Univerzitet Singidunum

Departman za poslediplomske studije

SAVREMENE INFORMACIONE TEHNOLOGIJE

MASTER STUDIJE

- Master rad -

Alati za Data Mining

Mentor
Prof. Dr Milan Milosavljević

Kandidat
Nikola Jovanović 410639/09

Beograd, 2011.

Rudarenje Podataka

Sažetak:

Svako preduzeće ili bilo kakva organizacija prikuplja velike količine podataka potrebnih za svoje poslovanje. Međutim, nas zanima, postoje li još nekakve dodatne informacije u tim već prikupljenim i poznatim podacima koje mogu biti korisne za poslovanje nekog preduzeća. Možemo li iz postojećih podatka pronaći još korisnih informacija?

Danas se koristi dosta alata različitih proizvođača za rudarenje podataka. Ali kako su zahtevi krajnjih korisnika zahtevniji, tako se vremenom ističu određene korporacije ili open-source programi koji su pristupačniji i prilagodljiviji krajnjim korisnicima. Data mining alati se koriste rasprostranjeno da bi se rešili realni problemi u inženjerstvu, nauci, poslu, itd. Kako se broj prodavaca alata za data mining povećava, tako je veći izazov odrediti koji alati su najbolje prilagođeni za određene vrste poslova, metoda i potreba. Ilustrovaćemo različite alate uz nekakvu klasifikaciju i globalnu analizu postojećih rešenja.

Drugi deo ovog rada će se pozabaviti jednim konkretnim alatom koji ćemo prikazati kroz nekoliko pokaznih primera.

Ključne reči – data mining, metode data mining, korišćenje data mining, alati za rudarenje podataka, softver za rudarenje podataka, WEKA, SPSS, WEKA, SPSS, Intelligent Miner, OLPARS

Data Mining

Abstract:

Any company or any organization collects a large amount of data needed for their business. However, we are concerned, there are still some additional information in those already known and collected data that may be useful for the operation of any enterprise. Can we find from existing data more useful information ?

Today it is used a lot of tools from various vendors for data mining. But how are the end users are demanding users, so over time highlight certain corporations or open-source programs that are more affordable and more flexible for end-users. Data mining tools are used widely to solve real problems in engineering, science, business, etc. As the number of vendors of tools for data mining is increasing, so is more challenging to determine which tools are best adapted for certain kinds of tasks, methods and requirements. We will represent different tools with some sort of classification and global analysis of existing solutions.

The second part of this paper deals with one specific tool that will show through several demonstrative examples.

Key words: data mining, methods of data mining, where we use data mining, data mining tools, data mining software, WEKA, SPSS, Intelligent Miner, OLPARS

SADRŽAJ

1. Metodologija istraživačkog rada ---- 1
1.1 Uvodne napomene 1
1.2 Predmet istraživanja 1
1.3 Hipotetički okvir 1
1.4 Ciljevi istraživanja 1
1.5. Metode istraživanja i tok istraživačkog procesa 2
2. Uvod ---- 3
2.1 Šta je Data Mining? 3
3. Poreklo data Mining-a (Istraživanje podataka) ---- 4
3.1 Poreklo 4
3.2 Tradicionalne tehnike su neefikasne 4
3.3 Izazovi u procesu IP-a 4
4. Data mining ---- 5
4.1 Upotreba Data Mininga 5
4.2 Metode rudarenja podataka 7
4.3 Najčešće korišćene metode: 7
4.4 Nove izvedene metode rudarenja podataka 13
5. Procesi ---- 16
5.1 Proces rudarenja podataka 16
5.2 Ko su učesnici u procesu Data Mininga? 17
5.3 Ko traži stručnjake u Data Miningu? 18
6. Primeri rudarenja podataka ---- 19
7. Alati za rudarenje podataka ---- 20
7.1 Poznati alati za rudarenje podataka 20
8. Poređenje alata za rudarenje podataka ---- 21
8.1 Poređenje alata za rudarenje po izradi određenih algoritama 21
9. IBM alat za rudarenje podataka SPSS Clementine ---- 25
9.1 Izgled Clementine ekrana 25
9.2 Primer korišćenja alata Clementine 30
9.3 Proces pravljenja C5.0 algoritma 30
9.4 Proces pravljenja GRI modela 37
10. Primena Clementine alata nad realnim podacima ---- 39
10.1 Ulazni podaci i njihov pregled 39
10.2 Klasifikacija podataka 41
10.3 Neuronske mreže 42
11. Zaključak ---- 44
12. Literatura ---- 45

1. Metodologija istraživačkog rada

1.1 Uvodne napomene

Istraživanje sprovedeno tokom pisanja ovoga rada je pokazalo da je rudarenje podataka važan analitički korak u otkrivanju znanja korišćenjem baza podataka. Moderne računarske komponente dozvoljavaju da u današnje vreme kompletan proces rudarenja dosta brže i jednostavnije funkcioniše. U ovom radu su detaljno analizirane neke metode rudarenja, radi sagledavanja njihovog mogućeg korišćenja na različitim slučajevima potreba.

Rad može da koristi manje stručnim korisnicima računara, ali i naprednijim korisnicima kao skraćeni prikaz načina rudarenja.

1.2 Predmet istraživanja

Predmet istraživanja je analiza i potpuno sagledavanje metodike upotrebe rudarenja podataka u današnjem poslovnom i privatnom sektoru. Prezentovanje osnovnih metoda kroz praktične primere. U toku ovog istraživanja prikazaćemo klasifikaciju i globalnu analizu postojećih rešenja. A zatim kroz konkretan alat, njegov prikaz i u njemu obrađeni određeni primeri.

1.3 Hipotetički okvir

Generalna ili opšta hipoteza: Priroda rudarenja podataka je evoluirala mašinskog učenja, veštačke inteligencije, preko baze podataka do današnjih softvera koji se sve više unapređuju, čime se teži da rudarenje postane dostupno i krajnjim korisnicima.

Posebna ili radna hipoteza: Dosta istraživanja i statističkih podataka ukazuju na ogromne količine podataka koji se ne koriste. U tom obilju podataka postoje skriveni podaci koji treba da budu otkriveni i funkcionalno iskorišćeni

Pojedinačna hipoteza: Informacije u savremenom društvu predstavljaju veoma cenjen resurs za kojim se povećava potražnja. Velike mega kompanije su to već odavno uvidele, i već uveliko traje utrkivanje ko će proizvesti bolji software za obradu informacija koji će biti dostupan svakome.

1.4 Ciljevi istraživanja

Naučni cilj ovog istraživanja je analiza i potpuno razumevanje prirode i funkcionalnosti data mining-a radi uspešne analize podataka i stvaranja novih informacija.

Konkretan cilj ovog istraživanja je pronalaženje odgovora na sledeća pitanja:

- Koje se različite metode rudarenja koriste i za koje slučajeve najčešće ?
- Kojim tehnikama se služe „rudari“ da bi došli do informacija ?
- Koji su procesi u rudarenju podataka ?

1.5. Metode istraživanja i tok istraživačkog procesa

Osnovne metode: Za potrebe istraživanja, izvršena je analiza metoda rudarenja podataka. Pomoću nekoliko kompetentnih izvora stručne literature, izvršena je podela metoda rudarenja, s ciljem da se rasvetli njihov značaj i njihova praktična primena na određenim tipovima podataka koji se analiziraju.

Statističke metode: Korišćeni su aktuelni statistički podaci vodećih svetskih kompanija koje se bave proizvodnjom i implementaciom softvera za rudarenje. Tokom rada je predstavljeno i nekoliko studija slučajeva odabranih primeraka metoda rudarenja.

Eksperiment: U cilju potpunog razumevanja strukture data mining-a biće prikazani određeni alati i njihova primena.

Tok istraživačkog procesa: Najpre je sprovedena analiza metoda rudarenja podataka koje se danas koriste. Pro tome su objašnjene njihove karakteristike i obrazloženo je za koje tipove podataka se koriste i u kojim slučajevima. Potom je izvršena i obimna analiza postojećih programskih rešenja dostupnih na današnjem tržištu. U sledećem delu su obrađeni procesi od kojih se sastoji jedno rudarenje podataka. Na kraju rada je dat pregled korišćene literature i stručnih izvora na kojima je deo istraživanja i osnovan.

2. Uvod

2.1 Šta je Data Mining?

Rudarenje podataka je proces pronalaženja novog i potencijalno korisnog znanja iz podataka, odnosno kao otkrivanje ili "rudarenje" znanja iz velike količine podataka. Bolji bi izraz možda bio "rudarenje znanja iz podataka", gde se rudarenje odnosi na ono što se pronalazi kao rezultat tog procesa.

Data Mining je prirodna evolucija tehnologije, a koja upotrebljava koncepte, metode i tehnike različitih disciplina kao što su baze podataka, statistika i veštačka inteligencija. Tehnologija baza podataka se razvila iz primitivnih u sofisticirane i moćne sisiteme baza podataka kakve danas poznajemo. Taj razvoj omogućio je beleženje i sakupljanje ogromnih količina podataka što je neizbežno dovelo i do potrebe obrade i analize tih podataka, a sve u svrhu dobijanja korisnih informacija i znanja. Današnja informaciona tehnologija omogućila nam je beleženje svake aktivnosti u obliku digitalnih podataka, a količina sakupljenih podataka u mnogim bazama meri se u terabajtima, a jedan terabajt iznosi hiljadu gigabajta ili 1 099 511 627 776 bajta. Ogromne baze podataka bogate su podacima, ali i siromašne informacijama koje su skrivene u sačuvanim podacima. Upravo je Data Mining taj koji pomaže da otkrijemo važne informacije i znanje utkano u podatke, poslovanju i nauci. Data Miningom se takođe otkrivaju odnosi, logičnost, pravilnost te bilo kakve strukture među podacima. Rudarenje podrazumeva organizaciju baza čišćenjem podataka kako bi se pristupilo znanju i sticanju istog na temelju postojećih podataka u bazama. Razvoj tehnologije, računara i interneta bitno doprinosi lakšoj organizaciji podataka, no da bi oni postali korisni, potrebno je njihovo pretvaranje u informacije i znanje.

Termin rudarenja često se izjednačuje sa dva različita procesa: otkrivanje i predviđanje znanja. Proces otkrivanja znanja implicira korisnikovo razumevanje izrečenih informacija za koje je bitno da su u čitljivom obliku. Predviđanje se odnosi na buduće događaje i u nekim pristupima može biti čitljivo i providno dok u drugim neprovidno.

Data Mining ne predstavlja Data Warehousing, SQL / Ad Hoc Queries / Reporting, Software Agents, Online Analytical Processing (OLAP), niti Data Visualization

3. Poreklo data Mining-a (Istraživanje podataka)

3.1 Poreklo

Poreklo samog pojma data mining potiče od prvih stvari koje su se pojavile sa pojavom računara. To su:

1. Mašinsko učenje - kao oblast veštačke inteligencije predstavlja skup paradigmi, algoritama, teorijskih rezultata i primena iz različitih oblasti veštačke inteligencije i evolucionih modela (u smislu pretraživanja) ali i drugih oblasti: statistike i verovatnoće (Bajesov klasifikator, raspodele i testovi) i drugih oblasti matematike, teorije izračunljivosti (mora postojati svest o složenosti problema i ograničenjima koja proističu iz toga), teorije (adaptivnog) upravljanja, informacione teorije, psihlogije i neurobiologije (neuronske mreže), filozofije, kognitivnih i drugih nauka.

2. Veštačka inteligencija - predstavlja mešavinu konvencionalne nauke, fiziologije i psihologije, sve u cilju da se napravi mašina koja bi se, po ljudskim merilima, mogla smatrati "inteligentnom".

3. Baze podataka – predstavlja kolekciju podataka organizovanih za brzo pretraživanje i pristup, koja zajedno sa sistemom za administraciju, organizovanje i memorisanje tih podataka, čini sistem baze podataka. Iz ugla korisnika, podaci su na neki način logički način povezani.

4. prepoznavnaje oblika (paterni)

3.2 Tradicionalne tehnike su neefikasne

Zbog količine podataka koje treba analizirati da bi se došlo do rezultata. Takođe zbog same dimenzionalnosti podataka. Raznosvrsnosti podataka i samom načinu distribucije podataka.

3.3 Izazovi u procesu IP-a

Interakcija sa čovekom (preko interfejsa), čime čovek kao živo biće može da dorađuje sam proces. Uklapanje u određeni model.

Elementi van granica modela (eng. *outliers*), što predstavlja nešto ne izračunljivo samom čoveku.

Interpretacija rezultata, Vizuelizacija rezultata, Skalabilnost (ne velike datoteke), Dimenzionalnost (veliki broj atributa), Složeni i heterogeni podaci (npr. multimedija,...), Kvalitet podataka (nedostajući i irelevantni podaci).

4. Data mining

4.1 Upotreba Data Mininga

"Ključ uspeha je da znaš nešto što niko drugi ne zna."[1]

U poslovanju Data Mining se najviše koristi na području marketinga, koji je usmeren sve više pojedinačnom kupcu – upravljanje odnosima s kupcima (eng. Customer Relationship Management – CRM) koje je usmereno stvaranju, održavanju ili poboljšavanju odnosa s kupcima. Cilj ove pojedinačne usmerenosti na kupca jeste pridobijanje novih kupaca i zadržavanje starih kupaca. CRM pokušava uvideti želje i potrebe kupaca, razumeti njihovo ponašanje te predvideti buduće ponašanje. Na ovom području Data Mining je posebno uspešno, naravno uz dovoljnu količinu kvalitetnih podataka o kupcu.

U marketingu, Data Mining se najviše upotrebljava za:

1. direktni marketing – npr., katalozi i ponude različitih artikala se šalju kupcima za koje postoji najveća verovatnoća odaziva; pozivi upućeni starijim generacijama (ugl. penzionerima) na prezentacije od 10-ak ljudi gde trgovački putnici prodaju različite proizvode po "tvorničkoj ceni samo za njih večeras". Npr. prodaju proizvode od merino vune, proizvode koje ne štete našem zdravlju.

2. izradu profila kupaca – utvrđuje se uzorak ponašanja kupaca da bi mu se kasnije poslala prilagođena ponuda.

3. segmentaciju – utvrđivanje grupa kupaca jednakih karakteristika.
4. istraživanje povezanosti prodaje različitih proizvoda – analiza kupovine se može upotrebiti za raspoređivanje artikala na policama.

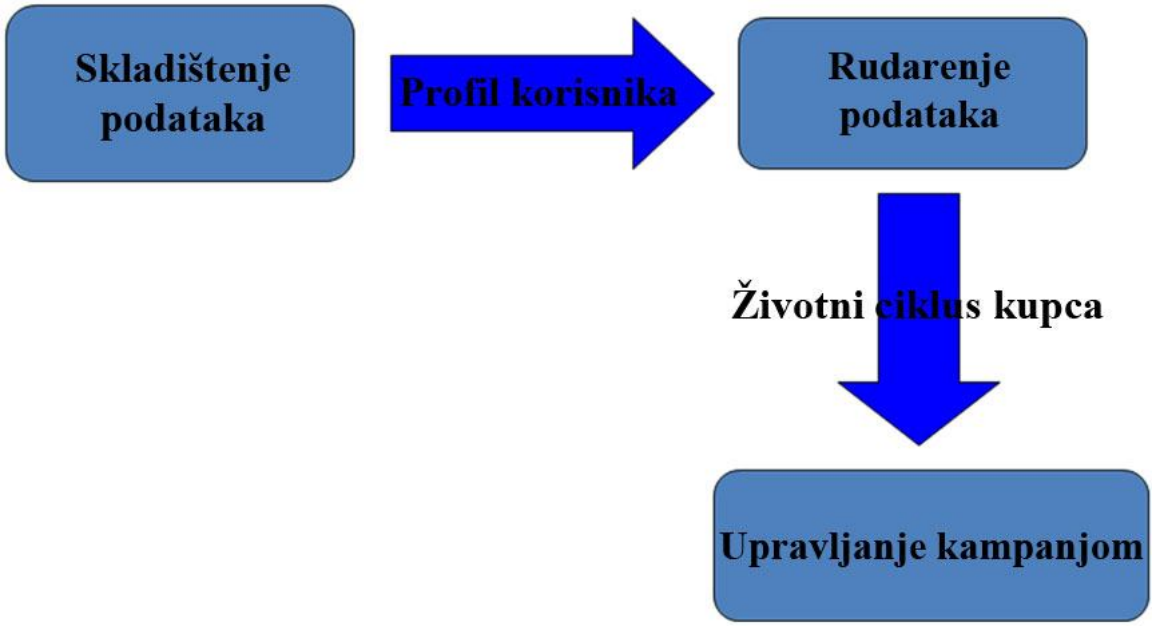
5. stimulacija kupovine drugih artikala istog preduzeća, odnosno veće količine istih artikala, što može nadoknaditi pridobijanje novih kupaca.
6. Kako ćemo ovo postići? Naš primer bi bio uz dva kupljena laptop računara dobijate na poklon laserski štampač koji je stigao kao gratis u to preduzeće (ali kupci to naravno ne znaju).

7. zadržavanje kupaca – ovo je puno jeftinije od pridobijanja novih kupaca.

Osim na području marketinga, Data Mining se koristi na području bankarstva (prognoze loših platiša kreditnih rata, prognoze vrednost deonica) i osiguranju (prognoze broja i vrednosti naknada za štetu, prognoze zloupotreba).

[1] **Aristotle Onassis**

Korišćenje DATA Mining-a u poslovnim sistemima (Slika 4.1):



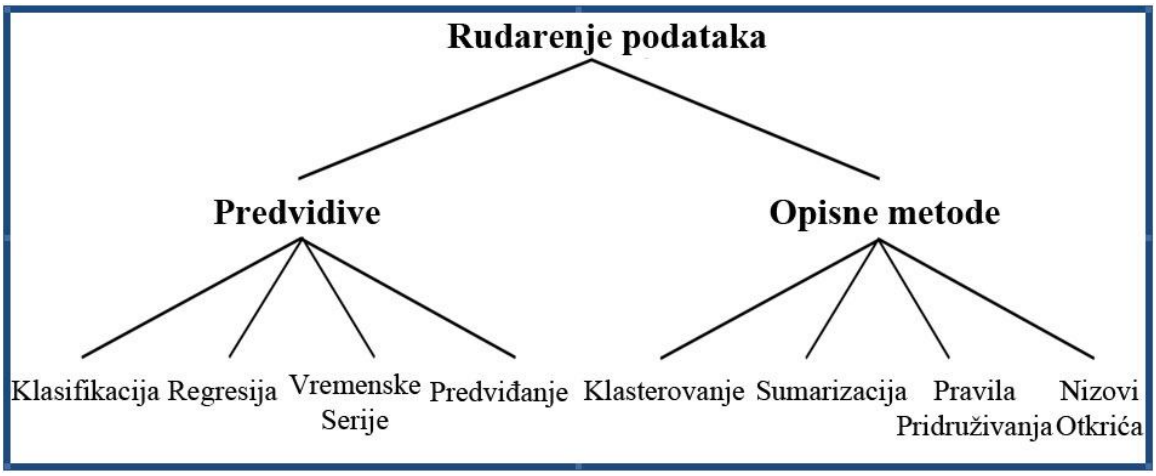
Slika 4.1 Upotreba data mining-a

Takođe razvijaju se rudarenja Weba (analiza podataka o poseti Web stranica i analiza puteva kojim posetioci dolaze do portala s podacima) i rudarenje tekstova (analiziraju se tekstovi i unutar njih traže uzorci i pravila koji se koriste pri kategorizaciji članaka po područjima, otkrivanju autorstva i sl.)

Masovnu upotrebu Data Mininga je omgućio razvoj IT industrije, jer je omasovljavanjem došlo do pada cena IT opreme koja je bitna za Data Minig, a tu spadaju tvrdi diskovi, operativna memorija i procesori. Sve je to omogućilo da obrada podatak bude brža, efektivnija, a samim tim i jeftinija.

4.2 Metode rudarenja podataka

“Razumeti znači uočiti šablone.” – Sir Isaiah Berlin



Slika 4.2 Metode rudarenja podataka

U rudarenju podataka koriste se metode već poznate iz statistike, matematike i računovodstva. Naročito treba istaći povezanost Data Mininga i statistike, treba znati da Data Mining ne zamenjuje statistiku, već služi kao njena nadogradnja.

4.3 Najčešće korišćene metode:

4.3.1 **Klasifikacija (Classification)**- omogućava automatsku klasificiju dokumenata npr. kada čitalac pronađe neki zanimljiv članak, pomoću ove metode može pronaći i druge slične članke.

I. Slogovi (vektori obeležja) koji se istražuju sadrže atribute(obeležja);
II. Traži se model za klase – funkcija koja preslikava atribute u neku od klasa
III. Klasifikacija preslikava podatke u predefinisane grupe tj. klase (kategorije)
IV. Cilj: prethodno neklasifikovane slogove dodeliti nekoj klasi što je moguće tačnije
V. Za testiranje tačnosti koriste se test podaci
VI. Naziva se i učenje sa nadgledanjem (sa učiteljem) jer su klase određene pre ispitivanja podataka (postoje prethodno klasifikovani uzorci)
VII. Prepoznavanje oblika – ulaz se klasifikuje na osnovu sličnosti sa predefinisanim klasama
VIII. Različito od predviđanja jer se ovde radi o trenutnom a ne o budućem stanju

Klasifikacija (primeri)

1. **Direktni marketing**

Cilj: smanjiti troškove slanja reklama određivanjem ciljne grupe potrošača koji žele da kupe nov proizvod

Pristup:

a. koriste se slični podaci sa prethodnih akcija istog tipa
b. zna se koji kupci su kupili a koji ne – klase su: (kupio, nije kupio)
c. o svakom potencijalnom kupcu se prikupljaju socijalni, demografski, ... podaci
d. Ove inforamcije se koriste kao ulaz u model

2. **Otkrivanje prevara**

Cilj: predupreditи moguće prevare kreditnim karticama

Pristup:

a. atributi iz (prošlih) transakcija kreditnim karticama i informacija o vlasnicima (kada, koliko, frekvencija kupovanja, ...)
b. Klase – prevara, regularna transakcija
c. Napraviti model za klase u transakcijama
d. Na osnovu modela otkriti prevare nadzorom transakcija

3. **Otkrivanje terorista**

Cilj: otkriti moguće teroriste na aerodromima

Pristup:

a. vizuelna kontrola pri ulasku na aerodrom
b. biometrijska kontrola putnika
c. pregled prtljaga skenerom
d. pregled ručnog prtljaga
e. pregled putnika skenerom
f. pokazivanje prtljaga pri ulasku u avion

4. **Odliv-priliv kupaca**

Cilj: otkriti da li će možda kupac otići kod konkurencije

5. **Katalogiziranje nebeskih tela**

Cilj: predvideti klasu zvezde ili galaksije na osnovu slika napravljenih telskopima

6. **Određivanje stetnosti virusa**

Cilj: otkriti način delovanja virusa na čoveka (žive organizme)

7. **Podela kandidata pri upisu na fakultet**

Cilj: upisati "prave" kandidate na određenu studijsku grupu

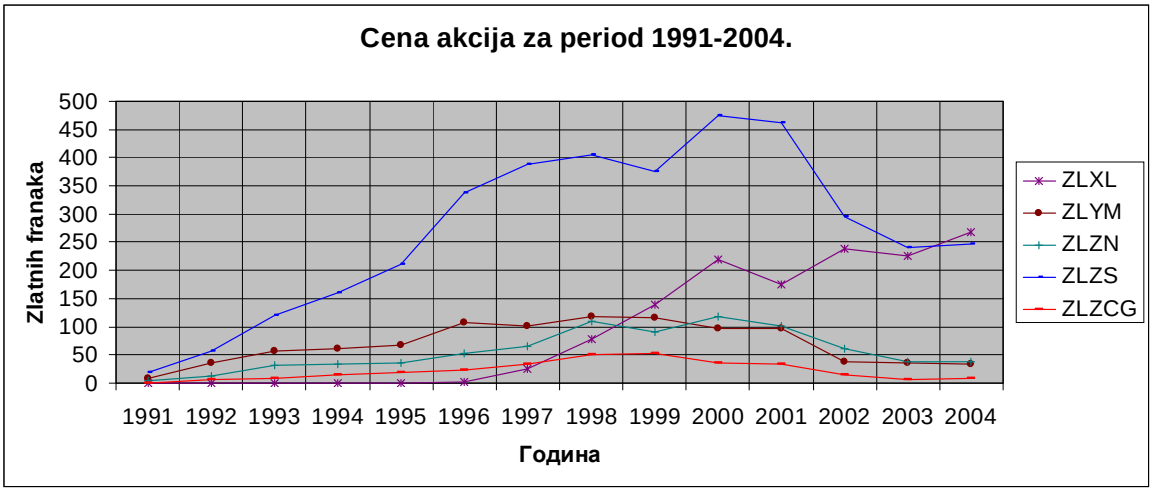
Regresija (Regression) - na osnovu vrednosti nekih promenljivih se predviđa vrednost promenljive čiji je domen skup realnih brojeva. Između promenljivih čije su vrednosti poznate i promenljive čija se vrednost predviđa postoje linearne ili nelinearne zavisnosti

4.3.2 **Regresija** (primeri)

1. Predviđa se količina prodaje novog artikla na osnovu utrošenih sredstava u reklamnoj kampanji
2. Predviđa se brzina vetra kao funkcija temeprature, vlažnosti, vazdušnog pritiska, količine oblaka
3. Predviđa se cena akcija na berzi u zavisnosti od vremenskog perioda i opštih ekonomskih prilika na tržištu

4.3.3 Analiza vremenskih serija (Time Series Analysis) - Istražuju se promene vrednosti atributa kroz vreme

Primer: odrediti u čije akcije bi investirali (Slika 4.3)



Slika 4.3 Vremenska serija

4.3.4 **Predviđanje (Prediction)**- Vid klasifikacije kod koga se predviđaju buduća stanja na osnovu prošlih i trenutnih stanja podataka. Ne spada u metode predviđanja. Razlika je u tome što metode predviđanja predviđaju trenutno a ne buduće stanje. Obuhvata pristupe koji ne spadaju u analizu vremenskih serija ili regresiju.

Predviđanje (primeri)

1. Predviđanje ponašanja glasačkog tela na izborima

Cilj: raspisati izbore u "pogodnom" momentu

Pristup:

a. istorija glasanja u nekoj oblasti
b. obrazovno-socijalna struktura stanovništva
c. podela teritorije na "pogodne" okruge
d. promena sistema glasanja

2. Predviđanje poplava

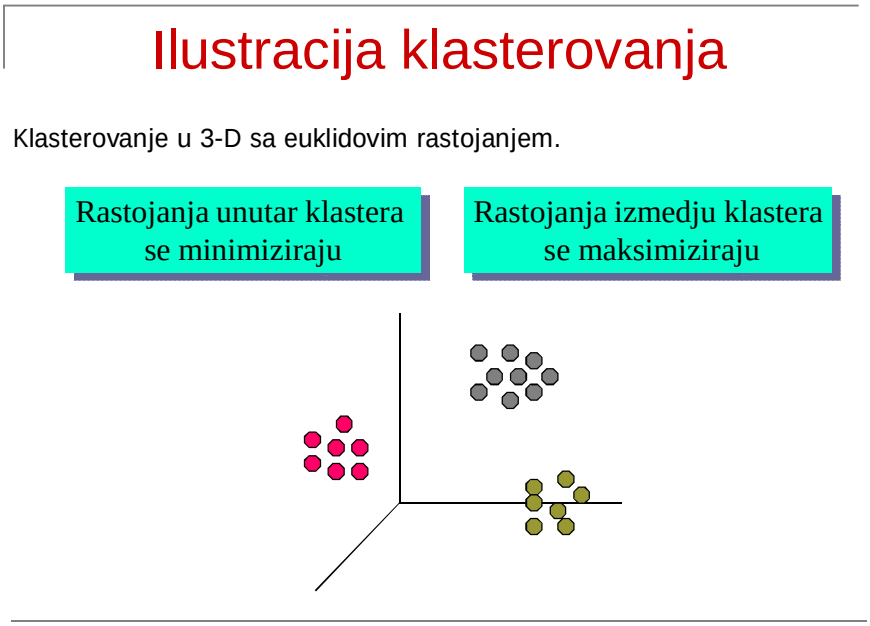
Cilj: predvideti na vreme poplave da bi se umanjile posledice

Pristup:

a. merenje vodostaja
b. merenje količine padavina
c. vremenska prognoza?

4.3.5 **Grupisanje-Grupisanje (Klastering)** - je razvrstavanje jedinki u grupe u kojima je postignuta njihova najveća sličnost.

Slični podaci (u odnosu na odgovarajuće atribute) se grupišu zajedno u klastere. Klasteri nisu unapred određeni. Obično je potrebno znanje eksperta da bi protumačio značenje formiranih klastera. Naziva se i učenje bez nadzora jer su klasteri nisu određeni pre ispitivanja podataka. Segmentacija – pojam koji se češće koristi kod baza podataka. Označava podelu podataka u disjunktne grupe (Slika 4.4).



Slika 4.4 Klasterovanje

Klasterovanje (primeri)

1. Upravljanje marketingom

Cilj: smanjiti troškove štampe kataloga u robnim kućama. Katalozi su namenjeni demografskim grupama u zavisnosti od atributa (zarada, lokacija, nivo obrazovanja, fizičke karakteristike - visina, težina, ...) potencijalnih kupaca

Pristup:

a. koristi se klasterovanje potencijalnih kupaca na osnovu određivanja vrednosti atributa

b. Na osnovu izlaza se formiraju posebni katalozi za određene grupe kupaca

2. Otkrivanje prevara

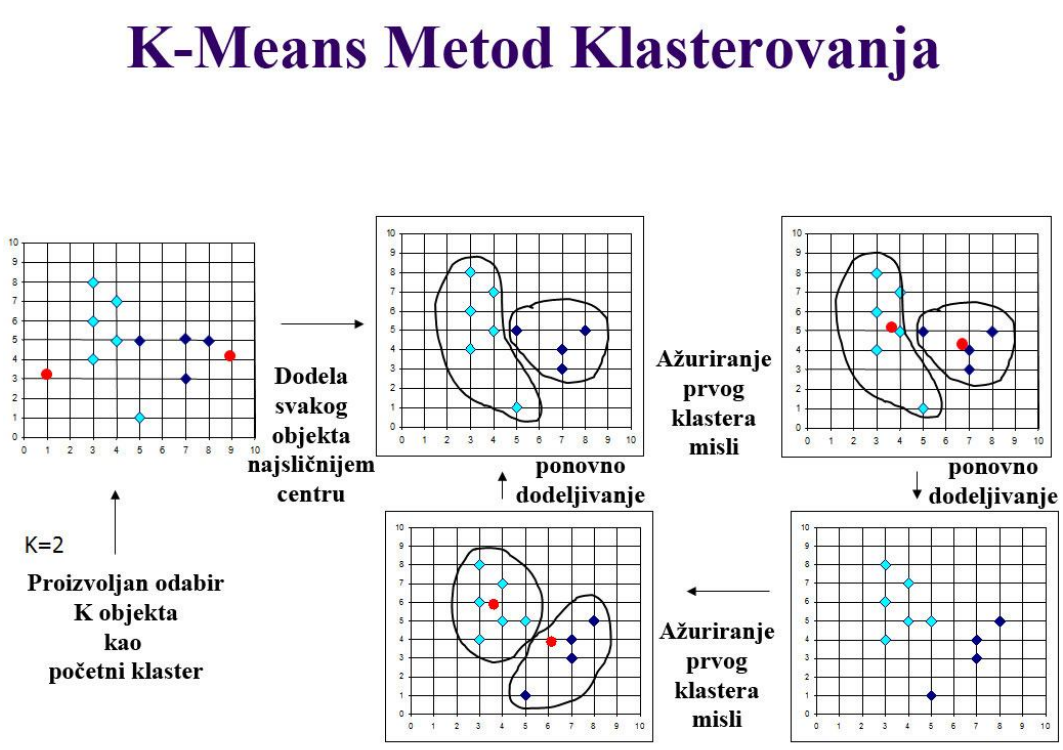
Cilj: predupredititi moguće prevare kreditnim karticama
Pristup:

a. registrovanje transakcija svih vlasnika kartica zajedno sa informacijama o godišnjem prihodu, kreditnoj sposobnosti, godinama života, ...

b. Broj prevara sa karticama je relativno mali u odnosu na broj transakcija primenom tehnika za otkrivanje anomalija se formira profil važeće transakcije za korisnika.

c. Svaka nova transakcija se poredi sa profilom korisnika

d. U slučaju velike razlike transakcija se označava kao moguće rizična, odn. potencijalna prevara



Slika 4.5 K-Means model

4.3.6 **Sumarizacija (Summarization)** - Preslikava podatke u podskupove sa pridruženim (jednostavnim) opisima. Izdvaja ili izvodi reprezentativne informacije iz baze. Naziva se i karakterizacija ili generalizacija. Iz podataka mogu da se izvedu i različiti sumarni prikazi (npr. numerički).

Sumarizacija (primer)
Načini poređena kvaliteta različitih univerziteta:

1. broj dobitnika Nobelove nagrade koji su studirali na tom univerzitetu
2. broj predstavnika u najvišoj državnoj vlasti koji su studirali na tom univerzitetu

Sumarizacija (primeri)

1. Bioinformatika: nalaženje grupe gena koja ima slične funkcionalnosti
2. Identifikacija Veb stranica kojima se pristupa zajedno
3. Razumevanje uticaja faktora na klimu
4. Raspoređivanje robe u supermarketima
5. Određivanje potrebne količine rezervnih delova u auto servisima (robe u magacinu prodavnica, ...)
6. Analiza potrošačke korpe

Cilj: Povećati prodaju kroz "promocije"

Pristup: uoči se zavisnost pri prodaji artikala, npr.

a. Hleb, keks, mleko
b. Pivo, hleb
c. Pivo, sok, pelene, mleko
d. Pivo, hleb, pelene, mleko
e. Sok, pelene, mleko

Zavisnosti:

{Pelene, Mleko} => {Pivo}, {Mleko } => {Sok}.

{Pelene, … } => {Pivo}

Pivo je posledica => može da se odredi šta treba preduzeti da bi se povećala njegova prodaja

Pelene su uzrok => određuje proizvod na koji utiče nedostatak pelena u prodaji

Pelene su uzrok i pivo je posledica => određuje proizvode koji treba da se prodaju uz pelene da bi se povećala prodaja piva

Kod sniženja cene pelena ne snižavati cenu piva

7. Preuređivanje sadržaja polica u supermarketima

Cilj: identifikovati proizove koji su zajedno kupovani od strane većeg broja kupaca

Pristup: uvesti bar-kod čitače na kasama da bi se prikupili podaci na osnovu kojih mogu da se uoče zavisnosti

Klasično pravilo

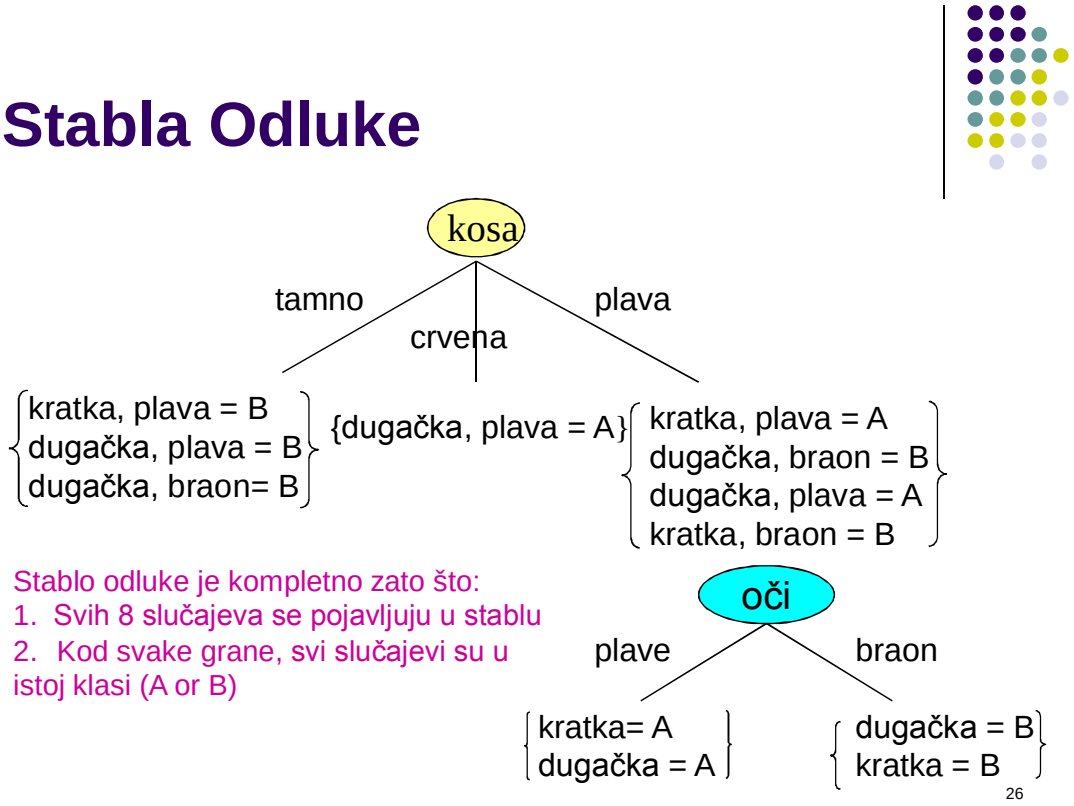
Ako kupac kupuje pelene i mleko, vrlo verovatno će kupiti i pivo

Posledica: blizu pored pelena staviti pakete sa pivom

4.3.7 **Pridruživanje (Association Rules)** - su pravila oblika AKO-TADA. Pokazuje verovatnost da jedan događaj veže uz sebe drugi događaj. Analiza pridruživanja (analiza veza, analiza afiniteta) označava otkrivanje obrazaca koji opisuju međusobno čvrsto povezane osobine podataka. Obično se predstavljaju preko implikacije ili osobina podskupova. Kako prostor koji se pretražuje raste eksponencijalno, cilj je izdvojiti najinteresantnije obrasce na najefikasniji način.

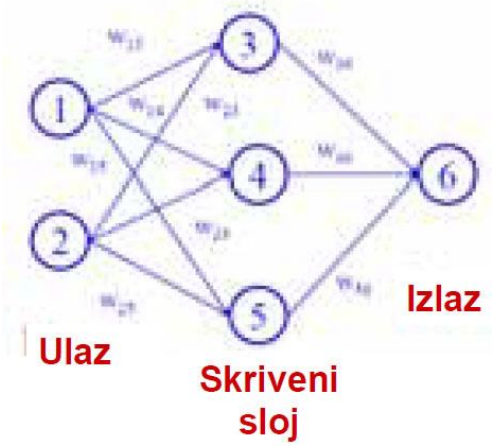
4.4 Nove izvedene metode rudarenja podataka

4.4.1 **Stabla odluke** - su serije pravila pomoću kojih se izdvajaju određene klase ili vrednosti. Mogu se podeliti kandidati za kredit na „dobre“ ili „loše“. Prva komponenta u stablu odluke jeste koren stabla koji suštinski predstavlja test koji treba uraditi. Suština tog testa je grananje stabla na podstabla od kojih svako predstavlja jedan od mogućih odgovora. Zavisno od algoritma svaki nod stabla može imati dva ili više podstabla. Svako podstablo vodi do noda koji predstavlja novu tačku odluke i tako se ide do dna stabla – do nodova koji se nazivaju listovi, koji predstavljaju završnu klasifikaciju i predstavljaju odgovor na postavljeno pitanje.



Slika 4.6 Primer stabla odluke

4.4.2 **Neuronske mreže -** deluju slično ljudskom mozgu. One se upotrebljavaju u analizi rizika i prognoziranju npr. vrednosti deonica. Rudarenje podataka temeljeno na ovoj metodi počinje «učenjem» mreže pomoću podataka za koje je poznata vrednost koju želimo prognozirati. Nakon toga naučeno znanje se proverava. Postupak učenja i provere ponavlja se sve dok rezultati provere ne budu zadovoljavajući. Problem neuronske mreže je to što ona funkcioniše kao crna kutija, pa tako ona daje odgovore, ali ne daje objašnjenja, pa korisnik ne mora strogo da veruje metodu neuronske mreže. Za upotrebu neuronskih mreža, potrebna je intenzivna obuka, jer se ona razlikuje od klasičnih statističkih metoda.



Slika 4.7 Primer neuronske mreže

Primer: Napraviti neuronsku mrežu koja izračunava rizik davanja kredita:

1) Ulazi: Primanja, veličina kredita, period kredita

2) Izlaz: Verovatnoća da klijent neće vratiti kredit [0,1]

3) Trening skup: Podaci o prošlim mušterijama sa poznatim rezultatima

Procene su inicijalne vrednosti težina. Nezavisne promenljive za prvi zapis se ubace na ulazne čvorove 1 i 2. Izlaz se uporedi i izračuna greška. Procene težina se ažuriraju prema odgovarajućem algoritmu (Back-propagation).

Ponavalj se sa sledećim trening zapisom, dok model ne konvergira (nema više značajnih promena težina).

Dobre strane neuronskih mreža:

1) Široke su primene, daju dobre rezultate u veliokom broju komplikovanih problem.

Loše strane neuronskih mreža:

1) Neuronske mreže ne pomažu razumevanju podataka

2) Ulazne i izlazne vrednosti najčešće moraju biti normalizovane u odgovarajuće intervale

4.4.3 **Genetski algoritmi** - su temeljeni na ideji imitiranja pravila biološkog razvoja. Koriste se za optimizaciju i mašinsko učenje pri rešavanju problema. Primena genetskih algoritama je moguća u slučajevima kada problem možemo opisati kao pretraživanje ili optimizaciju proizvoljnih podataka, te poznajemo način mogućeg merenja uspešnosti svakog pojedinog rešenja.

Karakteristike algoritma:

- model
- davanje prednosti nekom kriterijumu radi izbora najboljeg modela
- tehnike za pretraživanje podataka

4.4.4 **Tekst rudarstvo (Text Mining)** - ponekad naizmenično nazivaju tekst data mining, odnosi se na proces izvođenja kvalitetnih informacija iz teksta. Kvalitetne informacioje obično se izvode kroz osmišljavanje uzoraka i trendova putem sredstava kao što su statistički uzorak za učenje. Tekst rudarstvo obično uključuje proces strukturiranja unosa teksta (obično raščlanjivanjem, uz dodatak nekih izvedenih jezičnih karakteristika i uklanjanje drugih, I kasnije ubacivanje u bazu podataka), iz čega proističu uzorci unutar strukturiranih podataka, te konačno procenu I tumačenje izlaza.

5. Procesi

5.1 Proces rudarenja podataka

Proces Data Mininga sastoji se od više faza koje jedna osoba samostalno ne može izvršiti. Te faze, odnosno zadatke, možemo posmatrati iz tri perspektive. Prva je problemska perspektiva koja je važna na samom početku i kraju Data Mininga. Sastoji se od odabira problema, njegovog definisanja, procene i primene znanja. Osoba koja obavlja ove zadatke može se jednostavno nazvati korisnik. Druga perspektiva je perspektiva podataka i obuhvata sve zadatke vezane za pripremu podataka za njihovo rudarenje, a obavlja je informatičar. Metodološka perspektiva, kao treća i poslednja, sastoji se od svih zadataka vezanih za analizu podataka, metode selekcije, implementacije, prezentacije i interpretacije rezultata. Osoba zadužena za te zadatke je stručnjak Data Mininga.

1. KORAK: ***Definisanje problema***

Ovo je prvi korak kod Data Mininga. Znači treba se definisati problem koji se želi rešiti. Ako se problem loše definiše, ne može se očekivati da će Data Mining u poslovnom smislu biti uspešan, iako rezultati analize mogu biti adekvatni.

Primer dobrog definisanja problema je primer banke s velikim brojem klijenata koji kasne s otplatama rata hipotekarnih kredita. Ovde banka želi ustanoviti da li je moguće od podataka o zaduženicima predvideti hoće li on kasniti s otplatama ili ne.

2. KORAK: ***Prikupljanje i priprema podataka***

Iskustva iz prakse pokazuju da je najveći trud u procesu Data Mininga potrebno uložiti u ovaj korak. Zavisno o kvalitetu i organizovanosti podataka u organizaciji, uloženi rad u ovoj fazi kreće se od 50-80% celokupnog truda u procesu Data Mininga. Izbor i priprema podataka su kritične aktivnosti za konačni uspeh projekta. Ovaj korak obuhvata izbor izvora podataka, integraciju podataka iz različitih baza podataka, njihovo čišćenje i dodatno sakupljanje podataka koji su potrebni, a nisu trenutno raspoloživi.

3. KORAK: ***Odabir metode Data Mininga***

U ovom koraku se ne može znati najbolja metoda, jer sve zavisi od prirode problema. Npr. neuronske mreže su najbolje obrazloženje neke pojave npr. vrednost kuće, a metoda stabla odlučivanja je pogodna za npr. analizu odobravanja kredita. Znači potrebno je više metoda rudarenja prvo eksperimentisati pa onda izabrati onu najbolju koja daje rezultate.

4. KORAK: ***Učenje***

Ovde metoda uči na podacima. Ovde postoji opasnost da se uz previše učenja dobiju rezultati koji odlično odgovaraju podacima na kojima je algoritam učio i izgradio model, a loše budućim podacima. A cilj modela je da se dobro ponaša na budućim podacima.

5. KORAK: ***Interpretacija rezultata***

U ovom koraku analitičar na kraju odlučuje da li pravilo zanimljivo, relevantno i upotrebljivo za organizaciju.

5.2 Ko su učesnici u procesu Data Mininga?

U procesu Data Mininga poželjno je da saradjuju:

1. informatičar – on brine o pripremi podataka
2. analitičar podataka – to je npr. statističar. On brine o izboru metode i metodološkoj interpretaciji rezultata rudarenja
3. stručnjak – on mora biti poznavalac poslovnog područja, koji definiše poslovni problem, odabirea relevantne podatke, interpretira rezultate i predlaže aktivnosti na temelju rezultata rudarenja
4. menadžer projekta - on planira i organizovano vodi projekat Data Mininga.

5.3 Ko traži stručnjake u Data Miningu?

Istraživanja su pokazala da većina firmi koje traže stručnjaka Data Mininga dolaze iz IT sektora (43%), marketinga (19%), kao i iz bankarskog sektora i osiguranja (17%).

Položaj u preduzeću na koji se zapošljavaju nije jedinstven, no najčešće je to menadžer (20%), stručnjak za Data Mining (16%), statističar i/ili izvršni istraživač (12%), marketinški analitičar (10%) te softverski inžinjer/arhitekt (10%).

Podjednaki broj firmi želi zaposliti osobe sa diplomom (40%) i magisterijem (40%). Prilično veliki broj poslodavaca traži i doktorat (20%). Vezano uz godine radnog iskustva rezultati pokazuju da su jednako traženi oni sa manje od 5 godina iskustva i oni sa 5 do 10 godina iskustva. Samo u 3% slučajeva zahteva se više od 10 godina iskustva.

Zadaci unutar procesa Data Mininga koji se očekuju od zaposlenih su najčešće prezentacija rezultata (29%), izgradnja novih modela (15%) i priprema podataka (13%). Znanje klasifikacijske metode rudarenja podataka je najtraženije (18%). Metode predviđanja i statističke tehnike su jednako zastupljene sa po 8%.

Znanje softvera koje je potrebno imati najviše odlazi na SQL (44%) i SAS (42%). Zanimljivo je spomenuti da specifični softver za Data Mining se vrlo retko spominje. Dodatne veštine su analitičke veštine (42%), konsalting (19%), projektni menadžment (16%), profilisanje korisnika (15%) i iskustvo u području (15%).



Slika 5.1 Upotreba data mining-a

6. Primeri rudarenja podataka

Data mining se koristi još i u:

1) Politici: Data Mining je metoda kojom je U.S. Army uspela identifikovati vođu napada na Twin Towers, 11.9.2001.; a tom se metodom takođe koriste CIA i Canadian Security Intelligence Service.

2) Igrama: već od 60-ih godina u nekim kombinatornim igrama poput šaha.

3) Poslovanju: pomaže u bržem donošenju poslovnih odluka zbog kontakata samo onih klijenata za koje postoji visoka verovatnost da će odgovoriti.

Jedna savremena informaciona tehnologija izaziva jaku asocijaciju na poimanje informacije kao resursa, jednako kao što su to sirovine i energenti do kojih se dolazi rudarenjem u rudnicima, bilo uglja, železne rude i sl. Reč je o rudarenju podataka.

Savremeni računari imaju velike mogućnosti, uz pojeftinjenje računarske opreme od kojih su za Data Mining najbitniji tvrdi disk, radna memorija i procesor. Osnovno poimanje računara vezano je uz podatke, ali uz podatke oni čuvaju i znanje, a iz tog znanja mogu proizvoditi nova znanja. Poznata izreka koja kaže da ko ima informaciju ima moć danas više nije adekvatna. Informacija nije dovoljna, nužno je znanje što učiniti s tim informacijama. Znanje je snaga, pa se kaže da će samo ona privreda biti uspešna u ovom veku koja će znati koristiti sadašnje i proizvesti novo znanje. To znanje je već u računarskim memorijama i bazama podataka ali je gotovo nevidljivo, jer je rasuto poput mrvica u moru brojeva i reči.

Koristeći dostignuća veštačke inteligencije, Data Mining oblikovalo se u tehnološki pokret i biznis. Data Mining je mnogo više od kopanja, nije cilj kopati, već odvojiti korisno od nekorisnog. Kod Data Mininga iz mase podataka treba izdvojiti male delove koji predstavljaju znanje, a onda dodatnom obradom stvoriti novo znanje, pa doći i do novih tehnoloških otkrića. U pojmovnom razlikovanju znanja od podataka, znanje se sastoji od činjenica, pretpostavki i verovanja, ali posebno važno i od načina kako ih upotrebiti.

Rudarenje podataka je znači automatizovano traženje informacija u mnoštvu podataka. Cilj ove metode jest pronalaženje pravila, odnosno uzoraka koji omogućuju pronalaženje veza između uzroka i posledica.

Ogromne baze podataka bogate su podacima, ali i siromašne informacijama koje su skrivene u pohranjenim podacima. Upravo je Data Mining to koje pomaže otkriti važne informacije i znanje utkano u podatke, doprinoseći donošenju odluka, poslovanju i nauci.

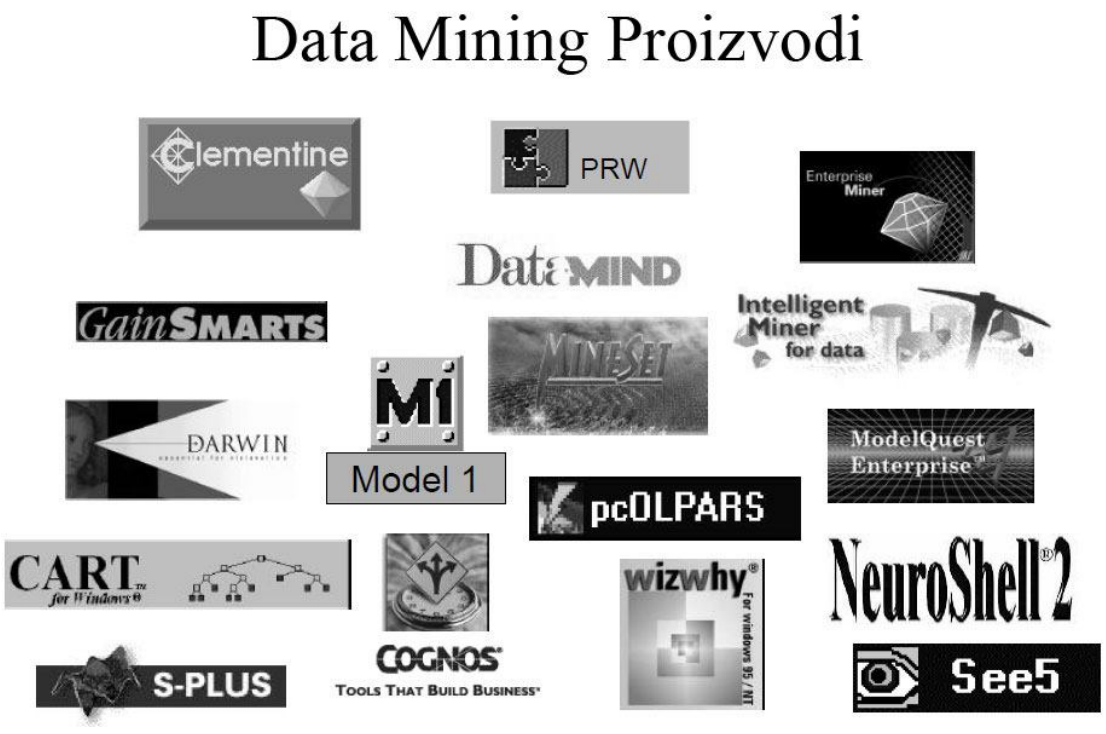
Proces Data Mininga ne može izvršiti samo jedna osoba, jer se on sastoji od više faza, 1. Problemska faza odnosno odabir i definisanje problema, procena zadnjeg i primena znanja. 2. Podatkovna faza: sve vezano za pripremu podataka za Data Mining; obavlja je informatičar 3. Metodološka faza: sve od analize, selekcije, prezentacije do interpretacije rezultata. Stručnjak za Data Mining je osoba zadužena za zadatke koji obuhvataju treću i poslednju fazu procesa rudarenja.

Ova metoda zahteva visok nivo stručnog znanja o metodama. Još uvek se razvijaju alati koji će omogućiti korištenje Data Mininga menadžerima i bez specijaliziranog stručnog znanja, te se radi na edukaciji korisnika ove metode.

7. Alati za rudarenje podataka

7.1 Poznati alati za rudarenje podataka

Tržište alata za rudarenje podataka je poslednjih godina postalo sve više popunjeno, sa više od 50 komercijalnih alata, mnogi od njih su izlistani na sajtu KDNuggets (http://www.kdnuggets.com/software/index.html).
Prema istraživanjima iz maja 2009. godine koje je urađeno na sajtu www.KDnuggets.com[2]. Alat koji je najviše korišćen u period od decembra 2008. Godine do maja 2009. godine bio je IBM-ov alat SPSS PASW Modeler (nekadašnji Clementine), drugi je bio Rapid Miner, treći SAS, četvrti Excel, peti SAS Enterprise Miner.



Slika 7.1 Popularni alati za rudarenje podataka

[2] Sajt koji se predstavlja zajednicu koja se bavi Data Mining-om, Analitčkim alatima, kursevima i dr. Vezano za Data Mining

8. Poređenje alata za rudarenje podataka

Opšte prihvaćeno mišljenje je da svi programi za Data Mining trebaju biti "open source" ali tako ne misle i proizvođači programa, jer bi time konkurencija imala uvid u određene prednosti nekog programa. Kao i u svakoj drugoj sferi programa tako i u Data Mining-u imamo licencirane i ne licencirane (open source) programe. Clementine je licencirani, dok je recimo Weka ne licencirani program.

Clementine	Integral Solutions, Ltd.	http://www.isl.co.uk/clem.html
Darwin	Thinking Machines, Corp.	http://www.think.com/html/products/products.htm
DataCruncher	DataMind	http://www.datamindcorp.com
Enterprise Miner	SAS Institute	http://www.sas.com/software/components/miner.html
GainSmarts	Urban Science	http://www.urbanscience.com/main/gainpage.htm
Intelligent Miner	IBM	http://www.software.ibm.com/data/iminer/
MineSet	Silicon Graphics, Inc.	http://www.sgi.com/Products/software/MineSet/
Model 1	Group 1/Unica Technologies	http://www.unica-usa.com/model1.htm
ModelQuest	AbTech Corp.	http://www.abtech.com
PRW	Unica Technologies, Inc.	http://www.unica-usa.com/prodinfo.htm
CART	Salford Systems	http://www.salford-systems.com
NeuroShell	Ward Systems Group, Inc.	http://www.wardsystems.com/neuroshe.htm
OLPARS	PAR Government Systems	mailto://olpars@partech.com
Scenario	Cognos	http://www.cognos.com/busintell/products/index.html
See5	RuleQuest Research	http://www.rulequest.com/see5-info.html
S-Plus	MathSoft	http://www.mathsoft.com/splus/
WizWhy	WizSoft	http://www.wizsoft.com/why.html

Slika 8.1 ***Neki alati i njihovi proizvođači***

8.1 Poređenje alata za rudarenje po izradi određenih algoritama

Na slici br. 11prikazano je poređenje alata po tome koliko opcija nude prilikom ulaznih i izlaznih parametara.

Ulaz podataka & Izlazni Modeli	Automatsko zaglavlje	Formati Snimanja Podataka	ODB	Ugrađeni drajveri za baze podataka	Pregled Izveštaja	Izlaz Korenog Koda
Clementine	√		√			√
Darwin		√	√			√
DataCruncher	√	√	√	√	√	
Enterprise Miner	√–		√	√	√–	√
GainSmarts	√	√		√	√	√
Intelligent Miner				√–		√
MineSet		√		√		
Model 1	√	√	√	√	√	√
ModelQuest	√			√	√	√
PRW	√	√	√		√	√
CART	√					
Scenario	√				√	
NeuroShell	√					
OLPARS		√				
See5	√–					
S-Plus	√		√		√	√
WizWhy	√				√	

Legenda

prazno	loše
√–	normalno
√	dobro
√+	odlično

Slika 8.2 *Opcije I/O*

Na slici 8.3 prikazano je poređenje alata po tome kakve su im performanse prilikom izrade analitičkih algoritama.

Algoritmi	Stablaodluke	Linijski/	Statistički	Višeljerno	Najbliži	Komšija	Vremenske	Pronalsaežreinjjee	sekvenci	K Means	Pravila	asocijacije	Kohonen
Clementine	√	√		√						√	√		√
Darwin	√			√	√								
Datamind													
Enterprise Miner	√	√		√			√			√	√		
GainSmarts	√	√+											
Intelligent Miner	√	√–		√			√	√		√+	√		
MineSet	√									√	√		
Model 1	√+	√		√						√			
ModelQuest	√	√			√		√–						
PRW		√+		√	√					√			
CART	√												
Cognos	√												
NeuroShell				√+			√–						
OLPARS		√		√	√					√			√
See5	√												
SPlus	√	√+					√			√			
WizWhy													

Slika 8.3 Performanse prilikom izrade algoritama

Na slici 8.4 prikazano je poređenje alata po tome kakve su im performanse prilikom izrade stabla odluke.

Stablo Odluke	"CART"	C5 or C4.5	CHAID	Drugi	Priors	Klasifikacija	Nedostatak podataka
Clementine		√				√	√
Darwin	√				√	√	√
Enterprise Miner	√	√–	√		√+	√	√
GainSmarts	√		√	√			√
Intelligent Miner				√			√
MineSet	√		√			√	√
Model 1	√		√				√–
ModelQuest		√–					√
CART	√+				√	√	√
Scenario				√			√
S-Plus	√						√
See5		√+				√	√

Slika 8.4 Izrada stabla odluke

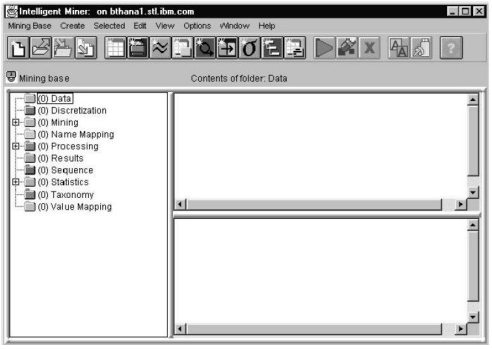
Na slici 8.5 prikazano je poređenje alata po tome koliko su upotrebljivi za krajnje korisnike.

Upotrebljivost Alata	Učitavanje i Obrada podataka	Izrada Modela	Razumevanje Modela	Tehnička Podrška	Ukupno
Clementine	√+	√+	√+	√+	√+
Darwin	√	√	√+	√	√
DataCruncher	√+	√+	√	√	√
Enterprise Miner	√	√	√	√	√
GainSmarts	√+	√	√	√	√
Intelligent Miner	√	√	√	√	√
MineSet	√	√+	√+	√	√+
Model 1	√+	√+	√+	√+	√+
ModelQuest Enterprise	√	√+	√+	√+	√+
PRW	√+	√+	√+	√+	√+
CART	√–	√	√	√	√
Scenario	√	√+	√+	√	√+
NeuroShell	√	√	√	√	√
OLPARS	√–	√	√	√	√
See5	√	√	√	√	√
S-Plus	√	√	√+	√	√
WizWhy	√	√	√+	√	√

Slika 8.5 Generalna upotrebljivost alata

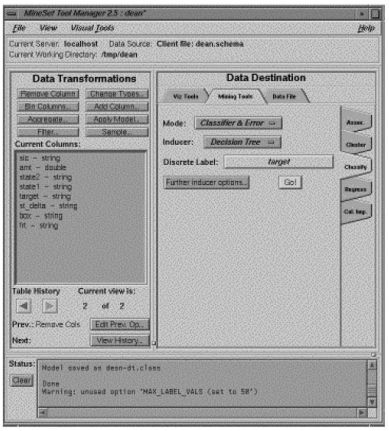
Ispod možete videti interfejs nekoliko različitih alata za rudarenje podataka.

Intelligent Miner: "Explorer" Interfejs



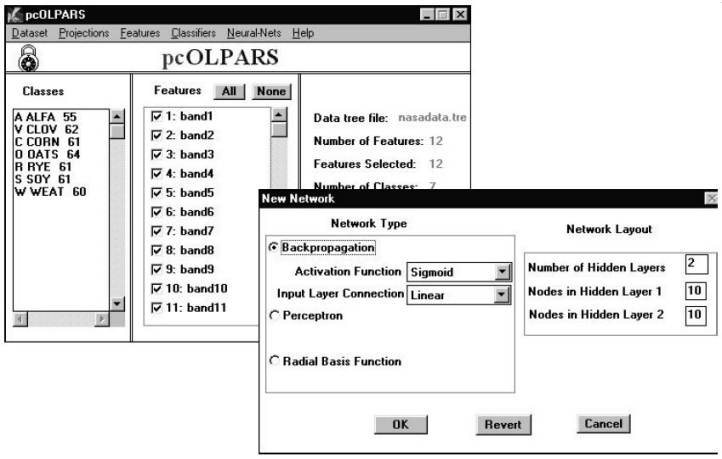
Slika 8.6 Izgled alata Intelligent Miner

MineSet: Interfejs



Slika 8.8 Izgled alata MineSet

OLPAR Interfejs

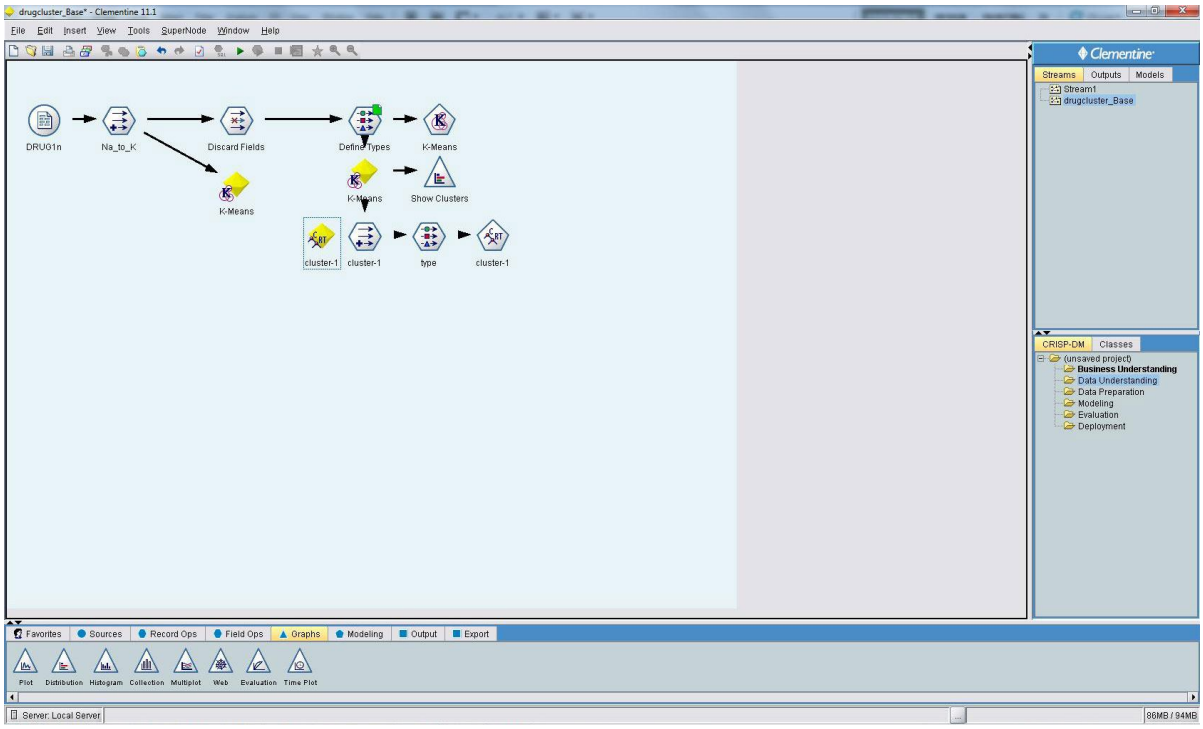


Slika 8.7 Izgled alata Olpar

9. IBM alat za rudarenje podataka SPSS Clementine

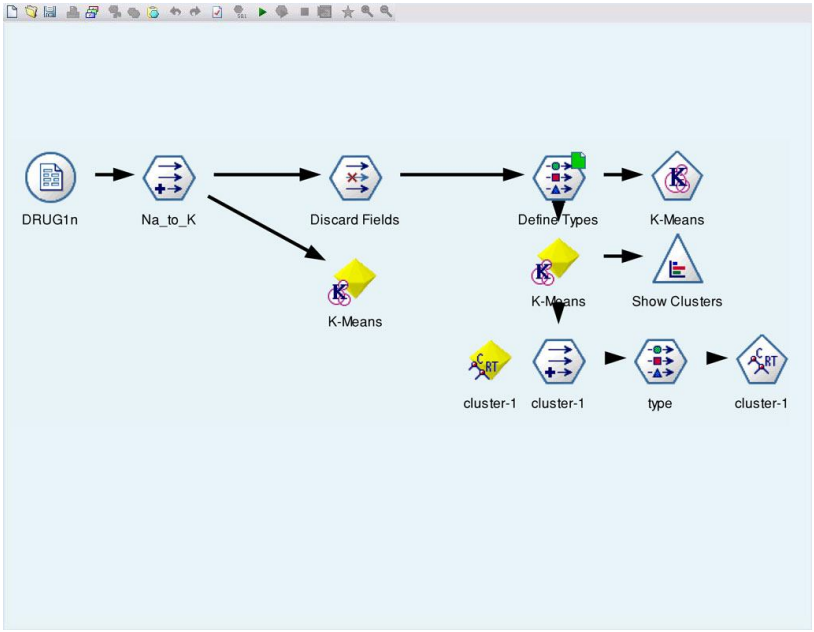
SPSS Clementine je alat za rudarenje podataka, čije okruženje omogućava da brzo razvijate prediktivne modele koristeći stečeno znanje. Vrlo lako možete implementirati u poslovne operacije da bi poboljšali donošenje odluka. Modeler podržava kompletan proces data mining-a, od samih podataka do boljih poslovnih rezultata. Aplikacija Clementine uključuje segmentaciju potrošača/ profilisanje za kompanije, detekciju prevara, kreditna bodovanja prognozu opterećenja za komunalne firme i predviđanje profita za preprodavce.

9.1 Izgled Clementine ekrana



Slika 9.1 Clementine ekrana

Površina u sredini je takozvana (Stream canvas) slika br.16. To je glavni deo aplikacije koji se koristi za sam rad u Clementine.



Slika 9.2 Stream canvas

Većina alata za unos podataka i modelovanje se nalazi u paletama(slika br. 17). U donjem delu se nalaze palete. Svaka paleta sadrži delove koji grafički prikazuju data mining procese, kao što su pristup i filtriranje podataka, izrada grafikona I modela.

Da bi ste dodali neki čvor na radnu površinu (Canvas) samo uradite „drag and drop“ ili „double ckick“ na određeni node[3] iz palete. Zatim ih možete povezati i napraviti stream. Stream predstavlja protok podataka.

Tako pod karticom “Sources” imate tipove podataka koji mogu da budu uvezeni u process rudarenja, kao što su: SQL Database[4], Var. File[5], SPSS File, Excel.

Rekord Ops paleta predstavlja operacije na data snimcima, kao što su selektovanje, integrisanje i dodavanje.

Field Ops nodovi vrše operacije nad poljima podataka, kao što su filtriranje,

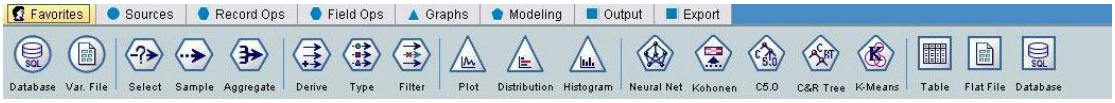
[3] Predstavlja operaciju koja će biti uradjena na podacima
[4] Vrsta baze podataka
[5] Predstavlja mogućnost uvoza podataka iz bilo kog tekstualnog fajla

izvlači nova polja, i određuje tipove podataka za data polja.

Graphs predstavlja grafičke prikaze podataka pre I posle modeliranja. Grafici uključuju plotove, histograme i Web nodove.

Modeling nod predstavlja algoritme za modelovanje dostupne u aplikaciji, kao što su Neuronske mreže, Stabla odluke i Klasterovanje.

Output nodovi predstavljaju različite vrste izlaza, grafikone i rezultate koji mogu biti pregledani u klementinu ili izvedeni direktno u neku drugu aplikaciju, kao što su SPSS ili Excel.



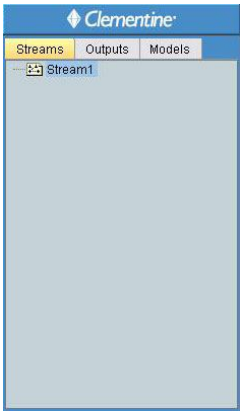
Slika 9.3 Izgled paleta

Clementine box gde se nalaze skraćenice za brži i jednostavniji proces data mining-a Slika 9.3.



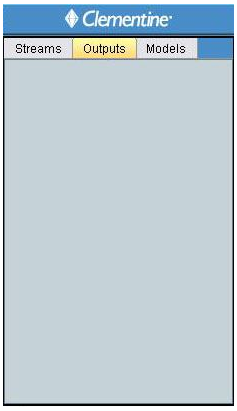
Slika 9.4 ToolBox

U gornjm desnom uglu se nalazi output i object menadžer. Ovi jezičci se koriste da bi videli i manipulisali raznim Clementine objektima.
Streams jezičak zadrži sve strimove korišćene u trenutnoj sesiji. Mogu se dodati zatvoriti i snimiti, kao i dodati u postojeći projekat.



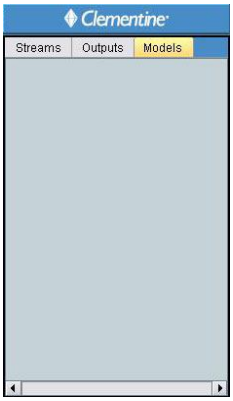
Slika 9.5 Streams

Output jezičak sadrži različite fajlove proizvedene u toku operacija u klementinu. Možete ih prikazati, preimenovati, zatvoriti jezičakele, grafikone i izveštaje izlistane ovde.



Slika 9.6 Output

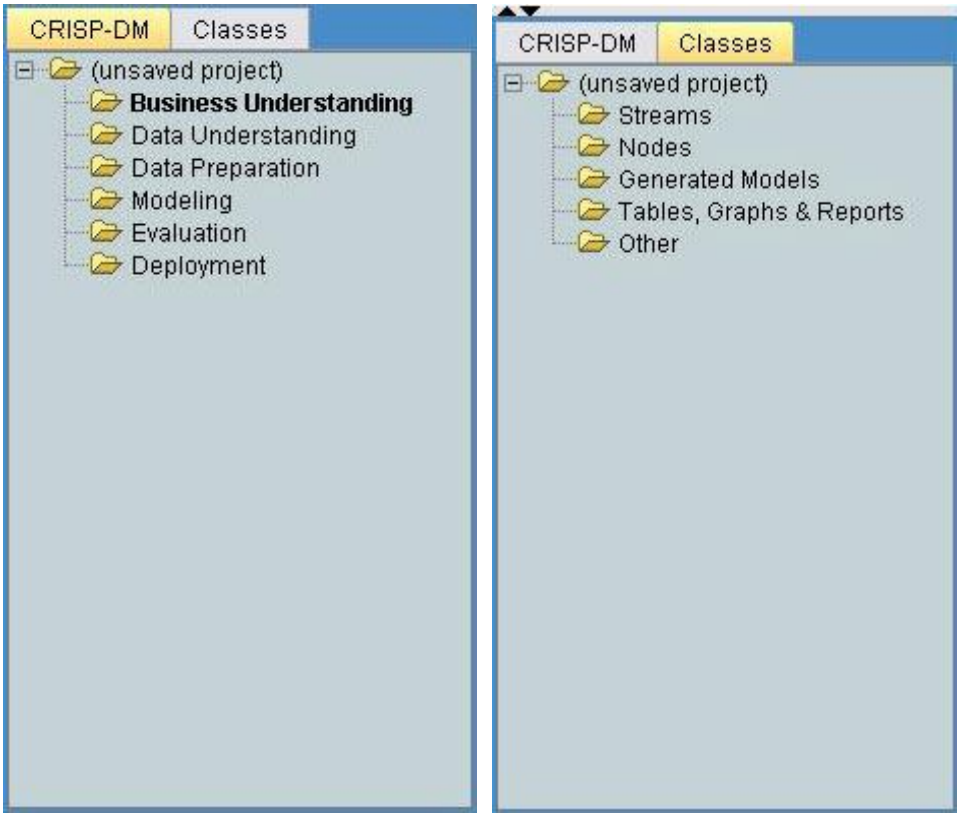
Models jezičak je moćan alat koji sadrži sve generisane modele (modeli koji su napravljeni u klementinu) za određenu sesiju.



Slika 9.7 Models

U donjem desnom uglu se nalazi project jezičak, koji se koristi da mi kreirali I manipulisali projektima.

Postoje dva načina da vidite projekte koje ste napravili u Klementinu. Klasni prikaz I CRISP-DM prikaz.



Slika 9.8 Project prikaz Klasni prikaz

CRISP-DM prikaz nudi mogućnost organizovanja projekata prema Cross-Industry Standard Process for Data Mining.

CRISP-DM predstavalja procesni model rudarenja podataka koji opisuje često korišćene pristupe od strane eksperta za rešavanje problema. Anketa sprovedena 2002, 2004 i 2007 godine pokazuje da je do vodeća metodoligija korišćena od strane data majnera.

Za one koji su stručni i one koji se prvi put bave rudarenjem, CRISP-DM je alat koji će vam pomoći da bolje organizujete svoj trud.

Klasni prikaz nudi mogućnost da organizujete svoj posao u Clementine-u kategorički. Po tipovima objekata koje ste kreirali. Ovaj prikaz je značajan kada radite popis podataka, strimova ili modela, itd.

9.2 Primer korišćenja alata Clementine

Ovaj primer se bavi fiktivnim podacima objašnjavajući artikle potrošačke korpe koji se kupuju obično zajedno, kao i lične podatke kupaca, na primer ID kupovne kartice.

Cilj je da se nađu grupe kupaca koji kupuju slične proizvode i koji mogu biti demografski karakterizovani po starosti, primanjima, itd.

Ovaj primer prikazuje dve faze rudarenja:

1. C5.0 (Predstavlja model Stabla odluke) uvodi profile kupaca identifikovanih grupa proizvoda.
2. Asocijacija i web display i prikazuje veze između kupljenih proizvoda.

9.3 Proces pravljenja C5.0 algoritma

1. Iz Source palete ubacite Excel u Canvas (Slika 9.9).

Sadržaj Excel kolona koji sam napravio kao primer:ID kupca(Brojevi), Vrednost cele kupovne korpe(Brojevi), način plaćanja (ček,keš,kartica), Pol (m,ž[6]), VlasnikStana (da,ne), prihodi, starost, ječam(Tačno,Netačno[7]), puter(Tačno,Netačno), sir(Tačno,Netačno), grašak(Tačno,Netačno), trešnja(Tačno,Netačno), paradajz sok(Tačno,Netačno), pljeskavica(Tačno,Netačno), slatkiš(Tačno,Netačno), kokakola(Tačno,Netačno), kukuruz(Tačno,Netačno), čokolada(Tačno,Netačno).

[6] Ž je u alatu predstavljeno kao F (Female)
[7] Netačno je u alatu predstavljeno kao F (False)



Slika 9.9 Biranje Excela iz palete Source

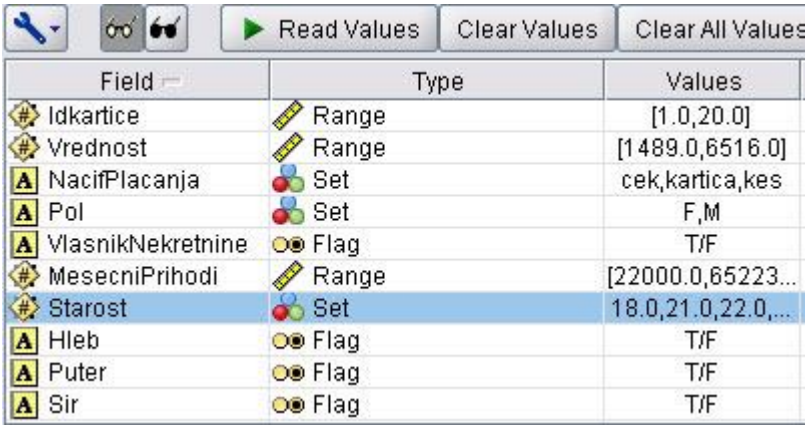
2. Desni klik na ikonicu, pa edit (Ovo može biti urađeno i double klik na nod), kliknite na dugme pored file box-a da nađete željeni fajl.



Slika 9.10 Biranje Excel željenog fajla

3. Kliknite na Types tab, zatim Clear All Values, zatim Read Values (Slika 9.11). Kao što možete vitedi na slici, postoje više tipova podataka kao što su (Range, Set, Flag, itd.)

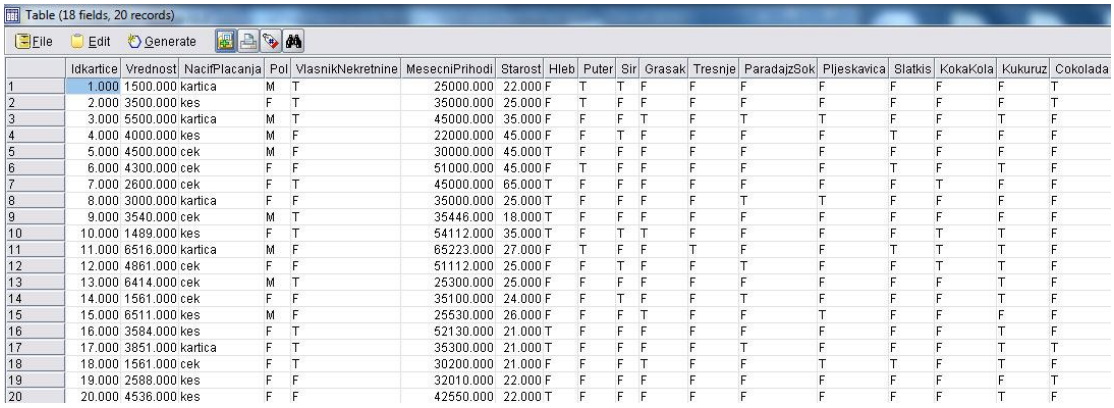
4. Tip polja starost promenite u tip polja Set, Apply, zatim Ok (Slika 25).



Slika 9.11 Učitavanje podataka iz Excel fajla

5. Iz palete Output ubacite Table nod, konektujte ga sa Excel nodom (Desni klik na Excel nod, connect, levi klik na Table nod).

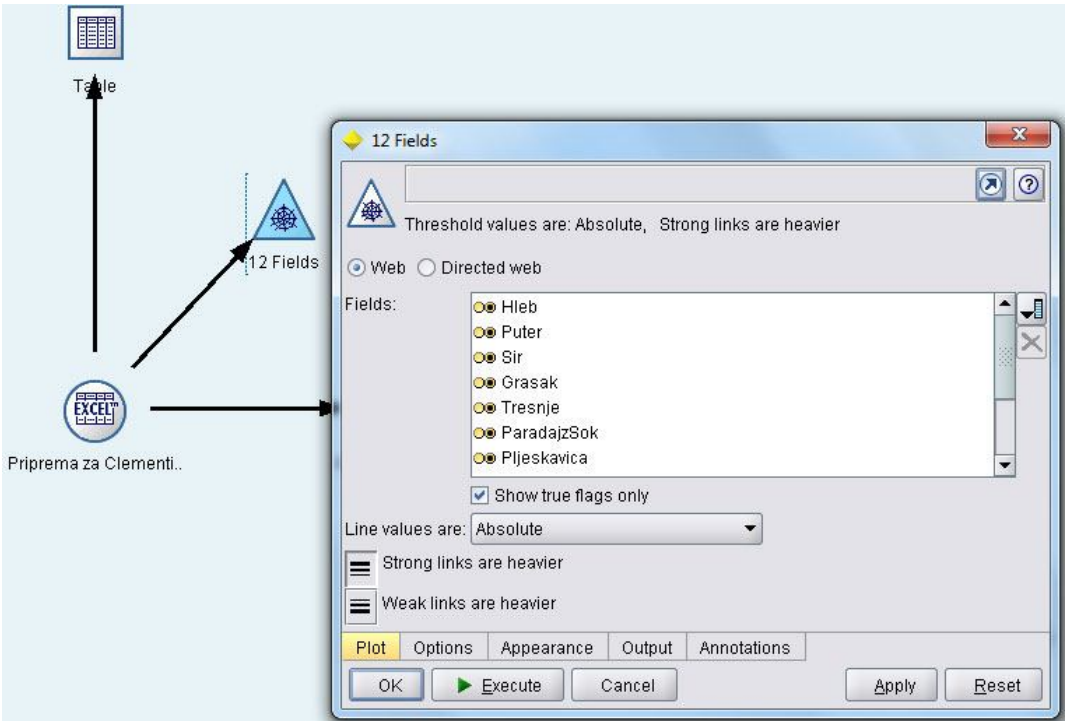
6. Desni klik na Table nod, Execute, da bi ste videli sadžaj podataka (Slika 26).



	Idkartice	Vrednost	NacifPlacanja	Pol	VlasnikNekretnine	MesecniPrihodi	Starost	Hleb	Puter	Sir	Grasak	Tresnje	ParadajzSok	Pljeskavica	Slatkis	KokaKola	Kukuruz	Cokolada
1	1.000	1500.000	kartica	M	T	25000.000	22.000	F	T	T	F	F	F	F	F	F	F	T
2	2.000	3500.000	kes	F	T	35000.000	25.000	F	T	F	F	F	F	F	F	F	F	T
3	3.000	5500.000	kartica	M	T	45000.000	35.000	F	F	T	F	T	T	F	F	T	F	
4	4.000	4000.000	kes	M	F	22000.000	45.000	F	F	T	F	F	F	T	F	F	F	
5	5.000	4500.000	cek	M	F	30000.000	45.000	T	F	F	F	F	F	F	F	F	F	
6	6.000	4300.000	cek	F	F	51000.000	45.000	F	T	F	F	F	F	T	F	T	F	
7	7.000	2600.000	cek	F	T	45000.000	65.000	T	F	F	F	F	F	F	T	F	F	
8	8.000	3000.000	kartica	F	F	35000.000	25.000	T	F	F	F	T	T	F	F	F	F	
9	9.000	3540.000	cek	M	T	35446.000	18.000	T	F	F	F	F	F	F	F	F	F	
10	10.000	1489.000	kes	F	T	54112.000	35.000	T	F	T	T	F	F	F	F	T	T	F
11	11.000	6516.000	kartica	M	F	65223.000	27.000	F	T	F	F	T	F	F	T	T	T	F
12	12.000	4861.000	cek	F	F	51112.000	25.000	F	F	T	F	F	T	F	F	T	T	F
13	13.000	6414.000	cek	M	T	25300.000	25.000	F	F	F	F	F	F	F	F	F	T	F
14	14.000	1561.000	cek	F	F	35100.000	24.000	F	F	T	F	F	T	F	F	F	T	F
15	15.000	6511.000	kes	M	F	25530.000	26.000	F	F	F	T	F	F	T	F	F	F	F
16	16.000	3584.000	kes	F	T	52130.000	21.000	T	F	F	F	F	F	F	F	F	T	F
17	17.000	3851.000	kartica	F	T	35300.000	21.000	T	F	F	F	F	T	F	F	F	T	T
18	18.000	1561.000	cek	F	T	30200.000	21.000	F	F	T	F	F	F	T	T	F	T	F
19	19.000	2588.000	kes	F	F	32010.000	22.000	F	F	F	F	F	F	F	F	F	F	T
20	20.000	4536.000	kes	F	F	42550.000	22.000	T	F	F	F	F	F	F	F	F	T	F

Slika 9.12 Učitavanje podataka iz Excel fajla

7. Iz palete sa grafikonima ubacite WEB, konektujte source fajl sa Web grafikonom, desni klik na Web Node i edit. Web node predstavlja grafički prikaz asocijacije između dva proizvoda. Predstavlja jačinu povezanosti između 2 ili više polja. Dok debljine linija na grafikonu predstavljaju jačinu tih međusobnih veza.

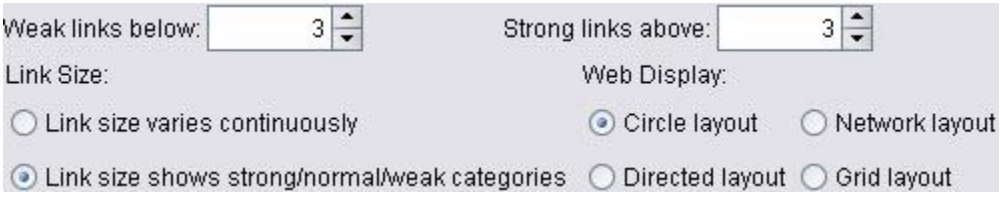


Slika 9.13 Web nod

8. Sa gornje desne strane vam se nalazi ikonica , obeležite sve proizvode koji su u kupovnim korpama, samo polja čiji su tipovi Set I Flag mogu biti prikazani u Web plotu, kliknite OK (Slika 9.13).

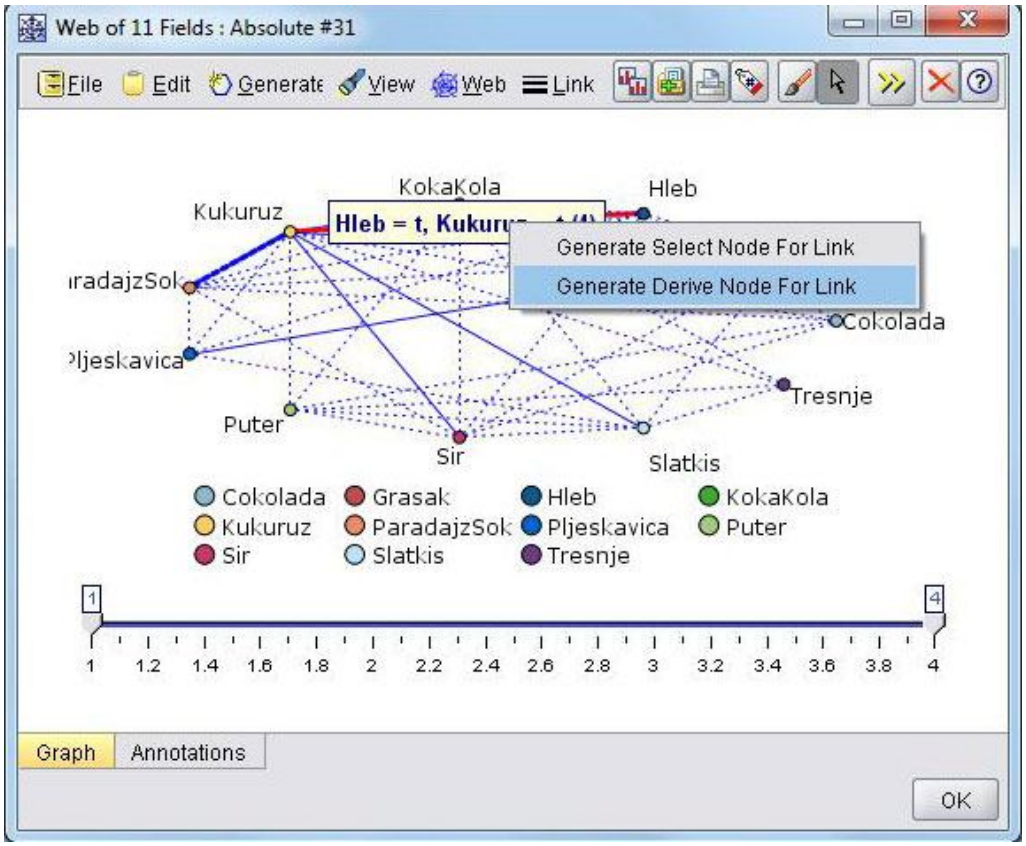
9. Štiklirajte Show true flags only (Slika 9.13).

10. Klik na Options tab, promenite Weak links below i Strong links above po tome kako želite da vam se grafikon pokaže. Prvi broj predstavlja osnovu iz koje se govori alatu da su slabe veze sve ispod vrednosti br. 3, strong links predstavlja sve vrednosti iznad zadatog broja. Ja sam stavio da mi Weak links bude 3, Strong links 3. To znači da će mi na grafikonu više podebljane biti konekcije između proizvoda čije su veze jače od 3 (Slika 9.14).



Slika 9.14 ***Opcije Web Noda***

11. Kliknite na radio dugme, Link size shows strong/normal/weak categories, da bi ste videli 3 različite jačine veza između proizvoda (Slika 9.14). Zatim Execute da bi ste videli grafikon (Slika 9.15).



Slika 9.15 Web grafikon

12. Možemo primetiti da se izdvajaju dve grupe korisnika na osnovu toga koje proizvode kupuju, oni koji kupuju Paradajz sok i Kukuruz, oni koji kupuju hleb i kukuruz (Slika 29).

13. Verovatno bi želeli da vidimo koji to korisnici kupuju ove vrste proizvoda. To je demografski profil. Ovo možemo uraditi tako što svakog korisnika koji kupuje određeni proizvod označiti sa „zastavicom“, za svaku grupu posebno. Zatim ćemo pomoću algoritma C5.0 napraviti profile bazirane na pravilima ovih „zastavica“.

14. Ajde da nađemo kupce koji kupuju kukuruz i hleb. Desni klik na podebljanu liniju između dva kukuruza i hleba, zatim Generate Derive Node for Link (Slika 29). Ovo će napraviti novi nod na radnoj površini u kome će biti proizvodi koje smo označili sa zastavicom.

15. Desni klik na source nod, connect sa izvedenim nodom (Možete povezivati nodove, tako što obeležite nod sa kog počinje veza držimo Alt + držanje levog klika).

16. Edit izvedeni nod, promenite Derive field u kukuruz i hleb (umesto T_T).

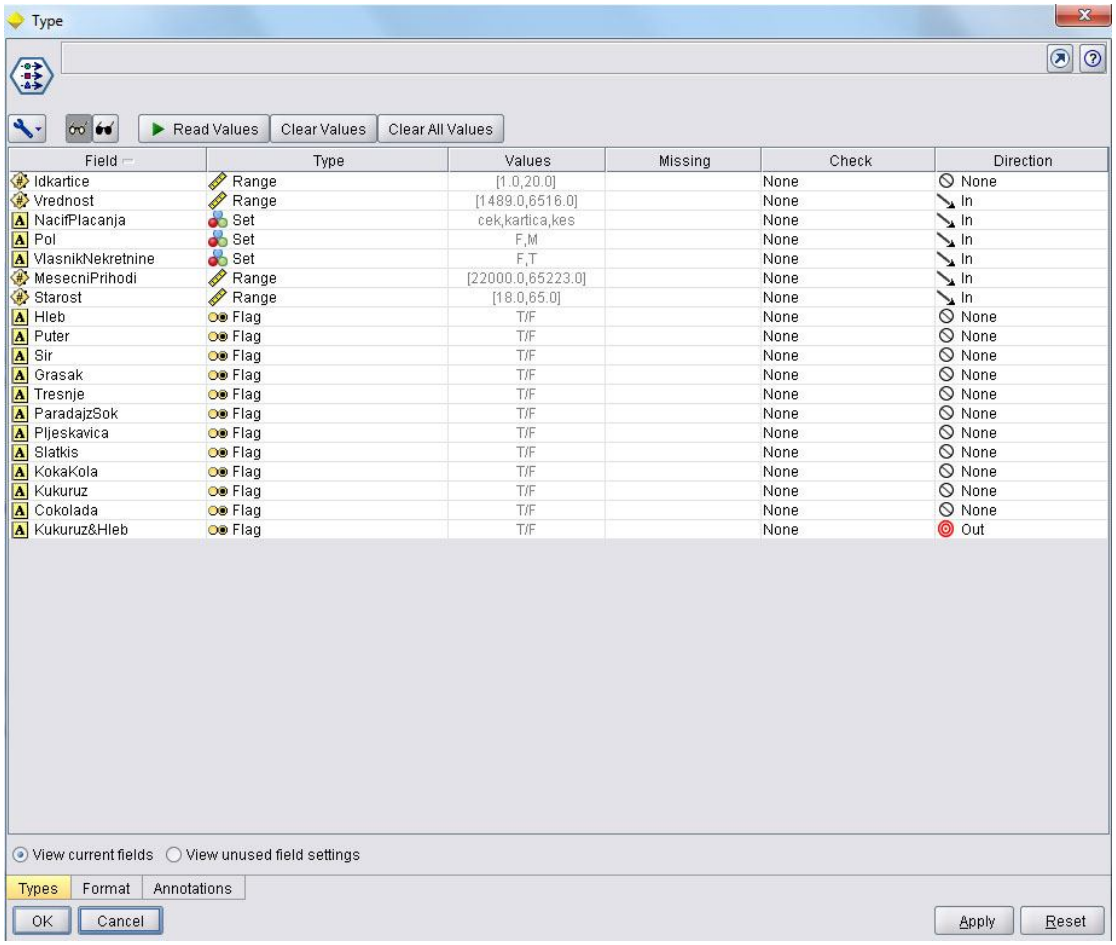
17. Iz Output palete postavite Table nod na radnu površinu. Povežite ga sa kukuruzom i hlebom. Desni klik pa Execute, kao poslednja kolona pojaviće se prikaz T za svakoga ko je kupio u prodavnici Kukuruz i Hleb zajedno (Slika 9.16).

Hleb	Puter	Sir	Grasak	Tresnje	ParadajzSok	Pljeskavica	Slatkis	KokaKola	Kukuruz	Cokolada	Kukuruz&Hleb
F	T	T	F	F	F	F	F	F	F	T	F
F	T	F	F	F	F	F	F	F	F	T	F
F	F	F	T	F	T	T	F	F	T	F	F
F	F	T	F	F	F	F	T	F	F	F	F
T	F	F	F	F	F	F	F	F	F	F	F
F	T	F	F	F	F	F	T	F	T	F	F
T	F	F	F	F	F	F	F	T	F	F	F
T	F	F	F	F	T	T	F	F	F	F	F
T	F	F	F	F	F	F	F	F	F	F	F
T	F	T	T	F	F	F	F	T	T	F	T
F	T	F	F	T	F	F	T	T	T	F	F
F	F	T	F	F	T	F	F	T	T	F	F
F	F	F	F	F	F	F	F	F	T	F	F
F	F	T	F	F	T	F	F	F	T	F	F
F	F	F	T	F	F	T	F	F	F	F	F
T	F	F	F	F	F	F	F	F	T	F	T
T	F	F	F	F	T	F	F	F	T	T	T
F	F	F	T	F	F	T	T	F	T	F	F
F	F	F	F	F	F	F	F	F	F	T	F
T	F	F	F	F	F	F	F	F	F	T	F

Slika 9.16 Web grafikon

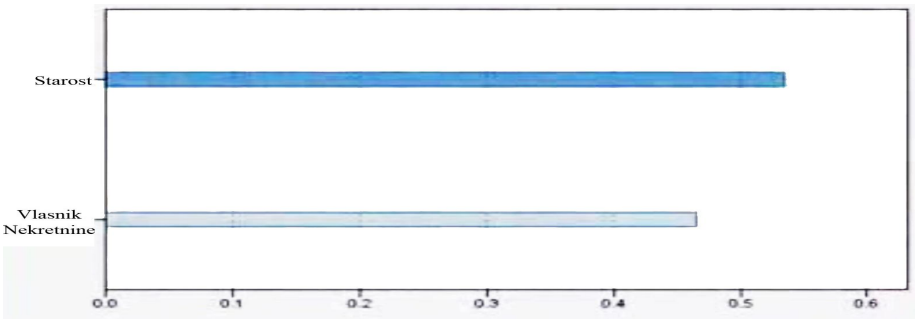
18. Iz palete Field Ops dodajemo Type nod i povezujemo ga sa nodom Kukuruz&Hleb.

19. Edit na type nod, za polje ID i polja svih proizvoda stavite NONE, za novo dobijeno polje Kukuruz&Hleb stavite Out (Slika 9.17).



Slika 9.17 Edit Type Nod

20. Dodajte iz Modeling palete, C5.0 nod (Algoritam za Klasifikaciju) i povežite ga sa Type nodom. Desni klik na C5.0 nod/Edit, kliknite na Rule set / Execute.



Slika 9.18 Browse C5.0

21. Model je stvoren u gornjem desnom uglu pod karticom models i zove se Kukuruz&Hleb. Desnim klikom na model/browse, dobijamo jasne demografske profile kupaca koji su kupili kukuruz i hleb (Slika 9.18) .

9.4 Proces pravljenja GRI modela

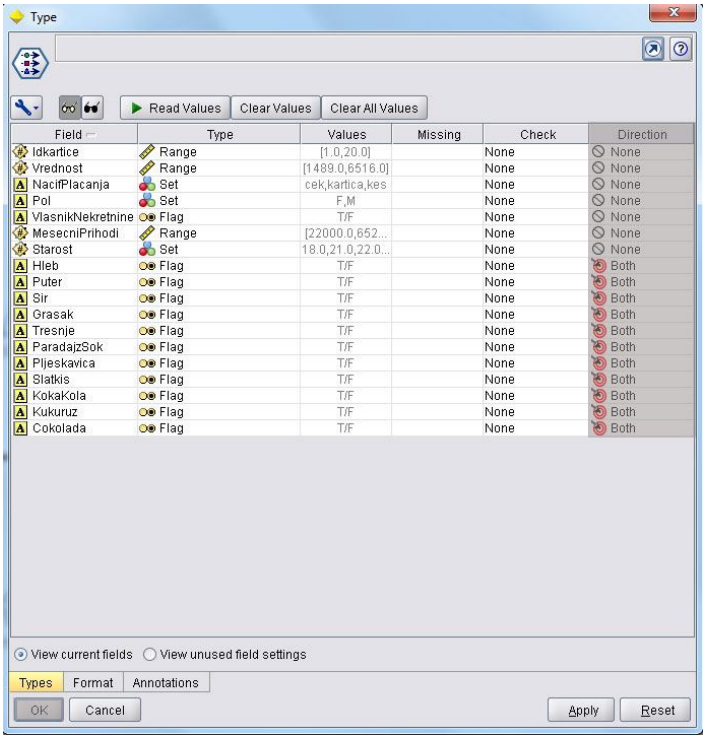
Sada ajde da nađemo vezu između kupljenih i potencijalno kupljenih proizvoda od strane kupaca koristeći GRI[8] (Generalized Rule Induction). GRI pronalazi asocijacije u podacima. Na primer, kupci koji kupe Gillete i after shave obično vole da kupe i kremu koja se koristi posle brijanja.

a) Dodajte Type nod I povežite sa source nodom.

b) Desni klik/edit, postavite tako da sva polja koja se tiču proizvoda budu Both, svi ostali None (Slika 9.19).

c) Dodajte iz modeling palete GRI nod I povežite sa type nod.

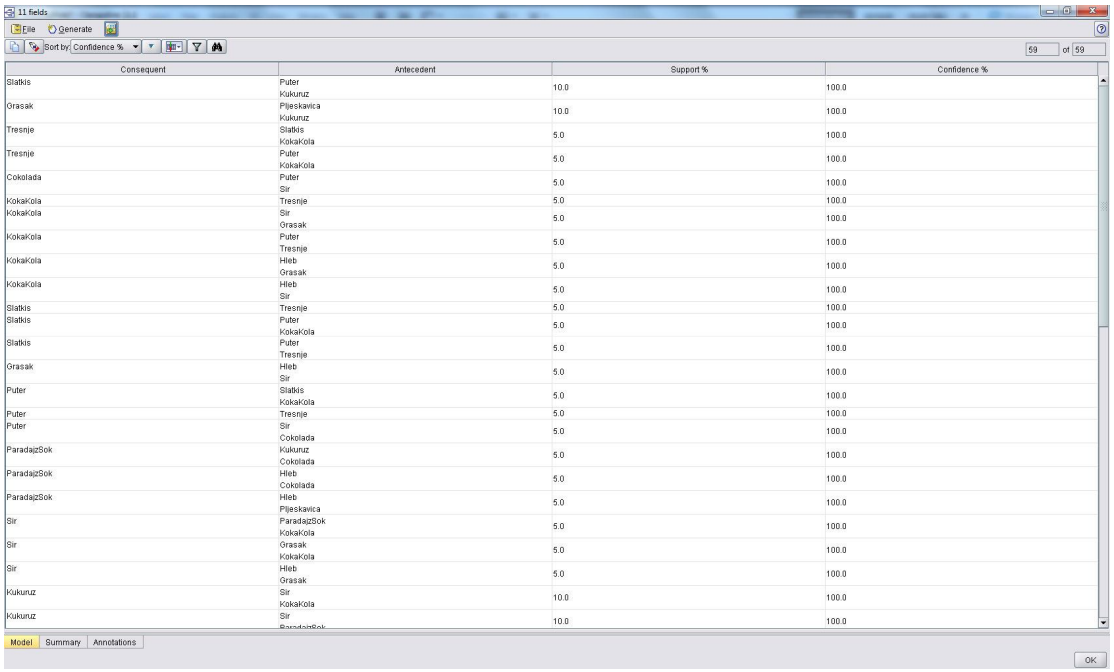
[8] Model Asocijacije



Slika 9.19 Type nod

d) Desni klik/edit na GRI nod, u tabu Model stavite da Maximum number od antecedents (prednost) bude 2/ Execute model.

e) Rezultat će biti model koji sadrži pravila asocijacije (Slika 9.20).



Slika 9.20 GRI model

Iz grafikona (slika 34) možemo videti primer kupaca hleba, koji su dosledni da kupe i kukuruz. Ako postavimo kukuruz pored hleba povećaćemo prodaju kukuza za 40 procenata. Ova informacija je 50% istinita.

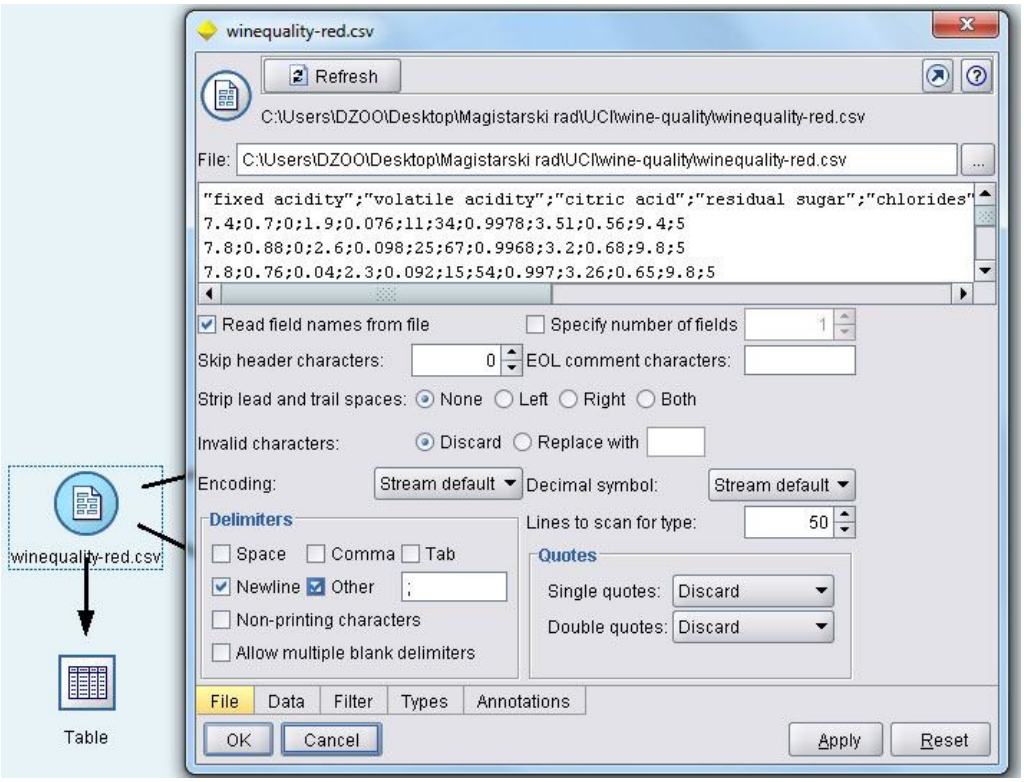
10. Primena Clementine alata nad realnim podacima

U ovoj studiji slučaja ćemo koristiti realne podatke koji su izvedeni iz hemijske analize vina. Prvi vlansik ovih podataka je Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy. Podaci su uzeti sa web adrese: http://archive.ics.uci.edu/ml/machine-learning-databases/wine/ .

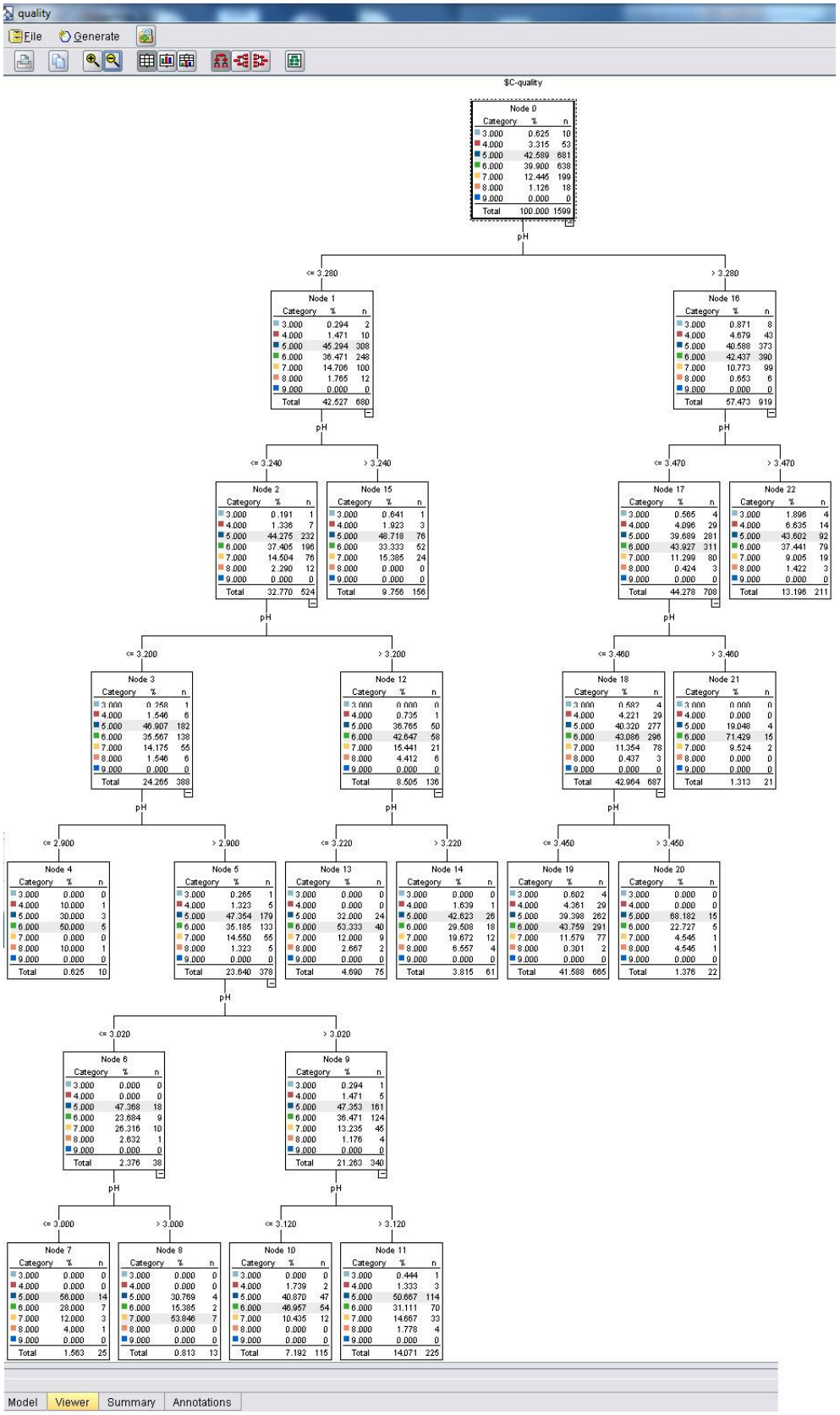
10.1 Ulazni podaci i njihov pregled

Na radnoj površini je napravljen Var.Node, i u njega importovan fajl sa podacima. Da bi ste pregledali podatke korišćen je table nod iz palete **output**.

Prilikom importovanja podataka obraćena je pažnja na format u kome se nalaze podaci. Naime svi podaci su izvornom fajlu podeljeni ne po standardu sa (,), nego sa (;). Na slici koja sledi možete primetiti da je u okviru grupe Delimiters, štiklirano polje u okviru grupe **other** i pored njega upisan u text box znak (;). Ovim smo stavili do znanja alatu da je u izvornom fajlu način odvajanja pojedinačnih podataka gore pomenuti znak.



Slika 10.1 Učitavanje podataka iz fajla

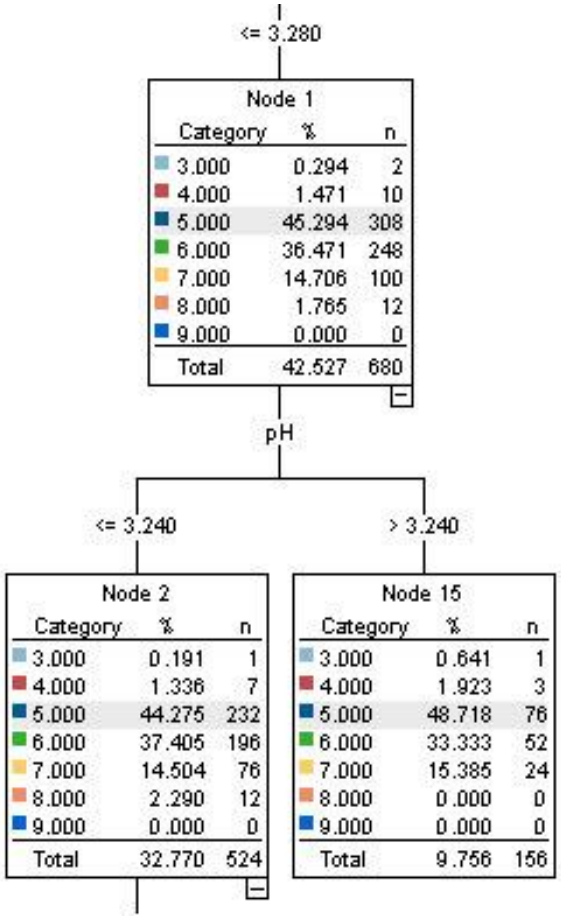


Slika 10.2 Kompletan izgled klasifikacije kroz Stablo odluke

10.2 Klasifikacija podataka

Da bi uradili klasifikaciju po bilo kome kriterijumu, potrebno je da na radnu površinu „ubacimo“ nod C5.0 iz **modeling** palete, koji predstavlja decision tree klasifikacije.

Klasifikacija je urađena tako što je kao cilj klasifikacije postavljen kvalitet vina, a kao ulazni parametar u algoritam Ph vrednost. Ovime smo dobili rezultat takav da je u svakom kvadratu pokazan broj vina određenog kvaliteta koji ima > ili < Ph vrednost od određene vrednosti (Slika 10.2).

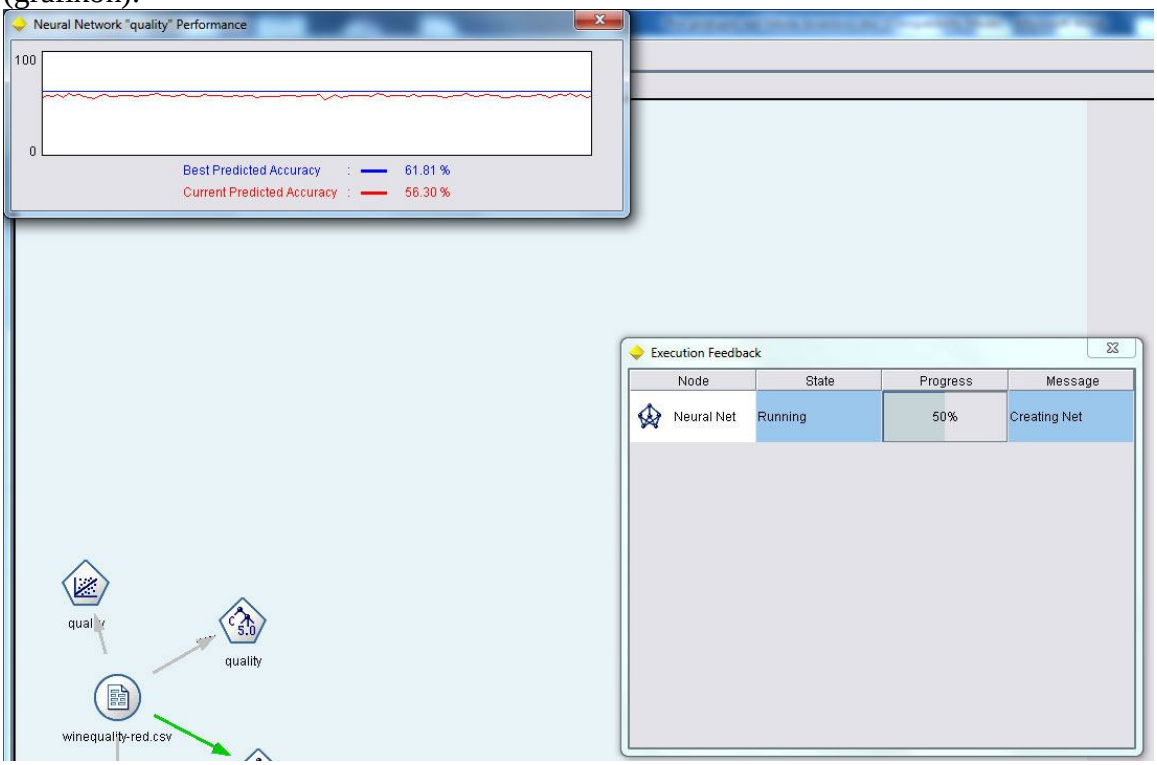


Slika 10.3 Izgled klasifikacije

Na slici 10.3 možemo videti da postoji 196 vina, kvaliteta 6 koja imaju manju ili jednalu Ph vrednost broju 3,24.

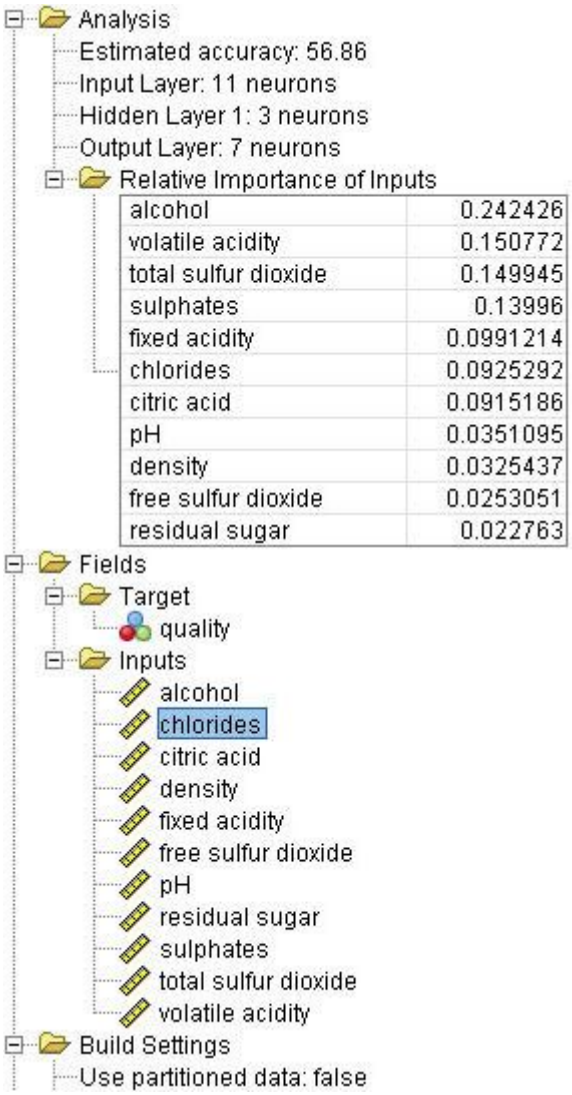
10.3 Neuronske mreže

Da bi saznali koji je od poznatih atributa „najvažniji“ da bi kvalitet vina bio određenog nivoa. Upotrebili smo iz palete modeling **Neural Net** nod. Preciznost ovog rešenja je dovedena na vrednost od 56,86 %. Na slici 10.4 možemo videti kako računar uz pomoć algoritama pokušava da nađe što idealnije rezultate (grafikon).



Slika 10.4 Proces formiranja neuronske mreže

Po rezultatima algoritma neuronsih mreža možemo sa sigurnošću od 56,86% zaključiti da od svih ponuđenih atributa, na kvalitet vina najviše utiče količina alkohola u njemu, zatim po važnosti slede isparljive kiseline (Slika 10.5).



Slika 10.5 Uticaj atributa na kvalitet vina

11. Zaključak

U današnje vreme se nalazi mnogo informacija za koje ni ne znamo da postoje tu pored nas. Imajući u vidu da je informacija danas najbitnija, moramo pronaći način da je nađemo I razumemo. Data mining je jedna oblast koja nam pruža da iz naizgled ne bitnih I nepotrebnih informacija saznamo neke nove informacije koje su nama veoma vredne.
U tome nam ima veliki udeo veoma brz razvoj današnjih tehnologija (procesori, tvrdi diskovi, internet I telekomunikacije).

Rudarenje podataka je znači automatizovano traženje informacija u mnoštvu podataka. Cilj ove metode jeste pronalaženje pravila, odnosno uzoraka koji omogućuju pronalaženje veza između uzroka i posledica.

Proces Data Mininga ne može izvršiti samo 1 osoba jer se on sastoji od više faza, tačnije tri.

Prva faza, tzv. problemska faza odnosno odabir i definisanje problema.

Druga faza pripremna faza: sve vezano uz pripremu podataka za Data Mining; obavlja je informatičar.

Treća faza, metodološka faza : sve od analize, selekcije, prezentacije do interpretacije rezultata. Stručnjak za Data Mining (eng.Data Miner) je osoba zadužena za zadatke koji obuhvataju treću i zadnju fazu procesa rudarenja.

Ova metoda zahteva visok nivo stručnog znanja o metodama. Još uvek se razvijaju alati koji će omogućiti korišćenje Data Mininga menadžerima i bez specijalizovanog stručnog znanja, te se takođe radi na edukaciji korisnika ove metode.

Clementine kao alat za Data Mining se pokazao u praksi kao jedan od najboljih alata sa velikim brojem funkcija koje su od neprocenjivog značaja za data mining.. Alati kao što su Rapid Miner, Weka i dr. Imaju takođe izuzetne karakteristike, ali nisu toliko u upotrebi kao naslednik Clementine alata (SPSS data modeler). SPSS PASW Modeler (ranije nazivan Clementine) ima prednost kao trenutni lider na tržištu, sa jakom distribucijom i podrškom za alat.

Neki alati imaju funkcije koji drugi nemaju ili bolje izrađuju određene modele, ali važi i obratno.

Što se tiče alata za data mining, stručnjak bi morao da zna više različitih alata, u zavisnosti od slučaja, a takođe bi morao da isproba rad svakog ozbiljnijeg alata da bi se uverio koje su mu prednosti a koje mane. Jer u zavisnosti od slučaja do slučaja moraju se birati alati koji najviše pogoduju za datom slučaju radi što temeljnije analize podataka i dobijanje što preciznijih informacija.

Clementine je alat koji nudi mnoštvo funkcija koje su u pozadini odlično usklađene sa samim algoritmima tih funkcija, veoma je korisnički opredeljen i lak za upotrebu, uz veliku preglednost samog toga rada. Videli smo neke osnovne funkcije u samom alatu koje možemo koristiti na mnoštvo slučajeva rudarenja podataka.

Moja preporuka je Clementine, u svakom slučaju kada se može primeniti.

12. Literatura

[1] Seyyed Jamaleddin Pishvayi - Customer Relationship Management, Tehran University, Spring 1383. Dostupno na: http://www.data-miners.com/companion/Chapter14-Example2.ppt

[2] Dr. Dubravko Ćulibrk - Poslovna inteligencija: Eksploatacija informacija i skladištenje podataka

[3] Prof. Dr Milan Milosavljević i Prof. Dr Angelina Njeguš - Predavanja SINGIDUNUM univerzitet, 2009/2010

[4] Bojan Ćirić, Poslovna Inteligencija, Beograd, 2006

[5] http://en.wikipedia.org/wiki/Data_mining 02.12.2011.

[6] Michael Berry and Gordon Linoff, Customer Relationship Management Through Data Mining, SAS Institute, 2000

[7] Michael Berry and Gordon Linoff, Mastering Data Mining, John Wiley & Sons, 2000

[8] IBM Corporation, International Technical Support Organization, Dept. QXXE Building 80-E2, 650 Harry Road, San Jose, California 95120-6099

[9] Doc. Nenad Mitić, Univerzitet u Beogradu, Matematički fakultet, 04.11.2011.

[10] http://archive.ics.uci.edu/ml/index.html 16.12.2011.

[11] Most Popular Data Mining Software. Dostupno na: http://www.the-data-mine.com/bin/view/Software/MostPopularDataMiningSoftware

[12] Tool Comparison. Dostupno na: http://www.datamininglab.com/TOOLCOMPARISON/tabid/58/Default.aspx

[13] Data Mining Tools Used Poll. Dostupno na: http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm

[14] Stranica kurseva, asistenta profesora na Iranskom univerzitetu. Dostupno na: http://webpages.iust.ac.ir/yaghini/Courses/Data_Mining_882/Data_Mining_882.htm

[15] An Evaluation of High-end Data Mining Tools for Fraud Detection, Dean W. Abbott, I. Philip Matkovsky, and John F. Elder IV, from the 1998 IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, October 12-14, 1998. Dostupno na: http://datamininglab.com/Portals/0/tool%20eval%20articles/smc98_abbott_mat_eld.pdf

[16] Help u okviru samog programa Clementine 11.1.