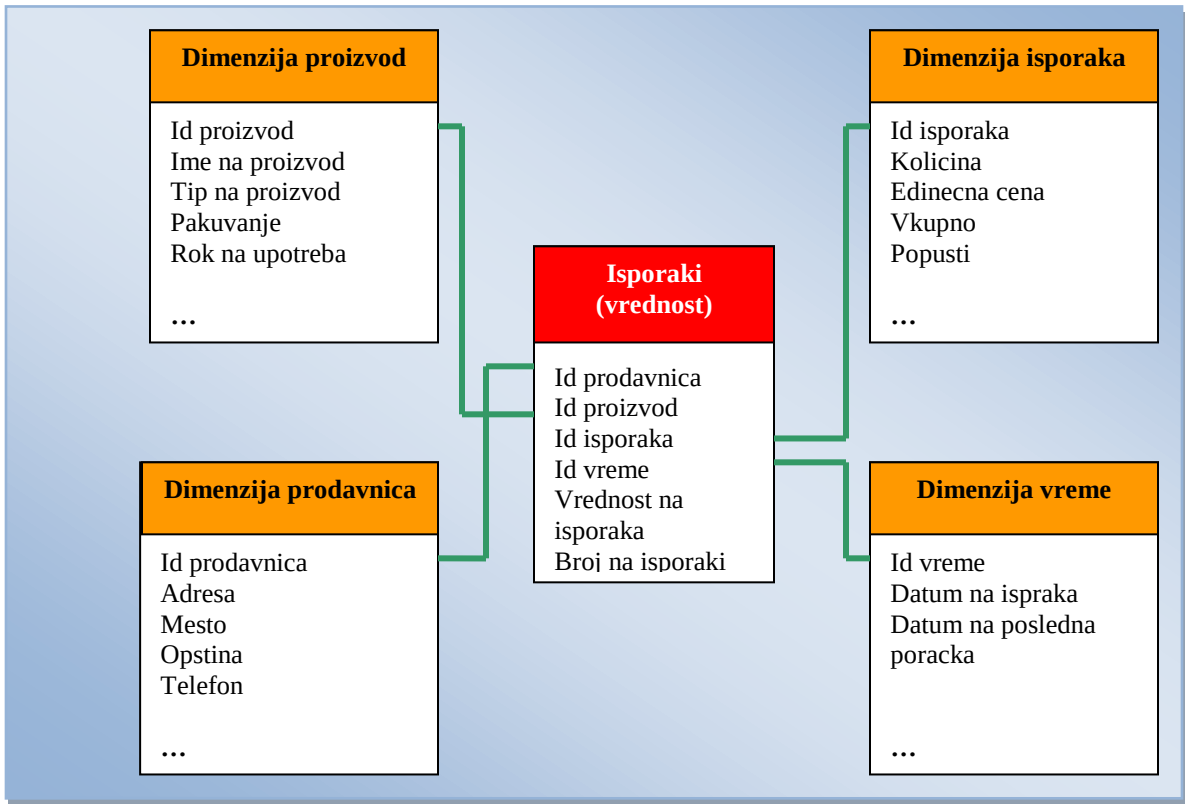
Ако има потреба, во оваа фаза, се проектираат раферентни табели што дополнително ја олеснуваат работата на базата на податоци. Во овој случај, пак, се проектираат референтните табели за да се олесни пристапот до податоците. Исто така, за полесен и побрз пристап до податоците може да се дозволи извесен степен на редундантност, што најчесто не се дозволува при проектирањето на релационите бази на податоци.;

- **Проектирање на полиња.** За оваа цел се врши:

- задавање име, тип, и големина на полето

- дефинирање на правила за внесување на податоци.
- проектирање на референтни табели
- правење на пробни податоци;
- **проектирање на менија**. Менијата се клучна компонента на добро осмислен систем. На корисникот треба да му се обезбеди некој систем на движење низ апликацијата. Важно е при проектирањето на менијата да се размисли за групирањето на командите.

Ваквата методологија за проектирање и изградба на складиште на податоци не значи дека тоа е **универзален пристап** при дизајнирањето на Data Warehouse, туку целта е да се покаже само еден од начините за негово проектирање на. Врз основа на претходно дефинираната методологија може да се дизајнира и еден поедноставен Data Warehouse, како што е прикажан на слика 16.



Слика 16. Пример за изградба на Data Warehouse

На сликата е прикажан е еден едноставен димензионален модел на складиште на податоци. Во централниот дел се наоѓа **вредносна табела** со нумерички вредности (вредност и број на испораки) и **клучеви** за врска со околните димензионални табели. **Димензијата производ** ги содржи сите информации за карактеристиките на производите. **Димензијата продавница** ги содржи информациите за продавницата до којашто се врши испораката. **Димензијата испорака** ги содржи информациите за производите што се испорачани, и тоа според која цена и кои збирни полиња за вкупната цена на испорачани производи, намалувања итн. Конечно, **Димензијата време** ги содржи информациите за точниот датум на испорака.

Ваквото складиште на податоци може да се развива во различни насоки, во зависност од барањата на корисниците за аналитички извештаи. Врз основа на гореприкажаниот димензионален модел може да се развие солиден Data Warehouse на кој можат да се продуцираат различни извештаи. Притоа, можат да се добијат информации за бројот и вредноста на испораките во определен временски период. Исто така, можат да се добијат информации за количините од поединечни продавници што се испорачани во определен временски период. Сето тоа може да се направи за поединечен производ или за вкупната количина производи. Оттука може да се заклучи дека ваквото потенцијално складиште на податоци може да биде од голема помош за оние кои донесуваат одлуки.

Најважните функции на Data Warehouse мораат да обезбедуваат што подобра поддршка за корисниците во текот на работата, како и да овозможуваат обединување на информациите потребни за успешно водење на деловната политика на компанијата. Информациите мораат да бидат лесно и брзо достапни за корисникот, а самата обработка на информациите да биде стабилна и временски усогласена со останатите потсистеми и апликации. Поставените барања за формирање специфични аналитички извештаи тесно се поврзани со нивото на коешто се внесуваат и обработуваат податоците. На база на основните информации може да се формираат аналитички извештаи што ќе го унапредат процесот на донесување одлуки. Поради тоа, развојот на добар Data Warehouse мора

континуирано да се движи напред, во согласност со идните потреби, што можат да се предвидат според реакциите на корисниците и деловните партнери.

4.Изградба на модел за Data Mining

При процесот на изградба на модел најважно е да се изнајде оптималната рамнотежа меѓу временската димензија на моделот и резултатите остварени со веќе изградениот модел. Притоа треба да се совладаат сите техники и принципи на изградба и да се земе предвид способноста за експериментирање со податоците, потребни за изградба на овој модел. При изградбата на моделот за Data Mining може да се тргне од одделни чекори што претставуваат најчест пристап при формирањето на моделот.

4.1 Дефинирање на целите

Во овој чекор, при изградбата на моделот за Data Mining, е потребно јасно да се постават **целите** и да се развијат процеси преку коишто ќе се постигнат проектираните цели. При тоа можат да се развијат два вида модели. **модели за предвидување,** со кои се пресметуваат целните вредности што претставуваат некоја идна активност; и **описни модели** преку коишто се создаваат правила, што се користат за нивно групирање во описни категории или сегменти.

Дефинирањето на деловниот проблем опфаќа[42]:

- идентификување на проблемот,
- мерење на потенцијалните трошоци и корисности,
- планирање на податоците.

При дефинирањето на проблемот е потребно да се определат условите од коишто зависи изнаоѓањето решение за проблемот. Тоа би значело да се определат правилните насоки при одбирањето и подготовката на податоците за дизајн и анализа како и вреднувањето и користењето на резултатите. Дефинирањето на деловниот проблем се спроведува преку разговор со експертите кои, со своето

[42] David Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining, MIT Press, Cambridge, MA 2001

знаење, го покриваат набљудуваниот делокруг на работа. Експертите треба да одговорат на одделни прашањата преку коишто ќе треба да се дознае повеќе за најважните подрачја од нивната работа, потоа некои информации потребни за работата и сл., односно да се дознае се за деловната околина и процесите од каде што тие доаѓаат

Потенцијалните трошоци и корисности се многу важни за деловното одлучување. Главни прашања, што треба да се одговорат при одлучувањето дали да се направи модел за Data Mining се следниве:

- колкаво ќе биде враќањето на инвестицијата за минималното ниво на точност што може да се смета за прифатливо;
- како враќањето на инвестицијата ќе се менува со зголемувањето на нивото на точност;
- колкава ќе биде корисноста од точните позитивни резултати;
- колкава ќе биде загубата од неточните позитивни резултати;
- колкава ќе биде корисноста од неисправните негативни резултати; и
- колкава ќе биде загубата од неисправните негативни резултати.

Data Mining е **итеративен процес**. Моделот постојано се вреднува и поправа се додека не се дојде до определеното ниво на точност за предвидување. Поради тоа се можни и корекции на претходно дефинираните цели, но сепак тоа треба да биде исклучок во случај ако нешто биде изоставено во текот на планирањето.

4.2 Пронајдување и избор на податоци

Добро избраните и подготвени податоци се клучен услов за изработка на ефикасен Data Mining модел. Ако во текот на изградбата на моделот се заклучи дека податоците не се соодветни, тогаш се настојува да се изнајдат нови податоци или да се модифицираат постојните. Во зависност од видот на моделот или во зависност од методот за Data Mining, определен вид податоци повеќе или помалку се применливи.

Податоците што се потребни за изградаба на Data Mining моделот можат да се поделат во две групи, односно:интерни и екстерни податоци.

Интерни податоци се оние податоци што се собрани од околината на некој деловен субјект. Тоа можат да бидат најразлични видови податоци за клиентите, маркетиншките активности, потоа податоци од трансакциските бази итн.

Екстерните податоци обично претставуваат листа на податоци што се собрани од специјализирани институции коишто се занимаваат со таа работа. Многу е важно да се напомене дека податоците кои се собрани од екстерни извори е најпрвин треба да се пречистат и дополнително да се подготват за да бидат соодветни за користење.

4.3 Подготовка на податоците

За реализирање на оваа активност е потребно добро да се познава подрачјето на деловниот проблем и неговите цели. Во секој случај, најдобар извор на информации за податоците и деловните процеси претставуваат експертите. Тие со своите искуства и знаења можат да укажат на најпотребните податоци коишто со својот квалитет ги задоволуваат потребите за изградба на модел за Data Mining. За најдобри резултати, при развојот на моделот, потребно е да се има **репрезентативна популација** на податоци[43]. Многу проектанати во почетокот настојуваат да се изврши анализа на податоците за да се види кои атрибути се добри за предвидување. Резултат од таквата анализа е **ранг листа** од најважните атрибути или променливи што се од особена важност за изградба на моделот. За изградба на моделот, исто така, многу е важно да се обезбедат валидни и по можност најнови податоци.

Определени методи за Data Mining имаат ограничувања во однос на големината на групата податоци или во однос на прецизноста на моделот што се сака да се изгради. Бројот на потребните податоци може да зависи од различни фактори. Така, на пример, ако се употребуваат повеќе променливи во моделот, тогаш е потребно да се обезбедат поголем број податоци. Колку поголем е бројот на податоците за определена променлива, толку поточно ќе биде предвидувањето за таа променлива. Во одделни случаи, при конструирање на групата податоци што

[43] Dorian Pyle, Data Preparation for Data Mining, Morgan Kaufmann, San Francisko, 2000

е потребна за да се изгради моделот, случајниот избор на податоците би бил сосема доволен. Друг начин е, при креирањето на таквиот збир податоци за моделирање, да се земе репрезентативен примерок според застапеноста на поединечните променливи кај целокупната популација. Трет начин на земање примерок од податоците е да се избере поголем број јавувања на поединечната променлива внатре во целокупната популација. Ова се користи кога нема доволно голема популација на податоци, па се настојува да се избере онаа појава за која моделот треба да предвидува со поголема точност.

Собраните податоци неопходно е да се **прочистат од грешки** и да се **надополнат со вредностите што недостасуваат**. Количината податоци зависи од потребите за Data Mining , но исто така, и од достапноста до податоците. Обично, количината достапни податоци е многу поголема од онаа што е потребна, бидејќи пред употребата тие податоци се агрегираат и сумираат во фазата на подготовка на податоците, во зависност од широчината на проблемот што се решава.

На пример, невронските мрежи имаат свој теоретски минимум на податоци што е потребен за успешно тренирање на мрежата. Тоа, најнапред, зависи од структурата на невронската мрежа т.е. од бројот на влезни и скриени единици во слоевите на невронската мрежа. Врз основа на податоците за бројот на неврони и слоеви на мрежата, може да се пресмета бројот на врските меѓу определени единици, и тоа со помош на формулата:

$$s*(v+2)+1$$

каде со s се означува бројот на скриени единици, а v го претставува бројот на влезни единици. Ако бројот на врски е поголем од бројот на податоци за тренирање на невронската мрежа, тогаш може да се случи т.н. overfitting, т.е. појава мрежата да запамети податоци за тренирање, што не е добро за крајниот резултат. За да се избегне тоа бројот на податоци за тренирање на мрежата мора да биде 10 пати поголем од бројот на врските. На пример, за 8 влезови и 2 скриени единици се добива 21 врска. Ако тоа се помножи со 10, се добиваат 210 податоци.

Кога станува збор за подготовката на податоците треба да се напомене дека самиот процес се одвива преку:

- **класификација**. Податоците можат да се поделат во две главни групи, и тоа квалитативни и квантитативни податоци. **Квалитативните** податоци употребуваат описни термини за различни вредности, а **квантитативните** податоци се опишани најчесто со нумерички вредности и обично се користат при развојот на моделот за предвидување. Освен ваквата класификација, постои и класификација на податоците на номинални, ординални, интервални и континуирани податоци. **Номиналните** се нумерички податоци што претставуваат категории и атрибути. **Ординалните** или рангирачки податоци се оние податоци што претставуваат категории што имаат релативна важност. **Интервалните** податоци се нумерички податоци кои имаат релативна важност и немаат т.н. нулта точка и, на крајот **континуирани** се оние податоци над коишто можат да се вршат основни аритметички операции;
- **земања примерок од податоците**. Земањето примерок од податоците (sampling) е креирање примерок на податоци што ќе има помалку податоци, отколку оригиналниот збир на податоците. Така креираниот примерок на податоци претставува т.н. **репрезентативен примерок,** конструиран со случаен (random) избор од целокупната популација податоци. Проверката дали помалиот примерок е доволно репрезентативен мора да покаже всушност дали вредностите се исти како и во оригиналниот збир на податоци односно дали тие имаат иста фреквенција на јавување;
- **чистење на податоците**. Чистењето на податоците ги опфаќа неисправните вредности на податоците и оние вредности што недостасуваат. **Неисправните вредности** на податоците се препознаваат ако излегуваат од предвидените рамки (кај нумеричките податоци) или ако изгледаат бесмислено (кај категоријалните податоци). Целта на **замена на вредностите што недостасуваат** е двојна: да се пополни празниот простор со најслични вредности и да се поправи

дистрибуцијата внатре во некоја променлива. Вредностите коишто недостасуваат се решаваат на неколку начини, односно[44]:

- ако поголем број записи во определено поле немаат вредност, тогаш тоа поле се исфрла од употреба,
- ако само за неколку записи нема вредност, тогаш тие записи можат да се отстранат,
- замена на податоците што недостасуваат со просечни вредности,
- замена на податоците што недостасуваат со определени константни вредности.

Податоците мораат да бидат достапни и чисти така да без некои поголеми промени можеме да ги користиме при моделирањето.

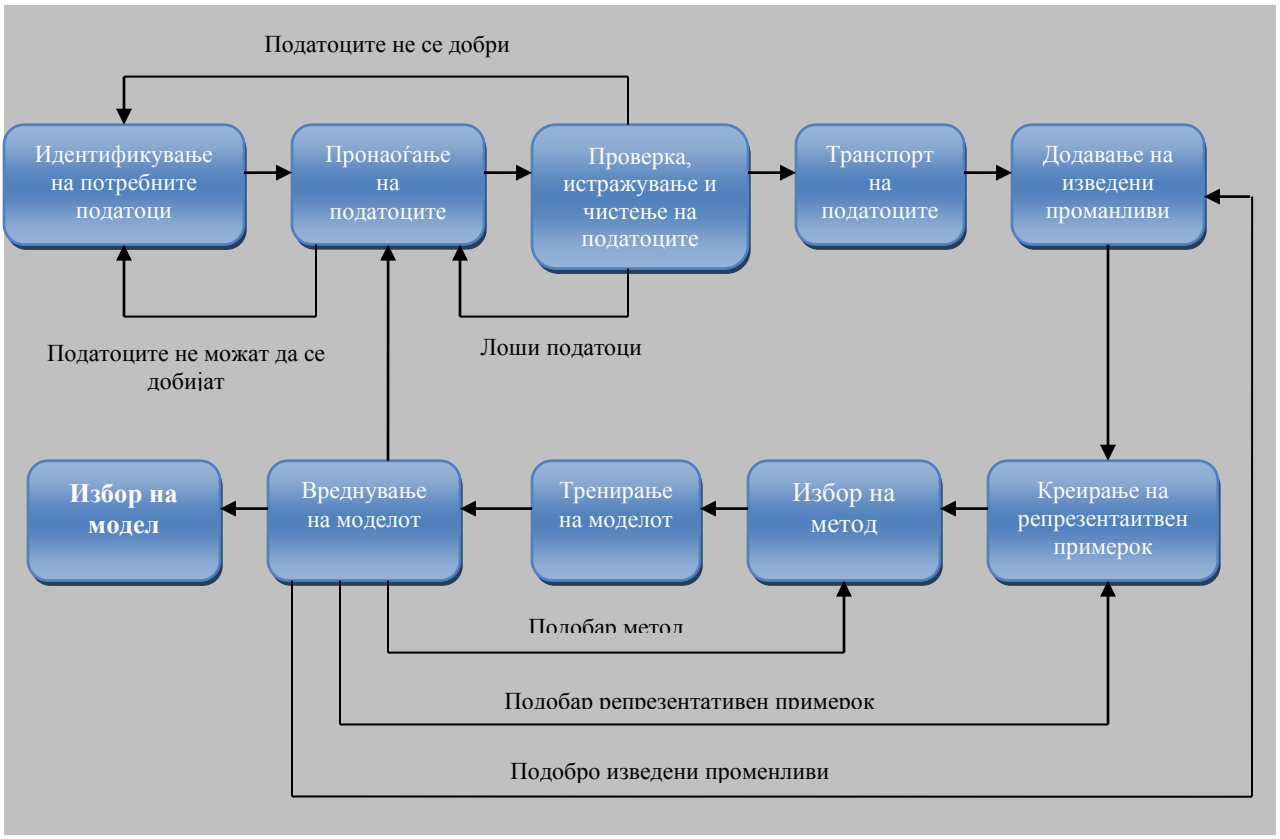
- **трансформација на податоците**. За потребите на Data Mining податоците мораат да се трансформираат во некој соодветен формат. Таквата трансформација може да започне уште во самото складиште на податоци. Data Warehouse е база на податоци каде што покрај информациите собрани интерно за време на деловниот процес, се собираат и екстерни информации од надворешните извори. Складиштето на податоци има ограничувања при користењето на новите извори на податоци, бидејќи тоа, по правило, се гради врз основа на постојните извори, така што со појавата на секој нов извор на податоци, складиштето мора значително да се менува. Податоците во Data Warehouse се импортираат по нивното чистење од какви било недостатоци или таквите податоци воопшто не се земаат предвид. Податоци што се особено детализирани од трансакциските информациски системи не се земаат предвид поради големата количина записи. Поради тоа преку агрегирање податоците се сведуваат на определено ниво што овозможува остварување на целите на Data Mining. Најчести случаи кога е потребно да се трансформираат податоците се сигурноста или заштитата на приватноста на податоците

[44] Dorian Pyle, Data Preparation for Data Mining, Morgan Kaufmann, San Francisko 2000

што се решава на тој начин што некои важни вредности на податоците се заменуваат со кодирани вредности.

4.4 Изградба на модел за Data Mining

Кога се располага со потребните податоци може да се донесе одлука кој модел за Data Mining може да се употреби: регресиона анализа, анализа на варијанса, стебло на одлучување, невронски мрежи и др. Бидејќи е можна примена на Data Mining за различни работи, некои методи или модели за Data Mining помалку или повеќе се применливи. Често е потребно да се изгради **повеќе од еден модел[45],** а подоцна да се определи кој модел е најдобар за решавање на одделен проблем. На слика 17 е прикажана изградбата на модел за Data Mining и положбата на податоците во однос на останатите процеси при изградбата на овој модел.



Слика 17. Изградба на модел за Data Mining

[45] The Technology Review Ten, MIT Technology Review, January/February 2001

Во процесот на избор и подготовка на податоците понекогаш е потребно да се проверат претходните процеси за да се види дали можат да се добијат некои квалитетни податоци. При вреднувањето на моделот може да се заклучи дека е потребно да се обезбедат подобри податоци, да се изнајдат дополнителни изведени податоци или да се формира примерок што порепрезентативно ќе ја претставува популацијата. Притоа, ваквото повторно враќање на претходните процеси има за цел подобрување на вредноста на моделот за Data Mining.

4.5 Вреднување на моделот

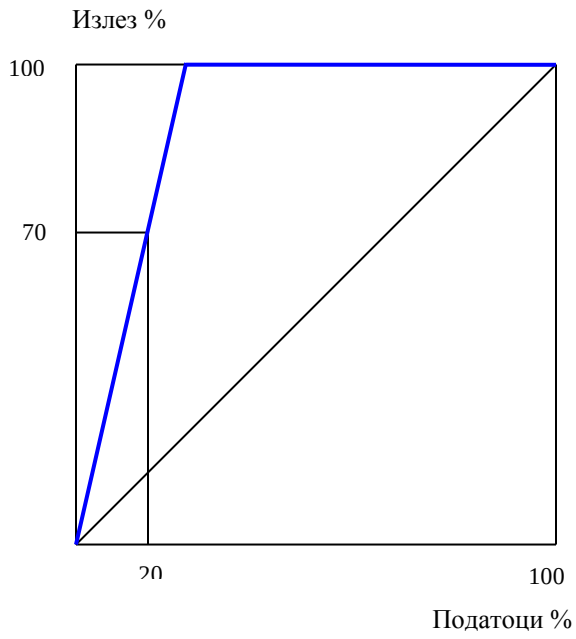
При процесот на Data Mining често се врши **експериментирање** со различни видови модели и алгоритми за Data Mining, преку менување на параметрите. Мерењето на перформансите и споредбата на повеќе модели може да се изврши со помош на **графиконот за кумулативна добивка**[46] или зголемување на корисноста, кога би се употребил моделот за Data Mining. Ако кривата на кумулативниот графикон е преблиску до основната линија, тогаш веројатно се работи за слаб модел. Во тој случај е потребно во моделот да се стават нови променливи, да се обезбедат нови податоци или да се избере друга техника за моделирање. Ако кривата е уште поблиску до основната линија, тогаш станува збор за **претренирање** на моделот (overfitting), односно тоа е појава кога моделот запаметува податоци за тренирање. Податоците што се употребени при изградбата толку се разликуваат од оние кои се употребуваат за предвидување, така што не можат истите да се применат. Во тој случај станува, пак, се покажува дека моделот не е способен доволно да генерализира и да биде применет во секоја ситуација. Таквата ситуација во секој случај треба да се избегнува. Кај невронските мрежи моделот може да се измени со намалување на бројот на јазли во скриениот слој, а кај стеблата на одлучување може да се намали бројот на гранките. Ако во графиконот се јави **рамна линија,** тоа значи дека постои силна корелација меѓу влезот во моделот и предвидениот излез. Тогаш, пак, станува збор за многу добар модел, што може да даде одлични резултати. Стабилен модел е оној што се

[46] www.spss.com

однесува подеднакво при примената на различни примероци на податоци, па врз основа на тоа постои доверба дека моделот ќе работи добро на кој било примерок на податоци.

Моделот секогаш треба да се вреднува со податоците што не биле користени при неговиот развој. Поради тоа, податоците можат да се поделат на податоци за моделирање и податоци за тестирање.

Точноста на моделот[47] се мери врз основа на предвидените излезни резултати и вистинските вредности. Ако се забележи отстапување во предвидувањето, тогаш треба да се изврши корекција на моделот. Со помош на кумулативниот графикон може да се прикаже рангираноста на моделот ако на истиот графикон се стават резултатите на секој поединечен модел.



Слика 18. Кумулативен графикон на корисноста од моделот

На хоризонталната оска се наоѓа процентот на користени податоци, а на вертикалната оска процентот на точни излезни резултати. Ако, на пример, за 20%

[47] www.goranclepac.com

податоци се добијат 70% сакани излези, тогаш може да се тврди дека моделот е добар. Притоа, оваа линија е далеку од основната дијагонална линија што означува предвидување без модел, па може да се каже дака моделот работи 3,5 пати подобро, отколку што се очекува. Колку површината помеѓу линијата на моделот и дијагоналната линија која означува предвидување без модел е поголема, толку моделот е подобар.

4.6 Имплементација на моделот за Data Mining

По верификувањето на моделот, следи негова примена за да се подобри деловното одлучување. Избраниот модел се имплементира преку апликации што поддржуваат динамичка измена на моделот. Влијанието врз основа на информациите може да се одвива на неколку начини[48]:

откривање на знаење,

преземање на одделни меркетиншки активности,

публикување вредни информации,

периодично предвидување,

исправка на податоците за подобрување на перформансите на моделот,

предвидување во реално време (употреба преку друга апликација).

По откривањето на некои врски, форми или модели меѓу податоците постојат неколку можност за нивна примена, односно:

- **презентација.** Шаблоните откриени со помош на Data Mining можат да се презентираат во документите во текстуален или графички облик. За разлика од текстуалните, за графичките прикази е потребно алатките за Data Mining да генерираат излез во соодветна форма. Така, презентираните информации претставуваат солидна основа за одлучување.
- **деловна интелегенција**. Шаблоните можат да се користат за прашалници над базата на податоци. За оваа цел алатките за Data Mining

[48] Peter Cabena, Pablo Hadjinian, Rolf stadler and Alessandro Zanasi, Discovering Data Mining:From Concept to Implementation, Prentice Hall, NY 1999

мораат да генерираат соодветен излез, на пример, дрва на одлучување во SQL облик, за да можат да се употребат при изработката на извештаите.

- **вреднување на податоците**. Откриените шаблони меѓу податоците имаат за цел нивно оценување и означување од базата за да се дефинира натамошното однесување. За остварување на оваа задача, исто така, е потребно да се има SQL.
- **системи за поддршка на одлучувањето**. Пронајдените шаблони се вклучуваат во системите за поддршка на одлучувањето за да станат нивни компоненти. Ова може да се реализира ако алатките за Data Mining се вклучат како компоненти на системите за поддршка на одлучувањето или како функција, односно потпрограма изработена во некој програмски јазик, на пример C# или Java.
- **мониторинг на деловниот процес**. Најголема корисност од пронајдените шаблони би била во мерењето или нормирањето на деловниот процес за да се откријат неправилностите во работењето и со тоа да се алармира системот, односно да се покренат активности за корекција. Тоа би значело откривање неправилности во што пократок временски интервал, а ова може да се реализира доколку алтките за Data Mining се вклучат како компоненти за мониторинг или од алатките за Data Mining се генерира излез во облик SQL, C# или Java модул.

III ДЕЛ
ОСНОВНИ МЕТОДИ И АЛГОРИТМИ ЗА DATA MINING

1.Методи и алгоритми што се користат за Data Mining

Развојот на информациските технологии е поттикнат од генерирањето не се поголемата количина податоци. Денес поголемите количини податоци се сместуваат во огромните бази на податоци, а некои од тие бази на податоци се достапни преку интернет сервисите. Кога станува збор за Data Mining може да се каже уште дека самите почетоци на Data Mining беа директно поврзани со некои статистички методи, па и Data Mining се нарекуваше статистичка анализа. Најпознатите производи кои се користат за таа цел произлегоа од најпознатите светски компании како што се **SAS** и **SPSS.** Иако и натаму, најчесто, се користат статистички методи сепак овие компании направија квалитативен чекор напред кон развивање методи за предвидување.

Со примената на секој метод можат да се изградат модели што ќе имаат свои добри и лоши карактеристики. Со комбинација на различните модели можат да се добијат подобри перформанси, отколку кога секој метод би се користел поединечно. Овде треба да се укаже и на позитивните ефекти што се добиваат од комбинацијата на Data Mining и знаењето на експертите. Таквиот процес се нарекува **Expert Mining**[49] или управување со знаење со помош на Data Mining. Трендовите покажуваат дека крајните корисници сакаат да користат модел преку термините што го дефинираат нивниот делокруг на работа. Позитивните карактеристики што треба да ги има секој метод се точноста, способноста за објаснување, брзината на одговор, преносливоста, компактноста, флексибилноста, можноста за комбинирање со други методи, толеранцијата во однос на комплексноста, независноста (од експерти), брзиот развој, лесното користење, едноставното користење на логичките релации како, и лесното користење на нумеричките податоци.

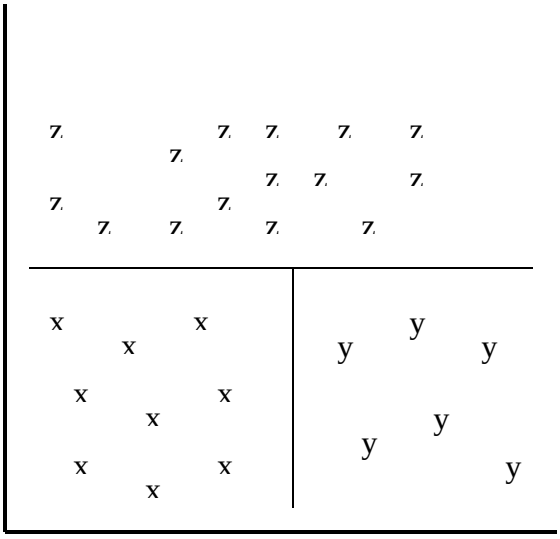
[49] www.data-miners.com

1.1 Кластерирање

Кластерирањето најдобро може да се објасни доколку се претпостави голема количина на податоци, а меѓу нив податоци што со своите особености ќе можат да се **групираат во различни групи**, секоја со свои специфични карактеристики. За квалитетно кластерирање мора да се располага со доволна количина податоци за некој објект. Објектот треба да ги има сите клучни описни елементи што можат еднозначно да го вбројат во определена група, односно **кластер**[50]. Во одделни случаи ако станува збор за клиент, тогаш еднозначните елементи би се нашле во збирот на неговите лични и демографски податоци и податоците за неговото однесување. Сето тоа мора да биде содржано во форма на запис во базата на податоци. Притоа, податоците се собираат од различни извори, најнапред од трансакциските бази на податоци и екстерните податоци од јавни извори.

Алатките што ги користат методите на кластерирање мораат да имаат способност за обработка на сите видови податоци (нумерички, текстуални, сликовити и др.). Алгоритмите за кластерирање мораат, на барање на корисникот, да пронајдат кластери што најсоодветно ќе ги опфатат со своите карактеритики определените групи податоци, на пример, потрошувачи со некои заеднички форми на однесување при купувањето. Техниките, односно методите што се користат за кластерирање можат да бидат: стандардни статистички методи, невронски мрежи, генетички алгоритми и стебла на одлучување. Кластерирањето сегментира определена група податоци на повеќе помали групи или кластери со слични карактеристики. Тоа значи дека податоците се групираат заедно, врз на основа на сличност на определените параметри. На слика 19 е прикажана една таква кластер анализа со помош на којашто податоците се сегментирани во определени групи, според некои нивни слични карактеристики.

[50]J. Han and M. Kamber, Data Mining:Concepts and Techniqeues, Morgan Kaufmann, San Francisko, 2004

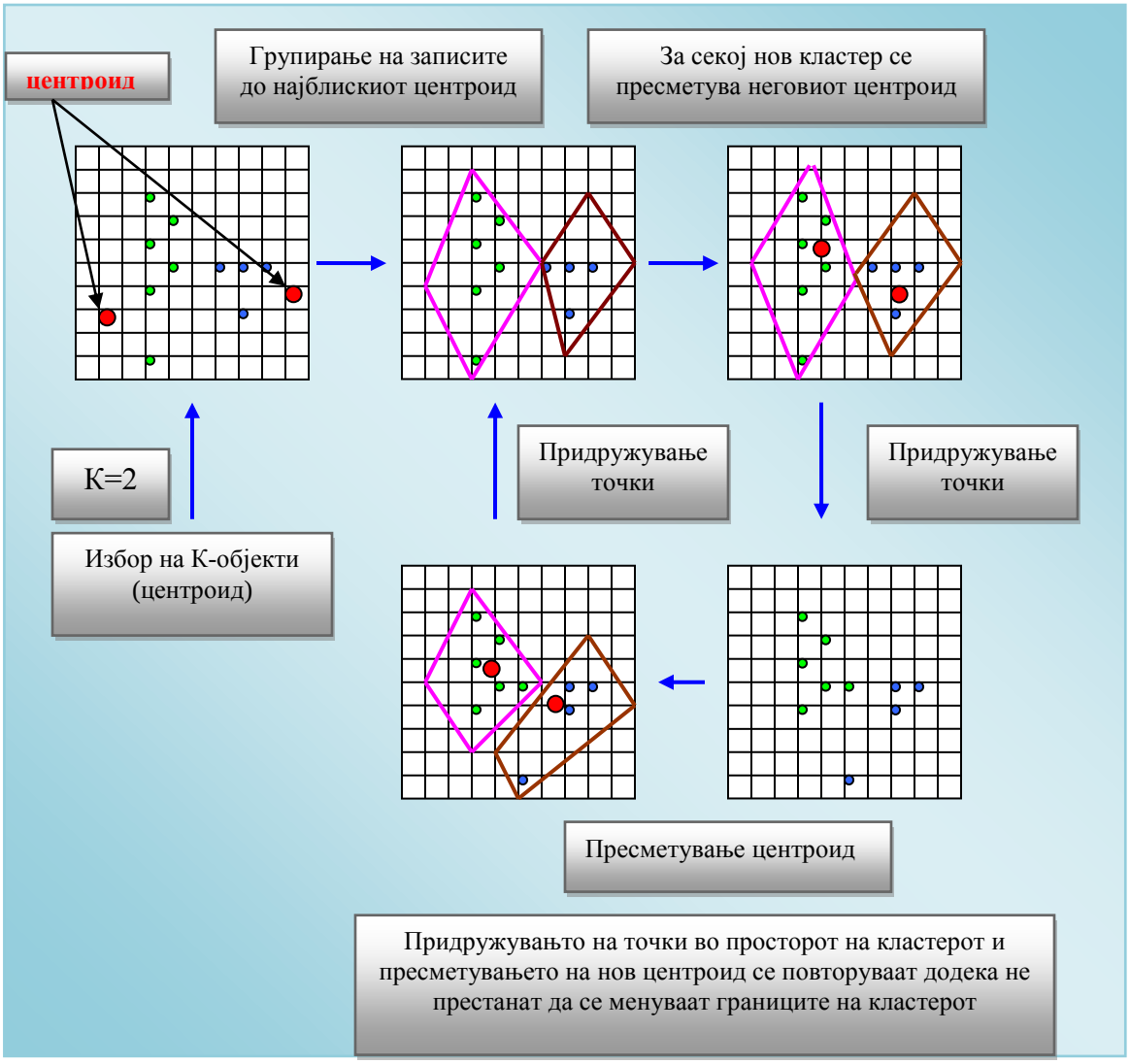


Слика 19. Податоци сегментирани со кластер анализа

Автоматската кластер анализа- припаѓа на групата техники на **неводен Data Mining**[51]. Најпознат алгоритам за кластерирање е кластер анализата **К-средина (K-mean)** на вредности на податоци. Најдобри резултати се постигнуваат со нумерички вредности на влезните податоци. K-s алгоритамот ја дели целата група податоци на определен број помали групи податоци или **кластери.** Бројот на кластерите се означува со "k". Секој поединечен податок (или запис на податоци) може да се прикаже како точка во повеќедимензионалниот простор, а бројот на димензиите зависи од бројот на полињата на поединечниот запис. Записите се групираат во поединечни кластери преку процес, што може да се повторува се додека определен запис не се групира во кластерот којшто според определените својства, му е најсличен. Во почетокот, со методот на случаен избор, се земаат неколку точки во просторот, т.н центроиди односно точки што ја означуваат средината на некој кластер. Потоа, за секој запис се пресметува оддалеченоста во однос на секој центроид и се определува кој центроид му е најблиску на дадениот запис. Со тоа секој поединечен запис е групиран во

[51] www.statsoft.com

определен кластер. Потоа, за секој нов кластер се пресметува неговиот центроид, така што се пресметува средината или просекот на секоја точка во рамки на поединечниот кластер, земајќи ги предвид сите димензии. Процесот на придружување на точки во просторот на определен кластер и пресметувањето на нов центроид се повторуваат се додека не престанат да се менуваат границите што ги ограничуваат поединечните кластери. Ваквиот процес на кластеризација е прикажан на слика 20.

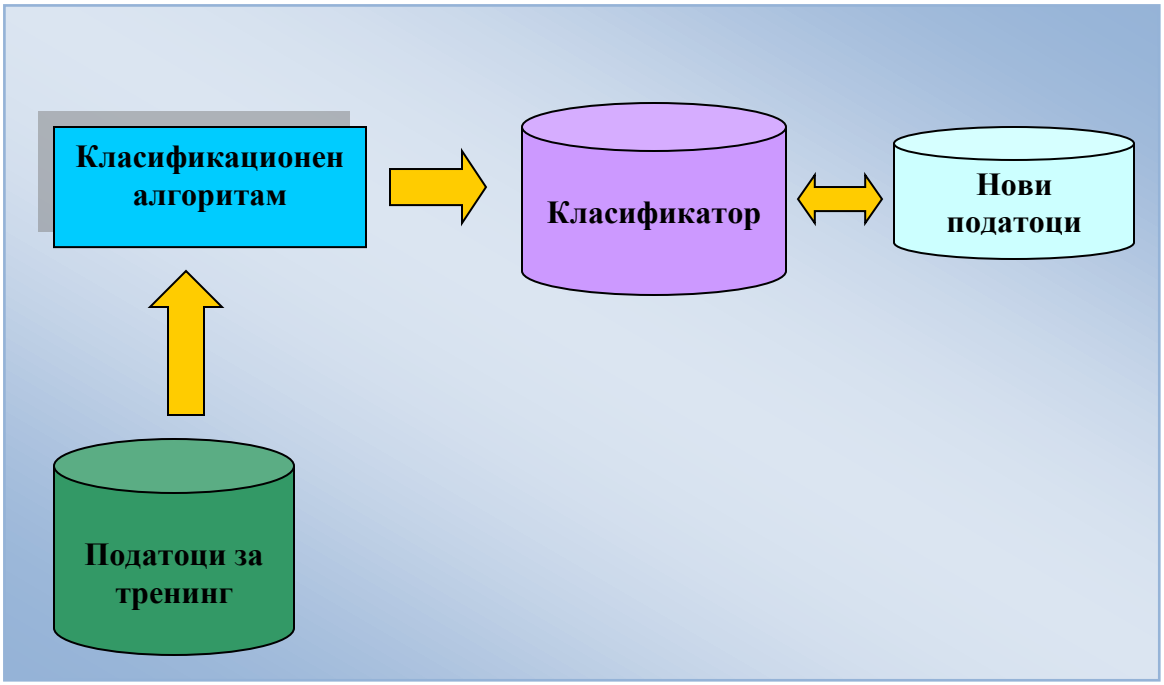


Слика 20. Процес на кластерирање

Резултатите од кластер анализата можат да послужат за проучување како кластерите се однесуваат при промена на влезните променливи. Автоматската кластер анализа најчесто се користи како подготовка за техниките на **воден Data Mining**. При изградба на стебло на одлучување, пак, центроидот на кластерот може да се користи како целна варијабла за определување на правилата со кои новите податоци можат да се групираат во поединечен сегмент или кластер. Кластер анализата се користи кога постојат многу примероци во податоците, па важно е да се издвои групата на примероци што најдобро го опишуваат однесувањето на набљудуваниот објект.

1.2 Класифицирање

Класифицирањето определува кој кластер (центроид на кластерот) е најблизок до карактеристиките на набљудуваната целна променлива (на пример, при класификација на клиенти). Тоа би значело дека со класифицирањето се истражуваат карактеристиките на клиентот и се стават во групи (кластери), во зависност од сличноста спрема главните карактеристики на кластерите, односно нивните центроиди. Според тоа, главната цел на класификацијата е претходно нераспоредените слогови да ги додели на некоја класа, што е можно попрецизно, при што за тестирање на прецизноста се користат тест податоци. Класификацијата ги проучува новите карактеристики на определен објект и го придружува кон претходно дефинираните класи. За класифицирање треба да постојат добро дефинирани класи за да може подоцна сето тоа да се примени при класифицирање на податоците што не се придружени кон некоја класа. На класифицирањето му претходи кластерирање со коешто се добиваат кластери. Класификацијата се разликува од предвидувањето, бидејќи во овој случај се работи за моментална, а не за некоја идна состојба. Така, на пример, со класифицирањето може да се определи во која група можат да се групираат акциите според вредноста, ризичноста или доходовноста, а потоа во која група можат да се групираат новите клиенти за кои се уште нема податоци за користењето на производите, определувањето на нивото на ризик при одобрување кредит итн. Процесот на класификација би можел да се прикаже како на слика 21



Слика 21. Процес на класификација

Потребно е, исто така, да се направи проценка колкава е можноста на класификаторот за генерализирање врз основа на податоците за тренирање. Тоа е потребно да се направи, бидејќи е можно да се јави проблемот на претренираност (overfitting) кога генерираниот класификациски модел добро се однесува на податоците за тренирање, но покажува слабост кон новите податоци (за тестирање). Тоа значи дека проблемот не е добро генерализиран. За тестирање мора да се користи нов примерок на податоци што не е користен за тренирање. На крајот, пак, се врши евалуација при што познатата класа се споредува со онаа што била предложена од страна на класификаторот.

Техниките за класифицирање се генетички алгоритми, невронски мрежи, стебла на одлучување и стандардни статистички методи.

1.3 Проценка и предвидување

Проценката и предвидувањето, за разлика од претходните две методи, генерираат **нумерички искажани вредности**. Така, овде, би можел да се процени кои акции, споредувајќи ги со саканите карактеристики, би можеле да се вклучат во соодветното портфолио на хартии од вредност, па во корелација со клучните показатели да се предвиди нивната идна вредност. Исто така, може акциите да се рангираат според берзанските индекси, па со тоа и според нивниот иден раст или пад на вредноста. Техниките за проценка и предвидување на ваквите идни движења се невронски мрежи и статистички методи.

Активноста на проценка се одвива преку **континуирани одговори**. Имено, според новите влезни податоци се врши проценка на вредноста на непознатите континуирани променливи. Врз основа на тоа, и резултатите од проценката се континуирани вредности. Историските податоци се користат за изградба на модел што ќе го објаснува однесувањето на набљудуваниот објект. Резултат од примената на моделот врз влезните податоци е предвидувањето на идното однесување. Пример за ова може да биде предвидувањето кој клиент или потрошувач може да ја напушти определената марка производ или услуга во некој иден временски период.

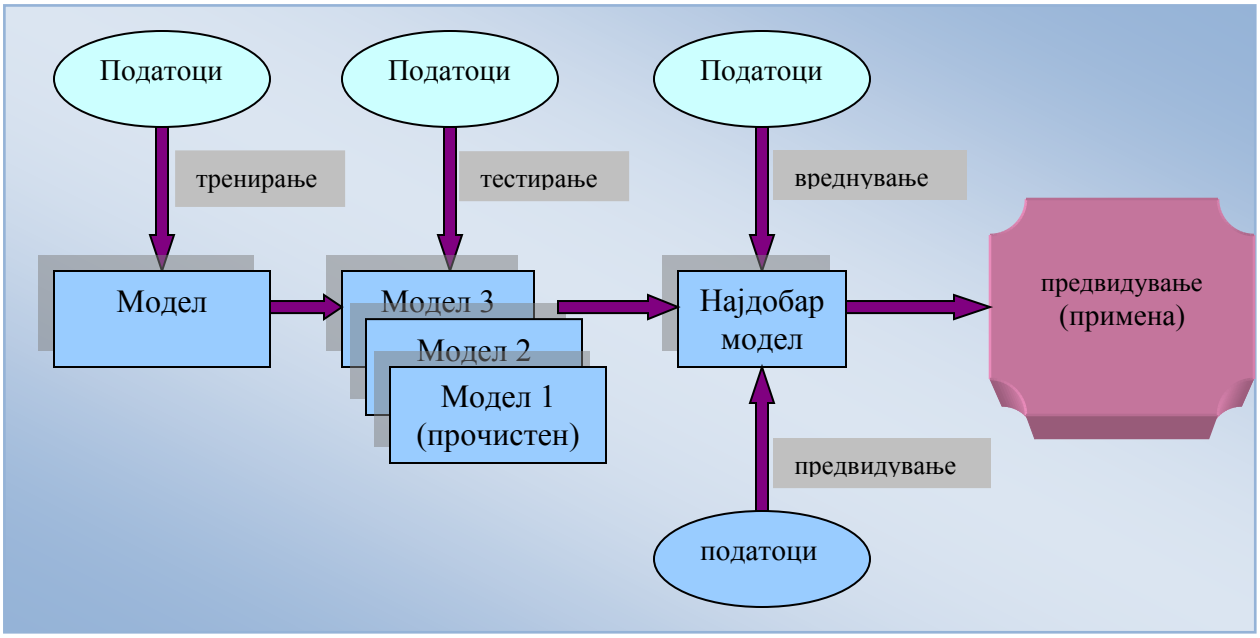
При изградбата на моделот за предвидување во почетокот е потребно да се најде доволна количина податоци со познати излезни резултати. Таквите познати примери се наменети за учење на моделот за податоците, односно тренирање на моделот.

Основни чекори во изградбата и примената на моделот за предвидување се[52]:

- моделот се тренира на делот што се однесува на податоците за тренирање, врз основа на што моделот се учи за однесувањето на податоците;

[52]Vasand Dhar and Roger Stein, Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice Hall, NY, 2002

- врз основа на делот од податоците за моделирање, што се нарекува група податоци за тестирање, моделот се обидува да генерализира за сите ситуации за да може да работи добро и со новите податоци;
- на третиот дел од податоците за моделирање, т.н. податоци за проценка, се врши споредба на повеќе различни модели, па врз основа на тоа се избира најдобриот модел;
- на крајот, моделот се применува на оние податоци што не биле употребени за моделирање. Врз основа на тоа се добива предвидување што ќе помогне во процесот на донесување одлуки.



Слика 22. Изградба и примена на моделот за предвидување

Моделот користи група податоци за тренирање за да ги препознае непознатите шеми и шаблони во збирот на податоци. Шаблоните се комбинација од општи шаблони во групата за моделирање и специфични шаблони во групата за тренирање. При примената на моделот, тој секогаш се потпира на временското поместување во процесот на моделирање, така што се применува на најсвежи податоци, многу понови од групата податоци за моделирање. Ако моделот се

користи неколкупати, тогаш треба да се провери неговата точност за да се знае кога би било потребно тој да се **редизајнира.**

Искуствата покажуваат дека за кој било проблем во класификацијата и предвидувањето постојат ситуации во кои, во иницијалната фаза, е избран еден метод, додека некој друг метод е избран во подоцнежните фази од проектот. Секој од методите е заснован на специфичен алгоритам и е соодветен за решавање на точно определен проблем и анализа на податоци. Исто така, секој метод има одделни позитивни ефекти од неговата примена, но и некои ограничувања. На крајот е важно да се напомене дека категоријалните податоци се поврзани со класификацијата, додека континуираните податоци се поврзани со предвидувањето (регресија).

1.4 Анализа на веројатност

Кога некоја емпириска дистрибуција апроксимира, т.е. следи определена теориска дистрибуција на веројатност, може да се употреби теоретското знаење за дистрибуцијата што се набљудува за да може да се добие одговор на одделни прашања за податоците. Тоа, најчесто, бара проценка на веројатноста. Анализата на веројатноста се заснова на веројатноста на набљудувањето на некој настан, односно таа претставува мерка за неизвесноста[53]. Притоа, таквата веројатност се искажува во мерни единици од 0 до 1, претставувајќи го опфатот на веројатноста на случување на настанот или на негово не случување. Ако веројатноста е 0, тогаш настанот не е можно да се случи, а ако веројатноста е 1, тогаш настанот сигурно ќе се случи. Тоа се двете гранични вредности на веројатноста. Постојат различни пристапи при мерењето на веројатноста, односно:

- **субјективен пристап**- што подразбира определен степен на субјективност дека настанот ќе се случи;
- **фреквенциски пристап**- што се заснова на броењето настани при повторување на експериментот;

[53] www.psychstat.missouristate.edu

- **a priori пристап**- претпоставува познавање на теоретскиот модел, т.е. дистрибуција на сите можни веројатности на некој настан;

Случајната променлива x е променлива што добива поединечни вредности со определена веројатност. Постојат два основни типа случајни променливи, и тоа:

- **дискретна или дисконтинуирана случајна променлива,** при што нејзината нумеричка вредност е претставена со цели броеви; и
- **континуирана случајна променлива,** при што нејзината нумеричка вредност е претставена со реални броеви.

Дистрибуцијата (рапределбата) на веројатноста го прикажува начинот на кој вкупната веројатност е распределена на поединечните вредности на случајната променлива. Секоја дистрибуција на веројатноста ја дефинираат одделни параметри, како што се, на пример просечната вредност, варијансата, коваријансата итн. Во зависност од типот на случајната променлива и дистрибуцијата може да се подели на дискретна и континуирана. Така, во дискретна дистрибуција на веројатноста можат да се вбројат биномна и Пуасонова распределба додека во континуирана дистрибуција на веројатноста спаѓаат нормалната распределба, χ^2 распределба, t-распределба, F-распределба.

Основа за пресметување на веројатноста за случување на некој настан се претходните појави на настанот. Значи, тоа се историски податоци врз основа на коишто се пресметува фреквенцијата на појавување на настанот.

1.5 Регресиона анализа

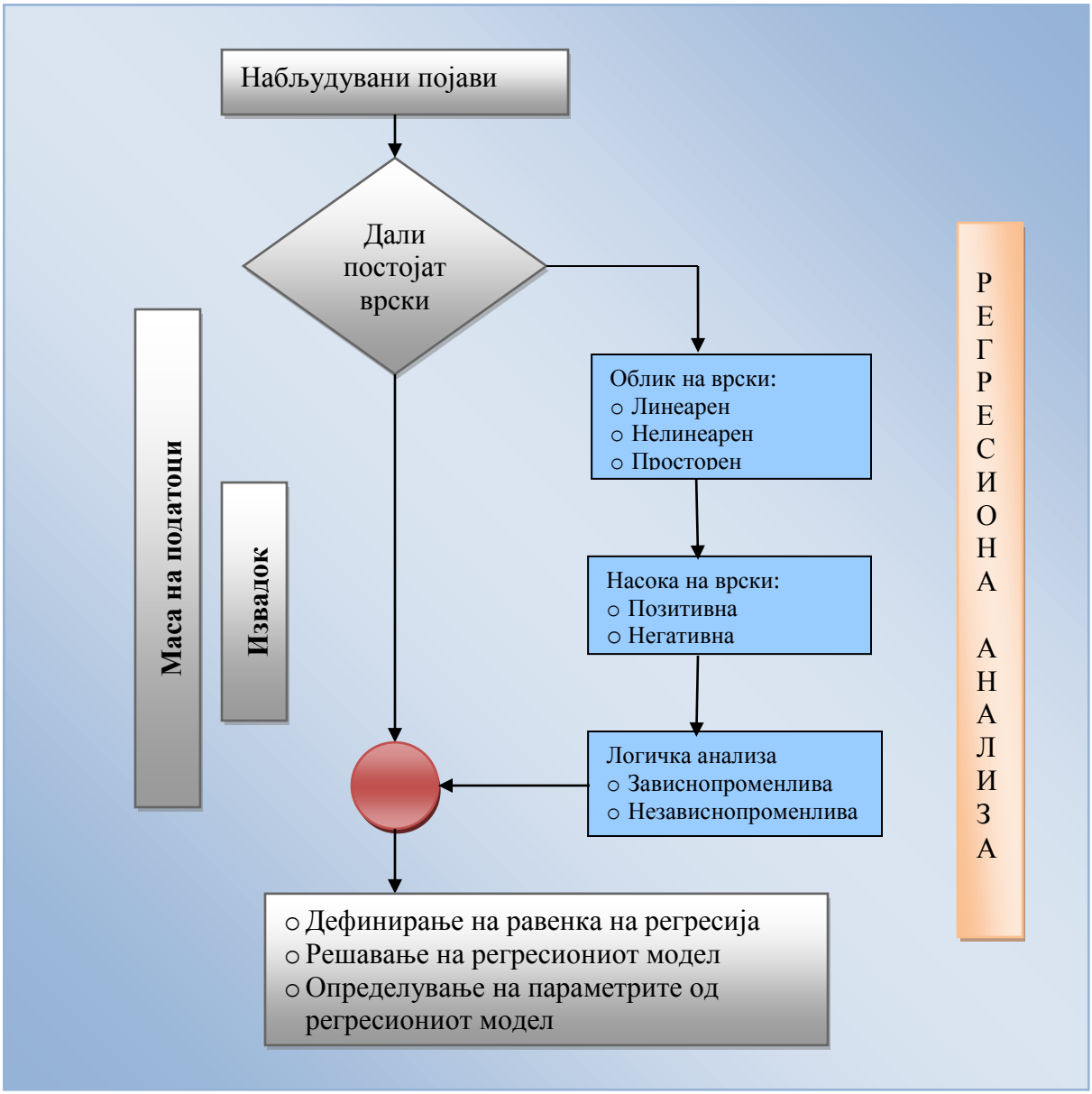
Овој метод се заснова на воочување на **врската меѓу две или повеќе променливи**[54]. Основна цел на регресионата анализа е да го идентификува постоењето на определена поврзаност меѓу две или повеќе појави или променливи, како и да го објасни обликот и насоката на тие врски. Со помош на регресионата анализа се согледува секој сегмент на промените на една или повеќе променливи врз промената на некои други променливи. Многу економски појави и процеси меѓусебно се поврзани и условени, па ваквите појави можат да се набљудуваат врз

[54]Wang G. C. and Jain C. L., Regression Analysis:Modeling and Forecasting, Institut of Business Forecasting, 2003

основа на регресионата анализа. Притоа се отстранува варијабилитетот на две или повеќе појави, односно се утврдува **регресиониот модел** на кого што најмногу му одговара квантитативното совпаѓање на варијациите од набљудуваните појави[55]. Таквиот регресионен модел или процес на регресиона анализа е прикажан на слика 23.

Но, треба да се има предвид дека често не е можно со истражувањето да се опфати целокупната популација на податоци, па поради тоа во регресионата анализа најчесто се испитува извадок и врз негова основа се носи одлука за зависноста меѓу одделните појави.

[55] Коста Сотироски, Статистика, Универзитет "Св. Климент Охридски", Економски факултет, Прилеп , 2004, стр 151



Слика 23. Процес на регресиона анализа[56]

Врз основа на претходниот алгоритам може да се заклучи дека за регресионата анализа еден од клучните услови е да се утврди обликот на набљудуваните појави. Таквиот облик на врска меѓу набљудуваните појави би можел да биде:

[56] Коста Сотироски-Статистика, Универзитет "Св. Климент Охридски", Економски факултет, Прилеп , 2004, стр152

- **линеарна врска**, односно праволиниска, ако на паровите оригинални вредности од обележјата на набљудуваните појави може да им се приспособи линеарна функција;
- **нелинеарна,** односно криволиниска, ако на паровите оригинални вредности од обележјата на набљудуваните појави може да им се приспособи нелинеарна функција;
- **просторна,** ако на паровите оригинални вредности од обележјата на набљудуваните појави може да им се приспособи просторна функција, во облик на рамнина или хиперрамнина.

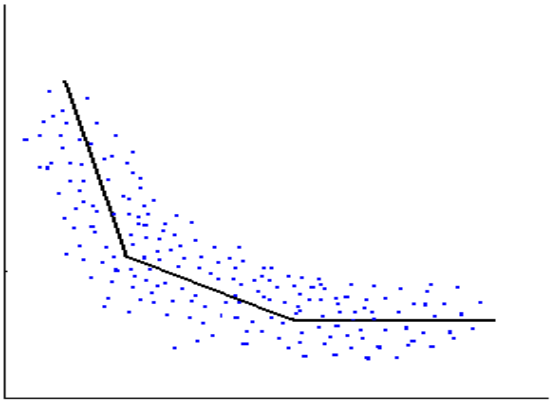
Исто така, насоката на врските може да биде:

- **позитивна врска** меѓу набљудуваните појави, што се реализира во случај кога растот на независната променлива во просек предизвикува пораст на зависната променлива; и
- **негативна врска** меѓу појавите кога растот на независната променлива во просек предизвикува опаѓање на зависната променлива.

Од графички аспект, регресијата е линија низ димензионалниот простор на вредностите на независните променливи. Вредностите на секоја променлива се прикажани како точки во просторот. Линијата на регресија го претставува најдобриот правец низ просторот од точки, односно правецот што ја минимизира просечната оддалеченост на точките од линијата[57]. Равенката на регресија може да има облик: **y = a + bx,** каде y е вредноста на зависната променлива или вредноста што треба да се придвиди, а x е вредноста на независната (позната) променлива. Врз основа на историските податоци се утврдува формулата за регресија, а со вклучувањето на познатите вредности се предвидува вредноста на целната променлива y.

На сликата 24 е прикажана линијата на регресија низ група вредности на податоците. Насоката на движење на линијата на регресија е определена од позицијата на точките во дводимензионалниот простор.

[57] www.spss.com



Слика 24. Линија на регресија

Линеарната регресија е статистичка техника што го мери односот меѓу две континуирани променливи-зависнопроменливи. Таа работи на принципот да се најде линија низ податоците што ќе ја минимизира грешката. Логичката регресија е слична на линеарната регресија . Разликата е во тоа што променливите се дискретни или категоријални. Логистичката регресија користи континуирани вредности за предвидување на категоријални резултати. При изградбата на моделот можат да се јават многу променливи за предвидување. Редуцирањето на бројот на променливите се врши со избирање на оние што имаат посебна важност за предвидувањето. Во почетокот сите променливи можат да се сместат во моделот, а потоа да се отстрануваат оние што се помалку важни за предвидувањето или оние за коишто е потешко да се предвидува.

1.6 Стебло (дрво) на одлучување

Резултатите што се добиваат со помош на стеблата (дрво) на одлучување се лесно разбирливи за луѓето. Откриеното знаење се прикажува во облик на стебло. Секое стебло на одлучување е збир на правила, разбирливи за луѓето во обликот **ако-тогаш (if-then)** правила**.** Главна предност на стеблата на одлучување е што со

својата структура и едноставни правила можат да објаснат на кој начин се доаѓа до заклучокот. Предноста на правилата се довербата на корисниците во моделот и разбирливоста на правилата. Притоа, откриените правила мораат да бидат конзистентни, односно не смее да постои контрадикторност меѓу нив. Ако правилата се развиваат независно едно од друго, тогаш можат да бидат контрадикторни.

Изградбата на моделот во облик на стебло на одлучување започнува од масата податоци што, со користење на одделни правила на гранење, се разделува на гранките на стеблото[58]. Овој процес продолжува се додека податоците не станат **хомогени.** Стеблата особено се корисни кај проблеми што имаат голем број влезни променливи. Со користењето на стеблата на одлучување се врши избор само на оние променливи што имаат клучна улога за решавањето на проблемот.

На коренот од стеблото податоците се **хетерогена маса**. Од таа маса на податоци стеблото генерира низа логички тестови (или прашања), а секој од нив ги дели записите во помалку ексклузивни групи. Секоја поделба креира гранки на стеблото. Поделбата на гранки има две цели: креирање групи што се хомогени и креирање групи што овозможуваат натамошна поделба, ако процесот продолжува. Поделбата продолжува се додека податоците не станат хомогени или се додека не е постигнат максимумот на јазли.

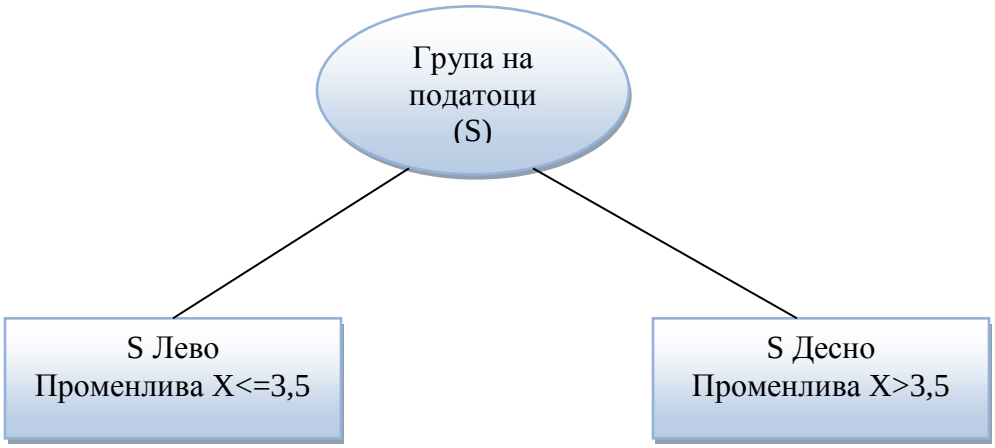
Некои алатки на Data Mining генерираат многу големи стебла на одлучување со голем број гранки. На таков начин се направени толку фини поделби, што и најмалите групи податоци од податоците за тренирање се групирани во некои гранки. Таквата детална поделба предизвикува **претренираност** кога моделот ја запаметува секоја појава од податоците во групата за тренирање, така што моделот не е применлив на нови податоци. Ова доведува до зголемување на грешката. Решението е во **осиромашување на стеблото (prunning)**[59] со исфрлање на некои поделби на крајот од гранките.

[58] www.thearling.com
[59]George M. Marakas, Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str 146

Постојат две основни примени кај стеблата на одлучување, и тоа: класификација и регресија (предвидување). Кај класификацијата записите се придружуваат на класите за да се дојде до определен степен на доверба во правилно дефинираните класи, додека кај регресијата се проценува вредноста на целната променлива. Класификациските стебла за одлучување се многу добри за идентификување на сегмент со посакувано однесување. Стеблата на одлучување имаат за цел секвенцијално да ги поделат податоците за да се минимизира разликата на зависната променлива. Стеблото на одлучување кај секој јазол, при некое разгранување, ја тестира променливата за да се утврди дали ќе се оди кон натамошно разгранување. Имено станува збор и за делење на просторот на вредноста на определена променлива. Со делење на одделните променливи се доаѓа до група или класа на определени променливи. Поделбата на просторот на вредноста на променливата може да биде **бинарна или мултикласна**[60]. Ако гранењето се врши на бинарна основа, тогаш поделбата е условена со одговор на прашањето-дали условот за гранење е задоволен или не (да или не, 0 или1). Гранењето тргнува од групата податоци или т.н. податоци за дефинирање на структурата на стеблото, па со одделни алгоритми се настојува да се поделат податоците на групи што имаат смисла.

Пример за едно такво гранење кај стебло на одлучување е кај **CART**[61]**,** што започнува од нехомоген простор на податоци за да ги подели податоците на хомогени групи. Алгоритамот најпрвин бара променлива што ќе се користи за поделба на податоците. Потоа, групата податоци се дели на две подгрупи:првата се податоци што се помали од т.н. праг на вредноста на појава на набљудуваната променлива и другата подгрупа се податоци што се поголеми од тој праг на вредности.

[60] Ramakrishnan R., Freytag J.C, Report based on a workshop Data Mining:The Next Generation, Dagstuhl seminar, 2005
[61] www.anderson.ucla.edu



Слика 25. Бинарно гранење на стебло на одлучување

Набљудувајќи го збирот на податоци S тој може да се замисли како група парови од влезови X и излези Y, односно S: $(X_1 Y_1)$, $(X_2 Y_2)$,…,$(X_n Y_n)$, каде X претставува низа на влезни променливи X= (X_A, X_B, X_C), а Y претставува класа на појава на целната променлива 1 или 0 (раст или опаѓање). Според тоа, збирот на податоци S е составен од нееднаква дистрибуција на појавата на излезната променлива, па таквата група податоци е нехомогена. Со помош на следната формула може да се пресмета **степенот на нехомогеност**[62]:

$$d(S) = \sum - \frac{N_{klasa}}{N_{\log 2}} \frac{N_{klasa}}{N}$$

каде S е збирот на податоци, N_{klasa} е број на јавувања на определена класа, а N е вкупниот број податоци. Ако збирот на податоци е потполно нехомоген, вредноста ќе биде 1, а ако групите потполно се хомогени, вредноста на d(S) ќе биде 0. По поделбата на почетната група податоци, алгоритамот ги бара сите можни поделби на двете дополнително формирани подгрупи, односно сите можни комбинации на

[62] J. Han and M. Kamber, Data Mining:Concepts and Techniqeues, Morgan Kaufmann, San Francisko, 2004

гранични вредности (прагови) на овие подгрупи. Таквиот праг за набљудуваната променлива може да биде 3,5, па поделбата се врши на левата страна со вредности помали или еднакви на 3,5 и на десната страна со вредности поголеми од 3,5. За секоја комбинација се пресметува **просечната вкупна нехомогеност (неред)[63],** користејќи ја формулата:

$$\text{Просечна вкупна нехомогеност} = [(N_{levo}/N)*d(S_{levo})] + [(N_{desno}/N)*d(S_{desno})]$$

Онаа комбинација гранични вредности (праг) што ја минимизира просечната вкупна нехомогеност ќе биде избрана како најдобра поделба. Истиот начин на поделба се применува и на гранките лево и десно.

Методите на стебло на одлучување се добар избор за Data Mining кога треба да се класифицираат записи или треба да се направи предвидување. Освен тоа, стеблата на одлучување можат да се користат за генерирање на лесно разбирливи правила што можат да се преведат во кој било програмски или природен јазик. Таквите правила можат лесно да се претворат во SQL или псевдокод.

Хибридниот пристап за изградба на модел за Data Mining ги издвојува стеблата на одлучување од тренираните невронски мрежи и ги комбинира со нив. Тренираната мрежа се користи како влез во алгоритамот за откривање на правилата.

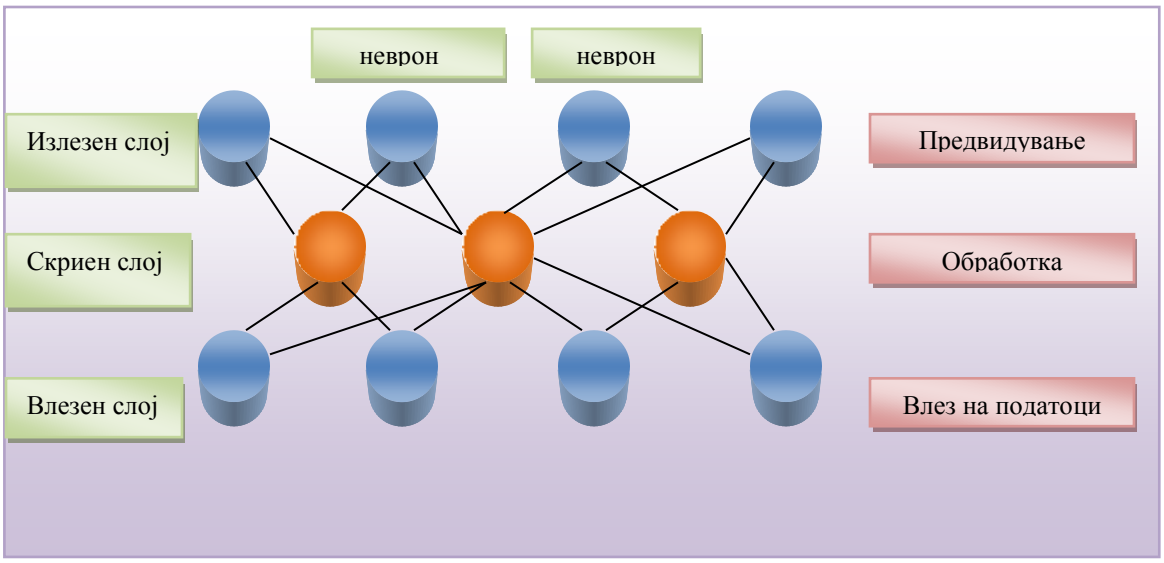
1.7 Невронски мрежи

Алгоритмите на невронските мрежи пробуваат да ја симулираат работата на **човечкиот мозок**. Тоа не значи дека невронските мрежи можат да мислат, туку нивна основна карактеристика е тоа што тие можат да "учат" врз основа на минатите податоци, да бараат врски меѓу податоците и да предвидат некоја идна состојба со извесно ниво на грешка. Невронските мрежи работат врз принципот на постојано презентирање парови на влезови и излези на нивните слоеви. Врз основа

[63] J. Han and M. Kamber, Data Mining:Concepts and Techniqeues, Morgan Kaufmann, San Francisko, 2004

на вредноста на грешката, пресметана како разлика помеѓу добиениот излез и точниот излез, се врши корекција на грешката. Невронските мрежи се поистоветуваат со "**црна кутија**"[64]. Од едната страна е влезот, а на излезот се предвидувањата, без потреба да се разбере што се случува внатре во "кутијата". Невронските мрежи се многу успешни при развојот на модел за класификација и предвидување кога има голема група податоци.

Невронските мрежи се состојат од неврони, врски помеѓу невроните, влезни функции, функции на активација. Збирот на влезните неврони се нарекува **влезен слој**, внатрешните скриени неврони го претставуваат **скриениот слој**, а излезните неврони го сочинуваат **излезниот слој**. Влезните и излезните неврони се поврзани со врски. Ваквата структура на една типична невронска мрежа е дадена на Слика 26.



Слика 26. Структура на Невронска мрежа

Во невронот j можат да влегуваат повеќе влезови x_{ij} што се множат со пондерите или тежинските фактори w_{ij}, по што се сумираат. Сумираната вредност

[64] George M. Marakas, Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str 132

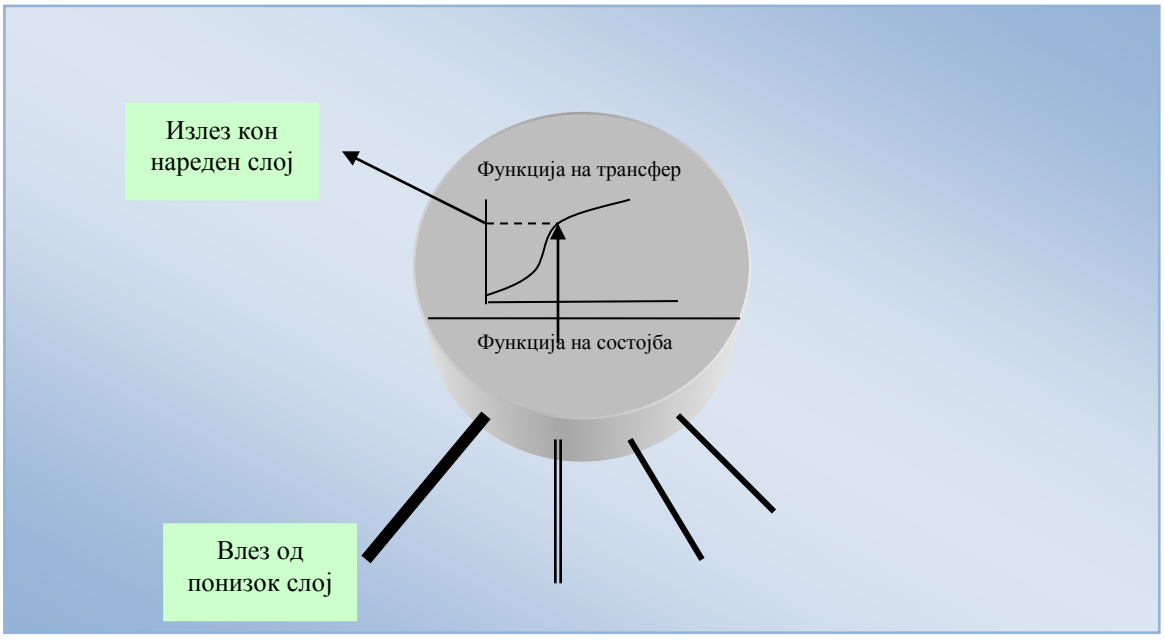
y се проследува преку активациската функција кон излезните неврони. **Внатрешната вредност на локалната променлива**[65] може да се прикаже како:

$$y = \sum w_{ij} x_j$$

Трансформацијата на сумираната вредност на локалната променлива y_i во излезна вредност x_i може да се прикаже како:

$$x_i = f(y_i) = f(\sum w_{ij} x_j)$$

натаму излезната вредност x_i од невроните води кон влезовите на други неврони или кон излезот од невронската мрежа. Структурата на еден таков неврон е прикажана на слика 27.



Слика 27. Типична невронска структура[66]

[65] J.G. Taylor , Neural Networks and their Applications, John Viley & Sons, NY, 1998

Чекорите при изградба на една невронска мрежа се:

- дефинирање на невроните и нивниот број,
- дефинирање на врските меѓу невроните за формирање на мрежата,
- иницијализација на тежинските фактори на мрежата,
- одбирање примерок за тренирање на мрежата,
- дефинирање на примерокот во облик на парови влез/излез,
- прилагодување на тежинските фактори со користење алгоритам за учење над групата податоци за минимизирање на грешките при предвидувањето.

Динамичкиот пристап на изградба на вештачки невронски мрежи укажува на два начина и тоа :

- се почнува со мрежа што не содржи скриени неврони (скриен слој), потоа мрежата се развива, додавајќи скриени неврони, се додека не се редуцира грешката до некое прифатливо ниво,
- почнува со комплексна мрежа, па се бараат врските што се сметаат за непотребни.

Постојат **два пристапа** при моделирањето на невронските мрежи. Имено може да се креира една мрежа што ќе покрива широк дијапазон правила. Недостатокот, во овој случај, е што за таквата мрежа е потребно да се обезбеди голема количина податоци за тренирање. Вториот пристап, пак, претставува хибриден метод со комбинирање на помали мрежи.

Невронските мрежи овозможуваат неколку **предности:** брз одзив, толеранција во однос на комплексноста, релативна независност од експерти, флексибилност и компактност. Недостаток на невронските мрежи е тоа што не можат да интерпретираат резултати. Невронските мрежи можат да обработуваат бинарни и мултикласни класификациски проблеми.

[66] George M. Marakas, Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, 137

За правилен развој на невронските мрежи е потребно придржување до определени **правила.** Влезовите во невронската мрежа често мораат да бидат во граници од -1 до 1, а тоа бара дополнителна трансформација на податоците. Не може да се работи со вредности што недостигаат во променливите, па тие треба да се надоместат со други вредности (проценка врз основа на други вредности во променливата). Невронските мрежи не работат добро со голема количина променливи, па за елиминирање на ова ограничување се користат стеблата на одлучување. Преку нив, всушност, се избираат најважните променливи што подоцна се користат за тренирање на невронските мрежи.

Механизам на тренирање. Тренирањето на невронските мрежи може да се дефинира како процес на **прилагодување на врските меѓу тежинските фактори**[67] на влезовите на секој неврон, така што мрежата треба да постигне што подобар резултат при предвидувањето на целната променлива. Постојат два начина на тренирање на невронските мрежи, а тоа се: надгледувано и ненадгледувано тренирање. Оптимизацијата на невронските мрежи за да се обезбеди подобро предвидување се постигнува со техниката наречена **backpropagation**[68]**,** што претставува техника за надгледувано обучување. Оваа техника работи врз принципот по добивањето на излезните резултати да се пресмета грешката меѓу нив и очекуваните или вистинските резултати. Кога ќе се добие грешката, тогаш одново треба да се вратиме назад низ мрежата и да се прилагодат врските, во зависност од тоа колкава била грешката, односно колкава била разликата. Враќањето назад низ мрежата продолжува се додека не се добие некое прифатливо ниво на грешка.

Прилагодувањето на тежинските фактори w или корекцијата на грешките се прави така што во некоја временска единица t (епоха или итерација при тренирањето на мрежата) врз основа на излезот од мрежата $y_i(t)$ и посакуваната вредност $d_i(t)$ се добива големина на грешка $e_i(t)$.

[67] www.neurosolutions.com

[68] George M. Marakas, Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str 134

$e_i(t) = d_i(t) - y_i(t)$

потоа се определува тежинскиот фактор[69] со помош на формулата:

$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_i(t)x_j(t)$

каде η е големина на чекорот (константна вредност) со којашто се напредува низ процесот на учење, $\delta_i(t)$ е големина на грешката (разлика меѓу вистинската и предвидената вредност), $x_j(t)$ е влезна вредност, а $w_{ij}(t)$ е вредноста на тежинскиот фактор во претходната временска единица t. Големината на грешката $\delta_i(t)$ може да се пресмета на излезот од невронот од грешката $e_i(t)$ или може да се пресмета како сума на грешки внатре во невронот.

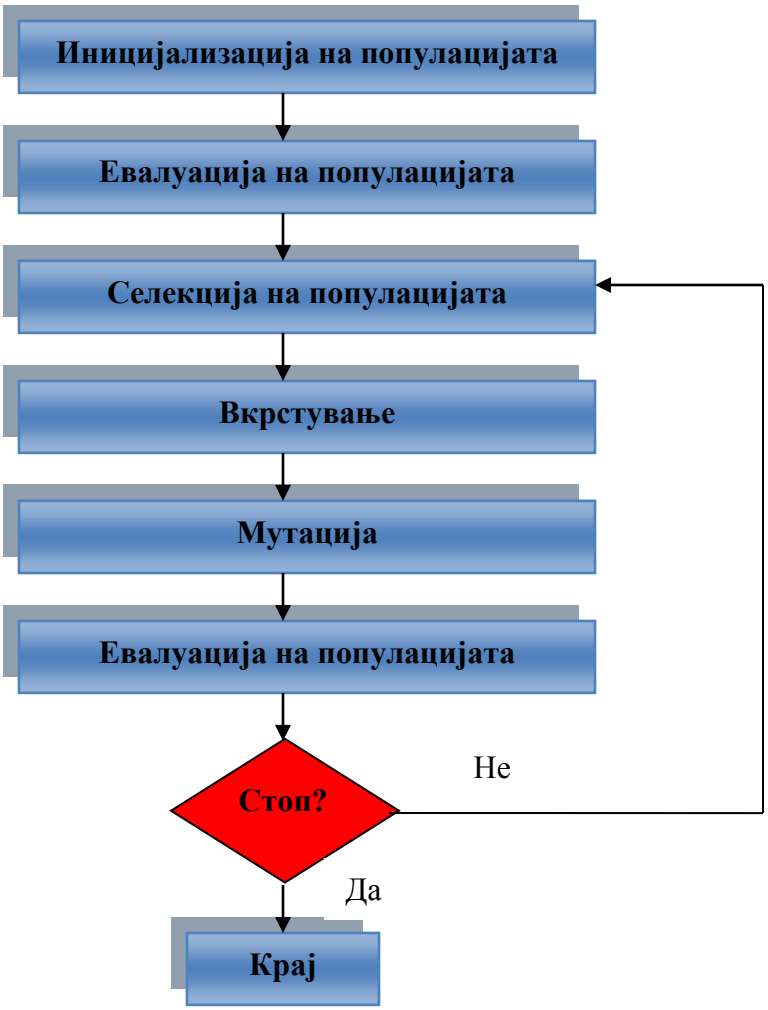
Ваквиот процес завршува кога ќе се достигне минимално ниво на грешка. Резултатот од тренирањето на мрежата е збир на тежински фактори или пондери што се користат во пресметките со кои се определува предвидувањето.

Најчеста примена на невронските мрежи е онаму каде што не се познати правилата за доаѓање до решение (black box), т.е. нема експерти или тие не се достапни; потоа онаму каде што променливите квантитативно можат да се изразат; или кога примената на статистички методи нема ефект, односно кога појавата не може да се претстави со некој линеарен модел; онаму каде однесувањето на појавата често е неизвесно и податоците се непотполни. Врз основа на истражувањата се покажало дека најчеста примена невронските мрежи имаат во производството и операциите (53,5%), финансиите (25,4%) и, во последно време, се почесто во маркетингот[70]. Невронските мрежи не се погодни за примена кога е потребно да се знае веројатноста на излезниот резултат, бидејќи тие не даваат податок за веродостојноста на системот

[69] www.neurai101.com
[70] Wong B.K., Selvi Y., Neural Network Application in Business, A review and analysis of the literature, 2001

1.8 Генетички алгоритам

Принципот на генетичкиот алгоритам (генетски алгоритам) се заснова на **Дарвиновата теорија** на природна селекција и прилагодување. Во оваа теорија како клучен факт се наведува дека организмите кои сакаат да преживеат мора да се прилагодат на условите на околината. Ова прилагодување на околината служи како основа за развој и користење на генетичките алгоритми. Според истата биолошка аналогија, најмала единица на генетичкиот алгоритам се нарекува **карактер (ген).** Тој претставува најмала информациска единица за проблемот и може да се смета како основа за градење одделно можно решение. Продолжувајќи со истата аналогија, серијата карактери (гени) што можат да претставуваат можно решение на проблемот се групираат како **стринг (хромозом).** Стринговите се претставени во компјутерската меморија како бинарни стрингови и тие можат да бидат "декодирани" од генетичкиот алгоритам за да се утврди дали некој стринг може да биде подобар при решавањето на определен проблем. На слика 28 е прикажан процесот на решавање определен проблем со помош на генетички алгоритам.



Слика 28. Тек на процесот на решавање проблем со помош на генетички алгоритам

Ваквиот проце на решавање определен проблем со примена на генетичкиот алгоритам го има следниов тек[71]:

- **иницијализација**. Првиот чекор од креирањето на генетичкиот алгоритам е развивањето иницијална популација на стрингови што можат да бидат можно решение на проблемот. Оваа популација на иницијална генетичка структура може да биде конструирана на начин

[71] George M. Marakas-Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, 142

што ќе ги претставува решенијата, случајно распоредени низ просторот на решенија;

- **евалуација.** Во фазата на евалуација секој стринг се декодира од страна на функцијата за декодирање и се оценува функцијата за детерминирање на тоа која генетска структура е добра, а која не е. Ваквата функција е слична на функцијата на критериум што често се користи кај голем број оптимизирачки техники (како на пример, линеарно програмирање) и се нарекува **Fitness** функција. Оваа функција претставува ограничување или барање за минимизирање или максимизирање при користењето на одделен ресурс;
- **селекција**. Кога веќе еднаш групата стрингови е евалуирана преку функцијата на критериум, генетичкиот алгоритам експериментира со нови солуции за преку нивно рангирање да се избере најдобрата алтернатива. Во суштина, гечетичкиот алгоритам започнува да "експериментира" со постојниот сет стрингови, преку комбинирање и пречистување на секој карактер внатре во стринговите;
- **вкрстување**. Продолжувајќи го процесот на рафинирање, фазата на вкрстување вклучува размена на генетичките информации (содржина на стринговите - карактери) меѓу селектираните стрингови (хромозоми). Оваа операција му дозволува на генетичкиот алгоритам формирање нови стрингови што ќе ги делат позитивните карактеристики, а ќе се елиминираат негативните карактеристики од иницијалните стрингови. Едно важно ограничување на оваа фаза е тоа што можат да се вкрстуваат само карактери што веќе постојат во популацијата стрингови;
- **мутација.** Крајниот чекор од овој процес е мутацијата. Фазата на мутација подразбира случајно менување на вредноста на карактерите од нивната постојна состојба во сосема нова состојба. Поголем број од мутациите, формирани во овој процес не се прифатливи и пожелни, но сепак тука ќе се јават и мутации на едно високопосакувано и бенефитно

ниво. Ваквиот процес на мутација овозможува откривање на одделни непознати подрачја во рамките на одлучувањето, а со тоа и зголемување на шансите за откривање на оптималното решение.

Со помош на генетичкиот алгоритми се настојува да се најде најдобар модел за предвидување. Кај секој модел се тестира способноста за предвидување. Врз основа на мерењето на моделот се доделува вредност што ја покажува неговата способност за предвидување. Потоа, моделот се споредува со други истородни модели. Добрата вредност за способноста на предвидување на моделот ги зголемува неговите шанси за прежививање и негово преминување во наредната фаза на тестирање. На крајот, се избира најдобриот модел.

Генетичките алгоритми се употребуваат за прилагодување на тежинските фактори или пондери. Нивната основна предност е заснована на употребата на паралелно процесирање и евалуациско прилагодување. Друга предност на генетичките алгоритми е тоа што, со нивна помош, често можат да се решат проблеми за коишто нема никаков предзнак за нивно решавање. Ако добро се опишат излезните компоненти на проблемот, со примена на **Fitness** функцијата, што ја користат генетичките алгоритми, тогаш може да се издвоиме едно или повеќе решенија на проблемот, што на друг начин би биле невидливи. Сепак, главната предност на генетичките алгоритми е нивната моќ да не даваат резултати преку комплексните алгоритми, туку преку релативно едноставен и широк, т.е. лесноразбирлив сет од концепти.

1.9 Нејасна (неизразена) логика

Неизразената логика се заснова на логичкиот пристап што датира од времето на старите Грци. Тоа е **метод не резонирање**[72] што дозволува делумно или нејасно опишување на правилата. Комбинирајќи го овој метод на резонирање со подрачјето на дигиталните процесори се доаѓа до една класа компјутерски апликации што можат да учат од сопствените "грешки" и можат да ги "разберат"

[72] George M. Marakas, Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str 127

нејасните ситуации, што се типични единствено за човечкото размислување. Овој метод работи во рангот на вредности од 0 до 1, каде што нулата претставува апсолутна грешка, додека единицата претставува апсолутна вистина. Важно е да се спомене дека Fuzzy логиката не значи давање некаков нејасен одговор, туку поточно прецизен одговор што математички е заснован во определен ранг на вреднсти. Уште повеќе што, овој метод може да работи со кој било степен на прецизност, во зависност од влезните податоци.

Честопати се јавува нејасност при разграничувањето на термините **неизразеност и веројатност**. И неизразеноста и веројатноста се движат во границите од 0 до 1, но сепак меѓу нив постои фундаметална разлика. Веројатноста укажува на можноста дека појавата има некое посебно својство, додека неизразената логика го дефинира степенот до кој ова својство е застапено во појавата. Пред појавата на неизразената логика можеше само да се нагаѓа за степенот на присуство на таквото својство, но сега може со сигурност да се тврди за присуството на тоа својство во појавата и да се споредува со останатите членови на групата. Со веројатноста може само да се детерминира присуството на некоја појава или својство во групата, додека со Fuzzy логиката може да се детерминира и каде се наоѓа некоја посебна вредност во континуитетот вредности на групата.

Неизразената логика или Fuzzy логиката припаѓа на "**expert mining**" методите и ги покрива задачите што опфаќаат донесување одлуки без модел и без доволна количина податоци за тренирање, но со издвоени лингвистички правила на експертите. Fuzzy логиката се комбинира со невронските мрежи. Механизмот на Fuzzy логиката служи за expert mining, а механизмот на невронските мрежи за копање по податоците во овој хибриден состав.

Системите за неизразената логика се користат во низа на апликации што можат да се сретнат во многу области од бизнис работењето. Fuzzy логиката дозволува моделирање што ги вклучува контрадикторностите без потреба за инволвирање на базата на знаење. Неизразените правила можат да бидат комплетно контрадикторни, а сепак да опстојуваат во хармонија. Во овој случај, пак, се доделува определен степен на толеранција на двосмислените одговори, што дава можност за постигнување на одделни компромиси, кога за тоа има потреба. За

разлика од традиционалните системи, засновани на правила, кај Fuzzy системите, доколку едно правило е грешно, грешката може да биде "компензирана" со правилноста на останатите принципи. Од друга страна, пак, употребата на ваквата Fazzy логика може да претставува сериозна пречка за **стабилноста и веродостојноста** на системот. Наспроти високата комплексност, овој метод не може да одговори кога правилата се точни и како такви треба да се употребат. Се уште не постои прецизен математички метод што докрај ќе ја верифицира коректноста на правилата, што ги нудат ваквите Fuzzy системи. Уште повеќе, високиот степен на комплексност на некоја појава, што се настојува да се моделира со користење на неизразената логика, може да доведе до губење на некои важни информации за појавата.

2.Класификација на методите за Data Mining

Data Mining може да се класифицира на **воден** (управуван) и **неводен** (неуправуван) Data Mining. Целта на управуваниот Data Mining е употреба на расположливите податоци за изградба на модел што опишува определена променлива во однос на останатите расположливи променливи. Неуправуваниот Data Mining служи за опишување и детектирање на врските, односно односите меѓу влезните променливи. Управуваниот Data Mining има **top-down** пристап, што се употребува кога се знае што се бара, а неуправуваниот Data Mining, **bottom-up** пристап има за цел пронаоѓање на некои шаблони меѓу податоците. Управуваниот Data Mining користи веќе познати примери и ги применува над непознатите податоци. Кај Data Mining некогаш не е важно како моделот работи, а некогаш дури за подобар поглед во податоците се бара полутранспарентен модел. Целта на предвидувањето е да се научи од минатото како да се употреби стекнатото знаење во иднина.

Методите за Data Mining според **функцијата** што ја извршуваат можат да се класифицираат на следниов начин:

- класификациски методи (дискриминативна анализа, логичка регресија, невронски мрежи и др.);
- регресиски методи (линеарна регресија, повеќекратна регресија и корелација, логичка регресија и др.);
- методи на кластерирање (автоматска кластер анализа);
- методи за визуелизација и истражувачка анализа на податоци;
- методи за моделирање на зависноста;
- методи за анализа на врските;
- методи за анализа на временските серии.

Исто така можна е и друга класификација на методите на Data Mining, и тоа :

- предиктивни (класификација, регресија, временски серии итн.),
- дескриптивни (кластеризација, правила на придружување, стебла на одлучување)

Методите за Data Mining можат уште да се поделат и на:

- статистички методи (регресија, анализа на веројатност, факторска анализа, кластеризација),
- методи на вештачка интелегенција (експертски системи, невронски мрежи, генетички алгоритми, неизразена логика итн.*)

3.Избор на метод

При изградбата на модел за Data Mining често се користат повеќе различни методи за Data Mining. Експериментирањето во изградбата е вообичаена постапка за да се открие способноста на поединечните методи што можат да придонесат за иградба на подобар модел. Претходно беше спомената можноста за оценување на методите со помош на графиконот за кумулативна добивка, при што можеше на истиот графикон да се нацртааат линиите што ги прикажуваат резултатите од

* Невронските мрежи, генетичките алгоритми, неизразената логика, исто така, спаѓаат во областа на т.н. soft computing, каде што е дозволена непрецизност, несигурност и делумна вистинитост и каде што не се инсистира на најдоброто можно решение, туку се прифаќа некое задоволувачко решение.

примената на различните модели. Притоа се избира оној модел чија линија зафаќа најголема површина со основната дијагонална линија.

Спредбата на успешноста може да биде меѓу модели изградени со помош на различни методи на Data Mining, а може да биде и меѓу модели развиени со помош на исти методи на Data Mining. Често, при развојот на моделот, темелно се експериментира со еден метод на Data Mining. Притоа можат да се добијат повеќе модели меѓу кои се врши избор на најдобриот модел од дадениот метод. Натаму, таквиот модел се споредува со останатите модели, добиени врз основа на други методи.

Освен кумулативниот графикон за споредба и избор на метод може да послужи и **матрицата за успешност во предвидувањето**[73] на некој метод. На едната страна на матрицата се наоѓаат вистинските вредности на податоците, а на другата страна предвидените вредности. Во пресекот помеѓу вистинските и предвидените вредности се гледа успешноста во предвидувањето во облик на номинална вредност или процент на точност. Така, може да се зборува за процент на точно предвидени излези или погрешно предвидени излези.

Вредноста на поединечните модели, изградени со помош на различни методи на Data Mining, се оценува врз основа на вредноста на превидувањето во однос на вистинската вредност на реалните резултати. Резултатите од поединечните методи можат да се сортираат и изберат на следниов начин:

- моделите што имаат висока точност на предвидување на излезот се први, а по нив доаѓаат моделите со ниска точност на предвидување;
- според трошочниот принцип, се врши избор на оние модели што претставуваат минимална загуба, во случај на погрешно предвидување.

Поединечните модели можат да се комбинираат за да се утврдат позитивните ефекти од таквото комбинирање. Комбинирањето на моделите може да се изврши на неколку начини, и тоа:

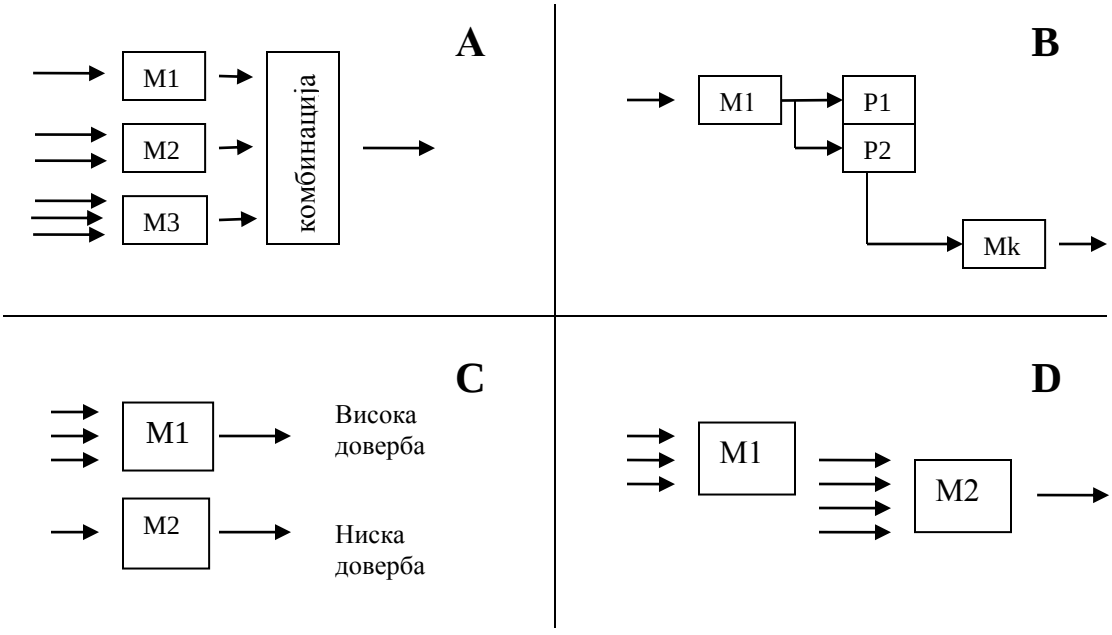
[73] www.goranclepac.com

A. **сегментно влезна** комбинација на модел, што користи различни модели за различни делови на влезните податоци. Само еден модел може да се користи за секој сегмент на влезните податоци. Секој модел врши свое предвидување, а потоа крајниот модел може да се состои од комбинација на повеќет помали модели за предвидување;

B. **моделирана сегментирана** комбинација на модел, што користи резултати од првиот модел за да се сегментираат влезните податоци, што потоа ќе се користат во друг модел за да се добие посакуваниот излез;

C. комбинација на модели со **корегирани грешки**, што ги користат резултатите од еден модел, а тие резултати се многу веројатни, за да се изгради друг одвоен модел. Ако довербата во моделот е голема, тогаш се користат резултатите од првиот модел, но ако не се користат резултатите од некој друг модел со пониска доверба;

D. комбинација на модели **управувана од податоци**, што ги користи резултатите на еден модел, како влез во друг модел.



Слика 29. Комбинација на моделите за Data Mining

Со комбинација на моделите можат да се добијат подобри рзултати, отколку кога предвидувањето би се вршело со поединечните модели. Пример за такво

комбинирање се различните техники, како што се: невронски мрежи и стеблата на одлучување. Важно е дека сите модели, при изградбата, користат исти податоци. Ако тоа не е случај, тогаш излезите мораат да се прилагодат за да се дојде на ниво на кое што можат да се споредуваат моделите.

Веројатноста за користење комбинација од два модела може да се пресмета на следниов начин:

- нека се користат два модели, модел А и модел Б. Доколку веројатноста за успешно предвидување (раст) на моделот А е 90%, а за пад на моделот е 10% и веројатноста за успешно предвидување (раст) на моделот Б е 80 %, а за пад е 20%, тогаш за комбинација од двата модели ќе се добие:
- веројатност за успешен модел = 0,9*0,8 = 0,72
- веројатност за пад на моделот= 0,1*0,2 = 0,02
- вкупна веројатност на двата модела = 0,72 + 0,02 = 0,74.

  Конечната вредност за комбинација на ваквите два модела за нивна успешност во предвидувањето (раст) или пад ќе биде:
- успешност = 0,72 / 0,74 = 0,97297 => 97,3 %
- пад = 0,02 / 0,74 = 0,027 => 2,7 %

Од претходните пресметки се гледа дека со комбинација на моделите се зголемува веројатноста за успешност при предвидувањето.

Времето кога може да се користи моделот за Data Mining најмногу зависи од веројатноста на моделот, па така имплементацијата на моделот може да се подели на три фази:

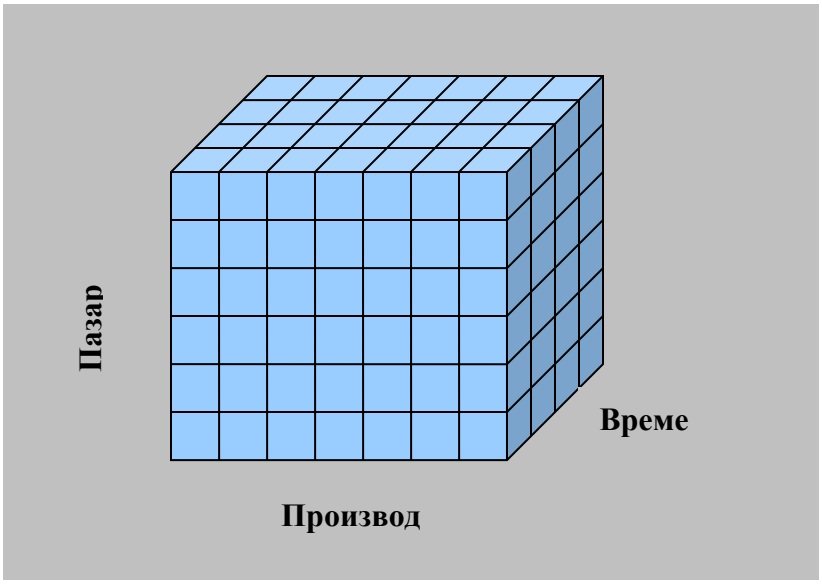
- Alpha – изградба на нов модел и примен во споредба со стариот модел. Резултатите се користат од стариот модел;
- Beta – почнуваат да се користат резултатите од новиот модел;
- Release – се отфрла стариот модел по согледувањето дека новиот модел има подобри перформанси.

4.Аналитичка обработка на податоците

Намалувањето на циклусот на одлучување, заради намалениот профит и критичната профитабилност создава потреба за обработка на голема количина квалитетни податоци што се сместени во складиштата на податоци и примена на соодветни аналитички алатки, во кои што можат да се вбројат и апликациите за аналитичка обработка на податоци т.н. OLAP. OLAP е кратенка од англискиот термин **On line analytical processing** и претставува технологија за брз, конзистентен и интерактивен пристап, односно манипулација со повеќедимензионалните податоци што доаѓаат од различни извори и се сместени во складиштето на податоци[74]. Складиштата на податоци мораат да бидат достапни на различните категории корисници за да ги задоволат нивните информациски потреби. Алатките за аналитичка обработка на податоците ги обработуваат податоците сместени во складиштата на податоци и се наменети да дадат кориснички преглед на големата количина информации, преку изведување на планирани или ad-hoc пресметки, пронаоѓање проблематични подрачја, проверка на претпоставките и плановите и извршување на некои други анализи. Со оглед на тоа дека се наменети за интерактивно користење, примарното барање за овие апликации е брзината. Алатките за OLAP, според својата природа, не се наменети за обработка на основните трансакциски податоци.

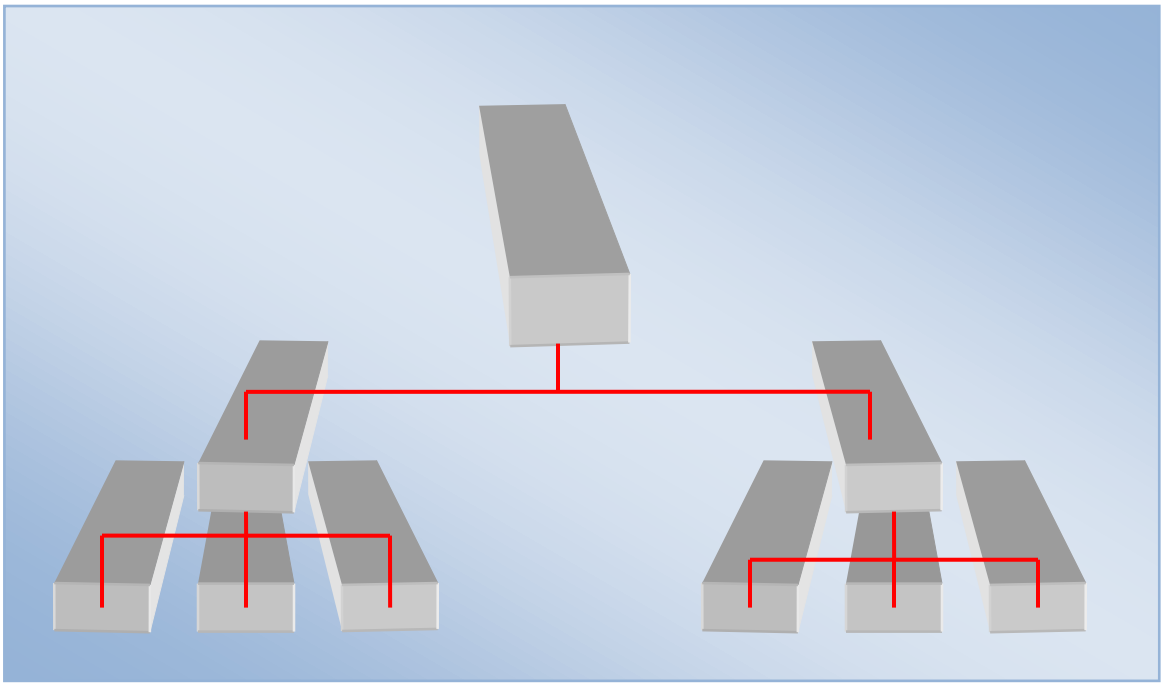
Димензионалната структура на податоците дава можност за нивна визуелизација и овозможува сликовито следење на деловните појави. Деловниот свет е **повеќедимензионален,** па користењето на повеќедимензионалните алатки го олеснуваат прикажувањето на податоците на крајните корисници. Таквата повеќедимензионална структура може да се прикаже во облик на коцка, каде што постојат три основни димензии, и тоа големина, време и сценарио, како и останати димензии (на пример, производи и услуги, проекти, локации, организациска структура, маркетинг сектори, купувачи, добавувачи и конкуренти, функции итн).

[74] www.olapreport.com



Слика 30. 3- D коцка на основните димензии

Најчесто, компаниите користат комбинација од основните димензии и две до пет споредни димензии. Основните диманзии формираат една 3 – D, коцка додека во комбинација со останатите димензии тие можат да доведат до десетици или стотици илјади комбинации, како што е прикажано на слика 31.



Слика 31. Комбинација на повеќе димензии

Во 1993 год E.F.Codd, пронаоѓачот на релационите бази на податоци, го воведе терминот On line analytical processing. Тој заклучил дека оперативните податоци и оперативните бази на податоци едноставно не се адекватни да ги дадат одговорите што типично ги интересираат менаџерите. Тргнувајќи од овој заклучок Codd развил сет од 12 правила за развој и користење на повеќедимензионалните бази на податоци за да помогнат во донесувањето одлуки во организацијата. Тие 12 правила што OLAP алатките мораат да ги задоволат, се[75]:

- **повеќедимензионален концептуален поглед**. Поради својата интуитивност и едноставност, мултидиманзионалниот поглед е поприроден начин на гледање на податоците, па OLAP алатките треба да ја задоволат оваа потреба
- **транспарентност**. OLAP мора да егзистира како архитектура од отворен тип што ќе овозможува примена на аналитички алатки според желбите на корсниците
- **достапност**. OLAP алатките мораат да дефинираат сопствена логичка шема за пристап до различните физички табели и да овозможат едноставен, кохерентен и конзистентен увид во податоците
- **конзистентно известување**. Известувачките перформанси не смеат да имаат деградирачко влијание ако се зголеми бројот на димензиите или количината податоци
- **клиент – сервер архитектура.** Компонентите што се задолжени за излезот треба да бидат доволно интелегенти, така што различните корисници да добиваат одговор на своите прашања, без вложување некои поголеми напори
- **општа димензионалност**. Секоја димензија на податоците мора да биде еквивалентна и во структурата, и во оперативните способности. Имено треба да егзистира само една логичка структура за сите димензии. Која

[75] George M. Marakas, Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003, str74

било функција што се применува на една димензија мора да може да се примени и на сите останати димензии

- **динамичко ракување со матричните податоци**. Физичката структура на OLAP серверот мора во целост да се прилагоди на специфичните аналитички модели што овозможуваат оптимално ракување со ретките податоци во матричните повеќедимензионални табели
- **повеќекорисничка поддршка**. OLAP алатките мораат да овозможат истовремено пристап, интегритет и сигурност на корисниците, кои манипулираат со исти податоци.
- **повеќедиманзионални операции**. Кај повеќедиманзионалните анализи сите димензии се третираат на ист начин. OLAP алатките мораат да ги пресметуваат податоците преку сите димензии, без да бараат од корисникот дефинирање на определена операција за пониските хиерархиски нивоа на податоците
- **интуитивна манипулација со податоците**. Манипулацијата мора да биде едноставна и да овозможува пристап директно, преку податоците во табелите
- **флексибилно известување**. Со употребата на OLAP алатките корисниците можат да манипулираат, анализираат, синтетизираат и прегледуваат податоци на кој било од нив посакуван начин
- **неограничени димензии и агрегација**. OLAP серверот мора да има способност вообичаените аналитички алати да работат барем со петнаесет димензии. Секоја од тие генерички димензии мора да дозволува неограничена агрегација и сумирање, дефинирано од страна на корисникот

OLAP технологиите стануваат се попопуларен комплементарен сегмент на информациските системи, особено во средните и големите компании. Со нивната примена се воведува ред и конзистентност во постапките на комплексната анализа и известување на ниво на поединечните делови или на целиот деловен систем, а нивната правилна имплементација претставува клуч за успехот во функционирањето на секој Data Warehouse. Користењето на OLAP се наметнува во

случај на деловни промени во компанијата, на пример, при реинженеринг на деловните процеси, интеграција или спојување со други компании, промена на раководството, дерегулација, замена на постојниот информациски систем и др.

OLAP серверите на апликациите на складишта на податоци им даваат можност за известување, предвидување на продажбата, анализа на производот и неговата профитабилност, маркетиншки анализи, што-ако анализи и др, односно апликации што бараат историски и изведени (пресметани) податоци. Како резултат на можноста за правење робусни калкулации, OLAP серверите ги трансформираат историските податоци во корисни изведени и предикативни информации. Тоа и овозможува на организацијата увид не само во историските податоци туку, што е уште поважно, и во идните услови на работење. Имено, потребно е да се креираат такви оперативни сценарија што ќе користат историски податоци и ќе ги претвораат во предикативни информации кои што во иднина, значајно ќе влијаат врз работењето на компанијата. Комбинираната анализа на историските податоци со проекциите за идното работење е критичен фактор за успех на современите деловни системи.

Како барања што се поставуваат над алатките за аналитичка обработка на податоците можат да се наведат[76]:

- способноста за обработка на голем обем податоци и за голем број корисници;
- конзистентните и брзи одговори на прашалниците за итеративна анализа
- интегрирани метаподатоци што ќе го поврзуваат OLAP серверот со складиштето на податоци;
- способноста за автоматско симнување од сумираните податоци на детални прегледи се до детални податоци во складиштето;
- калкулациски мотор, што вклучува робусни математички функции за пресметување на изведени податоци (агрегација, матрични пресметки, повеќедимензионални и процедурални пресметки);
- интеграцијата на историски, проектирани и пресметани информации

[76] Hammer J.,Advances in online analytical processing", Data and Knowledge Engineering, Elsevier 2003

- способноста за брзо развивање, прилагодување и економска исплатливост;
- робустен сигурностен систем што ќе го контролира пристапот до податоците и корисничкото управување;
- способност за широк опфат на различните извештаи и аналитички алатки што овозможуваат комуникација со различни корисници;

Трансакциските информациски системи, складиштатата на податоци и OLAP апликациите се разликуваат според многу нивни карактеристики, што се прикажани во табела 2.

	Трансакциски информациски ситеми	Складиште на податоци (Data Warehouse)	OLAP апликации
Цел	Поддршка на транасакциите во деловниот процес	Добивање на информации (сумарни извештаи) за поддршка на одлучувањерто	Добивање на информации (анализи) за поддршка на одлучувањерто
Подрачје на интеракција со кореиснците	Трансакција	Цела база на податоци	Цела база на податоци
Корисници	Службеници и пониски нивоа на менаџери	Менаџери	Аналитичари и менаџери
Тип на извештаи	Основни, едноставен приказ на моменталнта сосотјба, според ограничените кориснички барања	Едноставни извештаи потпомогнати од пред-планирачките процедури на основа на малиот број кориснички барања	Итеративни, компаративни аналитички обработки, ad-hoc навигација и испитување преку спуштање по хиерархиското ниво на димензиите
Форма на извештаи	Фиксни, предефинирани форми	Форма на извештаи управувана од корисникот	Форма на извештаи управувана од корисникот
Приоритеит	Перформанси и расположливост	Перформанси	Флексибилност и корисничка автономија
Пристап			
Пристап кон базата	Предефиниран воглавном статички	Предефиниран и динамички	Динамички
Фреквенција на ажурирање	Ажурирање на податпоци во реално време	Периодично ажурирање со воведување на нови податоци	Ad - hoc
Тип на пристап	Читање- Внесување	Само читање	Читање- Внесување
Пристап до податоците	Пристап до поединечен слог	Пристап до, поединечен слог или група на слогови	Пристап на група на слогови

Податоци			
Карактеристики на податоците	Трансакциски податоци	Историски и детализирани податоци	Аналитички податоци
Степен на деталност на податоците	Трансакциски детали	Пречистени и сумирани податоци	Сумирани, агрегирани и пресметани податоци
Структура на податоците	Нормализирана	Нормализирана и денормализирана	Димензионална, хиерархиска
Количина на податоци	Гигабајти	Гигабајти / Терабајти	Гигабајти
Имплементација			
Време потребно за развој на проектот	Долго време – повеќе месеци или години	Долго време – повеќе месеци или години	Кратко време денови / недели
Потреба за промена на системот	Ретка	Средна	Честа
Големина на инвестицијата	Средни до високи трошоци	Средни до екстрамно вискои трошоци	Минимални до средни трошоци

Табела 2. Разлика меѓу OLTP, Data Warehouse и OLAP

Типична примена на OLAP апликациите може да се најде кај:

- анализата на продажбата и пазарите,
- генерирањето финансиски извештаи,
- генерирањето извештаи за тековното работење,
- билансирањето
- планирањето и буџетарењето
- генерирањето контролинг извештаи
- анализата на профитабилноста според производи, купувачи, пазари, региони, погони, релации или сектори,
- анализата на квалитетот.

Барањето за брзо и ефикасно ракување и ажурирање на големите количини ретки податоци води до повеќе алтернативи на OLAP. Постојат неколку опции каде што податоците можат да се чуваат, и тоа :

- релациски бази на податоци, ако се работи за многу голема база на податоци,
- поделба на повеќедимензионалните бази на податоци на апликациските корисници, особено за средни оптимизирани бази на податоци,

- корисничка платформа во кој било формат, за мали бази што е карактеристично многу брз пристап до податоците

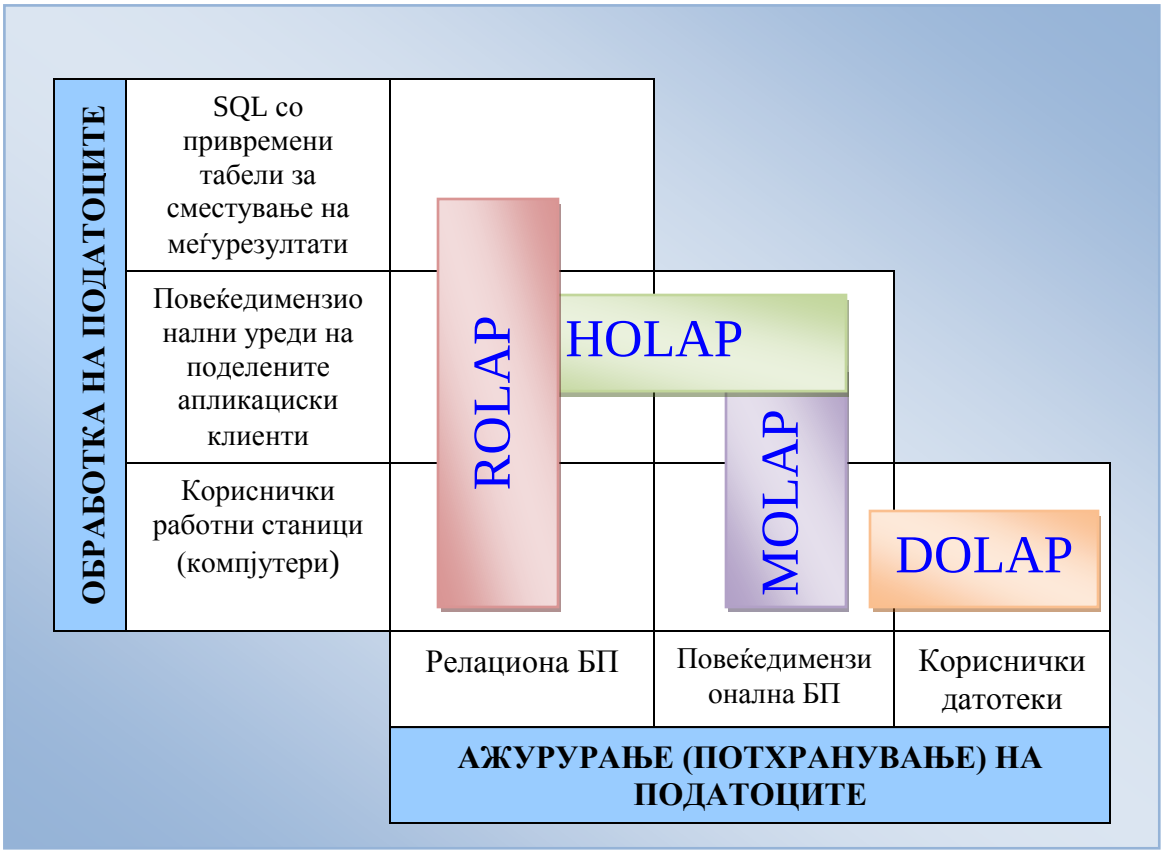
Активните повеќедимензионални податоци можат да се ажурираат (потхрануваат) во :

- работната меморија. што има многу брз, прилагодлив пристап и релативно ограничен капацитет,
- сопствен датотечен / табличен систем, што е средно брз, прилагодлив, ефикасен пристап, т.е. со средно ниво на капацитет,
- релацискатa база на податоци, што е бавен неефикасен пристап, и бара сложен дизајн и прилагодување, но со голем капацитет,

Врз основа на ова, OLAP алатките можат да се поделат во неколку основни групи, односно[77]:

- мултидимензионален OLAP – MOLAP,
- релациски OLAP – ROLAP,
- десктоп OLAP – DOLAP,
- хибриден OLAP – HOLAP.

[77] www.cs.columbia.edu



Слика 32. Комбинација на OLAP алтернативите во зависност од обработката и ажурирањето на податоците

Релацискиот OLAP се раководи од концептот дека податоците треба да бидат сместени во релациски бази што ги карактеризира јасна и конзистентна структура, подобро решавање на меморискиот простор за ажурирање на податоците, поголема отвореност кон стандардните релациски трансакциски бази и достапност до апликациите со кои што се служат крајните корисници. Прикажувањето на податоците на крајните корисници треба да биде во повеќедимензионална форма. Врз основа на овој концепт подоцна се разви димензионалното моделирање. Тој е наменет за ракување со големи бази на податоци (над 100GB) со многу големи димензии (повеќе од 500.000 членови) и сложен избор на атрибути. Како и кај повеќедиманзионалниот концепт, и овде се бара посебна единица за аналитичка обработка на податоците, но исто така е потребно да постои и администратор на релациската база на податоци, што ќе води до интерактивната трослојна архитектура. Овој концепт го карактеризираат

особено лоши перформанси, неадекватна аналитичка функционалност, високи трошоци за имплементација и ретко кога е добар избор.

Милтидимензионалниот OLAP се заснова на идејата дека повеќедимензионалниот увид во податоците е поконзистентен, што има свој одраз на деловната логика, па заради тоа е порелевантен за практично донесување одлуки во однос на табеларно ориентираните погледи во стандардните реалациски системи и системите засновани на датотеки. Повеќедимензионалниот поглед на коцката и/или хиперкоцката ги сегментира податоците во ќелии што се сместени во дескриптивно релевантни интерсекции. Повеќедимензионалните администратори ги разместуваат категоријалните вредности во димензии, а потоа димензиите во мултидимензионална структура – хиперкоцка. Мултидимензионалната коцка во логичкиот модел има аналогна позиција на позицијата на табелата или полето во релацискиот модел. Обработката на повеќедимензионалните податоци бара специфицирање на висок степен логички операции меѓу компонентите на повеќедимензионалната структура.Овој пристап го карактеризира едноставна обработка на определени логички релации и операции како што се:

- дефинирање на односите родител – дете меѓу димензиите и конструкција на димензионални хиерархии, преку географски, организациски, временски и други важни организациски концепти
- лесно изведување матрични пресметки, што дозволува деловите од хиперкоцката да бидат обработени во многу кратко време
- рангирање на повеќедимензионалните ќелии, што овозможува пофокусиран опис, известување и анализа
- ротирањето на мултидиманзионалните коцки дава различни погледи на исти податоци, без дополнителни зафати во оригиналните детални податоци
- аграгирање и деагрегирање, односно приказ на сите степени на хиерархија.

Основните предности на MOLAP се добрите перформанси, функционалноста, пристапот читање – внесување, поддршката за истовремен пристап на голем број корисници и малите трошоци за имплементација. Но, сепак,

MOLAP покажува лоши карактеристики, ако треба да се обработат големи количини податоци, а исто така и интеграцијата со трансакциските бази на податоци и апликации е многу лоша.

Десктоп OLAP се сместува на корисничките компјутери и го карактеризираат одлични перформанси и функционалност, како и добро поврзување со OLTP системите, но на него е можно да се обработат само многу мали количини податоци.

Хибридниот OLAP е комбинација меѓу ROLAP и MOLAP. Тој користи релациски пристап кон деталните податоци и повеќедимензионалната база на податоци. Ова решение се смета за најдобра опција, бидејќи ги комбинира предностите на ROLAP- добро ракување со големи количини на податоци и MOLAP – повеќедимензионална обработка и погледи.

IV ДЕЛ
ПРИМЕНА НА DATA MINING ВО ДЕЛОВНИОТ СВЕТ

1.Области на примена на Data Mining

Брзиот прогрес на информациските технологии, заедно со способноста за складирање големи количини научни или бизнис податоци во складиштата на податоци, како и технологиите за управување со базите на податоци, во комбинација со компјутерските технологии, компјутерските методи и брзината на процесирањето ја отвараат можноста за креирање податочно ориентирани модели. Користењето на софистицираните компјутерски аналитички методи претставува основа за сите истражувачки активности и нивна комерцијализација од страна на водечките компании на IT пазарот. Веќе се развиени иновативни методологии што користат алатки од областите како што се: традиционалната статистика, операционите истражувања, регресијата и предикативните модели, вештачката интелегенција, но од сите нив посебно се издвојува полето на Data Mining. Прашањето што се поставува пред научниците или менаџерите не е како да се добијат повеќе информации или како да се дизајнира подобар информациски систем, туку што да се прави и како подобро да се искористат информациите кои што се добиваат со веќе постојните информациски системи. Предизвикот е: како да се искористат корисните информации, врз основа на историските податоци и да се разбере минатото, односно како да се предвиди и да се влијае на иднината, преку поефикасното донесување на одлуките. Истражувањата во областа на Data Mining и системите за поддршка на одлучувањето (DSS) претставуваат одговор на овој предизвик. Data Mining и DSS заедно претставуваат аналитички информациски технологии што обезбедуваат унифицирана платформа за оптимална комбинација од складирани податоци и човечки управувана анализа. Data Mining алатките се користат за развивање специфични апликации. Овие апликации можат да бидат комбинација од знаење за некоја област или бизнис домен и собраните податоци или алатките што се користат за подобрување на одлучувањето, како Data Mining и DSS. Во контекст на овие апликации, Data Mining може да се користи од развојот

на предиктивни и дескриптивни модели до заштита и подобрување на човековата околина.

1.1 Примена на Data Mining во маркетинг

Data mining се употребува во многу области каде што се работи со големи количини на податоци за од нив да се откријат некои неочекувани шаблони или модели, односно да се детектираат врски меѓу на прв поглед неповрзаните податоци. Можеби најпознатата употреба на Data Mining е во **маркетингот,** каде што тој се користи за откривање на шаблони за однесувањето на потрошувачите, дефинирање на профитабилни сегменти на потрошувачи, превенција од губење на лојалните потрошувачи, проценка на ризикот за кредитирање, пазарните истражувања, идентификување на погрешното однесување на пазарот итн. Притоа, целта на Data Mining не е само да ги идентификува таквите шаблони во купот податоци, туку да открие неочекувани врски и шаблони помеѓу податоците што натаму би биле основа за учење за однесувањето на потрошувачите[78].

Во овој дел ќе бидат споменати неколку аспекти на интеракција на Data Mining со маркетингот за подобрување на бизнис работењето на компаниите.

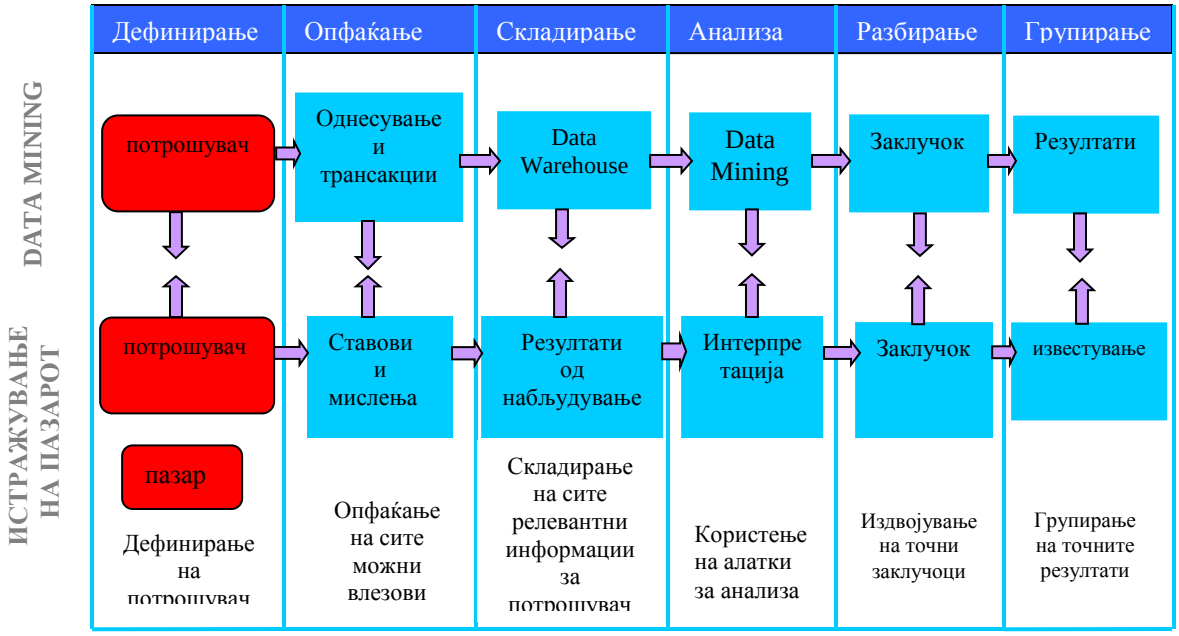
Првата интеракција на Data Mining е со **истражувањето на пазарот[79],** што може да доведе до синергетски ефекти на секој од овие процеси, при нивната меѓусебна комбинација. Во бизнис маркетингот денес пазарното истражување е фокусирано кон учење за потрошувачите, клиентите, конкуренцијата и пазарните трендови. Комбинацијата на овие две дисциплини обезбедува единствена можност за анализирање на разликата што се јавува помеѓу дефинираните планови и реалноста. Денес, истражувањето на пазарот и Data Mining се потпираат на одвоени внатрешни инфраструктури. Со приближувањето на овие две дисциплини се отвора можност за интеграција на технологиите што се уште не се доволно интегрирани. Таквите технологии подразбираат прибирање на податоци, управување со податоците, складирање на податоците, анализа на податоците и изготвување

[78]Ramakrishnan R., Freytag J.C, Report based on a workshop Data Mining:The Next Generation, Dagstuhl seminar, 2005

[79] Kenneth E., Scionti R., Page M., The Confluence of Data Mining and Market Research for Smarter CRM, Article, May 2006

извештаи. Интеграцијата меѓу овие две дисплини може да резултира во апликации од типот на управувањето со проекти или управувањето со знаење. Конвергенцијата на Data Mining и истражувањето на пазарот може да биде најдобро илустрирано преку испитување на секој одделен чекор од овие процеси. За таа цел ќе бидат генерализирани вообичаените чекори за секој од овие процеси, а тие вклучуваат[80]:

- јасно дефинирање на потрошувачот или групата потрошувачи,
- опфаќање на сите информации што треба да бидат собрани,
- складирање на информациите онаму каде што треба да бидат управувани и зачувани
- анализирање и испитување на информациите,
- разбирање на донесените заклучоци,
- групирање и разделување на заклучоците за потребите на организацијата



Слика 33. Data Mining и истражување на пазарот

[80] Kenneth E., Scionti R., Page M.,, The Confluence of Data Mining and Market Research for Smarter CRM, Article, May 2006

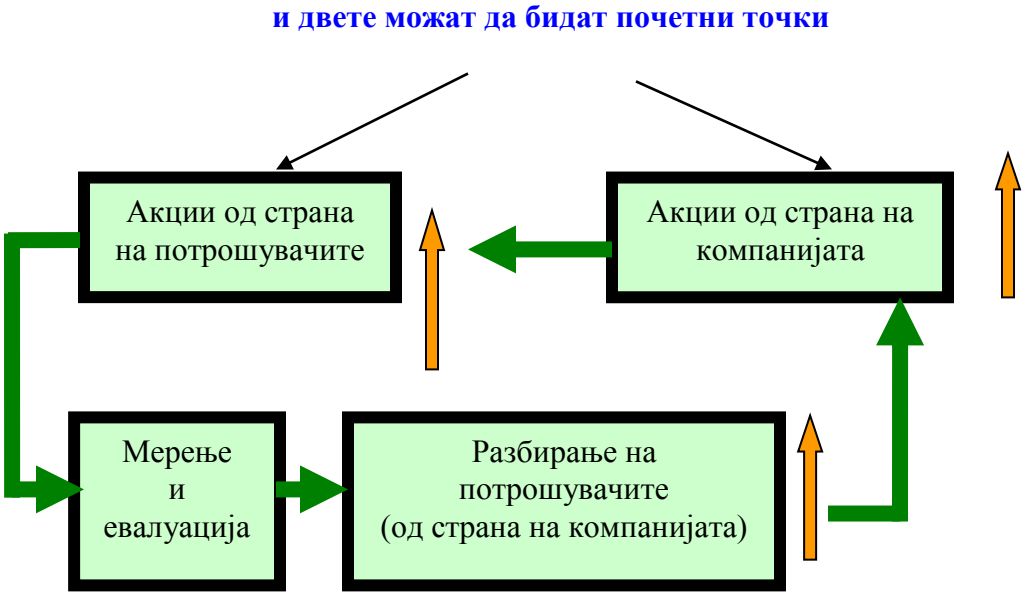
Data Mining и истражувањето на пазарот се уште не се обединети во испитувањето на околината на компаниите за да се постигнат подобри синергетски ефекти. Сепак, денес постојат примери на некои од водечките компании што ги имаат конвергирано овие две дисциплини за некои ad – hoc истражувања. Како една од најограничувачките бариери на конвергенцијата на овие две дисплини претставува **организациската структура, култура и самата инфраструктура** на компаниите. Во многу компании Data Mining и истражувањето на пазарот се употребуваат во различни структурни делови од бизнис работењето. Исто така, и физичката поделеност ја спречува нивната интеракција и кооперација. Организациската култура на компаниите придонесува за целосна поделеност на овие два процеса. Така, на извршно ниво, постои тенденција донесувањето на одлуките да биде врз основа на интерната аналитика на компанијата или врз основа на пазарните истражувања. Комфорот на оние кои донесуваат одлуки кон некој пристап и нивната неиновативност при применувањето нови и современи методи придонесуваат овие две дисцплини да функционираат сосема одделно. Конечно, внатрешната структура на компаниите придонесува за поделеноста меѓу овие две многу значајни дисциплини за успешно работење.

Со доближувањето на Data Mining и истражувањето на пазарот, односно со нивното надополнување ќе се овозможи интеграција на технологијата со која што компанијата располага, и тоа онаква каква што не била присутна. Сепак, можеби интеграцијата на овие две дисплини не е секогаш пожелна стратегија за определена компанија. Затоа само оние компании што можат да си дозволат значајни инвестиции во сферата на информациските технологии можат да очекуваат поволни резултати од ваквата интеграција.

Управувањето со односите со купувачите (Customer Relationship Management – CRM) претставува процес со кој што се менаџираат интеракциите помеѓу компаниите и нивните потрошувачи[81]. Тоа би значело дека разбирањето на однесувањето на потрошувачите од страна на компаниите е сржта на CRM.

[81]Ramakrishnan R., Freytag J.C, Report based on a workshop Data Mining:The Next Generation, Dagstuhl seminar, 2005

Ваквиот процес на разбирање на потрошувачите може да биде прикажан како што е на слика 34.

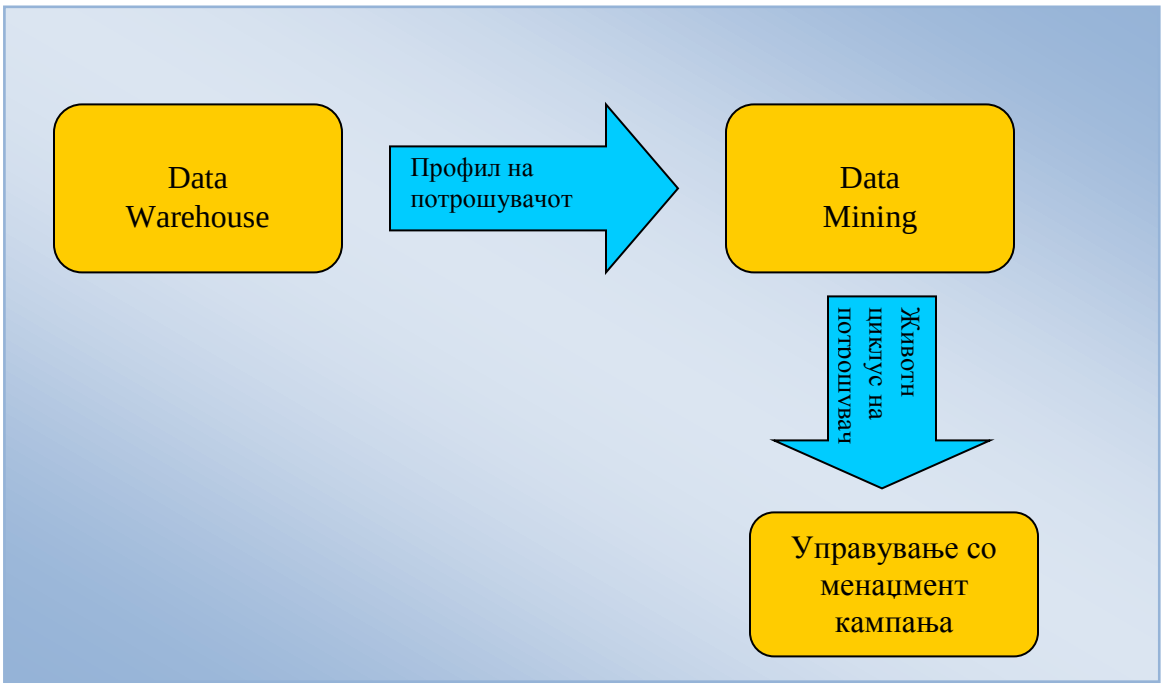


Слика 34. Основен CRM циклус[82]

На трансакциите што се однесуваат на потрошувачите се повеќе им се дава значење од страна на компаниите и на тој начин се поставува барањето за ефективно користење на ваквите податоци. Успешните компании мораат навреме да реагираат на барањата и преференциите на потрошувачите. Пазарот не е можно да чека од одговор на компанијата на промените во околината, па потрошувачите кои компанијата ги имала во еден момент, можат да се изгубат во следниот момент. Со цел да најде коректен начин за интеракција со своите потрошувачи, компанијата треба да го автоматизира процесот: вистинска понуда – на вистински потрошувач – во вистинско време – преку вистински канал[83]. Тоа би значело дека веќе не е можно да се чека на некаков сигнал од потрошувачот, па дури потоа да се презема акција, туку обратно.

[82]Ramakrishnan R., Freytag J.C, Report based on a workshop Data Mining:The Next Generation, Dagstuhl seminar, 2005
[83]Alex Berson, Stephen Smith, and Kurt Thearling-Building Data Mining Applications for CRM, NY, Workshop 2000

Data Mining дисиплината, што се занимава со откривање на нови непознати шаблони и врски меѓу големите количини податоци, претставува добра алтернатива за управувањето со односите со потрошувачите. Сепак, Data Mining претставува само еден дел од серијата чекори што можат да се преземат за одредување на врските помеѓу компаниите и нивните потрошувачи. Начинот на кој Data Mining влијае врз бизнис работењето на компаниите повеќе е условен од бизнис процесите, отколку од Data Mining процесите[84]. Главни корисници на ваквите CRM софтверски апликации се маркетинг менаџерите, кои настојуваат да го автоматизираат процесот на интеракција со своите потрошувачи. Таквата интеракција би можела да се прикаже како што е претставена на слика 35.



Слика 35. Data Mining во CRM

Притоа, првата задача е дефинирањето на пазарните сегменти, што бара значајна количина проспективни податоци за потрошувачите и нивните навики за

[84]Jim Davis, Ellen Joyner, Successful Customer Relationship Management, SAS Institut, Inc., 2000

купување. Од теоретски аспект, колку има поголем број податоци толку е подобро, но од практичен аспект, големата количина податоци може да доведе до потешкотии во проценката за тоа кои податоци имаат особена важност за овој процес. Натаму, маркетинг менаџерите користат Data Mining апликации со кои што се автоматизира пронаоѓањето шаблони од големата количина податоци што можат да бидат добри предикатори за идното однесување на потрошувачите. По Data Mining процесот, маркетинг менаџерите треба да ги вклучат добиените резултати во нивната менаџмент кампања за јасно да ги дефинираат маркетиг сегментите.

Употребата на Data Mining во управувањето со односите со потрошувачите може да помогне за детерминирање на однесувањето што е специфично за определен потрошувач или група потрошувачи или детерминирање на потрошувачите кои би можеле да го следат однесувањето од претходно откриените шаблони од постојните потрошувачи.

За многу компаниии примарно значење за нивниот раст има **привлекувањето нови потрошувачи**. Традиционалниот пристап за привлекување нови потрошувачи вклучуваше комбинација на **масовен маркетинг** (реклами во весници, билборди итн.) и **директен маркетинг** (телемаркетинг, mail итн.)[85]. Во овој случај, пак, преку ваквата маркетинг кампања се настојуваше да се влијае на новите потрошувачи. Во случајот на традиционалниот директен маркетинг привлекувањето нови потрошувачи беше далеку поедноставно, отколку со масовниот маркетинг. Локалните сервиси на компаниите располагаат со голема количина информации за идните потрошувачи, па тие можат демографски да се сегментираат (според возраст, пол, интерес за определени марки производи итн.). Маркетинг менаџерите демографски ги селектираат потрошувачите за кои се заинтересирани и, врз основа на податоците од продавачите, составуваат листа на потрошувачи кои би можеле да одговараат на дадените карактеристики. Иако маркетинг менаџерите имаат големо искуство при изборот на соодветен демографски критериум во однос на потрошувачите, сепак, големата количина

[85]Olivia Parr Rud, Data Mining Cookbook, Modeling Data for Marketing, Risk and Customer, Relationship Management, John Viley & Sons, New york, USA, 2001

податоци може да предизвика проблеми при одвивањето на овој процес. Комплексноста што е предизвикана од зголемувањето на податоците за секој потрошувач, како и зголемувањето на самиот број потрошувачи може да стане сериозен проблем за маркетарите при донесувањето одлуки за идните потрошувачи.

Data Mining може да помогне многу во овој процес, но самото купување и поседување на ваквите Data Mining алатки не значи и решавање на сите проблеми поврзани со привлекувањето нови потрошувачи. Затоа маркетинг менаџерите треба да ги комбинираат алатките за Data Mining на потенцијалните потрошувачи кои се генерирани со помош на Data Mining, со понудите за кои луѓето би биле заинтересирани. Донесувањето одлука за тоа која понуда би била интересна за определен сегмент потрошувачи, всушност, е умешноста во маркетингот.

Разликата меѓу Data Mining апликациите за привлекување нови потрошувачи и останатите Data Mining апликации во маркетингот е вклучувањето податоци што ќе послужат за градење предикативни модели. Туку, пред се, се мисли на големата количина податоци за потрошувачите кои немаат никакви посебни интеракции меѓу себе, за разлика од податоците на постојните потрошувачи кои се во некаква меѓусебна корелација. Првиот чекор при креирањето Data Mining модел за привлекување нови купувачи треба да води кон избор на тоа за какво однесување на потрошувачите компанијата е заинтересерина и нивото на грануларност. Нивото на кое што ќе функционира ваквиот предикативен модел, во голема мера, ќе влијае врз понудата што компанијата ќе је креира за новите потрошувачи. Целната променлива на Data Mining софтверот ќе биде предвидување на типовите реакција кај потенцијалните потрошувачи на маркетинг кампањата на компанијата на ниво што претходно било дефинирано. Моделот ќе може да се употребува на оние потрошувачи кои позитивно ќе одговорат на ваквата кампања на компанијата.

Data Mining може да најде примена и во анализата на **потрошувачката кошница.** Сопствениците на малопродажните центри постојано ги интересира структурата на нивните потрошувачи и производите што тие ги купуваат и користат. Методот на потрошувачка кошница, најчесто, се користи во

малопродажните центри, а основа за таквата анализа се податоците од сметките што се издадени на терминалите за наплата. Најчесто вака складираните податоци се наоѓаат во две релациски табели. Првата служи за регистрација на бројот на сметката, времето, датумот и часот, додека во другата релациска табела што е поврзана со релациски клуч, се содржани податоци за карактеристиките продадени производот и количината на производи. Во овој процес учествуваат уште низа на други табели, но за ваквата анализа ова се оние табели што се најпотребни. Со примената на Data Mining, при анализата на потрошувачката кошница можат да се откријат скриени шаблони за детектирање на законитостите за однесувањето на потрошувачите во малопродажните центри. Одлуките што можат да произлезат, како резултат од примената на овој метод можат да се засноваат на агресивната рекламна кампања за определен производ, насочена кон точно дефиниран сегмент потрошувачи. Врз основа на законитостите, откриени со овој метод, може да се направи распоред во продавницата на производите што се сметаат за "компаративни". Исто така, откако резултатите ќе се генерираат со помош на овој метод, може да влијаат на одобрувањето рабати, намалување на цените и останати бенефиции за потрошувачите. Според тоа, информациите, добиени со овој метод, менаџерите можат да ги искористат како показатели од стратегиска важност, бидејќи на ваков начин се предвидува идното однесување на потрошувачите. Битно е да се спомене дека овој метод нема примена само во малопродажните центри, туку и во големопродажбата, банките и некои други финансиски институции.

1.2 Примена на Data Mining во банкарството и финансиите

Data Mining е корисна опција секогаш кога постојат големи количини податоци од кои можат да се извлечат одредни корисни заклучоци. Постои правило дека нешто е вредно да се научи, ако откриеното знаење вреди повеќе, отколку што сме потрошиле на неговото откривање. Таквото знаење може да биде вредно само доколку се зголеми добивката со намалувањето на трошоците, а со тоа да се зголеми приходот итн. Така, и Data Mining може да се користи за управување со трошоците и приходите за зголемување на добивката. Постојат низа примери за

употребата на Data Mining во **банкарството и финансиите**[86]. Овде може да се спомене користењето на Data Mining во банките при креирањето листа на примероци за однесувањето на клиентите. Врз основа на информациите за сите трансакции на сметката на клиентот, всушност клиентите можат да се групираат во определени групи од таквите сегменти на клиенти на банката, таа ќе може да изработи понуда на банкарски услуги што се идентификувани како пожелни. Во тој контекст може да се креира систем за управување со банкарските сметки на корисниците. Со користењето на ваквиот систем би можела да се направи симулација како промената на политиката на банката влијае на идните преферниции и лојалноста на клиентите.

Можеби најинтересна е примената на Data Mining при креирањето **портфолио на хартии од вредност**[87] со различна ризичност. Некои од задачите на Data Mining за креирање такво портфолио на хартии од вредност и берзанско работење би можеле да бидат следниве:

- рано воочување на индикаторите што би можеле негативно да влијаат на вредноста на акциите,
- автоматизиран избор на хартии од вредност со посакувани карактеристики,
- прикажување и пребарување на хартиите од вредност на многу побрз и поточен начин,
- воочување на идните трендови на пазарот на хартии од вредност,
- откривање на ефектите и влијанието на еден пазар врз друг,
- заштита на тргувањето од невнимателни и незаконски дејствија.

Пример за Data Mining може да биде филтрирањето на различните хартии од вредност со помош на невронските мрежи. Моделот на невронска мрежа би можел да изврши проценка на најдобрите хартии од вредност што би значело да се избере помала количина највредни акции од големата количина на прв поглед еднакво

[86]Auroop R. Ganguly and Amar Gupta, Data Mining Technologies and Decision Support System for Business and Scientific Application, University of South Florida, 2004
[87]Ramakrishnan R., Freytag J.C, Report based on a workshop Data Mining:The Next Generation, Dagstuhl seminar, 2005

ризични акции. Притоа, како влезни променливи во ваквиот модел би можеле да бидат цената на акциите, големината на дивидендата, претходните приходи, состојбата во билансот итн. Пример за употреба на ваква Data Mining апликација е Recon системот што дава помош во изборот при купувањето акции. Врз основа на историските податоци, Recon врши предикција на процентот на враќање на средствата по три месеца. Во овој случај, системот препорачува кои акции да се купат, но не и во колкава количина. Примената на Data Mining во финансиите не претставува само следбеник на финансиските трендови, туку водечка апликација за мултидимензионални временски серии, како што се берзанските временски серии. A. Cowan, водечкиот економист на U.S Department of the Treusry, неведува дека аргументите што одат во прилог на ваквите апликации ќе овозможат големи вложувања, а со тоа стекнување предности на компаниите што користат апликации од овој тип на полето на финансиите.

Data Mining, исто така, наоѓа голема примена и во електронското банкарство, при користењето кредитни картички од страна на клиентите. Примената на Data Mining може да биде во привлекувањето нови корисници на кредитни картички, анализа на ризикиот од кредитирање, сегментирање на корисниците на кредитни картички во зависност од нивната платежна моќ итн.

Според тоа може да се забележи дека примената на Data Mining во иднина ќе има многу големо влијание врз банкарското работење и финансиите на финансиските институции. Развојот на Data Mining апликации за банкарското работење и финансиите може многу да придонесе за унапредување на работењето на сите финансиски институции.

1.3 Примена на Data Mining во производството

Производителите на производи и давателите на услуги се повеќе настојуваат да ги унифицираат и стандардизираат своите производи. Поради тоа, тие се повеќе се обидуваат да ги предвидат карактеристиките на своите производи со цел да дизајнираат производ кој ќе ги задоволи барањата на нивните потрошувачи. Примената на Data Mining означува градење предиктивен модел со кој што врз основа на историските податоци за производите што веќе компанијата ги

произведува и забележаното однесување на потрошувачите, можат да се добијат информации за оние производи што компанијата ќе им ги понуди на своите потрошувачи во иднина. Имено, во последно време успешноста на секоја компанија зависи од нејзината способност за иновација и продуцирање нови софистицирани производи. Исто така, на производителите им е потребно да имаат претстава колкав ќе биде бројот на потрошувачи за нивниот производ, па врз основа на тоа да го дефинираат обемот на производство од што во голема мера ќе зависи и самата цена на производот.

Еден од примерите за апликативна примена на Data Mining е Cassiopee системот што беше развиен како дел од заедничката соработка меѓу General Electric и SNECMA. Ваквиот систем беше применент во неколку авионски компании за да се дијагностицираат и предвидат проблемите што можат да се јават во врска со Boeing 737. За да се дијагностицираат и предвидат ваквите проблеми, овој систем го користеше методот на кластеризација. Успешноста на примената на овој систем, а со тоа и апликативната примена на Data Mining ја потврдува и првата награда во Европа што ја доби овој систем како најиновативна апликација.

1.4 Примена на Data Mining во откривањето грешки

Кај секоја компанија **грешките во работењето и неправилностите** можат до доведат до енормно зголемување на нејзините трошоци. Data Mining може многу да помогне во предвременото откривање на ваквите грешки и спречување за нивна идна појава. Така, со анализа на претходните трансакции, нивото на грешки во идното работење може да се сведе на миминимум. Банките, на пример, со анализа на претходните трансакции, преку откривањето на непознати шаблони од податоците, можат да го детерминираат идното неправилно работење. Имено, со користењето на FAIS[88] системот, владата на САД беше во можност да ги идентификува оние финансиски трансакции што упатуваа на можно девијантно однесување, а особено кога станува збор за перењето пари.

[88] Fayyad U., Piatetsky G. and Smith P., From data Mining to Knowledge Discovery in DataBases, AI magazine, 2000

1.5 Примена на Data Mining во телекомуникации

Телекомуникациските компании, најчесто, водат детална евиденција за телефонските разговори што ги вршат нивните клиенти. Со анализа на ваквите извештаи, односно со едноставно користење на шаблоните, откриени со помош на Data Mining, телекомуникациските компании можат да дефинираат сегменти на потрошувачи, и врз основа на таа сегментација, да развијат соодветен ценовен систем за своите потрошувачи, како и да ги планираат и насочат своите идни промотивни активности.

Пазарот на телекомуникациски услуги, денес, се повеќе созрева и ја признава важноста на проактивниот менаџмент на врската со корисниците, фокусирајќи се на постојаната грижа за корисниците, односно како да се задржат вредните корисници и како тие да станат попрофитабилни за компанијата. Data Mining може да се примени во предвидувањето на тоа кои корисници најверојатно ќе ја откажат услугата на компанијата во блиска иднина, што, најчесто во телекомуникацискатa индустрија се нарекува "**churn" (откажување на претплатата) предвидување**[89]. Откажувањето на претплатата, всушност, е најголемата грижа на телекомуникациските компании, пред се, поради карактеристиките на телекомуникацискиот пазар, како заситување проследено со динамични промени. Во тој контекст, како што се заситува телекомуникацискиот пазар, придобивањето на нов корисник станува се поскапо, отколку да се задржи постојната база корисници. Исто така, динамичните промени на пазарот, промените кај конкуренцијата, унапредувањето на технологиите и регулацијата можат да се создадат поголеми можности корисниците да ја напуштат постојната компанија и да преминат во друга конкурентска телекомуникациска компанија. Со користењето на Data Mining може да се изгради модел што ќе се потпира на историските податоци за корисникот и што ќе генерира листа на корисници за кои постои голема веројатност дека ќе ја напушатат компанијата . Ваквиот модел може да вклучува предвидување за веројатноста на откажување на претплатата за секој

[89] Seyyed Jamaleddin Pishvayi, Data Mining Technique for CRM, Tehran University, 2005

корисник и класификација на најверојатните корисници кои ќе ја откажат претплатата, наспроти оние кои нема да ја откажат.

Првичното дејствување на компанијата, за да го спречи корисникот да ја напушти, е да го разбере неговото однесување и да се обиде да ги најде шаблоните на однесување при откажување на претплататата и напуштањето на компанијата, за да може компанијата однапред да преземе соодветни чекори. Однесувањето на корисниците може да се разбере и од едноставните прашања или OLAP анализите, или пак едноставно од искуство. Меѓутоа, Data Mining треба да ја преземе улогата за идентификување на однесувањето на корисниците поврзано со откажувањето на претплатата, што е скриено во големите податочни сетови, како транскации од разговорите. Соодветен модел на Data Mining може да го идентификува шаблонот на однесување при откажувањето и да подготви пристојни понуди за корисниците за да го спречи нивното напуштање на компаниските услуги.

Според тоа, телекомуникациите се една од областите каде што Data Mining може да најде многу голема примена, а и самата примена е условена од природата на дејноста каде што се генерираат големи количини податоци. Па, кога станува збор за големите количини на податоци, тогаш употребата на Data Mining апликациите станува неминовност.

1.6 Останати области на примена на Data Mining

Data Mining наоѓа своја примена и во голем број други области. Некои прашања можат да послужат како корисна рамка за погодноста и применливоста на Data Mining во определени области, а тие се:

- која е крајната цел од примената на Data Mining;
- која е природата на процесите каде што би се користел Data Mining;
- колку податоците се интензивни во дадената област, колкава е нивната количина, структура и квалитет;
- колку чини прибирањето на податоците;
- дали оние кои се вклучени во Data Mining процесот го прифаќат Data Mining и дали располагаат со соодветни знаења;

- дали корисноста од примената на Data Mining апликацијата е поголема од трошоците што се направени за нејзина примена;
- кое е нивото на зрелост на Data Mining технологијата, релевантна за овој домен, итн.

Така, на пример, примената на Data Mining во **природните науки** вклучува користење на Data Mining во молекуларната биологија, биоинформатиката, биохемијата и медицинските истражувања. Со помош на Data Mining, најчесто, се анализираат податоците што се генерирани преку некои експерименти (како на пример експерименти во областа на микоробилогија, потоа во некои медицински истражувања, сегментирање на гени итн). Притоа, главната цел на Data Mining е откривање некаква корелација меѓу одделните генетски особини, медицинските третмани или реакцијата на пациентите. Особено употребата на Data Mining апликациите е голема во фармацевтската индустрија и биотехнологијата, најверојтно поради комплексноста на проблемите во овие области, но исто така и поради потенцијалните големи приходи на компаниите од оваа област со креирањето на иновативни производи.

Примената на Data Mining впечатливо е забележителна при проучувањето на молекулот на RNA[90]. Имено, групирањето на молекулите на RNA подразбира нивна внимателна селекција за споредување, преку откривање шаблони за нивното однесување. Така, на пример, потребно е групирање на молкули на RNA што делат вообичаена комбинација фактори на транскрипција. Ваквата анализа може да се направи со помош на алгоритми за кластерирање.

Во последно време се повеќе се вршат испитувања што се однесуваат на функцијата на гените кај човекот, познато како **генетски инженеринг**. Data Mining игра главна улога во овој процес[91]. За таа цел, најпрвин суровите податоци од микробиологијата треба да бидат трансформирани и нормализирани во податоци над кои што ќе може да се врши обработка и споредување. Така трансформираните податоци треба да бидат кластерирани во кластери на гени што меѓусебно ќе

[90] Ramakrishnan R., Freytag J.C, Report based on a workshop Data Mining:The Next Generation, Dagstuhl seminar, 2005

[91] Hakenberg J., Bickel S., Plake C., Brefeld U., Leser U., Systematic Feature Evaluation for Gene Name recognition, BMC Bioinformatic, 2005

корелираат. Многу популарен пристап за ваква кластерзација е автоматската кластер анализа (K-means). Откако еднаш кластерите ќе бидат дефинирани, функционалното запишување на гените во кластерите треба да биде декодирано во оригиналната форма. Ваквото декодирање на информациите го вршат генетските стручњаци. Потоа се врши дескрипција на билошките процеси, молекуларните функции и локацијата на ќелиите. Во идеален случај, гените од секој кластер треба да имаат исти карактеристики, т.е. да бидат исто функционално запишани. Како единствен проблем во оваа ситуација може да се јави бројот на податоци, т.е. особено кога станува збор за кластерирање на огромен број гени, но сепак тоа може да се надмине со намалување на бројот на опсервации и етапно решавање на проблемот.

Откривањето на **еволутивните разлики**[92] меѓу различните видови, исто така, е една од областите на примена на Data Mining. Напорите во оваа област доведоа до определени генетски пронајдоци кај различни видови (шимпанзо, глушец, кокошка, куче и неколку видови риби). Со помош на Data Mining беше дадена можност за компарација на гените на различните видови (како, на пример, помеѓу цицачи и риби), а со тоа дојде до откривање на некои скриени аспекти од еволуцијата на видовите низ илјадници години. Некои Data Mining методи, што се употребуваат во оваа област, вршат откривање на наједноставните шаблони што овозможуваат повеќекратен увид во кратките последователни секвенци што се нарекуваат **seed (зрно).** Голем број Data Mining апликации се засноваат токму на оваа идеја.

Data Mining своја примена наоѓа и на полето на **астрономијата, главно,** при анализа на сателитските снимки или анализа на снимките од телескопите. Научна анализа на визуелните податоци не е можна без нивна редукција, а редукцијата би можела да се врши преку класификација, кластеризација, каталогизација, делење на податоците итн. Во овој случај, врз основа на голем број сателитски снимки со огромна резолуција, се врши класификација на небеските тела (како, на пример, некои ѕвезди, галаксии, ѕвездани јата итн.). Таквата класификација се врши врз

[92] The GeneOntolgy Consortium, Genome research 200

основа на 40 атрибути за секое небеско тело, а атрибутите се задаваат од страна на астрономите, бидејќи голем дел од небеските тела што се наоѓаат на сликите се нејасни. Како пример за таква апликација може да се наведе системот **JARtool** [93](Jat Propulsion Labaratory, USA), чијашто задача е детекција на многубројни мали вулкани на Венера врз основа на сателитските снимки. Притоа, тренирањето на ваквиот систем се врши со помош на слики на кои што сигурно постојат такви мали вулкани. Друг таков пример е системот на NASA, **Quakefinder** што служи за детекција на тектонските активности на земјата врз основа на сателитски снимки[94]. Тој функционира врз основа на забележување на мали поместувања или промени на сликите, при што ја определува насоката и силата на поместување на тлото.

Во последно време **спортот** е една од областите каде што Data Mining наоѓа честа примена. Имено, врз основа на суровите податоци во складиштето на податоци (кој играч шутирал, процент на шутови, резултат од натпреварот итн) можат да се добијат шаблони од податоците за стратегијата на тимот за следните натпревари, како и да се изврши избор на оптималниот тим, во зависност од тактиката (напад, одбрана). Така, на пример, екипата на Boston Celtics го користи **IBM Advanced Scout** софтверот за Data Mining[95]. Овој софтвер функционира, користејќи ги податоците со кои што располага NBA (National Basketball Assotiation) за секој натпревар (како, на пример, податоци за шутирања, скокови, додавања итн.). Така, со помош на овој софтвер беа откриени шаблони што беа превидени од страна на тренерите на екипата. Имено беше откриено дека кога во игра беше центарот на екипата на New Yorks Knicks', Patrick Ewing процентот на шут на противничката екипа од рекетот бил доста низок. Па, врз основа на вакви и слични информации, тренерите на Boston можеа да ја постават стратегијата за игра против Knicks'.

Од претходната елаборација може да се заклучи дека Data Mining може да најде примена во најразлични области почнувајќи од природните науки, па се до забавните игри и спортски натпревари, и не само што можат да најдат примена

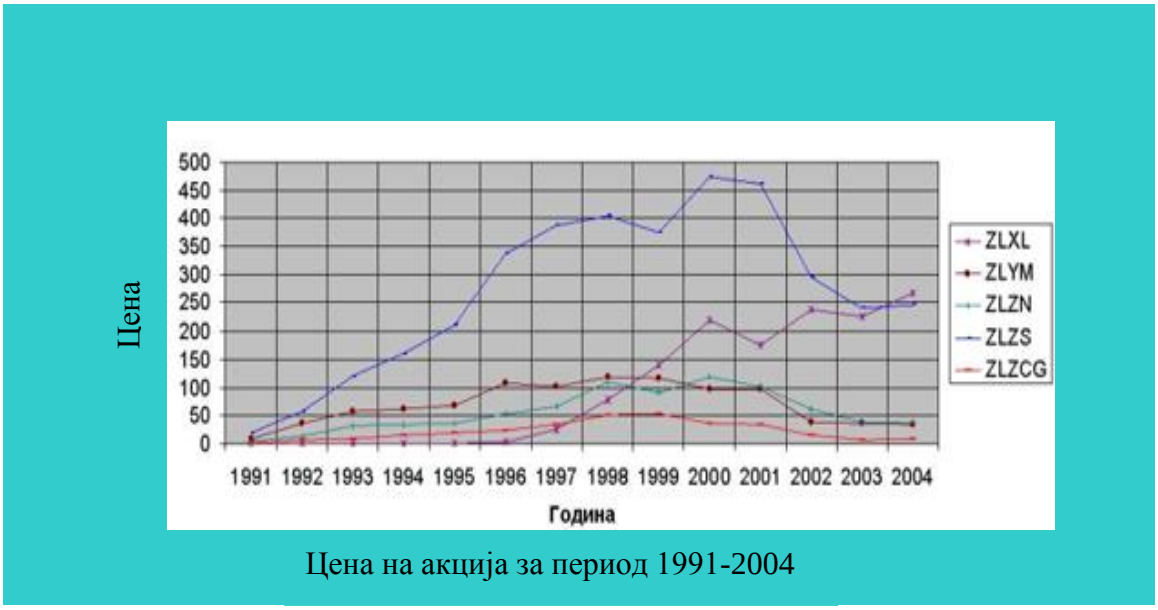
[93] McCuistion J.D., and Birk R., NASA Tachnical Report, 2002
[94] Potter C., Klooster S., Steinbach M., Tan P., Kumar V., Journal of Geophysical Research, 2003
[95] Bhandari I., Colet E., Parker J., pines Z., Advanced Scout:Data Mining and Knowledge Discovery in NBA Data, Data Mining and Knowledge Discovery 2002

ваквите Data Mining апликации, туку и ефектите од нивната примена би можеле да бидат долгорочни.

2.Специфични области на примена на Data Mining

2.1 Data Mining на временски серии

Со методот на временски серии се предвидуваат некои идни непознати вредности врз основа на временските променливи предикатори. Слични како и кај регресијата, и овде се користат познати резултати што се основа за предвидувањето. Моделот мора во пресметките да ги вклучи различните својства на појавата во текот на временскиот период, распоредени во некаква временска хиерархија (различно дефинирани временски периоди (како пет или седумдневна работна недела, месец, кваратал итн.). Според тоа, со помош на анализата на временските серии се испитуваат промените на вредноста на променливата во текот на некој определен временски период. Таков пример за временски серии е даден на следната слика, на која што е прикажано движењето на цената на акциите во периодот од 1991 до 2004 година. Врз основа на ваквата анализа би можело да се донесе одлука во кои акции е најдобро да се инвестира, ако се земат предвид историските податоци за вредноста на акциите во некој изминат временски период.



Слика 36. Временски серии

Најголем број од ваквите апликации користат комбинација на алгоритми што имаат специфичен начин на имплементација и што се прилагодени на потребите на корисниците. На пример, многу продавачи нудат верзија на CART или CHAID дрва на одлучување, со нивна зголемена способност за паралелна работа на повеќе компјутери.

2.2 Текст Data Mining

Текст Mining означува откривање на претходно непознати информации со автоматска екстракција на информациите од пишани извори. Клучен елемент во овој процес е поврзувањето на така издвоените информации за да можат заедно да сочинуваат некоја форма на нови факти или хипотези што понатаму ќе бидат испитувани со помош на некои поконвенционални методи[96]. Текст Mining е нешто сосема различно од обичните операции на пребарување, како, на пример, Web пребарувањето. При пребарувањето корисникот вообичаено бара нешто што веќе му е познато, а е напишано од некој друг. Во Текст Mining целта е да се откријат досега непознати информации, нешто што се уште никој не го знае или нешто што не е напишано од некого. Текст Mining е една од варијациите на класичен Data

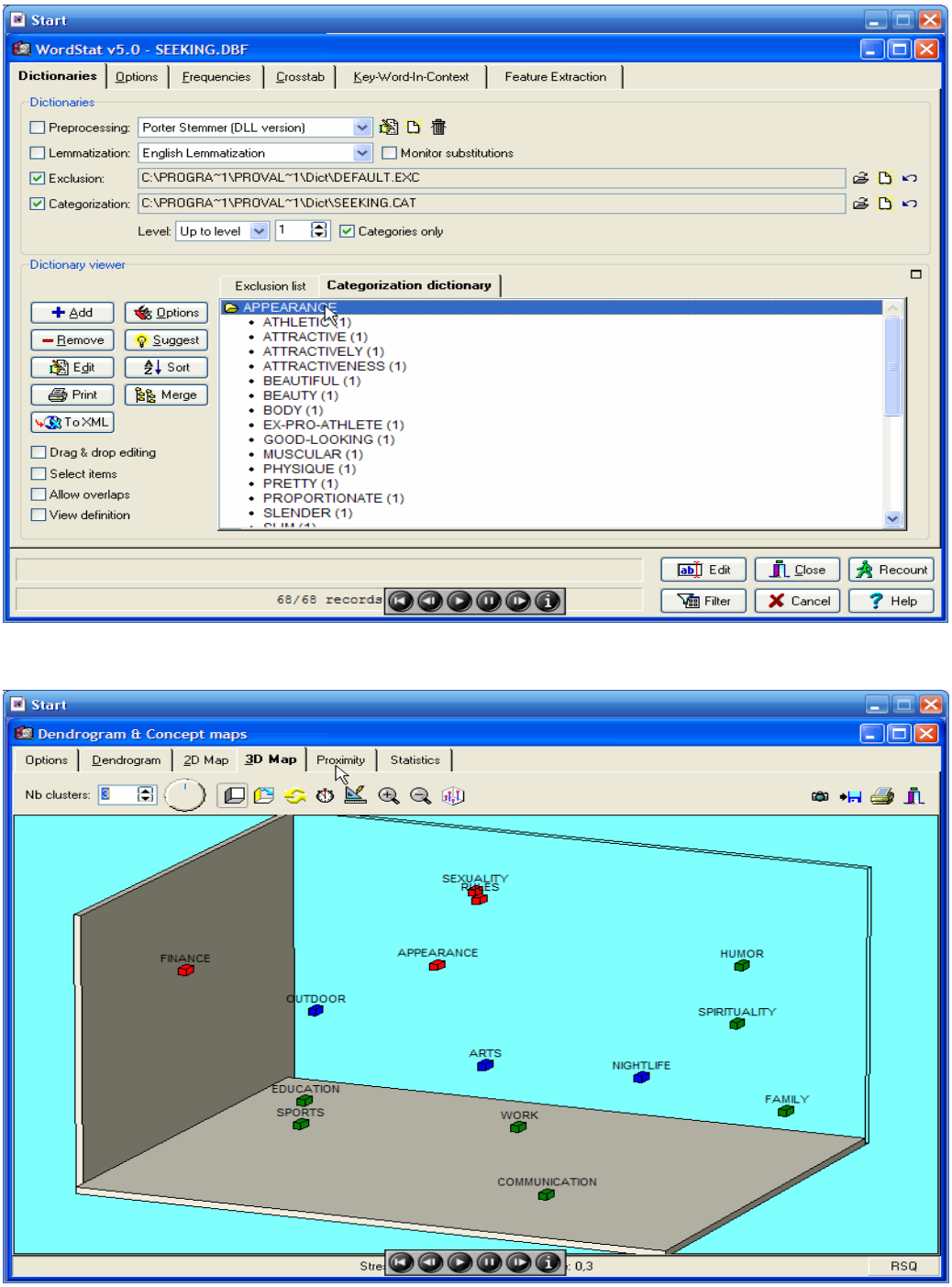
[96] Marti Hearst, "What Is Text Mining", Article In New York Times, October 17, 2003

Mining, што настојува да открие непознати шаблони од огромните бази податоци.

Разликата меѓу класичниот Data Mining и Текст Mining е во тоа што во Текст Mining шаблоните на податоци, што се откриени, повеќе се врз основа на текстот напишан на природен јазик, отколку од структуираните факти од базите на податоци. Базите на податоци се дизајнирани процесот на издвојување на податоци од нив да се врши автоматски, со помош на соодветни алатки. Се уште не постојат алатки што можат да читаат текст на ист начин како човеот и на ист начин да го интерпретираат прочитаниот текст. За да се овозможи тоа е потребно најпрвин да се симулира работата на човечкиот мозок што ќе биде основа да се напише програма која што би го читала текстот на ист начин како и луѓето. Ова поле на истражување се нарекува **компјутерска лингвистика**[97]**,** каде што е направен голем прогрес во откривањето мали шаблони на податоци при текст анализата. Постојат програми што со определена точност можат да издвојат информации од некој текст со однапред дефинирана структура. На пример, постои програма за читање на CV-то на кандидатите што издвојува само одделни делови (како името и презимето, адресата, работните вештини) и сето тоа го прави со извесна точност. Сепак, ова не може да се нарече класичен Текст Mining, бидејќи претставува едноставно издвојување на информациите од некој пишан текст. Текст Mining, за разлика од овој пристап, треба да открива некакво ново знаење што, всушност, е и самата цел на Data Mining.

На слика 37 е претставен интерфејсот на еден таков Текст Mining едитор, WordStat v5.0, што дава можност за повеќе начини на прикажување на податоците (табеларно, графичко тродимензионално итн)[98].

[97] www.sims.berkeley.edu
[98] www.provalisresearch.com



Слика 37. Текст Mining алатката WordStat v5.0

Сепак, фундаменталните ограничувања на Текст Mining се тоа што не може да се напише програма која што во целост ќе интерпретира текст во доволно долго време и тоа што голем дел од информациите што се потребни не се во соодветниот

текст формат. Доколку се напише програма што би ги детектирала новите зборови кои што се јавуваат, тогаш постои голема веројатност дека ќе се пропушти согледувањето на некои клучни релации меѓу постојните термини.

2.3 Web Data Mining

Техниките за откривање шаблони во **Web Mining** се потребни, бидејќи корисниците се соочуваат со проблеми поврзани со изнаоѓањето соодветни информации, неможност за креирање ново знаење врз основа на расположливите информации на Web, различната реакција на содржината и презентацијата на податоците на Web и откривањето на потребите на купувачите или корисните.

На Web страниците се создаваат огромни количини информации, при што голем дел од нив можат да бидат многу корисни. Меѓутоа, вистинската моќ и значење на овие информации доаѓа до израз дури тогаш кога тие информации ќе се поврзат со други бази на податоци, како што се автоматизираните системи за купувачите, системите за известување и истражувачките системи. Во оваа ситуација главниот проблем е во поврзувањето на податоците, бидејќи обично постојат голем број различни системи, односно различни бази на податоци. Токму во оваа област техниките за Web Mining можат да дадат голем придонес. Web Mining системот може да ги интегрира сите расположливи извори на податоци, без разлика на видот на базата на податоци во која што се наоѓаат, бидејќи поседува механизам за пречистување и филтрирање на податоците што потоа се користат за неопходните анализи.

Web Mining открива значајни деловни трендови и врски преку интеграција на Internet и Intranet деловните информации и останатите податоци. Исто така, техниките за Web Mining можат да бидат користени при решавање проблеми поврзани со преголемата количина информации.

Web Mining претставува користење на Data Mining техниките за автоматско откривање и екстракција на корисни информации од Web документите. Процесот на Web Mining може да се декомпонира на следниве подзадачи:

- **пронаоѓање ресурси**. Тоа подразбира пронаоѓање на саканите Web документи,

- **селекција на информации и предпроцесирање**, што опфаќа автоматска селекција и претпроцесирање на информации од саканите Web документи,
- **генерализација, којашто** опфаќа автоматско откривање на заеднички шаблони на различни Web страници,
- **анализа**, што опфаќа валидација и/или интерпретација на откриените шаблони.

Web Mining, исто така, може да се дефинира и како откривање и анализа на корисни информации на WWW. Имено, тој се дели на три истражувачки области, во зависност од тоа кој дел од Web-от се истражува, а тие области се [99]:

- **откривање содржини (Web Content Mining),** што претставува откривање корисни информации од Web содржините, податоците и документите. Истражувањата, направени во оваа област, можат да се поделат во два различни пристапа: **пристап заснован на агенти и пристап заснован на бази на податоци**. Пристапот заснован на агенти им помага на корсниците во пронаоѓањето релевантни информации и нивно филтрирање врз основа на профилот на корисникот. Пристапот заснован на бази на податоци формира модел на податоци на Web страницата и го интегрира со многу пософистицирани прашалници, отколку што се клучните зборови врз основа на коишто пребарувачите го вршат пребарувањето;
- **откривање структури на врски на Web (Web Structure Mining)**. Овде се настојува да се открие фундаменталниот модел на **структурата на линкови** на Web. Моделот мора да биде заснован на топологијата на хиперлинкови, со или без опис на линковите. Овој модел може да биде користен за категоризација на Web страниците, а исто така и за генерализација на информации, како што се сличност и врски помеѓу различните Web страници. Web Structure Mining може да биде користен и за откривање авторизирани сајтови;

[99] Alampije Veljovic, Nenad Dulanovic, Otkrivanje obrazaca u koriscenju Web-a, IDEASoft, Beograd, 2005

- **откривање шаблони при користењето на Web (Web Usage Mining),** што се обидува да им даде смисла на податоците, генерирани во Web корисничките сесии или податоци за однесувањето на корисникот. Додека Web Content Mining и Web Structure Mining користат реални и примарни податоци на Web, Web Usage Mining открива секундарни податоци, изведени од интеракцијата меѓу корисниците во текот на нивната работа на Web. Оваа област користи податоци од пристапните логови на Web серверите, логови на Proxy серверите, кориснички профили, регистрациски податоци, кориснички сесии или трансакции и кои било други податоци што настануваат како резултат на интеракцијата меѓу корисниците.

Овде треба да се нагласи дека меѓу наведените категории не постои граница што прецизно ги раздвојува. Нивното разгледување е можно само од аспект на опфатот. Локалниот дијапазон опфаќа индивидуални Web сајтови, додека глобалниот се однесува на целиот Web простор. Опфатот на Web Content Mining и Web Structure Mining во пристапот заснован на агенти е глобален, додека опфатот на Web Usage Mining и Web Content Mining во пристапот заснован на бази на податоци е локален[100].

Во трагањето по знаење низ огромната количина податоци, најновите алатки за откривање корисни шаблони употребуваат софистицирани техники од интелигентните информациски системи, Data Mining, психологијата и теоријата на информации. На пример, системот **WEBMINER**[101] воведува воопштена архитектура за Web Usage Mining. WEBMINER автоматски открива асоцијативни правила и секвенцијални шаблони од пристапните логови на серверот. Тој може да биде користен за различни видови анализи, како што е, на пример, анализата на најпосетените Web локации. Но, дизајниран е и нов алат за Web Data Mining, а тоа **WebLogMiner,** што користи комбинација на Data Mining техниките и OLAP. Примената на Data Mining техниките на лог фајловите овозможува откривање на

[100] Alampije Veljovic, Nenad Dulanovic, Otkrivanje obrazaca u koriscenju Web-a, IDEASoft, Beograd, 2005
[101] www.cit.org.yu

интересни шаблони во лог фајловите што можат да се користат за преструктуирање на сајтовите, нивно ефикасно дизајнирање, определување на локација за реклами и зголемување на заинтересираноста кај специфичните купувачи за купување на одделни видови производи.

Според тоа, со помош на специјализираните алатки за Web Data Mining што користат веќе постојните методи за добивање на потребните податоци од големите складишта на податоци, се добиваат показатели што имаат значење за некоја конкретна ситуација. Со примената на ваквите алатки се доаѓа до вистински, точни и егзактни информации за начинот на користење на Web сајтовите. Таквите анализи можат да се вршат со помош на денешните Web Mining системи. Резултатите од овие анализи се користат за унапредување на електронското работење, креирање маркетинг стратегии и зголемување на заинтересираноста на посетителите за Web страницата на компанијата што применува ваков вид истражување. Токму во ова е најголемиот придонес на Web Mining за успешното работење на Web страниците.

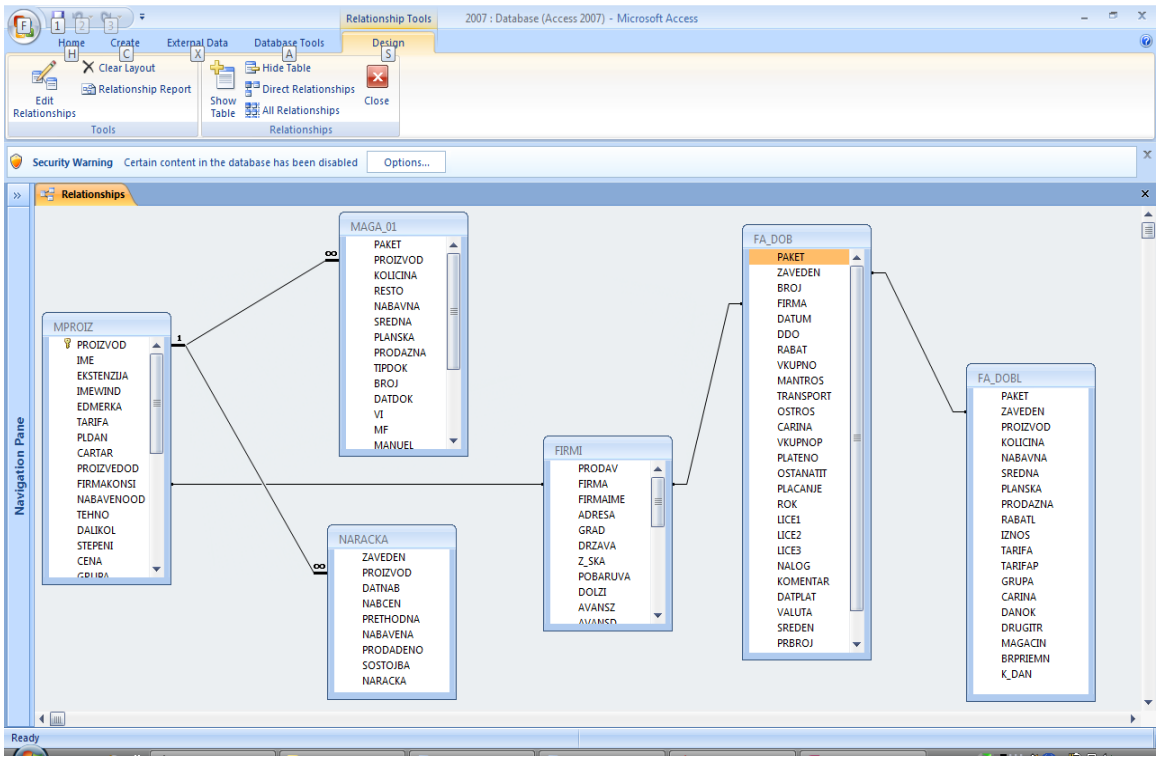
V дел

АСПЕКТИ НА ПРИМЕНА НА МЕТОДИТЕ ЗА DATA MINING ВО ДЕЛОВНОТО РАБОТЕЊЕ НА КОМПАНИИТЕ

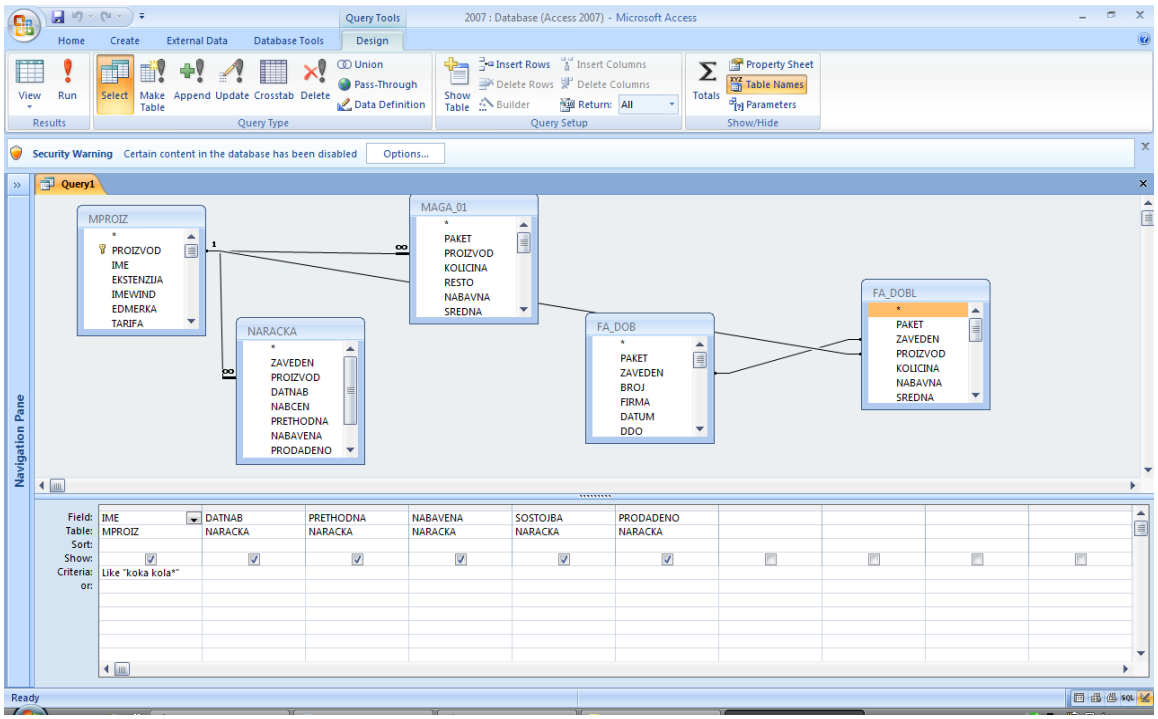
Постојат бројни начини на примена на методите за Data Mining. Притоа, со примената на секој од овие методи може да се изгради модел којшто може да има како позитивни, така и негативни карактеристики. Во секој случај, важно е таквиот модел да ги поседува следниве позитивни карактеристики: точност, способност за објаснување, брзина на одговор, преносливост, компактност, флексибилност, можност за комбинирање со други методи, толеранција во однос на комплексноста, независност (од експерти), брз развој, лесно користење, лесна употреба на логичките релации, како и едноставно користење на нумеричките податоци. Во овој дел од трудот ќе биде прикажано како, со помош на некој од методите за Data Mining може да се реши проблемот за предвидување на продажбата на едно трговско друштво, врз основа на историските податоци. Тоа би претставувало пример за предвидување на трендот на продажба.

Бидејќи во овој дел намерата е да биде прикажана примената и значењето на некој од методите за Data Mining натаму ќе биде прикажан еден едноставен и разбирлив пример за изградба на модел за Data Mining за предвидување на трендот на продажба.

Уште во почетокот е потребно да се обезбеди извор на податоци, што во нашиот случај би го претставувало складиштето на податоци (Data Warehouse) на деловниот субјект Алијанса. Изгледот на ваквата релациска база на податоци, во поедноставен облик, каде што се прикажани само оние табели што се од особена важност за решавање на дадениот проблем, е прикажан на следната слика.



Податоците што се сместени во складиштето на податоци најпрвин треба да се подготват за да може да се примени соодветниот метод за нивна обработка. Од складиштето на податоци иницијално треба да се извлечат само оние податоци што се потребни за да се изврши предвидувањето на продажбата. Потребните податоци се издвојуваат со помош на прашалник (Query), при што по нивното издвојување од базата е потребно да бидат конвертирани во Excel формат. Прашалниците се многу погодна алатка за пребарување на огромните складишта на податоци, особено кога станува збор за релациска база на податоци. Ваквата релациска база на податоци е изградена во Access, така што користењето на прашалникот е многу лесно при издвојувањето на потребните податоци. Формата на прашалникот, во овој случај, со кој што се издвојуваат податоците за производот koka kola(како што се датум на набавка, претходна состојба во магацинот, набавена количина на производот, и состојба во магацинот по извршената продажба) е прикажана на следната слика.



По подготовката и издвојувањето на податоците може да се започне со креирање модел за Data Mining со помош на методите, односно алгоритмите за Data Mining.

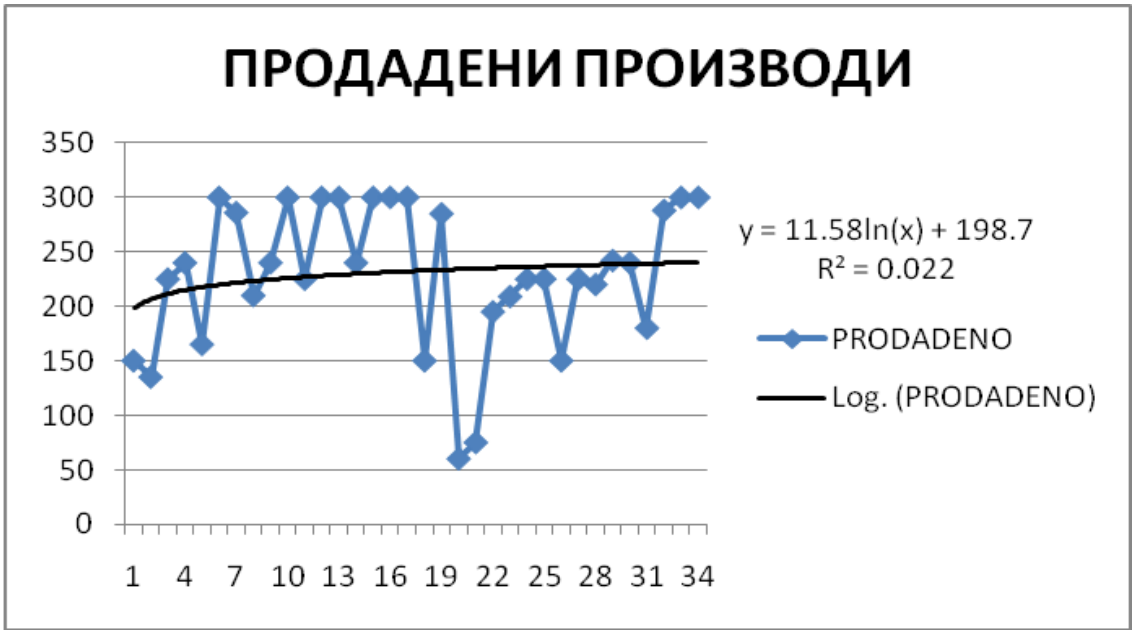
Во примерот за изградба на модел за Data Mining за предвидување на продажбата ќе биде користена алатката NeuralWorks Predict на компанијата NEURALWARE, а таа алатка работи во Excel окружување. Имено, станува збор за многу современа алатка што е произведена во 2005 година и е користена од страна на голем број познати светски реномирани компании. Со помош на ваквите невронски мрежи, на едноставен начин, може да се изгради мрежа, односно модел, при што ќе можат да се вршат промени во моделот во кој било степен од неговиот развој. Ваквата алатка користи и волшебници (Wizard) за изградба на потполно функционална невронска мрежа. Исто така, ваквата алатка има исклучителни визуелни карактеристики, па во текот на развојот на моделот можат да се следат неговите битни карактеристики, и тоа преку различни графикони и податоци што динамички можат да се менуваат на работната површина.

Невронската мрежа се смета како некој вид црна кутија, бидејќи не можат да се видат процесите внатре во моделот. Ваквиот недостаток на оваа алатка е решен, во извесна мера, со тоа што е можен прекин во развојот на моделот, промена на параметрите и повторно продолжување со развојниот тренд на моделот. Оваа невронска мрежа функционира врз основа на backpropagation техниката, што значи дека со паровите влезови и излези мрежата го пресметува сигналот на грешка кој што потоа, се враќа назад низ мрежата, како би се промениле тежинските пондери во внатрешните слоеви на мрежата.

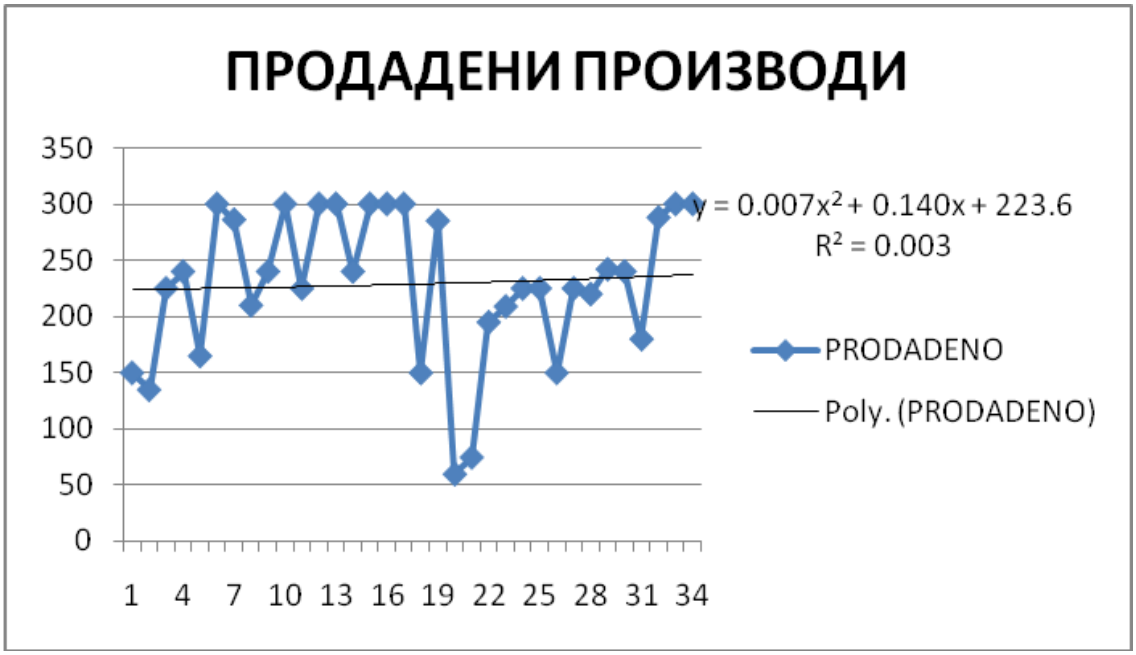
Невронските мрежи, најчесто, се употребуваат во случај кога примената на статистичките методи нема ефект или кога појавата не може да се претстави со некој од познатите модели или трендови, односно кога однесувањето на појавата често е неизвесно или ако постојат непотполни податоци. Исто така, за примената на невронските мрежи е потребно променливите да можат да се изразат квантитативно. Доколку продажбата, во овој случај, се претстави на графикон, тогаш може да се види дека тешко е да и се придружи некој од теоретските трендови на развој на појавата, па поради тоа тешко е да се откријат законитостите според коишто се одвива продажбата. Придружувањето не некој од најчесто употребуваните трендови на продажбата на компанијата Алијанса е прикажано на следните неколку графикони.

Графикон 1. Линеарен тренд

Графикон 2. Експоненцијален тренд



Графикон 3. Логаритамски тренд



Графикон 4. Квадратен тренд

Од прикажаните графикони може да се забележи дека никој од употребените трендови не може добро да ја презентира продажбата во минатиот период. Тоа најдобро го покажува коефициентот R, односно колку неговата

вредност е поблиску до 1, толку набљудуваната појава одговара на теоретскиот тренд што и се придружува на појавата, а колку е поблиску до 0 толку набљудуваната појава не одговара на теоретскиот тренд што и се придружува на истата појава. Во овој случај вредноста на R за кој било од трендовите е приближно еднаква на 0, што значи на вака набљудуваната продажба воопшто не и одговара ниту еден од теоретските трендови. Ваквиот проблем може многу лесно да се реши со примена на невронските мрежи.

Самото користење на невронската мрежа се одвива во неколку чекори. Имено најпрвин се врши тренирање на податоците. Тренирањето на мрежата е процес преку којшто се настојува да се определи оптималната вредност на параметрите на мрежата, односно пондерите. Пондерите се менуваат врз основа на два начина на учење на мрежата:надгледувано (supervised learninig) и ненадгледувано (unsupervised learning). Кај надгледуваното учење на невронската мрежа континуирано се задаваат парови на вредности од влезните променливи и целните излезни променливи, при што се настојува да се изнајдат најоптималните параметри на мрежата. Кај ненадгледуваното учење, пак, на мрежата и се задаваат само вредностите на влезните променливи, при што таа сама врши класификација на влезовите. Бидејќи со овој пример е дадено предвидување врз основа на познати вредности на влезовите и излезите, тогаш во овој случај ќе стане збор за мрежа заснована на принципот на учење на мрежата под надзор. Мрежата учи на вредноста на влезните променливи и грешката-разлика меѓу вистинските вредности на целната излезна променлива и предвидените вредности. Со ваквиот процес на учење се настојува да се постигнат подобри перформанси на мрежата со намалување на грешката до најмала можна мерка според дадените услови, односно претпоставки. Правилото за учење на мрежата е начин на којшто се менуваат пондерите или тежинските врски меѓу невроните за да се минимизира грешката.

Бидејќи оваа невронска мрежа функционира во Excel окружување потребно е податоците од релациската база на податоци, што беа издвоени со помош на прашалник, да бидат сместени во Excel табелата како што е прикажано на сликата. Предвидувањето ќе биде извршено за производот koka kola, и тоа ќе се врши врз основа на податоците за изминатите две години, односно за 2006 и 2007 год. Со

помош на прашалник се издвоени податоците врз основа на коишто ќе се предвидува продажбата за идниот период. Тие податоци се сместени во следната табела.

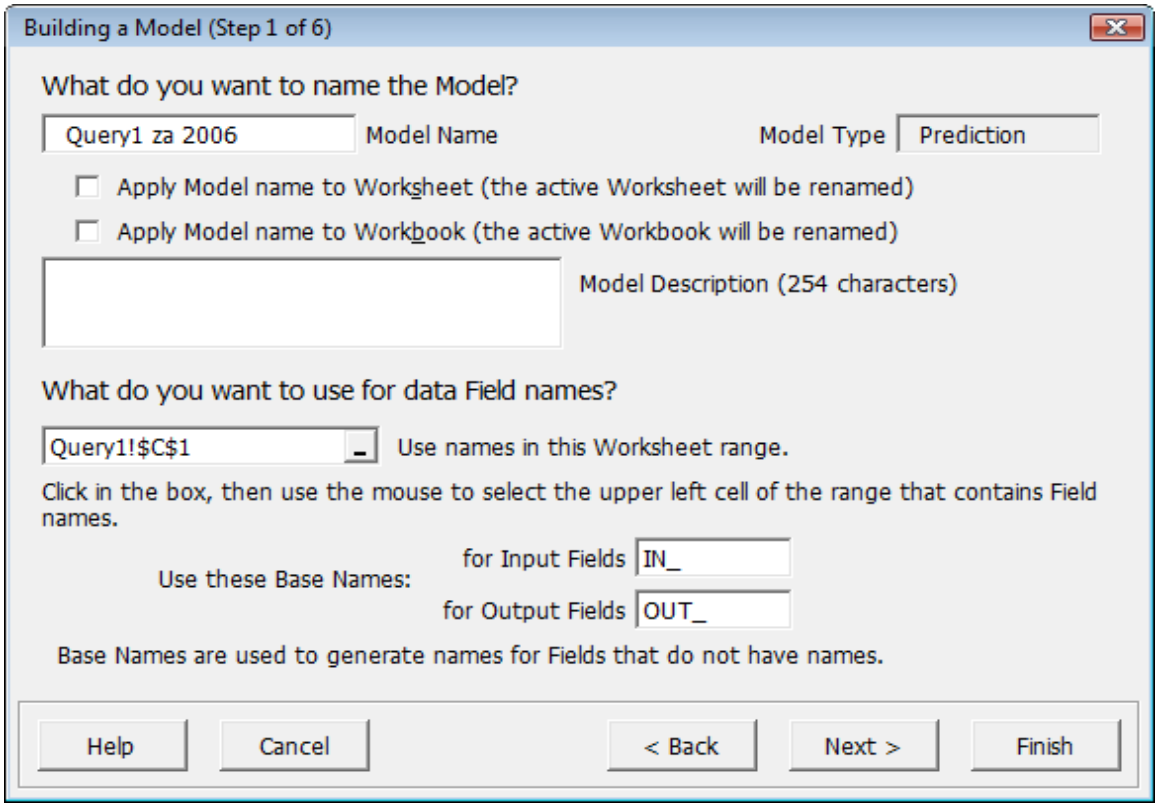
IME NA PROIZVODOT	DATNAB	PRETHODNA	NABAVENA	SOSTOJBA	PRODADENO
KOKA KOLA 0.5l MARVAL 15/1	01-Jan-06	0	150	0	150
KOKA KOLA 0.5l MARVAL 15/1	25-Mar-06	0	135	0	135
KOKA KOLA 0.5l MARVAL 15/1	31-Mar-06	0	225	0	225
KOKA KOLA 0.5l MARVAL 15/1	10-Apr-06	0	300	60	240
KOKA KOLA 0.5l MARVAL 15/1	20-Apr-06	15	150	0	165
KOKA KOLA 0.5l MARVAL 15/1	27-Apr-06	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	12-May-06	0	300	14	286
KOKA KOLA 0.5l MARVAL 15/1	01-Jun-06	0	225	15	210
KOKA KOLA 0.5l MARVAL 15/1	14-Jun-06	15	225	0	240
KOKA KOLA 0.5l MARVAL 15/1	30-Jun-06	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	18-Jul-06	0	225	0	225
KOKA KOLA 0.5l MARVAL 15/1	27-Jul-06	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	19-Aug-06	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	09-Sep-06	0	300	60	240
KOKA KOLA 0.5l MARVAL 15/1	15-Sep-06	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	15-Sep-06	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	19-Oct-06	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	18-Nov-06	0	150	0	150
KOKA KOLA 0.5l MARVAL 15/1	01-Jan-07	0	285	0	285
KOKA KOLA 0.5l MARVAL 15/1	01-Feb-07	0	149	89	60
KOKA KOLA 0.5l MARVAL 15/1	05-Feb-07	59	150	134	75
KOKA KOLA 0.5l MARVAL 15/1	17-Feb-07	119	150	74	195
KOKA KOLA 0.5l MARVAL 15/1	27-Feb-07	59	150	0	209
KOKA KOLA 0.5l MARVAL 15/1	15-Mar-07	0	225	0	225
KOKA KOLA 0.5l MARVAL 15/1	11-Apr-07	0	225	0	225

KOKA KOLA 0.5l MARVAL 15/1	23-Apr-07	0	210	60	150
KOKA KOLA 0.5l MARVAL 15/1	12-May-07	75	150	0	225
KOKA KOLA 0.5l MARVAL 15/1	19-May-07	0	300	80	220
KOKA KOLA 0.5l MARVAL 15/1	12-Jun-07	80	225	63	242
KOKA KOLA 0.5l MARVAL 15/1	30-Jun-07	18	225	3	240
KOKA KOLA 0.5l MARVAL 15/1	23-Jul-07	3	240	63	180
KOKA KOLA 0.5l MARVAL 15/1	08-Aug-07	63	225	0	288
KOKA KOLA 0.5l MARVAL 15/1	29-Aug-07	0	300	0	300
KOKA KOLA 0.5l MARVAL 15/1	11-Sep-07	0	300	0	300

Изградбата на моделот и тренирањето на претходно претставените податоци невронската мрежа го врши низ шест чекори. Низ тие чекори се врши дефинирање на паровите на влезови, дефинирање на излезните целни променливи и самото тренирање на податоците.

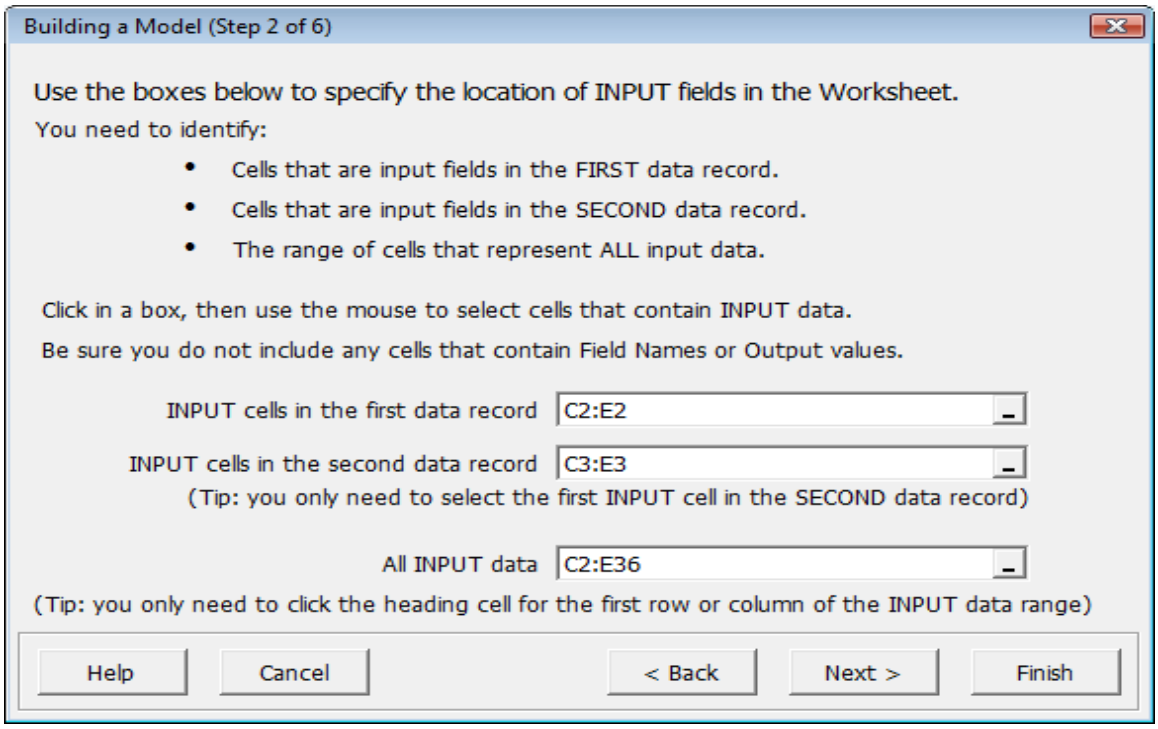
Чекор 1

Во првиот чекор се внесуваат основните карактеристики на моделот, како што се: описот на проблемот што се набљудува, името и типот на моделот и се дефинираат на паровите на влезови и излези.



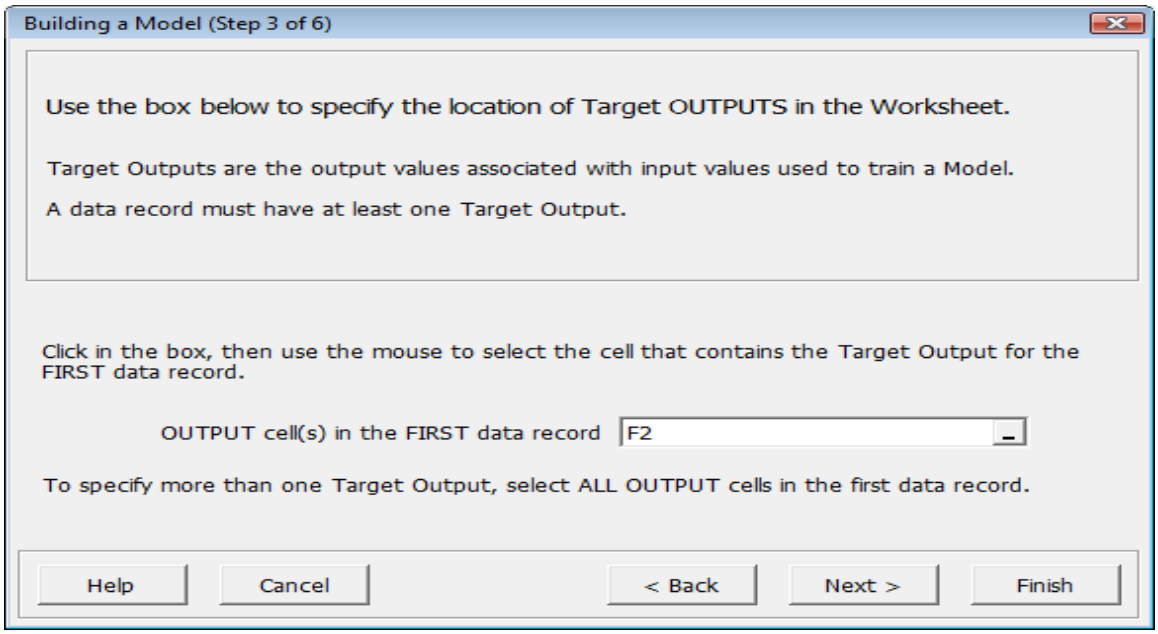
Чекор 2

Вториот чекор веќе претставува задавање вредности на влезните променливи. Најпрвин се задаваат вредности за паровите влезови на првиот слој од невронската мрежа, а потоа и за наредниот слој и, на крајот, се дефинира вкупниот број парови на влезни променливи. Таквиот начин на внесување на податоците е потребен затоа што мрежата функционира врз основа на backpropagation техниката, која што подразбира враќање назад низ мрежата за да се прилагодат тежинските фактори. Прилагодувањето на тежинските фактори се врши за да се овозможи што е можно поточно предвидување.



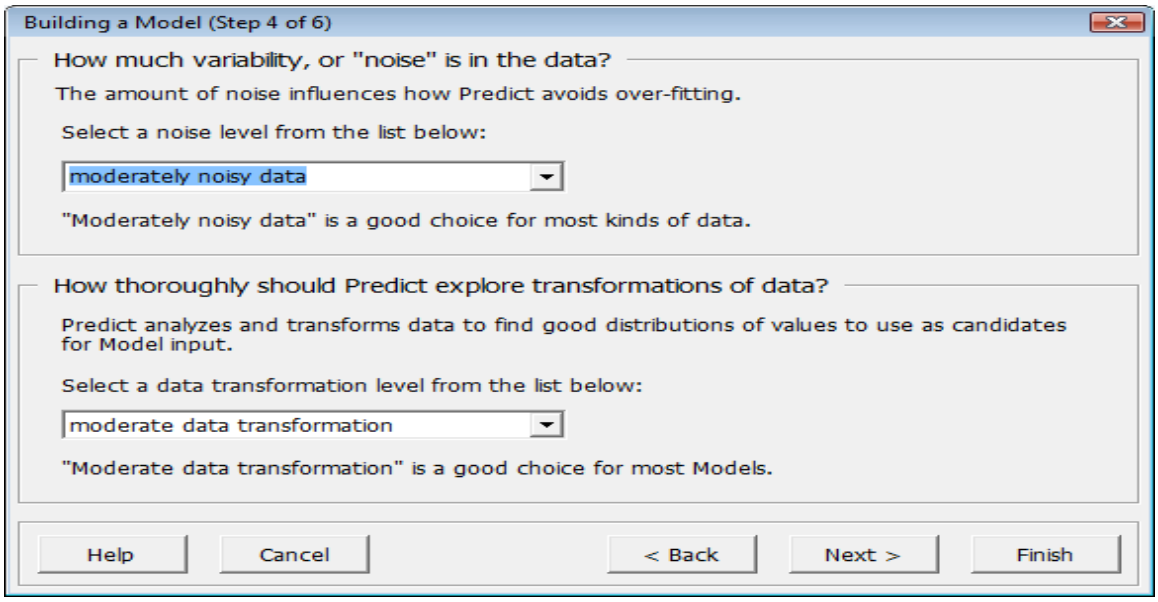
Чекор 3

Третиот чекор овозможува внесување на паровите излезни променливи што се потребни за да се изврши споредување на вистинските вредности на излезните променливи и предвидените вредности на излезните променливи. Преку споредбата на тие вредности се утврдува отстапувањето меѓу вистинските и предвидените вредности на целната променлива.

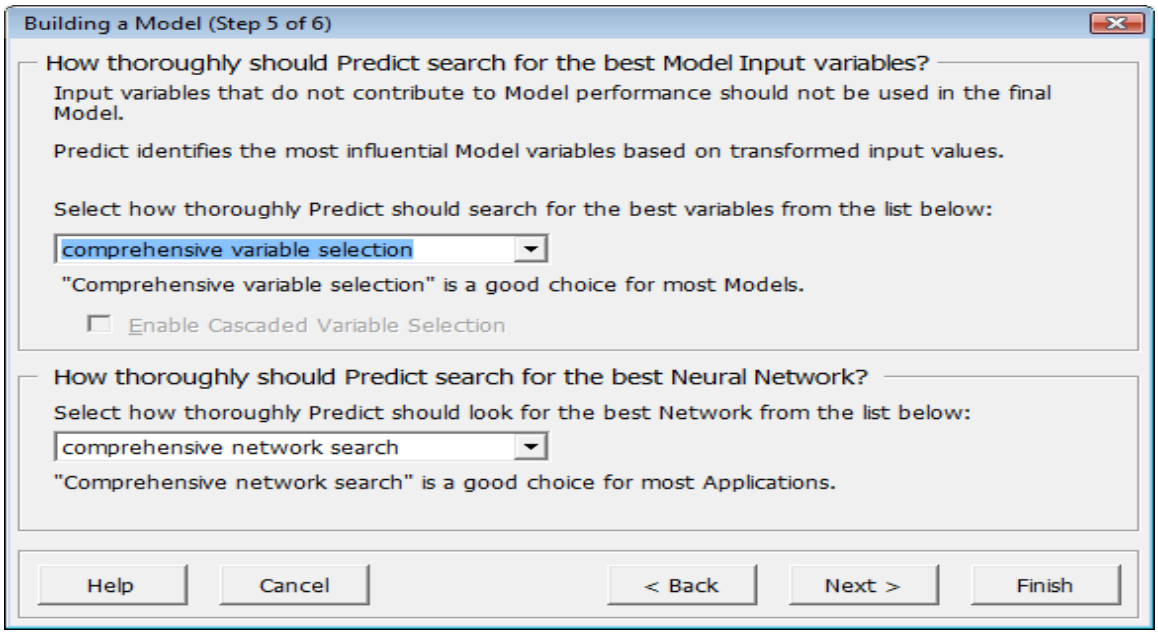


Чекорите 4 и 5 се повеќе од техничка природа и со нив се дефинира варијабилитетот на влезните променливи, т.е. податоците, потоа начинот на трансформација на податоците и изборот на целната променлива. Најчесто, притоа, се употребуваат однапред зададените параметри од страна мрежата.

Чекор 4

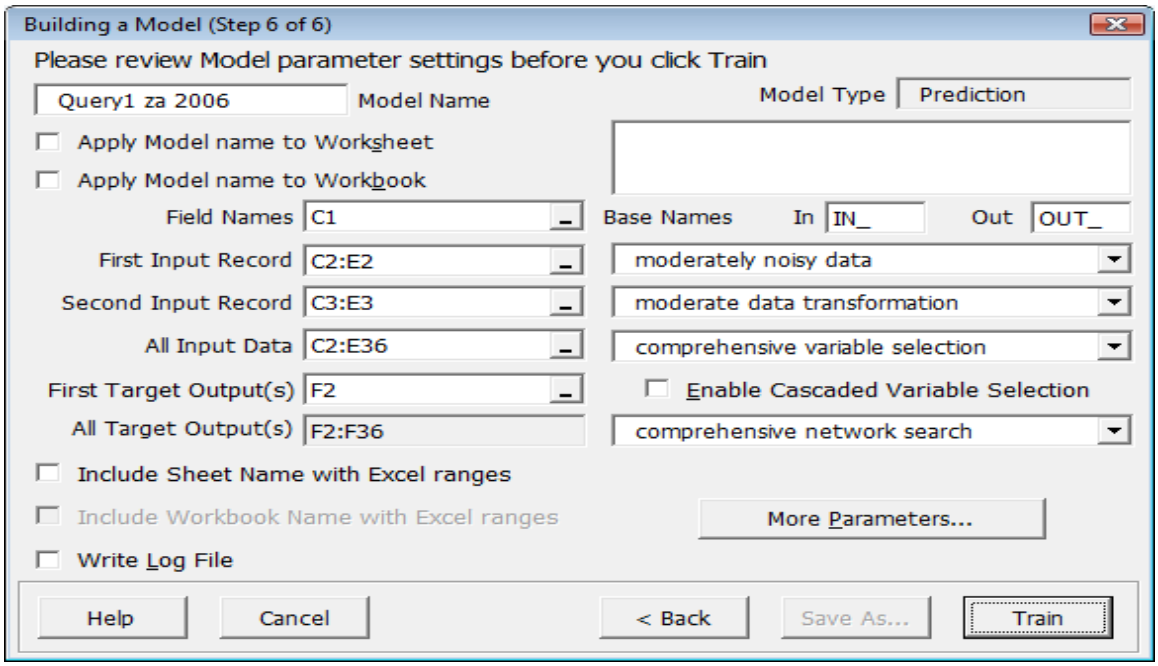


Чекор 5



Чекор 6

Во чекорот шест се добива преглед на се она што било внесено како влезни податоци, при што во овој чекор е можна модификација на претходно внесените параметри, доколку за тоа има потреба.

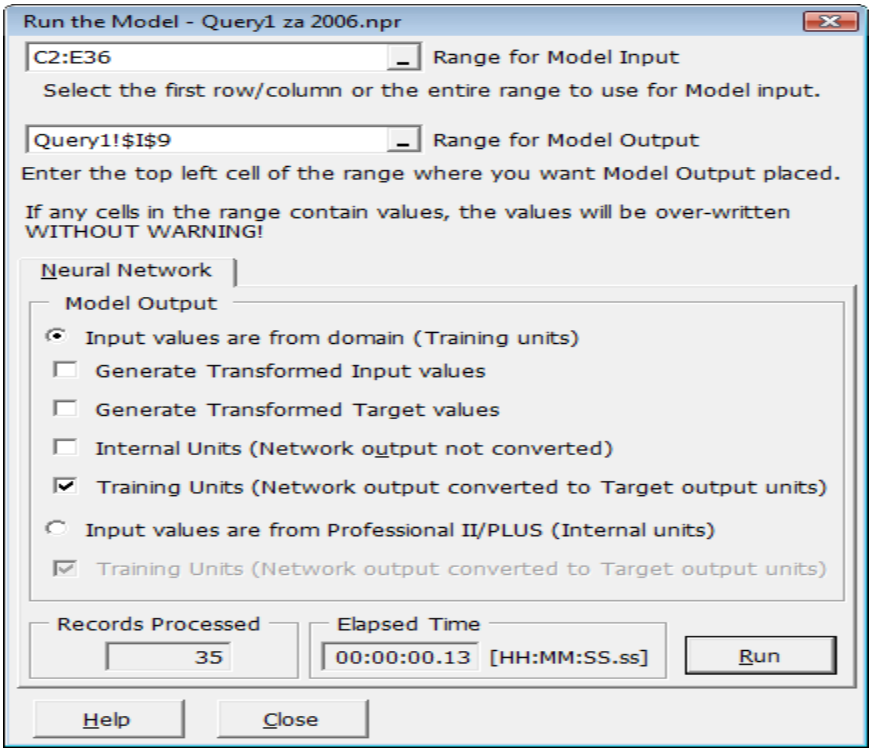


По шестиот чекор, со кој завршува изградбата на моделот, потребно е да се изврши тренирање на податоците со коешто, всушност, се врши прилагодување на тежинските фактори за да се минимизира грешката. Откако ќе се изврши тренирање на податоците, следи постапката на нивно тестирање. Со тестирањето на податоците се добива следната табела од која што можат да се извлечат голем број на информации за вака изградениот модел.

MODEL	R	Net-R	Avg. Abs.	Max. Abs.	RMS	Conf. Interval (95%)	Records
All	0.99731	0.921139	2.806053	30.5438	5.940002	12.04467	35
Train	0.99739	0.916044	2.787557	30.5438	6.561311	13.56879	24
Test	0.99731	0.921139	2.806053	30.5438	5.940002	12.04467	35

Од табелата што се добива по тестирањето на податоците, како најважен показател, може да се издвои вредноста на R, што претставува коефициент на корелација меѓу тренираните и тестираните податоци. Овој коефициент на корелација се движи во границите од 0-1 и може да се каже дека претставува клучна вредност којашто многу кажува за карактеристиките на вака изградениот модел. Вредноста на R покажува дали ваквиот модел може да се примени на податоците што се дадени. Колку вредноста на R на тренираните податоци е поблиску до вредноста на R на тестираните податоци, тоа значи дека моделот може добро да генерализира, па таквиот модел може да послужи за предвидување. Во овој случај, споредувајќи ја вредноста на R на тренираните податоци, што изнесува 0,99739 и вредноста на R на тестираните податоци што изнесува 0,99731, може да се заклучи дека овие вредности се многу блиски, па ваквиот модел добро генерализира и може да се предвидува продажбата на компанијата Алијанса за некој иден период.

Кога ќе се утврди дека моделот добро генерализира, тогаш може да се изврши негово стартување и да се реализира предвидувањето врз основа на историските податоци за изминатите две години.

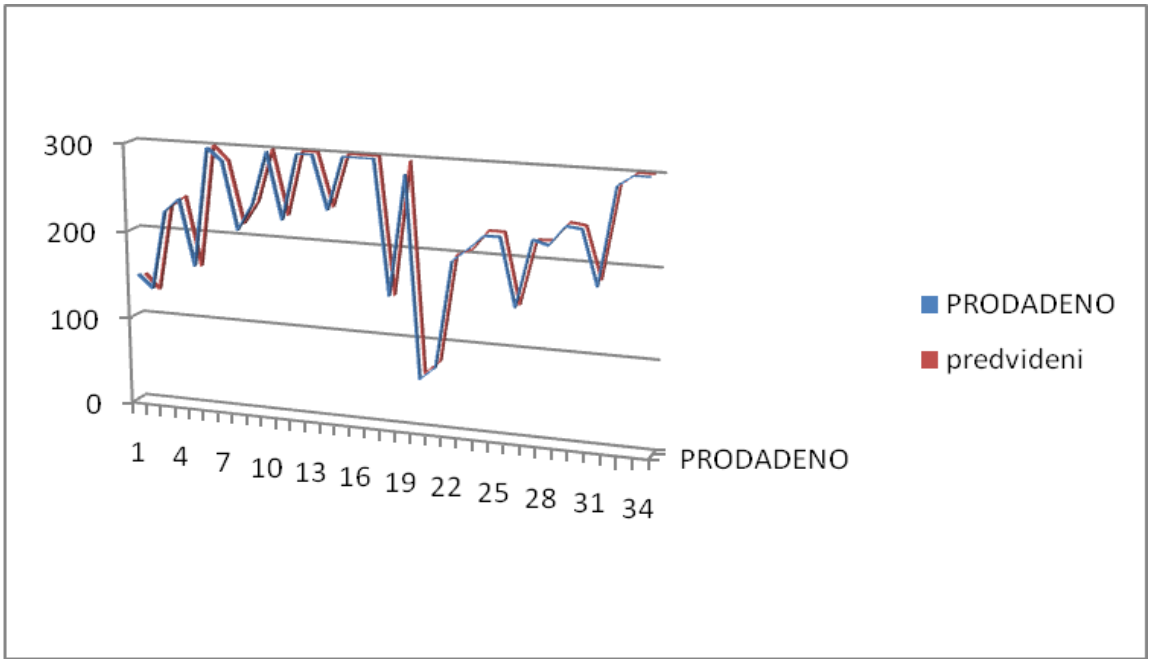


На тој начин се врши споредба меѓу вистинските вредности на излезната променлива и предвидените вредности на истата излезна променлива. Врз основа на отстапувањата меѓу овие вредности невронската мрежа учи, враќајќи се наназад и поправајќи ја грешката, преку прилагодување на тежинските фактори. Таквото отстапување меѓу вистинските и предвидените вредности може да се види и од следната табела.

PRETHODNA	NABAVENA	SOSTOJBA	PRODADENO	predvideni
0	150	0	150	146.9250031
0	135	0	135	130.0094604
0	225	0	225	227.0630341
0	300	60	240	240.219574
15	150	0	165	161.5819244
0	300	0	300	299.8052979
0	300	14	286	283.882843
0	225	15	210	213.6309662

15	225	0	240	240.0526581
0	300	0	300	299.8052979
0	225	0	225	227.0630341
0	300	0	300	299.8052979
0	300	0	300	299.8052979
0	300	60	240	240.219574
0	300	0	300	299.8052979
0	300	0	300	299.8052979
0	300	0	300	299.8052979
0	150	0	150	146.9250031
0	285	0	285	296.8951721
0	149	89	60	62.56642532
59	150	134	75	77.47583008
119	150	74	195	198.0269318
59	150	0	209	205.5336761
0	225	0	225	227.0630341
0	225	0	225	227.0630341
0	210	60	150	148.5244598
75	150	0	225	221.1418457
0	300	80	220	222.0917511
80	225	63	242	242.6233368
18	225	3	240	240.4075623
3	240	63	180	183.3834381
63	225	0	288	286.4053955
0	300	0	300	299.8052979
0	300	0	300	299.8052979

Оттук може да се изврши споредба меѓу вистинските и предвидените вредности на продажабата од страна на невронската мрежа. Најдобро би било тоа да се претстави преку графикон за да се види колкави се отстапувањата меѓу вистинските и предвидените вредности. Дадениот графикон покажува дека ваквите отстапувања се минимални, што би значело дека овој модел би можел да послужи за предвидување, затоа што добро генерализира.



Графикон 5. Однос меѓу вистинската и предвидената продажба

Врз основа на вака изградениот модел, односно врз основа на заклучокот дека овој модел може добро да генерализира се врши предвидување за некој иден период. Предвидувањето ќе се врши за претходна количина производи во магацинот, и тоа кога има 50 производи и остварена набавка од 250 производи.

PRETHODNA	NABAVENA	SOSTOJBA	PRODADENO	predvideni
50	250	0		296.8262329

Од податоците за вредноста на продажбата во изминатиот период, невронската мрежа предвидува продажба на производот за некој иден период од 296 производи. Од табелата се гледаме дека претходната состојба, плус набавената количина, изнесува вкупно 300 производи, додека невронската мрежа дава предвидена продажба од 296 производи. Тоа укажува на фактот дека до периодот на следната набавка во магацинот ќе има 4 производи што не се продадени. Од друга страна, доколку наместо 250 се набават 246 производи, тогаш во магацинот нема да се јави натрупување на залихи што врзуваат голем дел од финансиските

средства на компанијата. Целокупната работна постапка е дадена во следната слика, како еден Excel-ов работен документ.

IME	DATNAB	PRETHODNA	NABAVENA	SOSTOJBA	PRODADENO	predvideni
KOKA KOLA 0.5l MARVAL 15/1	01-Jan-06	0	150	0	150	146.9250031
KOKA KOLA 0.5l MARVAL 15/1	25-Mar-06	0	135	0	135	130.0094604
KOKA KOLA 0.5l MARVAL 15/1	31-Mar-06	0	225	0	225	227.0630341
KOKA KOLA 0.5l MARVAL 15/1	10-Apr-06	0	300	60	240	240.219574
KOKA KOLA 0.5l MARVAL 15/1	20-Apr-06	15	150	0	165	161.5819244
KOKA KOLA 0.5l MARVAL 15/1	27-Apr-06	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	12-May-06	0	300	14	286	283.882843
KOKA KOLA 0.5l MARVAL 15/1	01-Jun-06	0	225	15	210	213.6309662
KOKA KOLA 0.5l MARVAL 15/1	14-Jun-06	15	225	0	240	240.0526581
KOKA KOLA 0.5l MARVAL 15/1	30-Jun-06	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	18-Jul-06	0	225	0	225	227.0630341
KOKA KOLA 0.5l MARVAL 15/1	27-Jul-06	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	19-Aug-06	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	09-Sep-06	0	300	60	240	240.219574
KOKA KOLA 0.5l MARVAL 15/1	15-Sep-06	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	15-Sep-06	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	19-Oct-06	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	18-Nov-06	0	150	0	150	146.9250031
KOKA KOLA 0.5l MARVAL 15/1	01-Jan-07	0	285	0	285	296.8951721
KOKA KOLA 0.5l MARVAL 15/1	01-Feb-07	0	149	89	60	62.56642532
KOKA KOLA 0.5l MARVAL 15/1	05-Feb-07	59	150	134	75	77.47583008
KOKA KOLA 0.5l MARVAL 15/1	17-Feb-07	119	150	74	195	198.0269318
KOKA KOLA 0.5l MARVAL 15/1	27-Feb-07	59	150	0	209	205.5336761
KOKA KOLA 0.5l MARVAL 15/1	15-Mar-07	0	225	0	225	227.0630341
KOKA KOLA 0.5l MARVAL 15/1	11-Apr-07	0	225	0	225	227.0630341
KOKA KOLA 0.5l MARVAL 15/1	23-Apr-07	0	210	60	150	148.5244598
KOKA KOLA 0.5l MARVAL 15/1	12-May-07	75	150	0	225	221.1418457
KOKA KOLA 0.5l MARVAL 15/1	19-May-07	0	300	80	220	222.0917511

KOKA KOLA 0.5l MARVAL 15/1	12-Jun-07	80	225	63	242	242.6233368
KOKA KOLA 0.5l MARVAL 15/1	30-Jun-07	18	225	3	240	240.4075623
KOKA KOLA 0.5l MARVAL 15/1	23-Jul-07	3	240	63	180	183.3834381
KOKA KOLA 0.5l MARVAL 15/1	08-Aug-07	63	225	0	288	286.4053955
KOKA KOLA 0.5l MARVAL 15/1	29-Aug-07	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	11-Sep-07	0	300	0	300	299.8052979
KOKA KOLA 0.5l MARVAL 15/1	20-Oct-07	0	1	1	0	30.54380035

MODEL	R	Net-R	Avg. Abs.	Max. Abs.	RMS	Conf. Interval (95%)	Records
All	0.997305	0.9211394	2.806053	30.5438	5.940002	12.04467	35
Train	0.99739	0.9160439	2.787557	30.5438	6.561311	13.56879	24
Test	0.997305	0.9211394	2.806053	30.5438	5.940002	12.04467	35

предвидување за иден период	**50**	**250**	**0**	**296.82623**

Заклучни согледувања

Новите информациски технологии со се посилните перформанси овозможуваат складирање на големи количини податоци , нивна обработка, како и нивно искористување за добивање информации што служат како поддршка во одлучувањето и даваат големи можности за унапредување на деловните процеси. Во современите услови на работење информациите претставуваат клучен ресурс и основа за стекнување конкурентска предност, а основна цел на секоја компанија е да ја зголеми својата конкурентска предност во однос на останатите компании што учествуваат на пазарот. Со тоа растечките потреби на компаниите за навремени и точни информации ја наметнуваат неопходноста од обезбедување, складирање и обработка на тие информации. За стекнување таква конкурентска предност компаниите применуваат бројни методи, техники и алатки.

Денес, секоја модерна светска компанија својот интерес го насочува кон инвестирање во информацискатa технологија. Познавањето на концептот за откривање знаење и примена на алатките и методите што за таа цел ги развила информацискатa технологија има се поголемо значење. Во постојните услови на пазарот кога понудата значително ја надминува потрошувачката, конкурентската предност на компанијата не ја прави само количината производи, туку и нивниот квалитет. А, во квалитетни производи и услуги, освен основните суровини, се вградува се повеќе и знаењето, со што им се додава нова и поголема вредност. На тој начин доаѓа до преминување од индустриска економија во економија на знаење.

Преминот во економија на знаење се карактеризира со неколку важни чекори:информацијата станува најважен ресурс, а информацискатa технологија прераснува во основна инфраструктура. Исто така, главни фактори што ги наметнуваат условите на работење се потрошувачите и крајните корисници кои ги купуваат само најквалтетните производи, па според тоа само производите во кои што е вградено ново знаење можат да го издржат притисокот од конкуренцијата.

Успешните компании, при креирањето на деловните стратегии, ги анализираат и утврдуваат елементите што битно придонесуваат за создавање конкурентска предност и овозможуваат нивен долгорочен раст и развој. Во вакви

услови на работење најголема влезна бариера, но истовремено и конкурентска предност претставува заедничкото знаење и вештините на вработените во компанијата. Така, организициите кои "учат" го стават знаењето во прв план, додека информацискатa технологија создава инфраструктура за управување со знаењето и предуслови за создавање организација што "учи".

Имено, се посложените услови на работење секојдневно генерираат се поголем број на деловни настани внатре во комапниите и надвор од нив, а добиените податоци се складираат во оперативни бази на податоци. Големината на таквите бази на податоци не овозможува пребарување во реално време, а извештаите што од нив се добиваат, обично, се наоѓаат на различни работни станици, во дводимензионален облик, па нивната содржина, брзина и начин на пристап честопати не одговараат на потребите на менаџерите. Поради тоа, се јавува потребата од нов начин на организирање на податоците, што се заснова на концептот на складишта на податоци (Data Warehouse).

Преку истражувањето во овој труд се прави обид да се прикажат основните концепти , архитектура и методологија на развој на проект за складиште на податоци, и како тој како заокружена целина тој ќе може да послужи за запознавањето на оваа нова дисциплина и можнстите што таа ги нуди во реалните деловни процеси . Овој концепт, во суштина, претставува континуиран процес на планирање и собирање податоци од различни извори, како и нивно користење, управување и постојано унапредување.

Самата структура на складиштето на податоци ја чинат два основни дела, а тоа се: податоците и механизмите за управување со податоците. Делот од складиштето каде што се наоѓаат основните податоци и агрегираните повеќедимензионални податоци е првиот дел, а механизмите за манипулација со податоците ги претставуваат процесите на екстракција, трансформација и дополнување на податоците, потоа постапките за аналитичка обработка на податоците, и на крајот, презентацијата на податоците.

Складиштето на податоци е дизајнирано за анализа на непроменливи податоци, односно е изградено и одржувано со дополнување на податоци за подолг временски период, и е изразено преку едноставни деловни термини и агрегирање

на податоците што се пренесени од трансакциските системи. Основната намена на складиштето на податоци му е поддршка на одлучувањето за повисоките хиерархиски нивоа на менаџментот, и тоа по пат на прашалници и анализа на специјално структуираните податоци. Во складиштето на податоци се наоѓаат податоци од различни временски периоди, од различни делови на трансакцискиот систем, од интерни или надворешни извори итн. Со сето тоа се елиминира временската дводимензионалност и временското ограничување на една година, што е карактеристично за трансакциските информациски системи. Сите овие карактеристики ги поседува складиштето на податоци на компанијата Алијанса, што е користено во овој магистерски труд за да се прикажат аспектите на примена на ваквите нови концепти во реалноста.

Технологијата за складирање на податоците иницира развој на алатки за нивна аналитичка обработка или т.н. OLAP. Ваквата технологија претставува се популарен сегмент на информациските системи и во средните и големите компании. Со воведувањето на ваквата технологија се овозможува ред и конзистентност во постапките за комплексна анализа и известување на ниво на поединечни делови или на деловниот систем во целина. OLAP серверите им овозможуваат на апликациите на складиштата на податоци способност за известување, предвидување на продажбата, анализа на производот и неговата профитабилност, маркетинг анализи итн. Со својата можност за вршење робусни калкулациски операции, OLAP сереверите ги трансформираат историските податоци во корисни податоци. Тоа, пак, овозможува увид не само во историските податоци, туку и во идните услови на работење.

Како главни предности на складирањето на податоците (Data Warehouse) и OLAP можат да се наведат:

- систематизацијата и усогласувањето на хетерогените податоци внатре во компанијата;
- брзото изведување на комплексни анализи во реално време, што не е можно да се спреведе на друг начин;
- заштеда во време и човечки потенцијали за подготовка на комплексни извештаи;

- можност за прилагодување на содржината на аналитичките бази на податоци за потребите на корисниците;
- ниски трошоци за одржување и употреба;
- точен увид во вистинските состојби на работењето според сите сегменти на деловниот систем;

Главниот акцент на овој труд, сепак, е ставен на откривањето на знаењето од аналитичките бази на податоци (Data Mining), односно на пронаоѓањето скриени трендови, шаблони, законитости и односи меѓу податоците. Еволуцијата на Data Mining е започната уште во периодот кога податоците за првпат биле складирани во компјутерите, а продолжува континуирано со унапредувањето на пристапот до податоците и генерирањето технологии што им овозможуваат на корисниците пребарување низ податоците во реално време. Процесот на Data Mining денес е можно да се спроведува исклучиво благодарение на моќните мултипроцесорски компјутери, технологиите за масовно собирање и складирање на податоците и алгоритамските техники за Data Mining. Така, во современи услови, со се поголемото користење на базите на податоци и новиот пристап за динамичка обработка на податоците од големите количини податоци се доаѓа до скриените податоци што се многу значајни за добивањето информации и откривањето знаење, засновано на податоците, односно создавањето нови деловни вредности за компанијата.

Целта на овој труд е да ја прикаже, преку примена на некој од методите за Data Mining применливоста на оваа релативно нова технологија во подобрувањето на работењето на компаниите. Со практичен пример, пак, е покажана соодветноста на некој од методите за Data Mining за решавање на определен проблем, како и значењето на примената на ваквиот метод. Примената на ваквите методи доаѓа до израз кога не постои можност да се употреби некој од класичните методи што беа користени порано за решавањето на овој вид проблеми, што и дава уште поголема важност на оваа дисциплина. Проблемите за коишто тешко се наоѓаше решение или што бараа сложени постапки и алгоритми за нивно решавање сега со појавата на современите методи и алатки за Data Mining, лесно и брзо се решаваат. Можноста за примена на ваквите методи, главно, зависи од достапноста и

доволната количина квалитетни податоци, а тоа е евидентно и од моделот за предвидување, изнесен во практичниот дел од трудот. Можеби одделни методи за Data Mining, со својата архитектура, се ограничени во можноста за анализа на податоците, но на сите ним заедничка им е потребата од квалтитени податоци. Се друго зависи од начинот на којшто работи определениот метод. Оттука се покажува дека таквите напредни технологии имаат докажана вредност при откривањето нови информации и знаење. Откриеното знаење повеќекратно ја зголемува употребната вредност на збирот информации за некое деловно подрачје.

Целта на методите за Data Mining е автоматизирање на процесот за донесување одлуки. Денес постојат мал број системи што се способни самостојно да донесуваат одлуки, независно од факторот човек. Се уште, и во процесот на Data Mining најголемо влијание има човекот. Data mining поседува методи и алгоритми за анализа на податоците, но сепак човекот е тој кој, по пат на експериментирање, доаѓа до најдобри резулатати. Натамошниот развој на методите за откривање знаење во аналитичките бази на податоци треба да овозможи примена на тие методи со минимална интервенција на експерти и приближување до корисниците, без потреба за нивно широко и техничко образование и едукација за методите и алгоритмите за Data Mining.

Исто така, во трудот се прикажани и некои пошироки аспекти на примена на откривањето знаење што е можно единствено преку интеграција на сите релевантни извори на податоци од окружувањето. Имено, Data Mining наоѓа примена во голем број области, и тоа не само во рашавањето на проблемите од сферата на економијата, туку и во низа други области, како што се медицината, астрономијата, спортот, генетскиот инженеринг итн. Ова ја потврдува повеќекратната примена на методите на Data Mining при решавањето најразлични проблеми од голем број области, и нивната корисност за унапредување на работењето воопшто.

Користена литература:

1. George M. Marakas, Modern data Warehouising, mining and visualisation, Upper Saddle River, New Jersey, 2003
2. Daniel T. Larose, An Introduction to Data Mining, John Viley & Sons, New York, 2005
3. Two Crows Corporation, Intrtoduction to Data Mining and Knowledge Discovery, Third edition, 2005
4. Joseph M. Firestone, Data Mining and Knowledge Discovery in Database, White Paper No. Two, 1999
5. Коста Сотироски, Статистика, Универзитет "Св. Климент Охридски", Економски факултет, Прилеп, 2004
6. Ramakrishnan R., Freytag J.C, Report based on a workshop Data Mining:The Next Generation, Dagstuhl seminar, 2005
7. Маргарита Јанеска, Менаџмент Информациски Системи, Универзитет "Св. Климент Охридски", Економски Факултет, Прилеп, 2004
8. David Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining, MIT Press, Cambridge, MA 2001
9. Auroop R. Ganguly and Amar Gupta, Data Mining Technologies and Decision Support System for Business and Scientific Application, University of South Florida, 2004
10. Seyyed Jamaleddin Pishvayi, Data Mining Technique for CRM, Tehran University, 2005
11. Kovalerchuk B., Data Mining for Financial Applications, Central Washington University, USA, 2004
12. Agosta L., Orlov L. M. and Hudso R., The Future of Data Mining:Predictive Analytics, Forrester brief, November 2003
13. U. Fayyard, G. Piatetsky, Shapiro, P. Smyth & R. Uthurusamy-Advances in Knowledge Discovery & Data Mining, Cambridge MA (The AAAI Press/The MIT Press) 2000
14. Singh, S.-Interactive Data Warehousing, Prentice Hall, New York, 2000

15. Alex Berson, Stephen Smith, and Kurt Thearling,Building Data Mining Applications for CRM, 2000
16. Imielinski T. and Mannila H., A Database Perspective of Knowledge Discovery, Comunication of ACM, 1996
17. Cary N. Prague and Michael R. Irvin, Access 2002 Bible, Hungry Minds, 2002
18. Dorian Pyle- "Data Preparation for Data Mining", Morgan Kaufmann, San Francisko, 2000
19. The Technology Review Ten, MIT Technology Review, January/February 2001
20. Zoran Skocir, Ivan Matasic, Boris Vrdoljak, Organizacija obrade podataka, Sveuciliste u Zagrebu, Fakultet elektrotehnike I Racunarstva, Zagreb, 2000
21. The First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, August 20-21, 1995
22. Nocke T., Schumann H., Meta Data for visual Data Mining, Institut of Computer Graphics, University of Rostck, 2002
23. J. Han and M. Kamber, Data Mining:Concepts and Techniqeues, Morgan Kaufmann, San Francisko, 2004
24. H. Zhang, Mining and Visualization of Associoation Rules Over Relational DBMSs.PhD thesis, Department of Computer and Information Science and Engineering, The University of Florida, 2000
25. R. Kimball, The Data Warehouse Toolkit:Practical Techniques for Building dimansional Data Warehoses, John Viley & Sons, New York, 1999
26. Mladen Varga, Baze podatatka, DRIP, Zagreb 1998
27. Olivia Parr Rud, Data Mining Cookbook, Modeling Data for Marketing, Risk and Customer, Relationship Management, John Viley & Sons, New York, USA, 2001
28. Jim Davis, Ellen Joyner, Successful Customer Relationship Management, SAS Institut, Inc., 2000
29. V. Ceric, M. Varga, H. Birolla, Poslovno racunarstvo, Znak Zagreb, 1998
30. J.G. Taylor, Neural Networks and their Applications, John Viley & Sons, New York, 1998

31. U. Fayyad, G.G. Grinstein and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, San Francisko, 2001
32. Alex Berson, Stephen Smith, and Kurt Thearling, Building Data Mining Applications for CRM, 2000
33. Panian Z., Klepac G., Poslovna Intelegencija, Masmedia, Zagreb, 2003
34. Imielinski T. and Mannila H., A Database Perspective of Knowledge Discovery, Comunication of ACM. 1996
35. Kohavi R., and Provost F., Application of Data Mining to Electronic commerce, Kluwer Academic, 2001
36. Ante Panjkota, Poslovna Informatika, Visoka Skola za Turisticki Menadzment, Sibenik, 2006
37. Freitas A., and Lavington S., Mining very large databases with parallel processing, Kluwer Academic Publishers, 2000
38. Barry M. and Linoff G., Mastering Data Mining:The art and Science of Customer Relationship Management, 2003
39. Vasand Dhar and Roger Stein, Seven Methods for Transforming Corporate Data into Business Intelligence, Prentice Hall, New York, 2002
40. E.G.Mallach, Decision Support and Data Warehouse Systems, McGraw-Hill, New York, 2004
41. Singh, S., Using Information Technology Effectively, Information & Management 24, 1993
42. Saban, Josip, Menadzerska Izvjesca, HGR 6/98, Zagreb, 1998
43. Inmon, W., H.- Building the Data warehouse, Second Edition, John Viley & Sons, New York, 1999.
44. Kambayashi, Y., K. tanaka, S. Yajima, Problem of Relational Database Design, Berlin Heidelberg New York, 1994
45. Bhandari I., Colet E., Parker J., pines Z., Advanced Scout:Data Mining and Knowledge Discovery in NBA Data, Data Mining and Knowledge Discovery 2002

46. Carlson C., Turban E., Decision Support System, Elsevier, 2002
47. Hammer J., Advances in online analytical processing, Data and Knowledge Engineering, Elsevier 2003-
48. McCuistion J.D., and Birk R., NASA Tachnical Report, 2002
49. Potter C., Klooster S., Steinbach M., Tan P., Kumar V., Journal of Geophysical Research, 2003
50. Wang G. C. and jain C. L., Regression Analysis:Modeling and Forecasting, Institut of Business Forecasting, 2003
51. The GeneOntolgy Consortium, Genome research 2001
52. Hakenberg J., Bickel S., Plake C., Brefeld U., Leser U., Systematic Feature Evaluation for Gene Name recognition, BMC Bioinformatic, 2005

INTERNET IZVORI

www.spss.com
www.anderson.ucla.edu
www.statsoft.com
www.efzg.hr
www.cfp.org
www.thearling.com
www.sas.com
www.goranclepac.com
www.attar.com
www.neurosolutions.com
www.worldstats.org
www.outerbay.com
www.data-miners.com
www.bluemartini.com
www.olapreport.com
www.teradata.com
www.ncr.com
www.provalisresearch.com
www.corporateinformation.com
www.uncstd.org
www.mindtools.com
www.e-europestandards.org
www.psychstat.missouristate.edu