Data Mining: Definicije i primena

Današnje baze i skladišta podataka se mere veličinama reda terabajta podataka. U toj masi podataka mogu se kriti informacije od presudnog poslovnog značaja. Postavlja se pitanje kako do tih informacija doći iz mora podataka koji ih sakrivaju?

Odgovor na to pitanje može dati Data Mining.

Data Mining se može bukvalno prevesti kao rudarenje ili kopanje po podacima i verovatno i nema boljeg naziva koji opisuje ovaj proces. U primeni Data Mininga počinjemo od bilo kakvog izvora podataka (biramo mesto za kopanje) i pokušavamo da pronađemo neke zavisnosti, veze ili pravila koja postoje među podacima. Jednom rečju pokušavamo iz ogromnog broja nestruktuiranih i razbacanih podataka da dobijemo neko znanje (grumen zlata) koje će nam koristiti u rešavanju nekog problema ili unapređenju nekog poslovnog procesa. Naravno u tom procesu, kao ni u rudarenju nije zagarantovan uspeh u pronalaženju znanja. Do korisnih informacija možemo doći jako brzo, a može se desiti da posle dugog i napornog kopanja ne pronažemo baš ništa što nama može biti od koristi. Sa druge strane možemo "slučajno" doći do nekog znanja koje možemo primeniti pri rešavanju nekih drugih problema.

Postoji jako puno definicija Data Mining – a, to je i logično, jer je Data Mining jako širok pojam, tako da praktično svaki autor ima svoju definiciju. Navešćemo par najčešćih:

1. " Data Mining je proces otkrivanja značajnih veza, paterna i trendova kopanjem kroz ogromne količine uskladištenih podataka, korišćenjem tehnologija prepoznavanja paterna, kao i statističkih i matematičkih metoda.
2. Data Mining je analiza opservacionih skupova podataka , koja služi da bi se otkrile neočekivane veze i da bi se podaci sumirali na takav način da je razumljiv i koristan vlasniku podataka.

Zadaci (Problemi) Data Mining – a

Najčešći i najpoznatiji zadaci Data Mining-a su:

- Redukcija
- Estimacija (Procena)
- Predviđanje
- Klasifikacija
- Klasterovanje
- Asocijacija

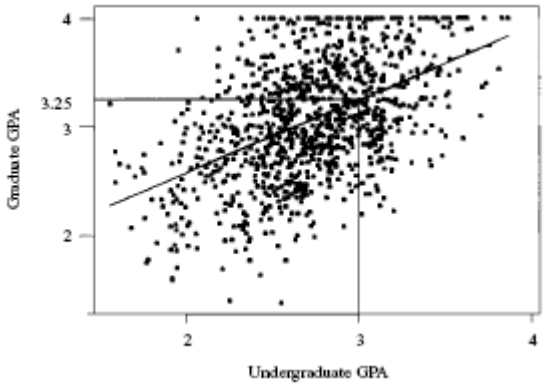
Redukcija

Redukcija predstavlja smanjivanje ili izostavljanje podataka koji nisu od značaja za istraživanje u cilju lakšeg uočavanja veza i zavisnosti između atributa ili objekata. Postoje mnogobrojne metode redukcije podataka. Postoje ručne redukcije i automatske redukcije (npr. Faktorska analiza).

Redukcija može biti redukcija atributa ili redukcija slučajeva (redova, zapisa). DO treba da uspostavi kompromis između želje da radi sa upravljivijim podacima i želje da sačuva tačnost podataka.

Estimacija

Estimacija (Procena, ocena) predstavlja procenu vrednosti određene (egzogene promenljive) na osnovu postojećih (endogenih) promenljivih koje su zabeležene u sistemu.



Na slici možemo videti tipičan primer linearne regresije, koja se koristi kao metoda estimacije u Data Mining-u. Dakle estimacija ili procena daje određeno pravilo ponašanja koje je izvedeno iz postojećih podataka.

Predviđanje

Predviđanje je sličan zadatak kao i procena s tim što se za modelovanje koriste najčešće složeniji, nelinearni modeli, veštačke neuronske mreže i slično. Ono što karakteriše predviđanje je to što se ocena vrši za stanje koje se očekuje u nekom trenutku u budućnosti. Ovo je značajan problem, jer model kojim opisujemo ponašanje i veze u sadašnjem sistemu ne mora važiti i za isti sistem u budućnosti (problem ekstrapolacije.)

Klasifikacija

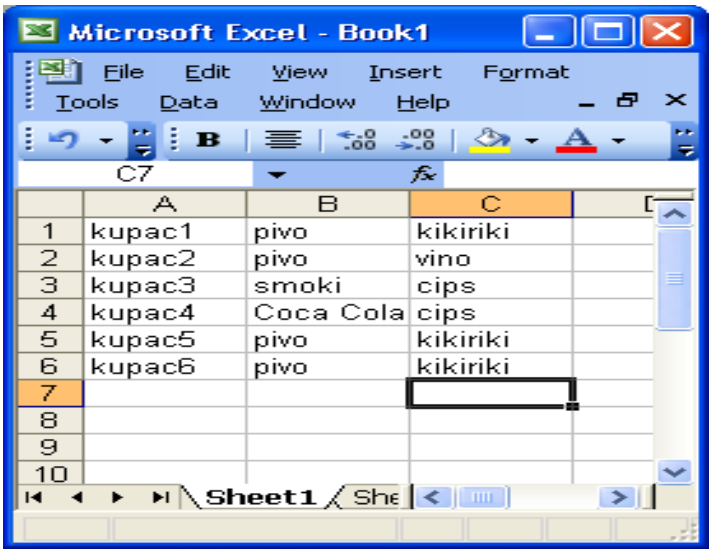
Klasifikacija predstavlja problem raspoređivanja elemenata u predodređene grupe ili klase. Elementi su opisani preko više promenljivih, gde jedna promenljiva (izlazna) označava klasu tog objekta (npr. promenljiva *Podoban za kredit* kod opisa klijenata ili *Bolest* kod opisa bolesnika. Problem klasifikacije je da generiše model koji će na osnovu opisa objekata (ulaznih promenljivih) odrediti klasu tog objekta (izlazna promenljiva).

Klasterovanje

Klaster predstavlja kolekciju elemenata koji su međusobno slični i koji su različiti u odnosu na elemente iz drugih klastera. U skladu sa tim, klasterovanje se bavi grupisanjem elemenata ili opservacija u klase sličnih objekata. Dakle, za razliku od klasifikacije, gde su grupe unapred definisane (apriori), a mi određujemo pripadnost nekog elementa grupi, ovde formiramo grupe aposteriori na osnovu sličnosti elemenata.

Asocijacija

Zadatak asocijacije u Data Mining-u je da pronađe pravila u bazi podataka. U poslovnom svetu asocijacija je poznata kao analiza afiniteta ili analiza potrošačke korpe gde je zadatak otkrivanje pravila po kojima se formiraju veze između dva ili više atributa. Za razliku od klasifikacije, kod zadatka asocijacije ne postoji unapred definisani izlazni atribut (atribut klase ili odluke), već svaki atribut može biti i kao uslov i kao posledica otkrivenog pravila. Asocijacija funkcioniše po principu ***IF atributA THEN atributB*** pravila čija se značajnost određuje na osnovu Podrške(Support) i Poverenja (Confidence). Podrška predstavlja procentualni udeo članova uzorka koji zadovoljavaju određeno pravilo (broj članova koji zadovoljava/ ukupni broj članova), dok Poverenje govori koji procenat članova koji poseduju atributA, poseduju i atributB.

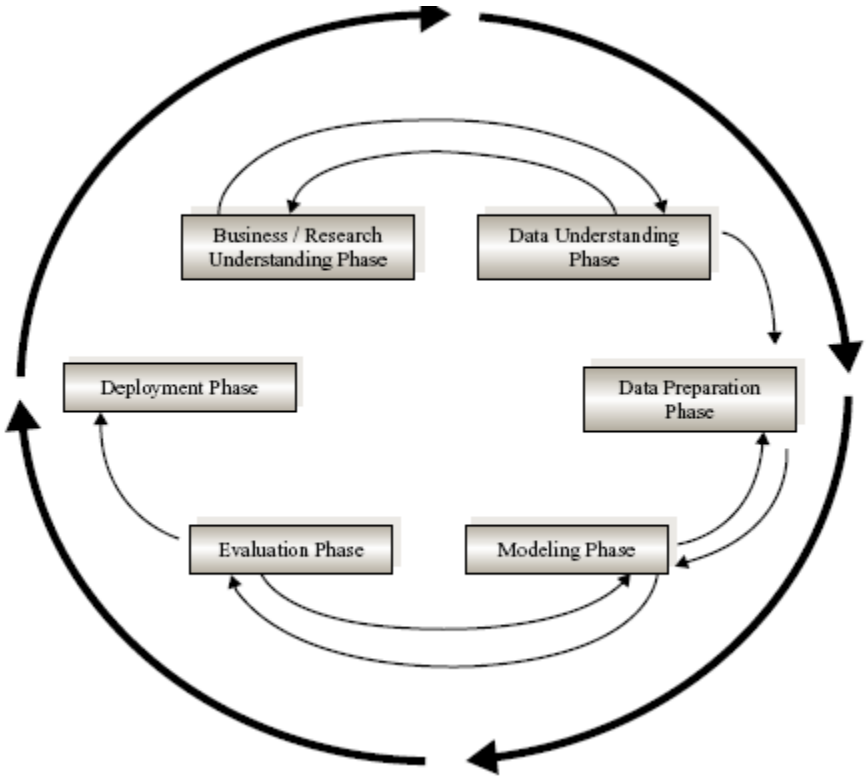


Iz tabele sa slike iznad možemo utvrditi pravilo da ukoliko kupac kupi pivo, kupiće i kikiriki sa podrškom od 66,67% i sa poverenjem od 75%.

CRISP–DM (CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING)

Crisp – DM predstavlja standard za primenu Data Mining-a koji nije obavezan u primeni data mining-a, može se prilagođavati konkretnom problemu po potrebi i odnosi se na generalno rešavanje problema u okviru poslovnih procesa ili istraživanja. Crisp-dm je razvijen 1996. godine od strane analitičara Daimler Srysler-a, SPSS-a i NCR.CRISP-a iz razloga što su mnoge kompanije pristupale problemu Data Mining-a nesistemski pokušavajući da preko noći postignu ogromne rezultate slučajnim pokušajima, što naravno nije davalo nikakve rezultate.

Prema Crisp–DM standardu, životni ciklus Data Mining projekta se sastoji iz 6 faza (slika ispod).



Crisp – životni ciklus Data Mining projekta

Sa slike se može primetiti da je redosled faza prilagodljiv konkretnom projektu. To znači da sledeća faza umnogome zavisi od rezultata prethodne faze. U zavisnosti od ponašanja i karakteristika modela i rezultata tekuće faze, određujemo koja faza nam je sledeća, a ponekad je neophodno ponoviti istu fazu ukoliko nismo dobili zadovoljavajuće rezultate. Iz toga sledi da Crisp predstavlja iterativno – inkrementalni postupak koji obezbeđuje rešenje konkretnog poslovnog ili istraživačkog problema i koji može otvoriti nova pitanja i probleme, koji se potom mogu rešavati korišćenjem istog procesa.

Faze CRISP – DM – a

1. BU - *Business understanding phase*.

Faza razumevanja poslovnog procesa ili faza razumevanja istraživanja u opštem slučaju ima 3 dela:

- Dobro razumevanje ciljeva projekta i zahteva projekta kao celine.
- Formulacija ciljeva i ograničenja u obliku problema Data Mining-a.
- Priprema početne strategije za ostvarenje tih ciljeva.

2. DU - *Data understanding phase*

Faza razumevanja podataka u opštem slučaju ima 4 dela:

- Prikupljanje podataka
- Upoznavanje sa podacima i njihovim vezama na osnovu jednostavnih analiza i prikaza (obično u grafičkoj formi: pite, zvezde...)
- Ocena kvaliteta i tačnosti podataka
- Opciono mogu se izabrati podskupovi podataka za koje smatramo da sadrže primenljive paterne za naš problem

3. DP - *Data preparation phase*

Faza pripreme podataka u opštem slučaju ima 4 dela i po pravilu ovo predstavlja najnaporniju i najtežu fazu CRISP – a.

- Priprema konačnog Data Set – a koji će biti korišćen u narednim fazama iz dobijenih sirovih podataka.
- Izbor promenljivih i njihovih atributa, koje želimo da analiziramo i za koje smatramo da će nam biti od koristi.
- Transformacija određenih promenljivih ukoliko je to potrebno.
- Prečišćavanje podataka i priprema za rad sa alatima za modeliranje.

4. M - *Modeling phase*

Kada su podaci prečišćeni i prilagođeni alatima prelazi se u fazu modeliranja. Koja ima 4 faze i koja je u principu zahteva najmanje rada ukoliko su prethodne faze dobro izvršene tj. ukoliko se dobro razume problem i ako su podaci adekvatno pripremljeni.

- Izbor odgovarajuće tehnike za modeliranje
- Podešavanje parametara modela
- Opciono primena drugih tehnika (uvek treba imati na umu da se različite tehnike mogu koristiti za rešavanje istih problema)
- Ukoliko dodje do problema, vraćanje na fazu pripreme podataka za konkretan algoritam.

5. E - *Evaluation phase*

Kada se izvrši algoritam modela I dobijemo rezultate prelazimo na fazu Ocena, gde ispitujemo tačnost I primenljivost rešenja, pre nego što ga uključimo u naš poslovni ili istraživački problem. Ukoliko su rezultati neodgovarajući ili neprimenljivi, vraćamo se na prethodnu fazu. Ova faza se takođe sastoji iz 4 dela:

- Ocena rezultata jednog ili više modela u smislu efektivnosti I primenljivosti (u slučaju loše ocene vraćamo se na prethodni korak).
- Ustanoviti dali se primenom modela zaista rešavaju problemi definisani u prvoj fazi.

- Ustanoviti da li neki detalji problema nisu dovoljno detaljno obrađeni.
- Donošenje odluke na osnovu rezultata.

6. D - *Deployment phase*

Konačno, kada smo dosli do zadovoljavajućih rešenja i doneli odluku na osnovu rezultata, analize i ocene modela, primenjujemo rešenje u realnom sistemu, na naš poslovni ili analitički problem.

- Ostvarivanje korisne primenue modela (Samo kreiranje modela ne znači da smo došli do rešenja problema. Data Mining utiče na sistem inkrementalno).
- Dati primer proste primene (Kreirati izveštaj)