



Univerzitet Singidunum

**MASTER STUDIJSKI PROGRAM
SAVREMENE INFORMACIONE TEHNOLOGIJE**

MASTER RAD

Primena Data mining u medicini

Mentor:

Prof. Dr Milan Milosavljević

Student:

Nevena Ilić

Br. indeksa: 410059/2015

Sažetak

Data mining je metoda pretraživanja podataka koja se munjevito razvila pre svega zbog razvoja računarske tehnologije. Razvitkom brzih računara dolazi do toga da je moguće efikasno pretraživati ogromne količine informacija. Zahvaljujući pretraživanju, skupovi podataka se struktuiraju i tako nastaju primenljive informacije i znanje koje se koristi u daljim istraživanjima.

Cilj rada je da predstavi pristup za kreiranje modela sistema Data mining kao i za njegovo unapređenje. Kroz studijski primer su ilustrovane metode i tehnike koje se koriste pri projektovanju modela za istraživanje podataka. U radu su prikazane metode elektronskog praćenja pacijenata korišćenjem savremenih programskih alata za podršku odlučivanju. Analizira se korišćenje ove tehnologije u zdravstvu, koja predlaže da se nadogradi postojeći informacioni zdravstveni sistem u cilju preveniranja obolevanja i smrtnosti. Ovakav projekat bi mogao da doprinese istraživanju zdravstvenog stanja populacije i otkrivanju potencijalnih maligniteta kod pojedinaca. Život nema cenu.

Ključne reči: istraživanje podataka, skupovi podataka, veštačka inteligencija, medicinska istraživanja, inteligentni sistemi

SADRŽAJ

1 Uvod.....	5
2 Data mining.....	6
2.1 Pojam istraživanje podataka	6
2.2 Cilj Data Mining	6
2.3 Zadaci (problemi) Data Mining	7
2.3.1 Redukcija	7
2.3.2 Estimacija.....	7
2.3.3 Klasifikacija	7
2.3.4 Klasterovanje	7
2.3.5 Asocijacija	7
2.4 Faze	8
2.4.1 Definisanje poslovnog problema	9
2.4.2 Priprema podataka	9
2.4.3 Modelovanje i implementacija.....	10
2.5 Tehnike i metode data mining.....	11
2.5.1 Stablo odlučivanja.....	11
2.5.2 Metoda najbližeg suseda (Nearest neighbor classification).....	13
2.5.3 Memorijski zasnovano rasuđivanje (Memory based reasoning)	14
2.5.4 Klastering (Clustering)	14
2.5.5 Analiza potrošačke korpe (Market Basket Analysis)	14
2.5.6 Pravilo indikacija (Rule indication).....	14
2.5.7 Metoda K najbližeg suseda (K Nearest neighbors)	15
2.5.8 Neuronske mreže (Neural networks)	15
2.5.9 Fazi logika (Fuzzy logic)	15
3 Neuronske mreže	16
3.1 Vrste neuronskih mreža	17
3.2 Podela neuronskih mreža	18
3.2.1 Podela NM prema vrsti veza.....	18
3.3 Vrste obučavanja neuronskih mreža	18
3.4 Primena neuronskih mreža.....	18
4 Fazi logika.....	19
4.1 Implementacija veštačkih neuronskih mreža i fazi logike - robotika	20

5 Statističke metode	21
5.1 Statistika maligniteta raka grlića materice	21
5.2 Statistika maligniteta melanoma	22
5.2.1 Rak dojke	24
6 Baze podataka	24
6.1 E-R dijagram	25
6.2 Specifikacija tabela	27
6.3 Velike baze podataka - Datawarehouse	30
6.3.1 Projektovanje VBP	30
6.3.1.1 <i>Primena fazi logike</i>	32
6.3.1.2 <i>Robotika i automatizacija</i>	32
7 Neophodna infrastruktura za DM	33
7.1 Projektovanje odgovarajuće baze	34
8 Primeri projektovani u softverskom alatu.....	35
8.1 Primer detekcije kritičnih grupa.....	35
8.2 Primer skrininga kancera grlića materice	36
8.3 Primer klasifikacije pacijenata za poziv na pregled.....	39
9 Opis softverskog paketa Veka	41
9.1 Instalacija	42
10 Zaključak.....	43
Reference	44

1 Uvod

Data mining je metoda pretraživanja podataka koja se munjevito razvila pre svega zbog razvoja računarske tehnologije. Razvitkom brzih računara dolazi do toga da je moguće efikasno pretraživati ogromne količine informacija. Zahvaljujući pretraživanju, skupovi podataka se struktuiraju i tako nastaju primenljive informacije i znanje koje se koristi u dalja istraživanja.

Cilj rada je da predstavi pristup za kreiranje modela sistema kao i njegovo unapređenje za elektronsko praćenje pacijenata korišćenjem savremenih programskih alata za podršku odlučivanju.

Kao primer će se analizirati korišćenje ove tehnologije u zdravstvu koja bi mogla da doprinese istraživanju zdravstvenog stanja populacije i otkrivanju potencijalnih maligniteta kod pojedinaca.

2 Data mining

Data Mining je proces istraživanja i analize, automatskim ili poluautomatskim sredstvima, velikih količina podataka radi otkrivanja smislenih modela i pravila. (M. Berry i G. Linoff)

IBM definiše Data mining kao “proces ekstrakovanja prethodno nepoznatih, valjanih i delotvornih informacija iz velikih baza podataka i korišćenje tih informacija za donošenje ključnih poslovnih odluka”.

2.1 Pojam istraživanje podataka

Istraživanje podataka (eng. Data mining) možemo definisati kao proces pronalaženja novog i potencijalno korisnog znanja iz podataka (Pyle, Business Modelling and Data Mining, 2003), odnosno kao otkrivanje ili "rudarenje" znanja iz velike količine podataka (Han & Kamber, 2001). Ili iskopavanje znanja iz podataka, gde se iskopavanje odnosi na ono što se pronalazi kao rezultat tog procesa. U suštini je svakako "otkrivanje znanju u bazama podataka".

“Istraživanje podataka je prirodna evolucija tehnologije, koja upotrebljava koncepte, metode i tehnike različitih disciplina kao što su baze podataka, statistika i veštačka inteligencija.” (M.P.B & M.V., 2016). Tehnologija baza podataka se razvila u sofisticirane i veoma jake sisteme baza podataka. Taj razvoj omogućio je sakupljanje ogromnih količina podataka i njihovo beleženje što je neizbežno dovelo i do potrebe da se isti obrade i analiziraju, sve u svrhu dobijanja korisnih informacija i znanja. Danas, razvojem informacione tehnologije, omogućeno nam je beleženje svake aktivnosti u obliku digitalnih podataka, a količina sakupljenih podataka meri se u terabajtima. Veoma velike baze podataka bogate su podacima, ali su siromašne informacijama koje su skrivene u sačuvanim podacima. Istraživanje podataka je upravo bilo to koje selektuje važne informacije i znanje koje se nalazi u podacima, doprinoseći donošenju odluka, poslovanju i nauci.

Istraživanje podataka može se definisati i kao proces podrške odlučivanju u kojem se traže uzorci informacija u podacima, tj. tehnika pretraživanja podataka kako bi se identifikovali traženi uzorci i njihove međusobne relacije. Najjednostavnije rečeno, tehnika za izdvajanje interesantnih, novih i potencijalno korisnih informacija ali i uzoraka sadržanih u velikim bazama podataka.

2.2 Cilj Data Mining

Osnovni cilj istraživanja podataka zapravo je otkrivanje skrivenih veza, predvidivih nizova i egzaktnih klasifikacija. Na taj način se iz velike količine sirovih podataka i veza koje se ne mogu odmah uvideti, definišu relacije, obrasci ili forme ponašanja što na kraju daje potrebne informacije iz početnih podataka.

2.3 Zadaci (problemi) Data Mining

Najčešći i najpoznatiji zadaci istraživanja podataka su sledeći:

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

2.3.1 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 kao što su ručne redukcije i automatske redukcije. Takođe, postoji redukcija atributa ili redukcija slučajeva tj. redova, zapisa.

2.3.2 Estimacija

Estimacija (procena, ocena) predstavlja procenu vrednosti određene na osnovu postojećih (endogenih) promenljivih koje su zabeležene u sistemu. Ona daje određeno pravilo ponašanja koje je izvedeno iz postojećih podataka.

2.3.3 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 Rizična grupa raka dojke, opisuje pacijenta sa anamnezom koja ukazuje da pripada rizičnoj grupi).

Problem klasifikacije je da generiše model koji će na osnovu opisa objekata (ulaznih promenljivih) odrediti klasu tog objekta (izlazna promenljiva).

2.3.4 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, za razliku od klasifikacije, gde su grupe unapred definisane (apriori), a mi određujemo pripadnost nekog elementa grupi, ovde formiramo grupe na osnovu sličnosti elemenata.

2.3.5 Asocijacija

Zadatak asocijacije u istraživanju podataka 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.

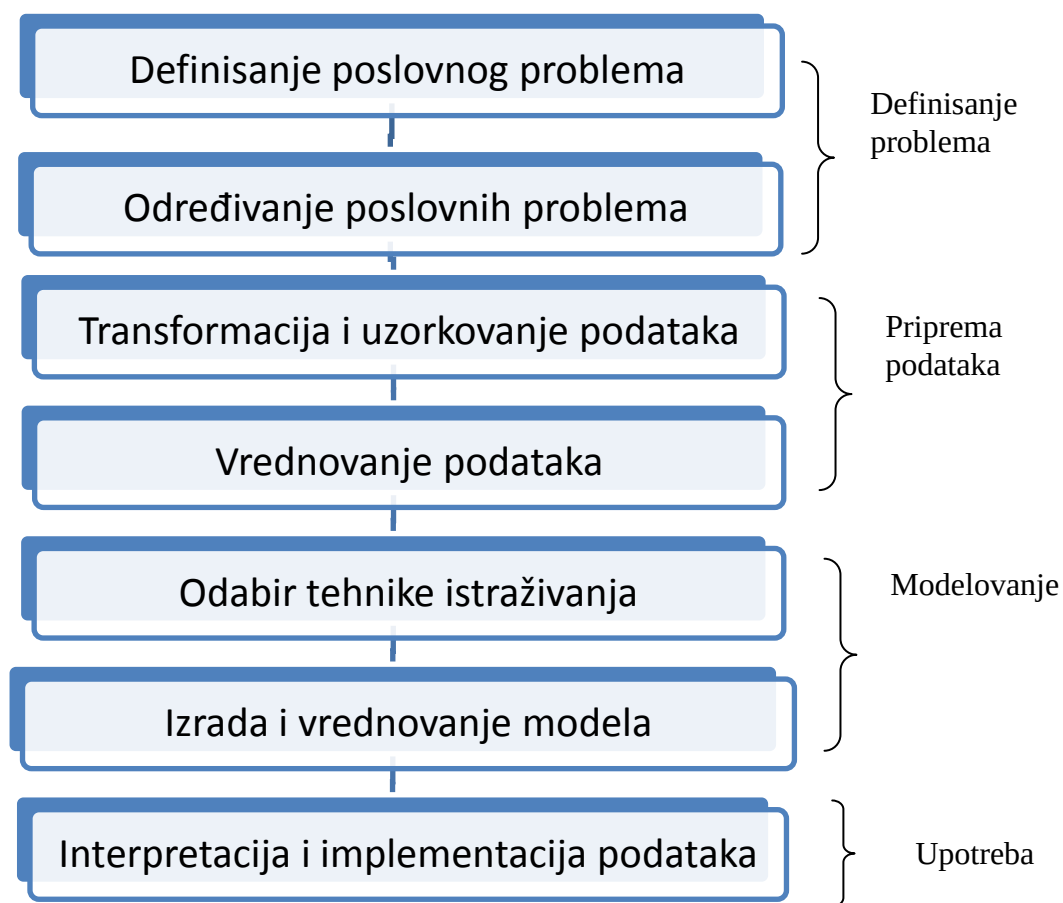
2.4 Faze

Proces istraživanja podataka sastoji se iz više faza koje izvršava više osoba. Te faze se mogu posmatrati iz 3 ugla.

Prvi je problemski koji je važan na samom početku i kraju istraživanja podataka. U sebi sadrži izbor problema, definisanje istoga, zatim procene i primene znanja. Osoba koja izvršava ove zadatke može se nazvati korisnik.

Drugi ugao je za prenos podataka a obuhvata sve zadatke u vezi sa pripremom podataka i njihovim istraživanjem, a obavlja je informatičar.

Treći i poslednji, metodološki, sastoji se od svih zadataka u vezi sa analizom podataka, metodama selekcije, implementacije, prezentacije i interpretacije rezultata. Stručnjak istraživanja podataka (eng. Data miner) jeste osoba zadužena za te zadatke.



Slika 1 Životni ciklus procesa istraživanja podataka

- Na samom početku se definiše poslovni problem.
- Zatim sledi priprema podataka koja a ona predstavlja, određivanje potrebnih podataka, transformaciju, uzorkovanje i vrednovanje istih.

- Treći korak je modelovanje, a obuhvata izbor metode istraživanja zatim izradu i vrednovanje modela.
- Poslednji deo odnosi se na implementaciju koja obuhvata interpretaciju i korišćenje rezultata.

Proces istraživanja podataka u svakom trenutku moguće vratiti se na neki od prethodnih koraka, to jest ima osobinu iterativnosti. Ovakav „skok unazad” pre se može nazvati pravilom nego izuzetkom, zato što je u istraživanju podataka veoma važno dobro definisati problem zatim dobro odabrati i pripremiti podatke, što je teško na samom početku uraditi kako treba. Tokom implementacije, dolazi do povećavanja znanja o samom problemu i podacima koji se koriste, i zato je takav „revidiran” opis poslovnog problema često bolji od početnog.

2.4.1 Definisanje poslovnog problema

Na samom početku u procesu istraživanja podataka nalazi se definisanje poslovnog problema i izražavanje samog problema u vidu pitanja na koja se na kraju tog procesa može odgovoriti. Pristup poslovnom problemu koji daje najbolje rezultate jeste analiziranje područja gde je istraživanje podataka već uspešno korišćeno. Kada se upoznamo sa uspešnim primenama istraživanja podataka, tek tada možemo odabrati područje koje je najkritičnije za našu poslovnu organizaciju ili oblast koja treba da se istražuje i koristi.

Ova faza je faza razumevanja poslovnog procesa ili faza razumevanja istraživanja. U opštem slučaju podrazumeva dobro razumevanje ciljeva i zahteva projekta kao celine, formulaciju ciljeva i ograničenja kao i priprema početne strategije za ostvarenje tih ciljeva.

U ovom se delu određuje koje će osobe učestvovati u projektu istraživanja podataka. Obično to su stručnjak za istraživanja podataka, informatičar koji bi trebalo da dobro poznaje baze i skladišta podataka date oblasti i naravno stručnjak iz oblasti koji je dobro upoznat sa mogućom primenom rezultata.

2.4.2 Priprema podataka

U ovom se delu određuju potrebni podaci, transformišu, uzorkuju i vrednuju. Najviše vremena će biti utrošeno u ovoj fazi, a obuhvata 60-90% vremena potrebnog za istraživanje podataka (Pyle, Business Modelling and Data Mining, 2003). Podaci za istraživanje bivaju predstavljeni u različitim oblicima, a najčešći su relacijske baze podataka ili skladišta podataka. Podaci koji će se koristiti mogu biti prikupljeni istraživanjem tržišta ili mogu biti eksterni izvori podataka. Zajedno, stručnjak za istraživanje podataka, informatičar i stručnjak iz istraživane oblasti, određuju koje podatke će upotrebiti za izradu modela.

Određivanje potencijalnih podataka - Podaci koji se najčešće koriste za istraživanje nalaze se u bazi klijenta u obliku transakcijske baze.

U ovom delu se takođe određuje koje će se promenljive izbaciti iz analize i koja će biti ciljna ili zavisna. Recimo, ako se radi analiza rizika za rak dojke, ciljna promenljiva će biti ona koja govori da li je pacijent imao prethodne nepovoljne rezultate na pregledima, starost i porodična anamneza.

Kao rezultat određivanja potrebnih podataka je popis promenljivih koje će se koristiti u izradi samog modela.

Transformacija podataka - Ovde se promenljive iz baza podataka transformišu u oblik koji je odgovarajući za istraživanje podataka. Podaci bi trebalo da budu u obliku tabele pri čemu se u kolonama nalaze promenljive (obeležja), a u redovima opažanja. Svaki red bi trebalo da opisuje podatak značajan za istraživanje (npr. kupca, proizvod, pacijenata...).

Na osnovu dostupnih promenljivih iz baza podataka izračunavaju se promenljive koje određuje stručnjak za procenu nekog rezultata u željenoj oblasti. Primeri takvih promenljivih su razlike između poslednjeg skrininga i današnjeg dana.

Uzorkovanje podataka - Za izradu samog modela verovatno i neće biti potrebno puno podataka, odnosno svi koji se nalaze u bazama, zato je neophodno da se upotrebi uzorkovanje podataka i na taj način se odabrala manja količina podataka za model. Nema jednoznačnog odgovora na pitanje koliko je podataka dovoljno za izradu modela zato što sam broj potrebnih podataka zavisi od algoritma.

Podaci za uzorak se dobijanju, najčešće, slučajnim izborom. Nakon što se odabere uzorak za izradu modela, potrebno je podeliti ga na dva dela

- Jedan deo za izradu modela i
- Jedan deo za testiranje modela.

Ovakav pristup je veoma važan za samo istraživanje podataka jer se na taj način može proveriti koliko je model efikasan sa podacima koji nisu korišćeni za njegovu izradu.

Vrednovanje podataka – Ovde je potrebno uraditi analizu postojanja neuobičajnih slučajeva vrednosti i nepotpunih podataka.

Neuobičajne vrednosti javljaju se skoro u svakoj bazi podataka, a radi se recimo o pacijentima koji ne vode računa o svom zdravlju, ne javljaju se na redovne preglede i ne odazivaju se na pozive od strane medicinske ustanove, a ulaze u zonu rizičnih grupa. Potrebno je odlučiti šta raditi s nepotpunim vrednostima. Moguće je napraviti analizu i sa njima ali treba izbaciti iz analize promenljivu koja ima mnogo nepotpunih vrednosti. Vrednosti promenljive mogu se takođe podeliti u kategorije.

Nepotpuni podaci najčešće se odnose na nepostojeće vrednosti, nejasno definisane podatke i netačne vrednosti. Ove pojave se mogu naći u svakoj bazi podataka. Nepostojeće vrednosti su česte, moguće je da se samo radi o tome da recimo za klijenta nemamo neke demografske podatke kao što je datum rođenja. Trebalo bi bolje proučiti podatke zato što je možda moguće nepotpune izračunati na osnovu nekih drugih podataka kao što je jedinstveni matični broj građana.

2.4.3 Modelovanje i implementacija

Kada su podaci sređeni i prilagođeni alatima prelazi se u fazu modelovanja, koja ima 4 faze i koja u principu zahteva najmanje rada ukoliko su prethodne faze dobro izvršene. A to znači da se dobro razume problem i da su podaci adekvatno pripremljeni. Vršiti se izbor odgovarajuće tehnike za modelovanje, podešavanje parametara modela, ili primena drugih tehnika (uvek treba imati na umu da se različite tehnike mogu koristiti za rešavanje istih problema).

U procesu istraživanje podataka mogu se koristiti sledeće metode: statistika, baze i skladišta podataka, veštačka inteligencija i vizualizacija. Postoji veliki broj različitih metoda istraživanja podataka ali se sve mogu podeliti u tri kategorije: otkrivanje, klasifikacija i predviđanje (Berry & Linoff, 2000). Metode otkrivanja predstavljaju postupke koji istražuju pravilnosti u podacima bez potrebe da unapred poznaju njihov odnos. Postoji puno metoda kojima se otkrivaju pravilnosti u podacima, neke ćemo navesti: segmentacija i asocijativna pravila.

Metode za klasifikaciju promenljivih najčešće se upotrebljavaju kako bi se predvidele kategorije, npr. da li će pacijent oboleti od gihta ili ne. Stabla odlučivanja se često koriste a klasifikaciju kao i regresija i neuronske mreže.

Metode za predviđanje vrednosti promenljivih koriste se kako bi se predvidele numeričke vrednosti. Koriste se i neuronske mreže, linearna regresija kao i metode vremenskih serija. Od samog cilja analize zavisi odabir metode.

Najčešće korišćene metode istraživanja podataka jesu metode klasifikacije, a najređe se koriste metode predviđanja vrednosti.

U ovoj fazi se obično proverava više metoda a veoma retko odlučujemo samo za jednu metodu. Nakon njihove primene i poređenja na određenim izabranim podacima sledi izbor metode za israživanje podataka.

2.5 Tehnike i metode data mining

Kako bi se brže i tačnije rešavali problemi, došlo je do razvoja velikog broja tehnika, algoritama i metoda data mining-a. Sve se one nazivaju – Data Mining tehnike.

One koje se najčešće primenjuju uglavnom su dobijene iz tri glavne oblasti: mašinskog učenja (neuronske mreže, stabla odlučivanja...), grane matematike (statistike, logike, fazi logike...) i baza podataka.

Algoritmi kao što su regresija i stabla odlučivanja preuzeti su iz statistike. Neki algoritmi su preuzeti iz oblasti mašinskog učenja kao što su neuronske mreže koje se izuzetno uspešno primenjuju kod klasifikacije i regresije a koriste se i za otkrivanje obrazaca ponašanja tokom analizi podataka.

Genetski algoritmi jesu još jedna tehnika koja se koristi za klasifikaciju mada je razvijeno dosta novih algoritama, metoda i softvera. Postoji i nekoliko skalabilnih verzija algoritama klasifikacije i klasterovanja koji koriste tehnike baza podataka, uključujući i Microsoftov algoritam klasterovanja. (Witten & Frank, 2005)

Data Mining tehnike možemo podeliti u dve grupe:

- 1) Tehnike za otkrivanje novih znanja (informacija)
- 2) Tehnike za predviđanje

2.5.1 Stablo odlučivanja

Kako bi se predvideli ishodi koriste se serije pitanja i pravila za kategorizaciju podataka.

Stablo odlučivanja nastaje grananjem a svaka grana jeste posledica ispunjenja uslova klasifikacijskih pitanja. Podatke će u podskupove podeliti svako pitanje. Ako ono ima dva odgovora, tada će kao posledica nastati dva podskupa (binarno stablo).

U našem primeru, će se uglavnom koristiti binarno stablo, puno ili sa jednim listom u zavisnosti od problematike i ishoda koji se traži u konkretnoj situaciji

Odlučivanje - zadaci – postavljanje problema odlučivanja – rešavanje problema odlučivanja – implementacija rešenja problema odlučivanja.

Koraci odlučivanja

- Određivanje ciljeva odlučivanja: u primeru koji je dat na slici, cilj je animirati osobe koje nisu redovne u praćenju osnovnih preventivnih pregleda, tj. staviti ih na lisu da im se treba uputiti poziv
- Definisane probleme odlučivanja, u našem primeru mogu biti životna dob, pol, da li su vršeni redovni kontrolni pregledi ili ne i to u zavisnosti od godina, jer za žene do 47. godine života je dovoljno da idu na godišnji pap test, dok za starije je neophodno da to bude na 6 meseci.
- Sagledavanje mogućih stanja sistema ili dela sistema: ako uzmemo skrining kancera ne samo reproduktivnih organa, recimo kancera dojke, tu bi vrlo veliki uticaj imala porodična anamneza (ova varijabla se može ispitivati i fuzzy metodom, jer bi u tom slučaju imali senzitivniju vrednost pri donošenju odluke).
- Sagledavanje i ocenjivanje mogućih posledica u smislu da se neki deo ne preskoči koji je relevantan za skrining, a opet ne treba da se ponavlja, eventualno treba ponavljanje smanjiti na najmanji mogući broj.
- Formulacija metode i odabir kriterijuma odlučivanja, mora biti usaglašenost sa ciljevima. U zavisnosti od vrste cilja, tj krajnjeg lista, moramo formirati čvorove binarnog stable. Kako imamo sveobuhvatni skrining, treba modelovati binarna stable sa što manje preklapanja, ako je to moguće.
- Ocenjivanje i izbor akcije
- Odluke
- Preduzimanje, praćenje i analiza efekata primene odabrane akcije

Pristup u našem studijskom primeru, zdravstvenom skriningu, je najčešće izabran pristup reaktivnog delovanja ("tradicionalni" pristup) – odnosi se na otkrivanje načina kako bi se evidentni problem mogao rešiti. Pošto je živ sistem, negde se pristupa i proaktivnim delovanjem ("savremeni" pristup) – odnosi se na istraživanje i prevenciju budućih sličnih problema, takođe obuhvata i istraživanje kao i iskorišćavanje povoljnih prilika.

Kreiranje stabla odlučivanja se radi ručno – na osnovu ekspertskog znanja tima stručnjaka iz medicine

Pitanje koje je ključalno u dizajnu stabla je "Koji atribut uzeti za grananje?" – Najbolje je uzeti onaj atribut koji deli početni čvor na što "čistije" potomke što je u našem primeru koren (Pol =m).

U listovima treba da bude zastupljeno što više instanci iste klase.

Prednosti stabla:

Mogu se grafički predstaviti i jednostavno interpretirati.

Mogu se primeniti kako na klasifikacione, tako i regresivne probleme.

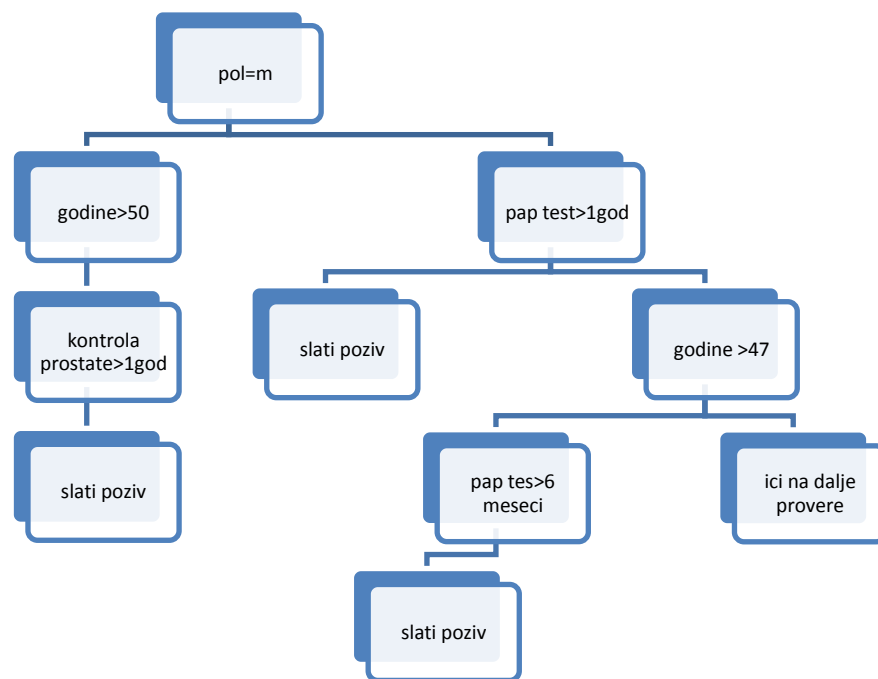
Mogu se primeniti i u slučaju da atributi imaju nedostajuće vrednosti.

Postoje i nedostaci:

Daju slabije rezultate (manje tačne predikcije) nego drugi pristupi nadgledanog mašinskog učenja.

Prilikom izgradnje stabla odlučivanja vrlo je bitno koje je se pitanje definiše kao prvo postavljeno. U našem slučaju je to pitanje o polu. Pitanje je bolji izbor kao prvo, toliko koliko će se njime bolje organizovati podaci, to jest koliko će se nakon toga stvoriti podskupovi koji su homogeniji. Modeli koji se oslanjaju na stabla odlučivanja algoritamski se razlikuju od drugih modela, jer zahtevaju obeležja pojedinih podataka na bazi kojih se kreiraju pitanja. Pored već pomenutog, binarnog, postoje još i heš stablo, B+, stablo prekoračenja... Sve vrste stabala odlučivanja se veoma primenjuju na relacijskim bazama podataka (npr. eskuel).

Primer stabla odlučivanja prikazan je na slici ispod u pitanju je binarno stablo.



Slika 2 Binarno stablo dela koji se odnosi na kancer reproduktivnih organa

2.5.2 Metoda najbližeg suseda (Nearest neighbor classification)

Metoda je jedna od najstarijih tehnika koja se koristi u DM a koristi se kako bi se klacifikovali podaci. Način rada ove tehnike je sličan ljudskom načinu razmišljanja, te je zbog toga ovo jedna je od najjednostavnijih metoda. U osnovi ove metode je traženje podataka koji imaju najsličinija

svojstva i poznato ponašanje. Podatak kod kog se otkriju najbližnja svojstva jeste najbliži sused a zato se i pretpostavlja da će se slično i ponašati. Metoda algoritma za određivanje ko je najbliži sused može raditi upotrebom principa euklidske geometrije u n-dimenzionalnom prostoru, te je jedan od najjednostavnijih načina. Kod ove metode svaka promenljiva ne nosi istu težinu važnosti.

Da bi pri upotrebi ova metoda bila što tačnija, neophodno je da se u bazi nađe što sličniji podatak, a za to su potrebne velike količine podataka.

Za razliku od ostalih tehnika, ovde, ne postoji proces učenja kako bi se kreirao model. Podaci koji se koriste za učenje je baš model. Pri pojavi novog podatka, algoritam analizira sve podatke u bazi, kako bi našao podgrupu slučajeva koja najbolje odgovara tom slučaju i na osnovu toga vrši predviđanje.

2.5.3 Memorijski zasnovano rasuđivanje (Memory based reasoning)

Memorijski zasnovano rasuđivanje je jedna od tehnika data m. koju koristimo za predviđanje i klasifikaciju. Od tehnike neuronskih mreža razlikuje se po tome što MBR uočava slične podatke. Međutim MBR ne utvrđuje pravila ili obrasce sličnosti kod podataka. Jedan primer bi bio, lekar koji ima pacijenta kod kog je uočio slične simptome bolesti kao i kod nekog od ranije i on će ipak na osnovu iskustva dati dijagnozu.

2.5.4 Klastering (Clustering)

Klastering je tehnika grupisanja, klasiranja. Ova tehnika omogućava grupisanje podataka po njihovoj sličnosti. Grupisanje jeste zapravo razvrstavanje, klasiranje elemenata u skupove, gde se uočava najveća sličnost podataka (segmentacija pacijenata – po starosti, polu, hipertenziji, gojaznosti...).

Postoje dva kriterijuma koje treba zadovoljiti kod podele:

- svaka grupa bi trebalo da bude homogen skup – slični podaci
- svaki skup bi trebalo da se razlikuje od ostalih skupova – značajne razlike u podacima

2.5.5 Analiza potrošačke korpe (Market Basket Analysis)

Metoda MBA se vrlo često naziva i grupisanje po sličnosti. Ovo grupisanje se odnosi na izbor proizvoda u potrošačkoj korpi. Vrš se otkrivanje grupe artikala koji se obično zajedno kupuju u jednom „pazaru”. Kada se izvrši analiza izbora proizvoda iz potrošačke korpe, može se utvrditi koja je verovatnoća da će kupac izabrati i proizvod B, ako je nakon prve kupovine izabrao proizvod A. Model se široko upotrebljava u tržišnim centrima i supermarketima. (Gračanin, 2009)

Ova metoda se može koristiti i pri analiziranju prodaje u marketima na potpuno različitim lokalitetima, dnevnim prodajama, godišnjim dobima, a sve sa ciljem da se povećala prodaja prilagođavanjem arsortimana i usluga.

2.5.6 Pravilo indikacija (Rule indication)

“Upotreba ove metode zasniva se na prolasku kroz bazu podataka upotrebljavajući logičke funkcije na varijablama, i računajući verovatnoću pojave takvog događaja, pojedinih zapisa, kako bi se došlo do skrivenih informacija. Kako bi se moglo doći do skrivenih informacija, potrebno je proći kroz što više mogućih međusobnih kombinacija varijabli (sve kombinacije), što drastično usporava i poskupljuje ovu metodu. Ako odbacujemo pojedine promenljive kao

nevažne, tada postoji mogućnost da nećemo videti vezu između pojedinih podataka i model učiniti manje tačnim. Osim s tehničke strane, pretraživanje sličnosti pojedinih podataka po svim promenljivim često daje ogroman broj povezanosti između pojedinih podataka, pa je ponekad potreban još jedan prolaz kroz dobijeni rezultat kako bi se izolovali oni zaključci koji su najinteresantniji.” (Gračanin, 2009)

Oni modeli koji se baziraju na upotrebi ove metode su se pokazali kao jedni od najtačnijih među dobrim modelima. Takođe su modeli koji su jednostavni za korišćenje.

2.5.7 Metoda K najbližeg suseda (K Nearest neighbors)

Ova metoda je poboljšana u odnosu na metodu najbližeg suseda tako što se ovde uzima u razmatranje ponašanje nekoliko sličnih suseda, tj. podataka, a ne samo jedan. Samim tim uz veći izbor, možemo tačnije predvideti ponašanje i svojstva pojedinog podatka. Metoda K najbližeg suseda je algoritam koji je lako implementirati.

2.5.8 Neuronske mreže (Neural networks)

Ova tehnika je osmišljena na osnovu analize delovanja ljudskog mozga. Po principu ljudskog mozga, nakon procesa učenja izvlače se određene pretpostavke na osnovu prethodnih zapažanja, i ove mreže predviđaju promene i dešavanja u sistemu nakon procesa učenja. Istraživanje podataka ovom tehnikom počinje „učenjem,” uz pomoć i na osnovu podataka koji su već poznati i nad kojima su uočena neka pravila. Takvi podaci su reporni za vrednosti koje želimo pretpostaviti. Posle učenja, znanje se proverava, sve dok rezultati provere ne budu na zadovoljavajućem nivou. Proces se svodi na sledeće: „Prvo se neuronskoj mreži daju određeni podaci za koje već znamo izlazne vrednosti. Na osnovu ovih podataka neuronske mreže prepoznaju obrasce i pravila. Zatim se na osnovu ovih obrazaca i funkcija istražuju gomile podataka koje preduzeća imaju u svojim bazama.” (Bramer, 2007)

Neuronske mreže su najsloženija metoda, što za upotrebu, što za primenu, međutim daje najtačnije modele. Nastale su pokušajima imitiranja rada i proučavanjem, pre svega ljudskog mozga i nervnog sistema živih bića.

Ova metoda je ključna za istraživanje podataka, zato joj je u ovom radu posvećen poseban deo.

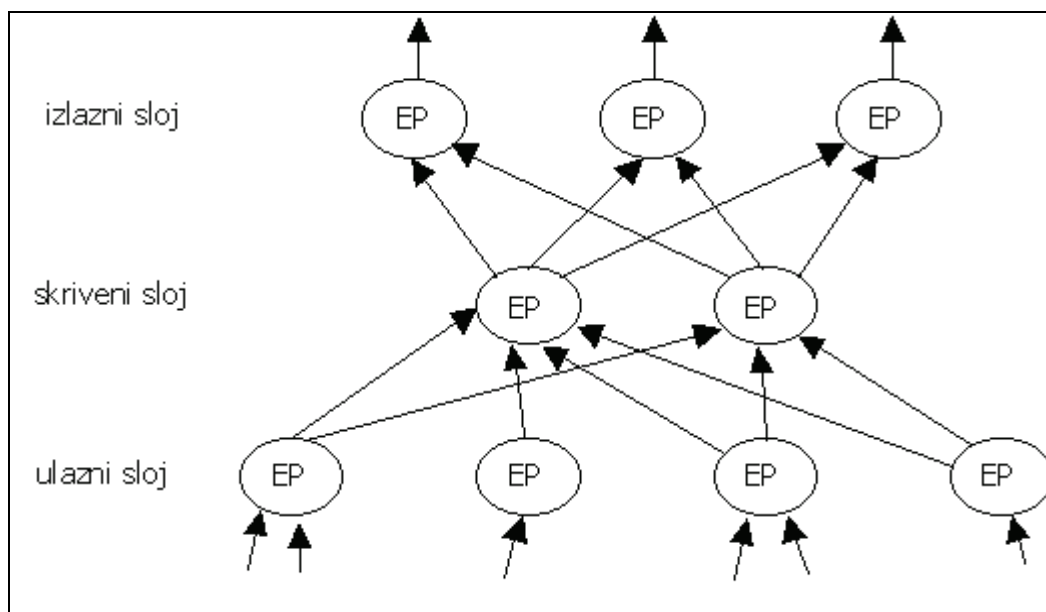
2.5.9 Fazi logika (Fuzzy logic)

Metoda u osnovi ima klasičnu logiku, koju je definisao Aristotel. Aristotelova, klasična logika zasnovana je na jasnim i precizno utvrđenim pravilima, i ona se bazira na teoriji skupova. Skupovi imaju jasno definisane granicne elemente, tj. neki element može ili pripadati skupu ili ne.

Kod fazi logike, nema jasno definisane pripadnosti elementa nekom skupu, nije isključiva kao klasična logika. Fazi se meri u procentima. Procenti su predstavljeni decimalama od 0 do 1 (100%). Ovakva logika je meka logika i bliska ljudskoj percepciji. Tehnika se realizuje kao simulacija ljudskog zaključivanja i rezona.

3 Neuronske mreže

Kao početak neuro-računarstva navodi se 1943. godina kada je objavljen članak Vorena MekKaleha (Warren McCulloch) i Voltera Pitsa (Walter Pitts) "Logički račun ideja svojstvenih nervnoj aktivnosti". Kibernetičar Norbert Viner i matematičar Džon von Njuman su imali teoriju da bi istraživanja u okviru računarstva, koja su imitacija radom ljudskog mozga, od izuzetne zanimljivosti. Frenk Rosenblat je otkrio jednoslojne neuronske mreže, nazvane perceptron. Računar sa perceptionom, je mogao uspešno da podešava težinske koeficijente, ali nije postigao bitne praktične rezultate. Krajem pedesetih godina 1957/8, Frenk Rozenblat i Čarls Vigman sa svojim timom saradnika uspevaju da konstruišu računar Mark što predstavlja prvi neuroračunar.



Slika 3 Šema neuronskih mreža

Kasnije, posle Rozenblata, Bernard Vidrou je sa svojim studentima (jedan od njih je Ted Hof, koji će kasnije stvoriti mikroprocesor) razradio novi tip "neurona"- ADALINE (ADaptivni LINearni Element) i odgovarajući zakon učenja. Većina od ovih ljudi su odustali od neuroračunarstva i otišli u srodne naučne grane. Sredinom šezdesetih postojala je teza da izgradnja neuronskih mreža nije pravi put do stvaranja računara kao integrisanu mašinu za smislene procese. Marvin Minsk i Seymour Papert u knjizi "Perceptrons" daju matematički dokaz da jednoslojna neuronska mreža "Perceptron" ne može da nauči funkciju iks-or, čak su išli i dotle da su tvrdili da dodavanjem više slojeva neurona problem iksora, neće biti rešen. Neuron, jedan, nema mogućnost da izvede iks-or funkciju, ali mreža od nekoliko neurona to itekako prevazilazi kao jednostavaniji zadatak. Početkom osamdesetih, DARPA, američka vojna Agencija za odbrambene istraživačke projekte, počinje da se interesuje za neuronske mreže i finansira njihovo istraživanje.

Sveo ovo dovodi da krajem 80-tih i početkom 90-tih, neuronske mreže i neuro računarstvo se uvodi kao predmet na nekoliko elitnih univerziteta u SAD. Danas se ova nauka, izučava na svim

tehničkim fakultetima, što kao poseban predmet, što kao deo nekog drugog predmeta. Iako, sada već postoji zvanična grana računarstva koja se bavi obim mrežama to je još uvek u ranoj fazi razvoja.

Neuronske mreže imaju veoma širok dijapazon primena u raznim oblastima.

3.1 Vrste neuronskih mreža

Neuronske mreže je moguće podeliti u dve grupe i to su veštačke i biološke. Kao predstavnika bioloških možemo navesti nervni sistem živih bića. Za veštačke neuronske mreže možemo reći da su po funkciji, strukturi, i obradi informacija slične biološkim, ali su zapravo veštačke tvorevine.

U računarskim sistemima, neuronska mreža predstavlja prilično povezanu mrežu elemenata koji zajedno obrađuju podatke. Njih su osposobili da izlaze na kraj problemima koji su se ranijim pristupom teško rešavani, kao što su govor i prepoznavanje oblika. Neuronske mreže imaju sposobnost da uče na ograničenom skupu primera što je jedna od njihovih važnijih osobina. Od matematičkih modela, dosta komplikovanije su biološke neuronske mreže, koje se koriste za veštačke neuronske mreže. Biološke neuronske mreže mogu predstavljati sisteme sastavljene od velikog broja jednostavnih elemenata za obradu podataka. Kod ovog tipa sistema je moguće prikupiti, memorisati i koristiti eksperimentalno znanje. Neuronska mreža je sistem sastavljen od više jednostavnih procesora (jedinica, neurona), svaki od njih ima lokalnu memoriju u kojoj pamti podatke koje obrađuje. Te jedinice su povezane komunikacionim kanalima (vezama). Ovim kanalima se razmenjuju podaci koji su obično numerički. Jedinice obrađuju samo svoje lokalne podatke i ulaze koje primaju preko konekcije. Ograničenja lokalnih operatora se mogu otkloniti tokom treninga.

Kako bi se konstruisao veštački sistem koji bi bio sposoban za istančano, možda i inteligentno izračunavanje, slično kao što to ljudski mozak rutinski izvodi, javila se želja za razvojem NM. Neuronske mreže nam mogu omogućiti razumevanje rada ljudskog mozga. Kolekcija matematičkih modela jesu veštačke neuronske mreže koje simuliraju neke od posmatranih osobina bioloških nervnih sistema i prepoznaju sličnosti sa prilagodljivim biološkim učenjem. One su sazdane od puno međusobno povezanih neurona koji su, kao što su i biološki neuroni, povezani vezama koje sadrže propusne, težinske koeficijente, koje liče na sinapse.

Kod bioloških sistema učenje se obavlja putem regulisanja sinaptičkih veza koje povezuju aksone i dendrite neurona. Preko treninga ili otkrića se uče tipični događaji i dolazi se do tačnih setova podataka ulaza - izlaza koji treniraju algoritam tako što ponavljaju a uz to i podešavaju propusne (težinske) koeficijente veza - sinapse. Ovakve veze mogu sačuvati znanje koje će biti potrebno za rešavanje datog problema. Uglavnom NM imaju različite vrste pravila za obučavanje, a time se na osnovu ulaznih podataka podešavaju koeficijenti veza između neurona. Može se reći da NM uče putem primera i mogu izvršiti generalizaciju nakon trening podataka. Njihov potencijal se može naći u njihovoj mogućnosti paralelne obrade podataka, tokom izračunavanja komponenti koje su nezavisne.

3.2 Podela neuronskih mreža

S postojanjem velikog broja realizacija NM postoji i puno podela. Njih možemo podeliti prema:

- broju slojeva,
- vrsti veza između neurona,
- vrsti obučavanja neuronskih mreža,
- prema smeru prostiranja informacija,
- prema vrsti podataka.

Podela neuronskih mreža prema broju slojeva je jedna od najopštijih podela NM:

- 1) jednoslojne,
- 2) višeslojne.

U današnje vreme se najviše izučava i primenjuje višeslojna NM koja pored ulaznih i izlaznih slojeva sadrži neurone na srednjim (skrivenim) slojevima.

3.2.1 Podela NM prema vrsti veza

Neuronske mreže se takođe mogu podeliti u odnosu na vrste veza odnosno arhitekture na:

- slojevite gde su neuroni raspoređeni tako da formiraju slojeve. Na ulaz jednog neurona se dovode izlazi svih neurona sa prethodnog sloja, a njegov izlaz se vodi na ulaze svih neurona na narednom sloju. Neuroni na prvom ulaznom sloju imaju samo jedan ulaz. Izlazi neurona sa izlaznog, zadnjeg sloja predstavljaju izlaze mreže.
- potpuno povezane kod kojih izlaz jednog neurona vodi ka ulazu svih neurona u mreži.
- i celularne kod kojih su međusobno povezani samo susedni neuroni. Bez obzira na lokalnu povezanost, signali se prostiru i na neurone van susedstva zbog indirektnog prostiranja informacija.

3.3 Vrste obučavanja neuronskih mreža

Obučavanje NM možemo pristupiti na tri načina:

- Nadgledano
- Nenadgledano
- Delimično nadgledano ono radi na način da mreža uči samostalno s tim da povremeno dobija ocenu prethodnog rada.

3.4 Primena neuronskih mreža

Na samom početku nezronske mreže su koristili naučnici računarskih i kognitivnih nauka kako bi napravili čulni sistem živih organizama. Ova oblast predstavlja veoma atraktivnu oblast istraživanja a postoje i brojne druge oblasti u kojima se koriste. Primenjuju se i za:

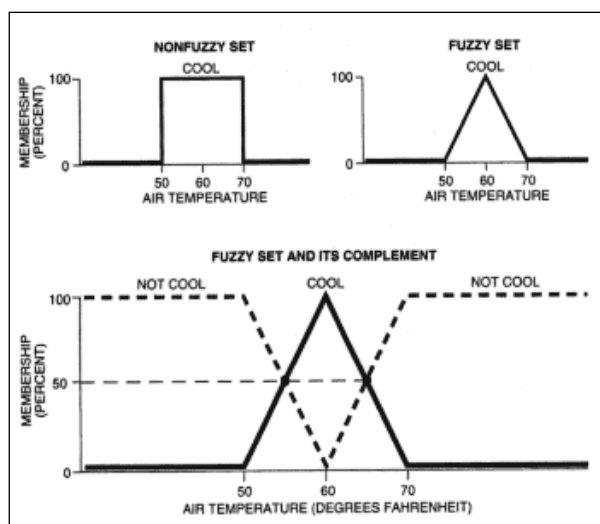
- prepoznavanje oblika,
- prepoznavanje rukopisa,
- prepoznavanje govora,

- finansijske i ekonomske modele,
- predviđanje kretanja cena na tržištu,
- upravljanje sistemima,
- upravljanje proizvodnim procesima,
- analizu električnih kola,
- psihijatrijske procene,
- kompresovanje podataka,
- naftna istraživanja,
- kriminološka istraživanja,
- analizu medicinskih testova,
- pronalaženje optimalnog rešenja,
- upravljanje robotima,
- u bioračunarskim sistemima.

4 Fazi logika

Fazi logikom se bavio Lotfi Zadeh 1965. godine, a za kontrolu sistema uveo ju je E. Mamdani. Tada je primećeno da ovaj pristup privlači puno pažnje. Za jednostavne sisteme ovakav pristup se pokazao kao veoma efikasan i jasno prilagođen ljudskom rezonovanju, za neke komplikovanije sisteme se pokazao kao veoma zahtevan. Kako bi se realizovao kontroler potrebno je puno resursa i vremenskih i intelektualnih.

Ova logika je konceptualno izuzetno prirodna, čak više nego što se to odmah može primetiti. Dešavaju se i situacije u kojima se znanje ne može predstaviti na apsolutno precizan način. Kako bi pravilno i uspešno predstavili ovo znanje potrebno je odreći se klasične, binarne logike koja govori da je nešto ili tačno ili netačno i koristiti fazi logiku.



Slika 4 Šematski prikaz fazi sistema

U klasičnoj teoriji skupova polazi se od stava da neki element x iz određenog, univerzalnog skupa X pripada ili ne pripada nekom konkretnom skupu A . Slično razdvajanje postoji u klasičnoj logici: iskaz je istinit ili lažan i isključuje se treća mogućnost. Pripadnost je uslovljena karakteristikom elementa, odnosno uslovom koji element skupa X treba da ispuni da bi pripadao skupu A .

U govoru se veoma često koriste izrazi koji mogu opisivati skup čije su granice nejasne, tako da se za neke elemente univerzalnog skupa ne može jednostavno zaključiti da li ispunjavaju uslov pripadnosti konkretnom skupu. Za takav izraz se koristi termin lingvistička promenljiva. Primer lingvističke promenljive jeste ljudi srednjeg rasta, velike zarade, brzi automobili, mala rastojanja, itd. Ako navedeni atributi, srednji, veliki, brzi, mala označavaju uslove koji elementi razmatranih skupova, ljudi, zarade, automobili, rastojanja, treba da ispune da bi se odredili konkretni podskupovi (ljudi srednjeg rasta, velike zarade, brzi automobili, mala rastojanja), onda je očigledno da nema dovoljno informacija da bi se to jednoznačno obavilo.

Fazi logika nije odgovor na sve tehničke probleme, ali za probleme kontrole, gde je važna brzina i jednostavnost implementacije, fazi kontroler je jak kandidat. Fazi kontroler skuplja izlaze iz modela, poredi ih sa referentim ulazima i odlučuje koji će biti ulazi u model, da bi se dostigao željeni cilj. Fazi kontroler omogućava korišćenje pravila za kontrolu modela, umesto diferencijalnih jednačina. Korišćenje pravila je prirodnije za ljude.

Fazi informacije su informacije koje uključujemo u kontroler, za koje primenjujemo pravila, mogu biti dobijena od eksperta za datu oblast, koji je proveo dosta vremena učeći kako je najbolje kontrolisati proces. U nekim slučajevima ne postoji ekspert, pa se mora proučiti dinamika modela (koristeći modelovanje i simulaciju) kako bi se odredila pravila za upravljanje. Npr. ako je brzina manja od željene, ubrzaj. Više opisuje: Ako je brzina manja od željene i ako se brzina brzo približava željenoj, onda malo ubrzaj.

Postoje i mane fazi logike. Npr. da li će ponašanja koje je uvideo ekspert i koje su korišćenje za fazi kontroler uključiti sve situacije koje mogu da se dese usled šuma, varijacije u parametrima modela. Da li ekspert može realno uočiti i predvideti probleme koje se mogu desiti usled nestabilnosti

4.1 Implementacija veštačkih neuronskih mreža i fazi logike - robotika

Robotika je jedna od novih tehnologija koja uz nove materijale, telekomunikacije, informatiku, biotehnologiju, genetski inženjering, nanotehnologiju, farmaceutske i svemirsku tehnologiju predstavlja okosnicu razvoja najrazvijenijih zemalja u 21. veku.

Roboti su smart mašine. Početak razvoja tih mašina možemo povezati s izumom tranzistora krajem 40-ih godina 20. veka. Mikroelektronička revolucija stvorila je tehnološku platformu za izradu sofisticiranih uređaja, koji su do tada postojali samo u sferi naučne fantastike.

Roboti, to jest, robotika u prvo vreme su služili za obavljanje teških, opasnih i zaglupljujućih poslova. Jedna od prvih i industrijskih primena ostvarena je na pokretnoj traci. Roboti su obavljali jednostavne radne operacije gotovo bez greške i tako su poboljšali kvalitet ali i produktivnost.

Savremeni roboti u velikoj meri doprinose razvoju nauke. Bez robotskih uređaja ne bi se moglo analizirati najskrivenija mesta na Zemlji (npr. velike morske dubine) niti bi se moglo proučavati udaljena svemirska prostranstva.

Vremenom roboti postaju sve "pametniji" i multifunkcionalniji. Tome je doprineo razvoj matematike, prvenstveno logike, digitalne i računarske tehnologije, veštačke inteligencije i dr.

5 Statističke metode

Statističke metode i algoritmi grupisanja mogu nam omogućiti da pronađemo model koji opisuje razlike u klasama ili konceptima za predstojeća predviđanja.

“Deskriptivna statistička analiza predstavlja skup metoda kojima se vrši izračunavanje, prikazivanje i opisivanje osnovnih karakteristika statističkih serija.

Deskriptivna statistička analiza obuhvata sledeće radnje:

1. Grupisanje i sređivanje statističkih podataka
2. Prikazivanje statističkih podataka
3. Izračunavanje i određivanje osnovnih mera statističkih serija.

Za numeričke statističke serije osnovne karakteristike su:

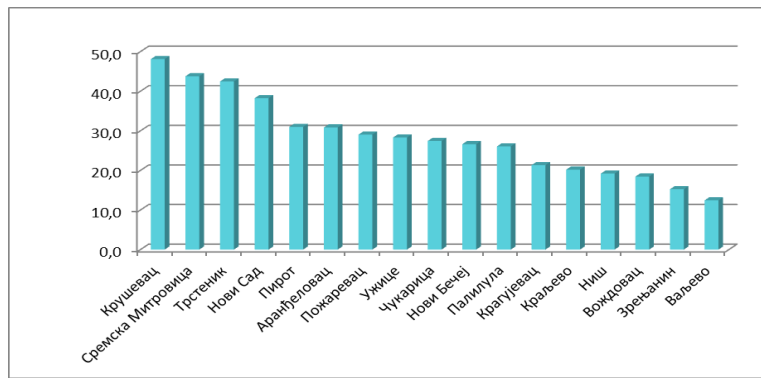
1. Srednja vrednost (mera centralne tendencije)
2. Mera varijacije (mere raspršenosti, disperzije)
3. Mera oblika i rasporeda” (Savić, 2005)

5.1 Statistika maligniteta raka grlića materice

Krajem 2015. godine završen je trogodišnji ciklus sprovođenja organizovanog programa skrininga raka grlića materice na teritoriji od 17 opština. U okviru ove tri godine u pozvano je na pregled 334.242 žena životne dobi od 25 do 64 godine, što čini 56,1% ciljne populacije.

PAP test je urađen kod 190.906 žena (32%). Od svih urađenih PAP testova 11.740 je bilo sa pozitivnim nalazima (6,1%).

Kolposkopija je urađena kod 31.473 žene, a potom 2465 biopsija kojima je patohistološki potvrđeno 112 karcinoma.



Slika 5 Obuhvat PAP testom

5.2 Statistika maligniteta melanoma

U Srbiji broj obolelih od melanoma nažalost stalno je u porastu. Zadnji zvaničan podatak koji imamo je iz 2007. godine kada je u Centralnoj Srbiji registrovano 410 osoba sa melanomom, dok je 5 godina ranije broj obolelih bio 316. Stopa godišnjeg porasta obolelih od melanoma u Srbiji iznosi oko 6%, što znači da će se broj obolelih praktično udvostručiti za 10 godina (Tabela 1). Ovakvu stopu porasta obolevanja, na svetskom nivou, beleži još jedino, rak pluća kod žena.

Smrtnost od melanoma, ne beleži nikakav značajan porast, po svim svetskim podacima. U Centralnoj Srbiji, u kratkom vremenskom praćenju od 5 godina, zabeležen je porast od 5%.

Godina	Žene	Muškarci	Ukupno
2002	152	164	316
2003	197	175	372
2004	215	177	392
2006	145	130	275
2007	214	196	410

Tabela 1 Incidenca melanoma u Centralnoj Srbiji

Godina	Žene	Muškarci	Ukupno
2002	73	63	136
2003	69	60	129
2004	86	70	156
2006	72	68	140
2007	93	80	173

Tabela 2 Smrtnost melanoma u Centralnoj Srbiji

Podaci za zemlje Evropske Unije za incidencu su sledeći: prosečna incidenca je najveća u Švajcarskoj a najniža u Grčkoj. Podaci za zemlje Evropske unije za smrtnost su sledeći: prosečna smrtnost najveća u Norveškoj a najniža u Grčkoj (Tabele 3 i 4).

Zemlje	Najniža incidenca	Prosečna incidenca	Najveća incidenca
Grčka	2,9		
Evropska unija		11,3	
Švajcarska			24

Tabela 3 Incidenca melanoma u Evropi

Zemlje	Najniža smrtnost	Prosečna smrtnost	Najveća smrtnost
Grčka	1,2		
Evropska unija		2,1	
Norveška			5,6

Tabela 4 Smrtnost od melanoma u Evropi

Zemlje	Najniži procenat	Prosek	Najveći procenat
Grčka	52		
Evropska unija		79,9	
Švajcarska			89

Tabela 5 Petogodišnje preživljavanje od melanoma u Evropi

Petogodišnje preživljavanje, nakon dijagnostikovanog melanoma u zemljama Evropske unije preživi prosečno oko 80% obolelih, najviše u Švajcarskoj a najmanje u Grčkoj. (Tabela 5).

Podatak o broju pacijenata sa melanoma koji prežive pet godina nakon ove dijagnoze za sada, mi nemamo u Srbiji. Ovaj podatak je važan iz razloga što ukazuje na efikasnost lečenja, odnosno povećanjem broja rano dijagnostikovanih melanoma doći će do značajnog porasta broja pacijenata koji nakon dijagnostike žive više od pet godina.

Na osnovu ovih podataka možemo uočiti sledeće karakteristike obolevanja od melanoma u Srbiji - incidenca je trenutno ispod Evropskog proseka, ali ima tendenciju stalnog rasta, smrtnost je iznad proseka i nemamo podatke o broju pacijenata koji žive više od pet godina nakon dijagnostikovanog melanoma.

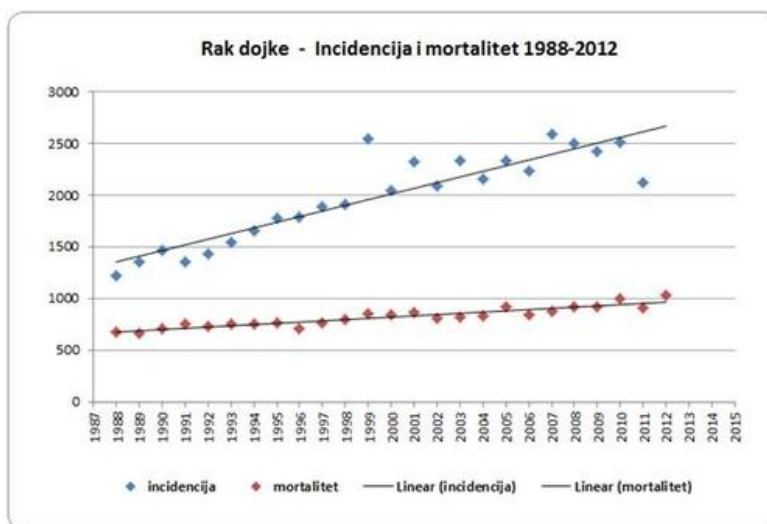
Ovi pokazatelji ukazuju nam da su mere primarne prevencije, skrininga i rane dijagnostike od ključne važnosti za uspešnu borbu protiv melanoma.

5.2.1 Rak dojke

Rak dojke je veoma česta ali i zloćudna bolest koja pogađa žene. Kod velikog broja novih slučajeva bolest bude otkrivena u fazi ranog raka, a kod manjeg broja kada se postavlja dijagnoza bolest već bude u uznapredovaloj fazi.

Prema podacima Registra za rak, Zavoda za javno zdravstvo Republike Srbije, godišnje od raka dojke oboli preko 2500 žena. Prema poslednjim dostupnim epidemiološkim podacima rak dojke je treći uzrok smrti u ženskoj populaciji u 2012. Stopa smrtnosti od raka dojke u Srbiji je jedna od najviših u Evropi, 2012. godine preko 1000 žena umrlo je od raka dojke.

Na sledećoj slici prikazano je kretanje incidencije i mortaliteta od raka dojke:



Slika 6 Statistički podaci o raku dojke

6 Baze podataka

Organizovan i uređen skup međusobno povezanih podatakaj jeste baza podataka. U njoj je jednostavno skladištiti podatke iste vrste, zatim njihovo pretraživanje i jednostavno manipulisane istim. Najkraće rečeno ona je skup organizovanih informacija koje se odnose na jednu određenu temu i koje se jednostavno mogu pregledati, pretraživati, menjati, sortirati, upoređivati. BP su postojale dosta pre upotrebe računara i postojale su u vidu papirne dokumentacije. Pojavom računara su dobile elektronski oblik a po pravilu, koristi veliki broj korisnika.

Podela podataka u bazi:

1. Ulazni podaci - opisuju transakcije
2. Izlazni podaci - odnose se na poruke i izveštaje
3. Radni podaci - obuhvataju i ulazne i izlazne podatke. Transakcija koja prebacuje jedno stanje u drugo prilikom obrade podataka.

Neka od svojstava baze podataka:

- One smanjuju ponavljanje podataka - podaci se unose samo jednom u bazu
- Podaci su nezavisni – nezavisni su od programa koji ih obrađuje
- Obezbeđuju pouzdanost podataka – ona se ostvaruje kontrolom pristupa
- Obezbeđuju integritet i tačnost podataka – podrazumeva se ispravnost i korektnost podataka.

Programi kojima je omogućen pristup i korišćenje podataka iz baze podataka nazivaju se sistemi za upravljanje bazama podataka (SUBP). Može se reći da se softverski sistem SUBP koristi za čuvanje i pretraživanje podataka. Ovi programi imaju zadatke oporavka baze podataka, upravljanja paralelnim izvršavanjem transakcija i zaštite podataka.

Postoji veliki broj alata za projektovanje baza podataka samo neki od njih su:

- SQL Server Management Studio
- MS Access
- MySQL
- Sybase Power Designer

6.1 E-R dijagram

Na sledećoj slici je predstavljen E-R dijagram skladišta podataka gde su predstavljeni sledeći informacioni zahtevi:

Lični podaci pacijenta, tj elektronski karton gde je primarni ključ LBO, ime prezime, pol, datum rođenja, adresa, mesto prebivališta, mesto rođenja, kontakt telefon i jmbg. Ovde je najzgodniji atribut za primarni ključ LBO pacijenta.

Pored ličnih podataka pacijenata, neophodno je pratiti i sve vakcine koje je primio, datum kada je primio svaku od vakcina.

Da bi se što bolje ispratilo zdravstveno stanje pacijenta, u cilju što boljeg skrininga, potrebno je pratiti porodičnu anamnezu, to jest sve bolesti u porodici. Takođe je, obavezno, praćenje svih hroničnih bolesti pacijenta, stadijum bolesti i od kada je bolest nastupila.

Neophodno je da se u skladište uvrsti katalog propisanih vakcina, koji sadrži šifru, naziv vakcine, dob u kojoj se daje vakcina, ako postoji propisana dob i na koje vreme se može davati ta vakcina.

Za katalog bolesti se prati šifra bolesti, naziv i latinski naziv.

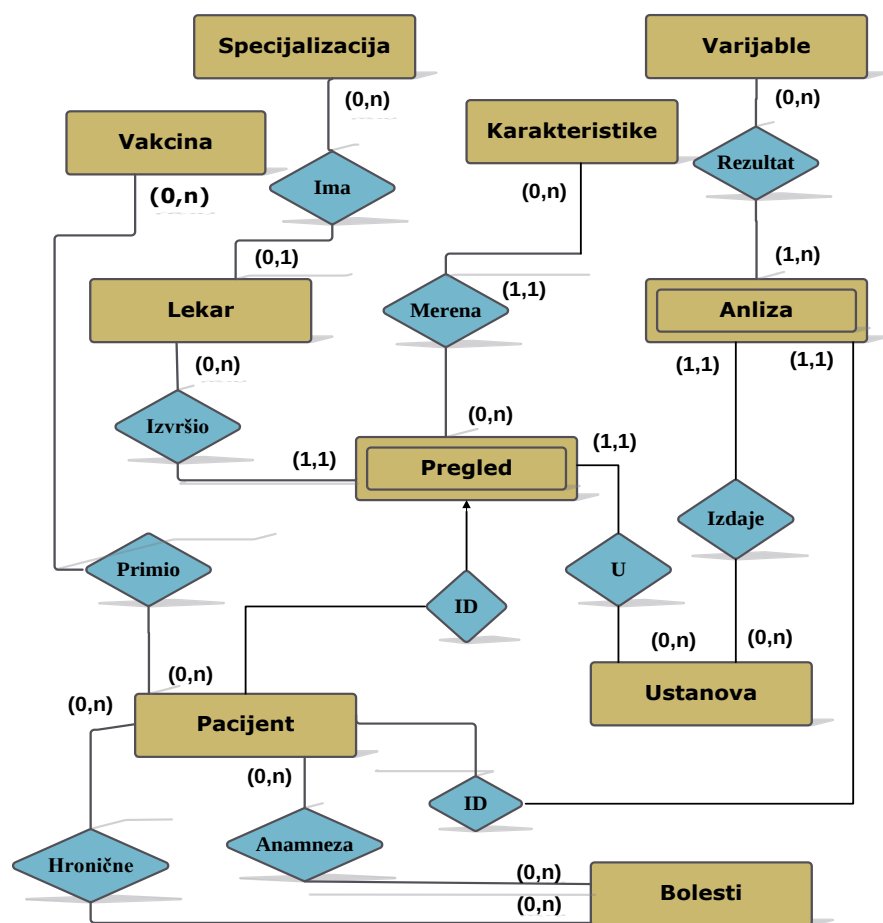
U skladištu podataka se prati mišljenja specijalista i lekara koji su vršili preglede nad pacijentom. Zbog toga je ubačen spisak lekara u zemlji, oni koji rade u državnim ustanovama ali i oni koji imaju licencu u privatnoj praksi. Za lekare se prati šifra lekara, matični broj građana, ime, prezime, adresa, mesto. Takođe je za svakog lekara potrebno pratiti koju specijalizaciju ima, ako je specijalista za neku oblast. Može se desiti da neki lekar nema specijalizaciju a opet da neki lekar ima specijalizaciju iz više oblasti. Za ove druge se prati samo specijalizacija koju on obalja

u svojoj lekarskoj praksi, tj na kom random mestu radi. Zbog ove evidencije u skladištu je neophodan katalog oblasti specijalizanata, koja ima šifru i naziv specijalizacije po nomenkleaturi.

Pacijentu se prate sve laboratorijske i druge analize na kojima je bio. Entitet analiza je ustvari laboratorijski izveštaj sa objedinjenim rezultatima svih analiza koje su tom prilikom rađene na nekom uzorku. U tom smislu analiza sadrži podatak na kog pacijenta se odnosi uzorak koji je uzet za analiziranje, datum kada su vršene analize, u kojoj ustanovi su vršene analize i rezultati za svaku od vršenih analiza. Entitet analiza poseduje i broj naloga za vršenje analize koji može biti deo primarnog ključa. Ovaj dokument dolazi od strane različitih ustanova, moguće je da se desi da različite ustanove daju isti broj naloga, te zbog toga je neophodno da se uspostavi identifikaciona zavisnost od pacijenta.

Ustanova ima svoju šifru, naziv, adresu, mesto, vrstu (laboratorija, dom zdravlja, roentgen...), oblik vlasništva (privatno, državna) i da li je još uvek aktivna (moguće je da je u međuvremenu zatvorena).

Da bi se pratili kontrolni i specijalistički pregledi mora se voditi evidencija o svakoj pacijentovoj poseti lekaru. Kako bi bila brža pretraga, najzgodnije je da pregled bude identifikaciono vezan sa pacijentom, tj. primarni ključ pregleda je LBO pacijenta i broj protokola pregleda. Pored primarnog ključa treba u svakom pregledu imati praćenje o datumu posete, lekaru koji je izvršio pregled, mišljenje lekara vrednosti karakteristika koje su tom prilikom merene, ako su merene. Karakteristike su na primer, visina, težina, krvni pritisak... Potrebno je uspostaviti šifarnik karakteristika u kom pored šifre i naziva, može stajati i opseg dozvoljenih vrednosti.



Slika 1 E-R dijagram podataka o pacijentima

6.2 Specifikacija tabela

PACIJENT:

LBO

Ime

Prezime

Adresa

Mesto_prebivališta

Datum_rođenja

Pol

Mesto_rođenja,

Kontakt_telefon

Jmbg

Primarni ključ: LBO

PREGLED:

LBO pacijenta

Broj protokola pregleda

Datum pregleda

Šifra lekara

Mišljenje lekara

LEKAR:

Šifra lekara

Matični broj građana

Ime lekara

Prezime lekara,

Adresa

Mesto

Telefon

Šifra specijalizacije

Primarni ključ: Šifra_lekara

SPECIJALNOST:

Šifra specijalizacije

Naziv specijalizacije

Primarni ključ: Šifra_specijalizacije

BOLEST:

Šifra_bolesti

Naziv

Latinski_naziv

Primarni ključ: Šifra_bolesti

ANAMNEZA:

Šifra_bolesti

LBO

Opis

Primarni ključ: Šifra_bolesti+LBO

HRONIČNE:

Šifra_bolesti

LBO

Stadijum

Od_kada

Primarni ključ: Šifra_bolesti+LBO

USTANOVA:

Šifra ustanove

Naziv

Adresa

Mesto

Vrsta ustanove

Oblik vlasništva

Status aktivnosti

Primarni ključ: Šifra_ustanove

ANALIZA:

LBO

Broj naloga

Uzorak

Datum analize

Šifra ustanove

Primarni ključ: LBO+broj_naloga

VAKCINA:

Šifra_vakcine

Naziv

Dob_primanja

Primarni ključ: Šifra_vakcine

PRIMIO:

Šifra_vakcine

LBO

Datum_primanja

Primarni ključ: Šifra_vakcine +LBO

REZULTAT:

LBO

Broj protokola pregleda

Broj naloga

Šifra varijable

Stanje varijable

Primarni ključ: LBO+Broj naloga+Šifra varijable

KARAKTERISTIKA:

Šifra_karakteristike

Naziv karakteristike

Opseg

Primarni ključ: Šifra_karakteristike

MERENA:

Šifra karakteristike

LBO

Broj protokola pregleda

Vrednost karakteristike

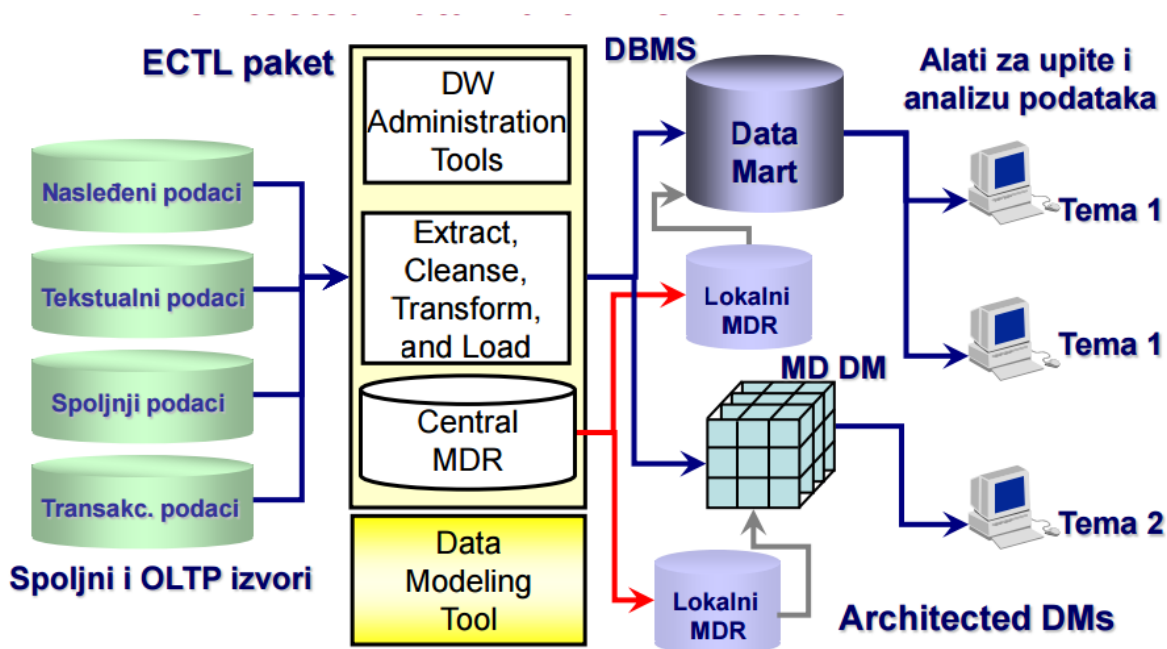
Primarni ključ: LBO+Broj protokola pregleda+Šifra karakteristike

Ovako isprojektovana baza bi trebalo da se isprojektuje u nekom od alata za projektovanje BP. Čuvanje podataka bi bilo relaizovano u DW.

6.3 Velike baze podataka - Datawarehouse

“Skladištenje podataka je metoda kojom se analizira i obrađuje velika količina podataka za podršku odlučivanju i upravljanju u preduzeću. Jedna je od metoda poslovne inteligencije.” (Wikipedija, 2017)

Arhitektura sistema se obično sastoji iz više slojeva i obuhvata servere VBP – sa odgovarajućim SUBP, područja između sistema, deo sa podacima (DSA), skladištenje operacionih podataka (ODS) – memorišu se podaci taktičkog nivoa iz operativne BP ili iz drugih spoljnjih izvora. Ovo je mesto operativne pripreme podataka za upis u sistem VBP sa softverskom podrškom za kreiranje, osvežavanje i korišćenje velike baze podataka.



Slika 7 Arhitektura jednog primera DV¹

6.3.1 Projektovanje VBP

Skladište podataka zapravo je centralizovana baza podataka koja sadrži sve relevantne podatke vezane za problematiku u standardizovanom formatu. Ovi podaci su formatirani u više organizacionih delova. Pojedinačni delovi baze podataka se razlikuju po strukturi i standardizaciji i mogu biti implementirani čak i pod različitim sistemima za upravljanje bazama podataka.

¹ Slika preuzeta sa mreže http://www.acs.uns.ac.rs/sites/default/files/1_BPA_DW_OsnovniKoncepti.pdf

Funkcije koje obuhvata projektovanje VBP

- ☐ KreirajBazuPodataka (Baza, ImeFajla, Veličina, MaksVeličina)
- ☐ IzvršiScriptuBaze (Baza, List)
- ☐ PrikažiListuIzvoraPodataka (Server)
- ☐ SnimiIzvorPodataka (IzvorPodataka)
- ☐ UčitajMetaPodatke (IzvorPodataka)
- ☐ SnimiPoslovneKljučeve (Kolona, Vrednost)
- ☐ IdentifikujHubTabele (Pravilo)
- ☐ IdentifikujLinkTabele (Pravilo)
- ☐ IdentifikujSatelitTabele (Pravilo)
- ☐ ModifikujSatelitKandidate (Kolona, Vrednost)
- ☐ KreirajHubTabele (Pravilo, SkladištePodataka)
- ☐ KreirajLinkTabele (Pravilo, SkladištePodataka)
- ☐ KreirajSatelitKandidate (Pravilo, SkladištePodataka)
- ☐ PrikažiVrednostMetaPodatka (SkladištePodataka, PrelaznaBaza)
- ☐ UčitajHubTabele (Pravilo)
- ☐ UčitajLinkTabele (Pravilo)
- ☐ UčitajSatelitTabele (Pravilo)
- ☐ KreirajDMProcedureTabele (ImeBaze, List)
- ☐ SnimiDataMartIzvorPodataka (SkladištePodataka)
- ☐ UčitajDVMetaPodatke (SkladištePodataka)
- ☐ SnimiPoslovneMere (Kolona, Vrednost)
- ☐ IdentifikujInicijalneDimenzije (Pravilo)
- ☐ SnimiKonačneDimenzije (Kolona, Vrednost)
- ☐ OznačiVremenskeDimenzije (Kolona, Vrednost)
- ☐ KreirajTabeluVremenskihDimenzija (StartDate, EndDate, DataMart)
- ☐ PrikažiIzabraneMere (DataMart)
- ☐ SnimiNačinAgregacijeMera (Kolona, FaktIme, Formula)
- ☐ KreirajTabeleDimenzijaMera (Pravilo, ImeTabeleMera)
- ☐ PrikažiDataMartMetaPodatke (DataMart)

- ☐ UčitajDimTabele (Pravilo)
- ☐ UčitajFaktTabele (Pravilo, FaktTabela)
- ☐ PrikažiListuIzvoraPodataka (Server)
- ☐ SnimiPolustrukturiraniFajl (Pravilo)
- ☐ UčitajSSMetaPodatke (TipFajla, ImeFajla, PrelaznaBaza)
- ☐ KreirajNovuSatelitTabelu (Pravilo, SkladištePodataka)
- ☐ UčitajSatelitTabeluSS (Pravilo, Fajl)
- ☐ KreirajNovuDimTabelu (Pravilo, DataMart, SatelitTabela)
- ☐ UčitajDimTabeluSS (Pravilo, DataMart, DimTabela)

6.3.1.1 Primena fazi logike

Kod fazi logike pripadnost elementa nekom skupu se meri u procentima zato što ona nema jasno definisanu pripadnost. Procenti su predstavljeni decimalama od nula do jedan (100%). U fazi logici, istinitost svake tvrdnje meri se u procentima.

U samom primeru pitanja koja bi bila zgodna za model fazi je starosna dob, gde bi procenat pomnožen sa 100 bila godina života, s tim da godine više od sto bi se računao i dalje kao sto.

Međutim ovaj atribut, životna dob ne mora biti ovako definisan jer kod nekih bolesti u zavisnosti od životne dobi imaju veći procenat rizika oboljevanja, od drugih perioda. Npr, kancer grlića materice se kreće od 24 do 37 po jednoj kvoti, od 37 do 50 po drugoj, od 50 do 55 je u pitanju treća kvota, dok je posle menopauze taj rizik minimalan, kao i pre 23. godine života.

Atribut pušački staž je takođe pogodan za primenu fazi mere, kada se analizira kategorija visokog rizika kancera pluća.

Rak debelog creva ima veći rizik kod populacije muškog pola, tako da i ovaj atribut može imati brojčanu vrednost od 0 do 1 za kategorizaciju ove bolesti. Opet u većini slučajeva, ovaj atribut je nebitan, ili pak prvi u stablu odlučivanja koji odbacuje pacijenta ako je npr. muški pol jer se meri rizik za kancer dojke ili kancer grlića materice.

6.3.1.2 Robotika i automatizacija

Roboti su smart mašine. U samom studijskom primeru za animiranje i kontaktiranje pacijenata koji treba da se jave na pregled bi prvo trebalo da se ostvari putem imejla, kao najjeftinijem načinu komunikacije. Međutim, ako se pacijent ne odazove u dogledno a i programirano vreme, ili ako pacijent nema imejl, neophodno je da se pacijent kontaktira putem telefona pomoću telefonskog automata. Ovde se podrazumeva da svi pacijenti moraju da ostave svoj kontakt telefon. Međutim, moguće je da dođe do promene telefonskog broja a zdravstvena ustanova ne bude obavешtena o tome, zbg toga postoji treća solucija a to je pismenim putem pozivati pacijente slanjem pozvinice na kućnu adresu.

Za realizaciju automatskog zvanja, bi trebalo pripremiti automatski snimljene poruke za telefonsko obaveštenje, gde bi u kratkoj poruci trebalo da stoji ime i prezime pacijenta i da treba da se javi u Dom zdravlja kojem pripada pacijent, kako bi odradio kontrolne preglede u cilju prevencije određene bolesti.

Primer:

“Poštovani, Petar Petrović, dužni ste da se javite u Dom zdravlja kome pripadate, radi kontrolnog pregleda, u cilju prevencije raka prostate. Molimo Vas, budite odgovorni prema svom zdravlju. Srdačan pozdrav Vam želi tim medicinskog osoblja.”

Ova forma je primenljiva na sve vrste bolesti, menjane imena i prezima, kao vrste bolesti se automatski menja od slučaja do slučaja.

Realizacija automatskog slanja mejla, bi trebalo sadržati obaveštenje, koje bi takođe kao i u telefonskoj poruci moralo da stoji ime i prezime pacijenta koji treba da se javi u Dom zdravlja kojem pripada pacijent, kako bi odradio kontrolne preglede u cilju prevencije određene bolesti.

Ova forma, takođe kao i mejl poruka, je primenljiva na sve vrste bolesti i sve pacijente, menja se ime, prezime, i vrste bolesti. Obaveštenja putem pisma su istovetne forme kao i mejl.

Tokom izrade modela, moguće je napraviti izmenu u formi poruka, gde bi se pored vrste bolesti navelo i da pacijent pripada rizičnoj grupi. Za ovu promenu je neophodno konsultovati psihologa. Postoji mogućnost da se pacijent uplaši, što može da dovede do kontraefekta ili do poželjnog efekta.

7 Neophodna infrastruktura za DM

Cena sistema nije velika jer već postoji određena infrastruktura, zapravo u većini opština postoji informacioni sistem koji ima program za elektronski karton. Izdaci bi bili cena modelovanja DM sistema i dodatna oprema za proširivanje baze podataka jer je ideja da se objedine i podaci koji se nalaze u privatnim lekarskim praksama. Takođe, postoji izvestan trošak na opremu koja bi bila zadužena za operativni deo posla, naime oprema koja bi automatski obavljala telefonske pozive, automatski slala mejlove i poštu.

S obzirom da je glavni cilj ovakvog projekta zdravija nacija uz preventivno delovanje, a preventiva je uvek daleko jeftinija, može se reći da svi navedeni troškovi ne mogu da se porede sa troškovima hemioterapije, zračenja, bolovanja...

Kao primer korišćenja neuronske mreže može da predstavlja klasifikaciju rizičnih grupa za određene vrste kancera, pre svega kancera koji su nažalost po statistici najučestaliji i najpogubniji, na našem podnevlju.

U bazi podataka imamo uzorke koji predstavljaju podatke o pacijentima, tačnije o pregledima pacijenata, oni su smešteni u tabelama pacijent, pregled, analize, rezultati, merene karakteristike. Zadatak je napraviti model rizičnih grupa pacijenata, po bolestima na osnovu analiziranih

karakteristika, vrednosti koje direktno utiču na konkretnu kancerogenu pojavu, tj. klasifikovati pacijente ocenama od 1-10 za svaku bolest koja se ispituje.

7.1 Projektovanje odgovarajuće baze

Karakteristike baze podataka:

- Baza više promenljivih;
- Broj instanci

Karakteristike atributa: celi brojevi, realni brojevi, logički podaci (yes/no), tekstualni podaci sa zadatom listom vrednosti, zadati intervali celobrojne vrednosti.

- Broj atributa: 12
- Zadatak: Klasifikacija
- Nedostajuće vrednosti Nema

Ulazni atributi:

1. Godine života
2. Pol
3. Pušački staž
4. Gojaznost
5. Porodična anamneza
6. Hronične bolesti
7. Prethodne disfazije koje se odnose na bolest
8. Urednost kontrolisanja
9. Socio-ekonomski faktor
10. Gojaznost
11. Fizička forma

Izlazni atribut (labela, tj. atribut po kome se vrši klasifikacija): 12. Kategorija

Međutim sedmi atribut, koji se odnosi na disfazije koje su vezane za konkretnu bolest ovde nije detaljno definisan. Zapravo suština neuronske mreže i jeste u prepletanju višeslojnih mreža, a ovaj sedmi atribut bi trebao da se sastoji od različitih atributa koji bi takođe trebalo da se iskristališe kao relevantni atribut za korišćenje mreže.

Recimo za kancer dojke je značajno da li su dojke policistične, da li je pacijent gojazan, da li se bavi sportom (ekstremnim, profesionalno ili rekreativno). Za kancer melanoma ništa nema težinu od ovoga što je navedeno za kategorizaciju rizične grupe kancera dojke. Takođe za kancer pluća ne utiče ništa od navedenog, niti igra ulogu pol pacijenta, dok kod kancera reproduktivnih organa, naravno da je pol prva selekcija.

Ovde bi trebalo uvideti da je dobar primer velika ekspertska baza, tj. da kvalitet modela zavisi od količine podataka i deskriptivnosti same baze podataka.

8 Primeri projektovani u softverskom alatu

8.1 Primer detekcije kritičnih grupa

REM Generated by v1.01 of e2gRuleWriter 03/06/2017 21:20 from:

```
RULE [Kancer grlica materice]
If [pol = muski] = false and
[porodicna anamneza pozitivna] = true and
[godine<41] = false
Then [Rizicna grupa kancera grlica materice] = "Rizicna grupa kancera grlica materice"
```

```
RULE [Kancer dojke]
If [pol = muski] = false and
[porodicna anamneza pozitivna] = true and
[godine>24] = true
Then [Rizicna grupa kancera dojke] = "Rizicna grupa kancera dojke"
```

```
RULE [Kancer prostate]
If [pol = muski] = true and
[porodicna anamneza pozitivna] = true and
[godine>50] = true
Then [Rizicna grupa kancera prostate] = "Rizicna grupa kancera prostate"
```

```
RULE [Kancer pluca]
If [promene na plucima] = true and
[pusac] = true
Then [Rizicna grupa kancera pluca] = "Rizk na kancer pluca"
```

```
RULE [Kancer melanoma]
If [porodicna anamneza pozitivna] = true and
[promena na mladezu] = true
Then [Rizicna grupa malignog melanoma] = "Rizicna grupa melanoma"
```

```
PROMPT [pol = muski] YesNo CF
""
```

```
PROMPT [porodicna anamneza pozitivna] YesNo CF
""
```

```
PROMPT [godine<41] YesNo CF
""
```

```
PROMPT [god>24] YesNo CF
""
```

PROMPT [policisticnost dojke] YesNo CF
""

PROMPT [poslednji PAP<1 god] YesNo CF
""

PROMPT [podlednji PAP >6 meseci] YesNo CF
""

PROMPT [porodjaj] YesNo CF
""

PROMPT [promena na mladezu] YesNo CF
""

PROMPT [promene na plucima] YesNo CF
""

PROMPT [pusac] YesNo CF
""

PROMPT [godine>50] YesNo CF
""

MAXVALS [pol = muski] 2
MAXVALS [Rizicna grupa kancera prostate] 3
MAXVALS [Rizicna grupa kancera dojke] 3
MAXVALS [Rizicna grupa kancera grlica materice] 3
MAXVALS [] 8

DEFAULT [Rizicna grupa kancera grlica materice] = "Nije rizicna grupa"

MINCF 80

8.2 Primer skrininga kancera grlića materice

Rak grlića je drugi najčešći karcinom kod žena posle karcinoma dojke. U zemljama EU i razvijenim zemljama sveta bolest je praktično iskorenjena. Predstavlja bolest koja pripada istoriji medicine. To je postignuto opštim društvenim naporom, masovnim organizovanim programima primarne i sekundarne prevencije.

U Srbiji, nažalost, godišnje umre između 300 do 900 žena od raka grlića materice. Incidencija obolevanja je između 12/100 000 i čak 35/100 000.

Postoje primarni i sekundarni metodi prevencije. Primarna prevencija je sprečavanje nastanka bolesti (na primer vakcinacija), a sekundarna je rano otkrivanje bolesti (na primer kolposkopija ili PAP bris). Za mnoge maligne bolesti primarne prevencije nema, jer se još uvek ne zna šta ih izaziva (za razliku od infektivnih bolesti).

Ako se ova podmukla bolest pgm otkrije na vreme, u premalignoj fazi, izlečiva je 100%.

Uz overenu zdravstvenu knjižicu ili u procesu organizovanog skrininga postupak je potpuno besplatan.

Rak grlića materice je jedan od retkih malignih promena kod kojih se zna šta ga izaziva. Izaziva ga Humani Papiloma virus (HPV). Njega ima u preko 80 različitih sero tipova (isti virus ali malo strukturno promenjen). HPV se deli na one serotipove sa izraženim onkogenim potencijalom i na manje onkogene. Za nastanak raka grlića materice je neophodno prisustvo HPV-a. Karcinom grlića materice se razvija isključivo pod uticajem HPV-a, ali prisustvo HPV ne mora usloviti pojavu karcinoma. Nema karcinoma bez HPV-a, ali ima HPV bez onkološkog ishoda.

HPV se prenosi polnim putem (seksualnim odnosima, analnim ili oralnim seksom).

Žene koje imaju povišen rizik za nastanak raka grlića materice:

- Žene koje su seksualno aktivne od ranoj mladosti (pre 16. godine)
- Žene koje su imale veliki broj seksualnih partnera
- Žene koje su imale odnose sa muškarcem koji ima HPV infekciju
- Žene kod kojih postoji anamneza seksualno prenosivih bolesti ili genitalnih bradavica
- Žene koje su pušači ili bivši pušači
- Žene kod kojih postoji odsustvo normalnog imunog odgovora (sve HIV pozitivne žene, koje su imale transplantaciju organa ili koje iz različitih razloga primaju imunosupresive)
- Žene koje dugo uzimaju oralne kontraceptive
- Žene koje ne odlaze redovno na kontrole
- I one kojima prethodno nisu redovno radile Papanikolau test (PAP test)

REM Generated by v1.01 of e2gRuleWriter 03/07/2017 07:52 from:

```
RULE [velika rizicnost 1]
If [pol] = "zenski" and
[porodicna anamneza] = visok rizik and
[godine zivota] = "23+" and
[HP virus] = true and
[dugotrajna oralna kontracepcija] = true and
[pregled>6 mesici] = true
Then [najvisa rizicnost] = "odmah pozivati"
```

```
RULE [visok rizik 2]
If [pol] = "zenski" and
[porodicna anamneza] = blazi rizik and
[vakcinacija HPV] = false and
[pusac ili bivsi pusac] = true and
[CIN2, CIN3] = true and
[pregled>6 mesici] = false
Then [najvisa rizicnost] = "kontrole svakih 6 meseci"
```

RULE [visok rizik 3]

If [pol] = "zenski" and

[godine zivota] = "23+" and

[vakcinacija HPV] = false and

[anamneza venericnih bolesti] = true and

[stupanje u seksualni odnos pre 16g] = true and

[pregled < 12 meseci] = true and

[los socio-ekonomski status] = true

Then [najvisa rizicnost] = "reodvan skrining na godinu dana"

RULE [visok rizik 4]

If [pol] = "zenski" and

[porodicna anamneza] = bez rizika and

[godine zivota] = "37+" and

[vakcinacija HPV] = false and

[dugotrajna oralna kontracepcija] = true and

[anamneza venericnih bolesti] = false and

[CIN2, CIN3] = true and

[pregled < 12 meseci] = true and

[los socio-ekonomski status] = true

Then [najvisa rizicnost] = "reodvan skrining na godinu dana"

RULE [grupa 2]

If [pol] = "zenski" and

[porodicna anamneza] = false and

[godine zivota] = "23+" and

[vakcinacija HPV] = true

Then [najvisa rizicnost] = "nije rizicna grupa"

RULE [grupa 3]

If [pol] = "zenski" and

[porodicna anamneza] = visok rizik and

[godine zivota] = "37+" and

[HP virus] = false and

[vakcinacija HPV] = false and

[pusac ili bivsi pusac] = true and

[dugotrajna oralna kontracepcija] = false and

[anamneza venericnih bolesti] = false and

[CIN2, CIN3] = false and

[stupanje u seksualni odnos pre 16g] = false

Then [najvisa rizicnost] = "nizak rizik"

RULE [grupa 4]

If [pol] = "zenski" and

[porodicna anamneza] = blazi rizik and

[HP virus] = true and

[pregled>6 mesici] = true
Then [najvisa rizicnost] = "slati poziv"

PROMPT [porodicna anamneza] YesNo CF
""

PROMPT [HP virus] YesNo CF
""

PROMPT [vakcinacija HPV] YesNo CF
""

PROMPT [pusac ili bivsi pusac] YesNo CF
""

PROMPT [dugotrajna oralna kontracepcija] YesNo CF
""

PROMPT [anamneza venericnih bolest] YesNo CF
""

PROMPT [CIN2, CIN3] YesNo CF
""

PROMPT [stupanje u seksualni odnos pre 16g] YesNo CF
""

PROMPT [pregled< 12 meseci] YesNo CF
""

PROMPT [pregled>6 meseci] YesNo CF
""

PROMPT [los socio-ekonomski status] YesNo CF
""

MINCF 80

8.3 Primer klasifikacije pacijenata za poziv na pregled

REM Generated by v1.01 of e2gRuleWriter 03/07/2017 08:48 from:

RULE [debeli crevo]
If [godine] >35 & <42 and
[porodicna anamneza visokog rizika] = "debeli crevo" and
[pacijent kontakt telefon] = true and

[pacijent ima aktivan mejl] = true and
[rizicna grupa] = "kancer debelog creva" and
[kontrola vise od godinu dana] = true
Then [kontaktirati] = "slati mejl" and
[kontrolisanje] = "kancer debelog creva"

RULE [kancer dojke]
If [pol] = "zenski" and
[pusacki staz] >10 and
[porodicna anamneza visikog rizika] = "dojka" and
[pacijent kontakt telefon] = true and
[rizicna grupa] = "kancer dojke" and
[kontrola vise od godinu dana] = true
Then [kontaktirati] = "telefonski kontakt" and
[kontrolisanje] = "kancer dojke"

RULE [kancer grlica]
If [pol] = "zenski" and
[pusac] = true and
[pacijent kontakt na kucnu adresu] = true and
[pacijent kontakt telefon] = false and
[pacijent ima aktivan mejl] = false and
[rizicna grupa] = "kancer grlica materice" and
[kontrola vise od godinu dana] = true
Then [kontaktirati] = "poziv na kucnu adresu" and
[kontrolisanje] = "kancer grlica materice"

RULE [leukemija]
If [porodicna anamneza visikog rizika] = "leukemija" and
[pacijent kontakt telefon] = true and
[pacijent ima aktivan mejl] = false and
[rizicna grupa] = "leukemija" and
[kontrola vise od godinu dana] = true
Then [kontaktirati] = "telefonski kontakt" and
[kontrolisanje] = "leukemija"

RULE [prostata]
If [pol] = "muski" and
[godine] >42 & <50 and
[pusacki staz] >10 and
[pacijent kontakt telefon] = true and
[pacijent ima aktivan mejl] = true and
[rizicna grupa] = "kancer prostate" and
[kontrola vise od godinu dana] = true
Then [kontaktirati] = "slati mejl" and
[kontrolisanje] = "kancer prostate"

```
RULE [rak kostiju]
If [pusac] = true and
[pacijent kontakt telefon] = true and
[pacijent ima aktivan mejl] = true and
[rizicna grupa] = "kancer kostiju" and
[kontrola vise od godinu dana] = true
Then [kontaktirati] = "slati mejl" and
[kontrolisanje] = "kancer kostiju"
```

```
RULE [rak pluca]
If [pusacki staz] >15 and
[porodicna anamneza visikog rizika] = "pluca" and
[pacijent kontakt telefon] = false and
[pacijent ima aktivan mejl] = true and
[kontrola vise od godinu dana] = true
Then [kontaktirati] = "telefonski kontakt" and
[kontrolisanje] = "kancer pluca"
```

```
PROMPT [pusac] YesNo CF
""
```

```
PROMPT [pacijent kontakt na kucnu adresu] YesNo CF
""
```

```
PROMPT [pacijent kontakt telefon] YesNo CF
""
```

```
PROMPT [pacijent ima aktivan mejl] YesNo CF
""
```

```
PROMPT [kontrola vise od godinu dana] YesNo CF
""
```

```
MINCF 80
```

9 Opis softverskog paketa Veka

WEKA je softver sa otvorenim kodom razvijen u programskog jeziku Java na Univerzitetu Waikato na Novom Zelandu, koji sadrži mnogobrojne algoritme automatskog učenja a namenjen je rešavanju različitih data mining problema. Algoritmi mogu biti direktno primenjeni na odgovarajući set podataka ili pozvani iz određenog Java programskog koda. WEKA sadrži alate koji omogućavaju klasifikaciju, regresiju ili klasterovanje podataka, utvrđivanje odgovarajućih pravila, preprocesiranje ulaznih podataka, statističku evaluaciju modela dobijenog primenom

metode automatskog učenja, kao i vizuelizaciju ulaznih podataka i rezultata obučavanja. Primena metoda automatskog učenja omogućava automatsku analizu velike količine podataka i utvrđivanje najrelevantnijih informacija koje mogu biti upotrebljene u različitim prediktivnim procesima. Primena ovih metoda takođe može doprineti bržem i tačnijem donošenju odluka. S druge strane, simbolika naziva leži u činjenici da je Weka ptica radoznale prirode bez mogućnosti letenja, koja predstavlja autohtonu vrstu na ostrvima Novog Zelanda i može se pronaći samo tamo.

9.1 Instalacija

Softver WEKA distribuira se u skladu sa GNU General Public Licence, može da radi na gotovo svakoj platformi a testiran je u različitim operativnim sistemima kao što su Linux, Windows, Macintosh, pa čak i PDA (engl. Personal Digital Assistant). WEKA poseduje četiri različita korisnička interfejsa, pri čemu su Explorer, Knowledge Flow i Experimenter grafički korisnički interfejsi dok upotreba četvrtog interfejsa podrazumeva unošenje tekstualnih komandi sa komandne linije.



Slika 8 Ulazni prozor alata Veka

10 Zaključak

Predviđa se da će uspešne organizacije biti samo one koje budu znale da iskoriste znanje koje sada poseduju i koje budu umele da generišu novo znanje. Znanje je zapravo nevidljivo i već sada se nalazi u memorijama računara i bazama podataka. Implementirani proizvodi data mining su uneli revolucionaran pogled na rešavanje problema upravljanja podacima poslovnim sistemima. Na ovaj način, kompanije mogu da u potpunosti da iskoriste sakupljanje podatke a samim tim se poboljšava poslovanje, donošenje poslovnih odluka, i unaprede izuzetno važnu komunikaciju sa klijntima.

Cilj ovog rada je da da predlog za upotrebu osnovnih mogućnosti, kada je data mining u pitanju. Razvojem ovakvog modela omogućeno je da se strategija očuvanja zdravlja podigne na viši nivo kao i kontrolnu strategiju prevencije tj. omogućavanje timskog rada medicinskih stručnjaka i IT tehnologije. Performanse ovako predloženog okvira modela razvijenog sistema bi mogle za relativno kratko vreme da budu sagledane.

Kako bi se pristup, koji je predstavljen u ovom radu, primenio u praksi, neophodno je da se razvije kompletan model. To podrazumeva dalji razvoj, potpunu analizu i izradu prototipa ali i konačne aplikacije ili, što je još bolje rešenje, razviti potpuno novi DM model koji će se bazirati na tehnikama i metodama koje su prezentovane u radu.

U trenutku pisanja ovog zaključka, u medijima se prezentuje e-servis koji podseća građane da im ističe neki dokument. Autor ovog rada, smatra da je daleko bitnije da postoji e-servis koji bi podsećao građane da im je vreme da idu na kontrolni pregled radi očuvanja svog zdravlja.

Bez obzira na napredne metode, tehnike i algoritme DM-a, bez obzira na ogromnu količinu podataka kojom možemo raspolagati, ipak je čovek u središtu svakog procesa istraživanja podataka.

Reference

1. Stojković, M. (2001). *Statistika*. Subotica: Ekonomski fakultet u Subotici.
2. Bach, M. P. (2005). Rudarenje podataka u bankarstvu. *STRUČNI RAD* (pp. 182-193). Zagreb: Zbornik Ekonomskog fakulteta u Zagrebu.
3. Berry, M. J., & Linoff, G. S. (2000). *Mastering Data Mining*. Chichester: , M.J.A., Linoff, G.S. Wiley.
4. Bramer, M. (2007). *British Library Cataloguing in Publication data*. London: Springer-Verlag London limited.
5. Filipović, L., & Papić-Blagojević, N. (2013). *Kvantitativne metode*. Novi Sad: Alfa-Graf NS.
6. Gračanin, V. K.-Š. (2009). *DATA MINING*. Kragujevac: Univerzitet u Kragujevcu.
7. Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. San Francisco,: Morgan Kaufmann Publishers.
8. Ilić, V. (2017, 1 22). *Veštačka inteligencija*. Retrieved from http://solair.eunet.rs/~ilicv/AI_index.htm
9. M.P.B, & M.V. (2016, 11 3). *profil-strucnjaka*. Retrieved from <http://www.skladistenje.com>.
10. *melanom-statistika.asp*. (n.d.). Retrieved from <http://orshospital.rs/melanom-statistika.asp>
11. *Ministarstvo zdravlja*. (n.d.). Retrieved from <http://www.zdravlje.gov.rs/downloads/Zakoni/Strategije/Nacionalni%20Program%20Srbija%20Protiv%20Raka.pdf>
12. Mogin, P., Luković, I., & Govedarica, M. (2004). *Principi projektovanja baza podataka*. Novi Sad: Fakultet tehničkih nauka .
13. Pyle, D. (1999). *Dana Preparation for Data Mining*. San Franciso: Morgan Kaufmann.
14. Pyle, D. (2003). *Business Modelling and Data Mining*. San Francisco: Morgan Kaufmann Publishers.
15. *rak-dojke-statistika/*. (n.d.). Retrieved from <http://www.onkologija.hr/rak-dojke/rak-dojke-statistika/>
16. Savić, M. (2005). *Poslovna statistika*. Subotica: Autor.

17. Simeunović, V. (2015). *INFORMACIONI MODEL I SOFTVERSKA PODRŠKA ZA PREDVIĐANJE USPJEŠNOSTI STUDIRANJA - doktorska disertacija*. Novi Sad: Univerzitet u Novom Sadu Fakultet tehničkih nauka.
18. Subašić, P. (1997). *Fazi logika i neuronske mreže*. Beograd: Tehnička knjiga.
19. *Svet kompjutera*. (2017, 1 25). Retrieved from Svet kompjutera: <http://www.sk.rs/2005/05/skpr01.html>
20. *Wikipedija*. (2017, 2 6). Retrieved from Wikipedija: https://hr.wikipedia.org/wiki/Skladi%C5%A1tenje_podataka
21. Witten, I. H., & Frank, E. (2005). *Practical Machine Learning Tools and Techniques*,. Elsevier Inc.