KOMPARATIVNA ANALIZA PREDIKTIVNIH TEHNIKA RUDARENJA PODATAKA

Ivanišević, Frane

Master's thesis / Diplomski rad

2016

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Split, Faculty of economics Split / Sveučilište u Splitu, Ekonomski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:124:131094

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Repository / Repozitorij:

REFST - Repository of Economics faculty in Split

zir.nsk.hr

UNIVERSITY OF SPLIT

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI



SVEUČILIŠTE U SPLITU

EKONOMSKI FAKULTET

DIPLOMSKI RAD

KOMPARATIVNA ANALIZA PREDIKTIVNIH TEHNIKA RUDARENJA PODATAKA

MENTOR: **STUDENT:**

izv.prof.dr.sc. Mario Jadrić **Frane Ivanišević**

Split, rujan 2016

SAŽETAK

Ovaj rad uspoređuje 4 različite prediktivne tehnike rudarenja podataka na 3 različita skupa podataka.Skupovi podataka su posebni te imaju različite kombinacije karakteristike: broj instanci,broj atributa i broj klasa.Skupovi podataka su predstavljeni i analizirani te iskorišteni za predviđanje klasa. Skupovi podataka preuzeti sa Interneta su čisti i spremni za analizu.Priprema skupova podataka obuhvaćala je samo odabir atributnih podskupova za izgradnju modela.. U teorijskom dijelu rada detaljno je opisano sve od pojma rudarenja podataka do 4 najpopularnija algoritma za klasificiranje klasa.

ABSTRACT

This thesis compares four diffrent predictive data-mining tehniques on three diffrent data set. These data are unique and have diffrent combination of the characteristics: number of instances,number of attributes and number of classes. Data set are introduced ,analyzed and used for class prediction.Data set downloadeded from UCI repository are clean and ready for analysis. Data set preperation includes attribute subset selection for building predictive model.In theoretic part of thesis there is detail description of data-mining proces to 4 most used classification algorithm

Contents

1.UVOD.....6

1.2PROBLEM ISTRAŽIVANJA 6
1.2 PREDMET ISTRAŽIVANJA.....8
1.3 CILJEVI ISTRAŽIVANJA 9
1.4 ISTRAŽIVAČKE HIPOTEZE:.....9
1.5 METODE ISTRAŽIVANJA 10
1.6 DOPRINOS ISTRAŽIVANJA 12
1.7. STRUKTURA DIPLOMSKOG RADA.....12

2.RUDARENJE PODATAKA.....13

2.1 INFORMACIJSKO DOBA 13
2.2. ŠTO JE RUDARENJE PODATAKA 14
2.3 METODOLOGIJA RUDARENJA PODATAKA.....16
2.4 KATEGORIZACIJA.....18
2.5 PREDIKTIVNO MODELIRANJE 20
2.6 KLASIFIKACIJA 22
2.7 EVALUACIJA MODELA.....25

4.METODE KLASIFIKACIJE.....31

4.1 STABLO ODLUČIVANJA.....31
4.2 NEURONSKE MREŽE 41
4.4 NAIVNI BAYESOV ALGORITAM 48

5.SKUPOVI PODATAKA 51

5.1 'BANK MARKETING' SKUP PODATAKA 52
5.2 'SQUASH-STORED' SKUP PODATAKA.....53
5.3 'NURSERY' SKUP PODATAKA.....55

6.KOMPARATIVNA ANALIZA I REZULTATI 56

6.1 WEKA-ALAT ZA RUDARENJE PODATAKA 58
6.2 KRITERIJI KORIŠTENI U OVOJ KOMPARATIVNOJ ANALIZI.....66
6.3 ANALIZA 'BANK' SKUPA PODATAKA.....66
6.4.ANALIZA 'SQUASH-STORED' SKUPA PODATAKA 71
6.5.ANALIZA 'NURSES' SKUPA PODATAKA 75
6.6.OSVRT NA TEZE 79

7.ZAKLJUČAK 80

8. POPIS SLIKA I TABLICA 81

9.LITERATURA.....83

1.UVOD

1.2PROBLEM ISTRAŽIVANJA

Rudarenje podataka je interdisciplinarno, skup disciplina, uključujući statističke sustave baze podataka, strojno učenje, vizualizaciju te informacijsku znanost.

Rudarenje podataka je analiza (često velikih) opservacijskih podatkovnih setova s ciljem pronalaženja neočekivanih veza ili prikaza podataka koji su za vlasnika podataka novi i korisni. Spomenuti veze i prikazi se često nazivaju modelima ili uzorcima. Oni se pak mogu izraziti kao npr. linearne jednadžbe, pravila, segmenti, grafovi, stablaste strukture i sl. Kada se govori o opservacijskim podacima onda se misli na one koji nisu prikupljeni s ciljem rudarenja. Najčešće se rudarenje podataka izvodi nad podacima koji su prikupljani zbog praćenja raznih transakcija ili operativnih događaja u specifičnim sredinama. Zbog toga se rudarenje podataka često naziva sekundarnom te je to i glavna razlika između nje i statistike.

Funkcionalnosti rudarenja podataka se koriste kako bi se specifirala vrsta uzorka koji se traži prilikom rudarenja. Rudarenje može biti klasificirano u dvije kategorije kao što je deskriptivna ili prediktivna. Deskriptivno rudarenje karakterizira prikaz osnovnih specifikacija podataka u bazi podataka, dok prediktivno preferira zaključivanje nad podacima kako bi bilo moguće izvesti predviđanja.

Prediktivno rudarenje podataka možemo podijeliti na klasificiranje i predviđanje u ovom radu ćemo uspoređivati različite klasifikacijske tehnike. Tehnike klasifikacije koje ćemo koristiti u radu su: stablo odlučivanja, neuronske mreže, SVM i Bayes ove mreže. To su tehnike koje spadaju pod nadzirano učenje pod klasifikaciju te za razliku od predikcije varijabla koju želimo predvidjeti je kategorija klase (diskretna ili nominalna). Sve 4 tehnike imaju istu svrhu a to je predviđanje klase te su samim time pogodne za usporedbu. Regresijske tehnike nisu pogodne iz razloga što svaki regresijski model koristi stohastičku varijablu i po tome se razlikuju od determinističkog modela

	Algoritmi strojnog učenja	
	Nenadzirano učenje	Nadzirano učenje
Kontinuirane	• Klaster i redukcija dimenzija	Regresija
Kategorijske	• Asocijacijska analiza • Markov modeli	Klasifikacija

Tablica 1. Generalna kategorizacija tehnika rudarenja podataka

Klasifikacija se smatra temeljenom zadaćom procesa otkrivanja znanja u podacima (Fayyad, Piatetsky-Shapiro i Smith, 1996.) te je u fokusu interesa ovog rada. Proces otkrivanja znanja u podacima sastoji se od nekoliko koraka, a priprema podataka koja obuhvaća čišćenje podataka i selekciju atributa oduzima 60% - 95% ukupnog vremena cijelog procesa (De Veaux, 2005.) te analiza i interpretacija rezultata. Selekcija atributa, kao najvažniji dio toga koraka, odnosi se na problem odabira onih atributa koji daju najveću prediktivnu informaciju s obzirom na izlaz.

Izrada klasifikacijske procedure iz skupa podataka za koji je poznata pripadnost slučajeva klasi naziva se diskriminacija ili nadzirano učenje (Michie, Spiegelhalter i Taylor, 1994.). Klasifikacija ima dva različita značenja. U jednoj situaciji moguće je raspolagati sa skupom instanci i cilj je utvrđivanje postojanja klasa ili klastera u podacima. U drugoj situaciji možemo znati koliko je točno klasa i cilj je utvrditi pravila na temelju kojih možemo klasificirati nove instance u postojeće klase. Prva situacija naziva se nenadzirano učenje, a druga nadzirano učenje. U ovom radu ćemo analizirati i usporediti tehnike nadziranog učenja.

Klasifikacija se sastoji od predviđanja određenog ishoda temeljenom na ulaznim jedinicama. Kako bi predvidili ishod, algoritam procesira *training* skup koji sadrži atribute i odgovarajući ishod, uobičajeno nazivan ciljni/prediktivni atribut. Algoritam pokušava otkriti vezu među

atributima kako bi bilo moguće predvidjeti ishod. Set podataka bez atributa koji se pokušava predvidjeti se naziva prediktivni skup podataka koji sadrži sve atribut osim onog kojeg želimo predvidjeti. Algoritam analizira atribute i proizvodi predikciju. Preciznost predviđanja definira kvalitetu algoritma.

Postoji niz tehnika i algoritama koji su razvijeni kako bi se riješili probleme klasificiranja. Također je razvijen niz alata koji su nam dostupni i pružaju nam niz klasifikacijskih tehnika kako bi pomogli pri donošenju odluke. Međutim nije jasno koju tehniku i koji algoritam koristiti pri posebnim okolnostima tj. određenim skupovima podataka . Donositelji odluka se suočavaju sa pitanjem: "Koja tehnika rudarenja podataka je najbolja za određeni set podataka ?" Osnovni fokus ovog istraživanja će biti učinak tehnika strojnog učenja na različitim skupovima podataka

Jedan od najtežih zadataka u cijelom 'KDD'(Knowledge Discovery from Data) procesu je izabrati pravu tehniku rudarenja podataka, kako komercijalni software alati pružaju sve više mogućnosti tako je i za odluku potrebno sve više znanja s metodične strane gledanja. Zaista, postoji veliki broj tehnika rudarenja podataka za znanstvenika koji želi otkriti neki model iz podataka. Tolika raznolikost može uzrokovati mnoštvo problema za znanstvenika koji često neznaju koje su uopće raspoložive metode kako bi se riješio određeni problem. Također u literaturi za rudarenje podataka nema zajedničke terminologije.

1.2 PREDMET ISTRAŽIVANJA

Iz prethodno navedenog problema proizlazi i predmet istraživanja ovog rada. Predmet istraživanja ovog rada će biti analiziranje prediktivnih tehnika rudarenja podataka na različitim skupovima podataka.

U ovom radu 4 tehnike predviđanja klasa(Stablo odlučivanja,Neuralne mreže,SVM i Bayes ove mreže) korištene su na 3 različite skupine podataka iz različitih domena. Promatrat će se rezultati tehnika na različitih podacima i njihova ovisnost. Prednosti i nedostaci tehnika će također biti objašnjeni

U teorijskom dijelu rada najprije će se objasniti osnovni pojmovi,a to su pojam, značenje te teorije rudarenja podataka. Istraživati će se 4 vrste tehnika i algoritama koji idu uz te tehnike te analizirati faktori koji utječu na odabir najbolje tehnike na određenim podacima. Također, opisati će se bitne značajke, podatci i informacije o spomenutim tehnikama, postupku pripreme podataka te

izvlačenju informacija iz podataka. Opisat ćemo cijeli postupak od pred procesiranja do interpretiranja rezultata.

U empirijskom dijelu rada istražiti će se značajke spomenutih tehnika i ishod njihovog korištenja s obzirom na određeni skup podataka. Rezultate ćemo usporediti i analizirati te u skladu s postavljenim istraživačkim hipotezama donijeti relevantne zaključke koji će se prethodno pomno analizirati. Na temelju dobivenih zaključaka moguće je definirati prijedloge za lakši izbor između prediktivnih tehnika. Podaci koji su potrebni za provođenje ovoga istraživanja će biti skinuti preko web stranica koje nude baze podataka za istraživanje. Na temelju istraživanja rada biti će moguće ocijeniti preciznost tehnika na određenim skupovima podataka, glavne značajke pojedinih tehnika i njihova ovisnost o skupu podataka te ćemo analizirati i usporediti rezultate svake pojedine tehnike tu će nam pomoći statistički testovi.

1.3 CILJEVI ISTRAŽIVANJA

Osnovni cilj istraživanja je identificirati tehniku koja ima najpreciznije rezultate predviđanja na određenim skupovima podataka koje ćemo koristiti u radu. Također, ciljevi su usporediti spomenute tehnike i analizirati rezultate te pokazati vezu između skupa podataka i korištene tehnike

Svrha ovog rada je dati teorijski uvid u problematiku izbora tehnike rudarenja te empirijski ispitati 4 tehnike rudarenja podataka na 3 različita podatkovna skupa, analizirati i usporediti dobivene rezultate.

Uzimajući u obzir navedene ciljeve istraživanja definirane su i glavne hipoteze istraživanja.

1.4 ISTRAŽIVAČKE HIPOTEZE:

Nakon što su postavljeni problem, predmet i ciljevi istraživanja postavit će se hipoteze koje će se ispitati i dokazati. Hipoteza predstavlja pretpostavku, koja se na temelju istraživanja potvrđuje ili odbacuje. U skladu sa definiranim problemom i predmetom istraživanja, te ciljevima istraživanja postavljene su sljedeće hipoteze:

H1.....Različiti skupovi podataka utječu na izbor najbolje tehnike za predviđanje

Priprema podataka je vrlo važna jer različite tehnike predviđanja u rudarenju podataka se ponašaju drukčije i ovise o pred procesiranju i transformacijskim metodama.

Postoji cijeli niz tehnika procesiranja skupova podataka koje pomažu otkriti prirodu podataka s ciljem lakšeg izbora najbolje tehnike predviđanja. Ovom hipotezom ćemo istražiti kako vrsta podataka utječe na preciznost tehnika

H2…Preciznost predviđanja tehnika rudarenja podataka se značajno razlikuju na istim skupovima podataka

Tehnike rudarenja podataka koriste različite algoritme za predviđanje te postoji određena razlika u preciznosti rezultata iako im je svrha jednaka. Ovom hipotezom ćemo istražiti kolike su razlike u preciznosti predviđanja na istom skupu podataka.

H3…. Ne postoji dominantna prediktivna tehnika rudarenja podataka

Istražiti postoji li dominantni najbolji algoritam koji ima najbolju preciznost na uređenim i čistim skupovima podataka kojeg možemo koristiti za predviđanje. Ovom hipotezom ćemo istražiti postoji li dominantna tehnika koju možemo koristiti pri predviđanju klasa

1.5 METODE ISTRAŽIVANJA

U cilju potvrđivanja odnosno odbacivanja postavljenih hipoteza, te u skladu sa različitostima zahtjeva pojedinog dijela rada koristit će se različite metodologije. Teorijski dio rada temeljit će se na prikupljanju i analiziranju relevantne stručne i znanstvene literature i podataka, te izvođenju novih spoznaja na temelju dosadašnjih empirijskih analiza.

U izradi teorijskog dijela koristiti će se sljedeće metode znanstveno-istraživačkog rada:

- **Metoda analize** – raščlanjivanje složenih pojmova, sudova i zaključaka na njihove jednostavnije sastavne dijelove te izučavanje svakog dijela posebno i u odnosu na druge dijelove. *U radu će se koristiti kako bi se opisale funkcionalnosti svake tehnike za sebe te usporedili s drugim tehnikama*
- **Metoda sinteze** - postupak znanstvenog istraživanja putem spajanja dijelova ili elemenata u cjelinu, sastavljanja jednostavnih misaonih tvorevina u složene i složenih u još složenije.

U radu će se koristiti kako bi se problematika koja se istražuje prikazala na što jednostavniji način.

- **Metoda dokazivanja i opovrgavanja** koristi se za potvrdu istinitosti nekih ranije definiranih stavova. Prilikom dokazivanja traže se pretpostavke koje određenu hipotezu trebaju dokazati. Suprotan slučaj je opovrgavanje, odnosno situacija kada se teze odbacuju ili opovrgavaju. *U radu će se koristiti kako bi se potvrdili ili opovrgnuli zaključci o određenim tehnikama predviđanja, odnosno korisnosti određenih tehnika s obzirom na vrstu podataka*
- **Metoda komparacije** - postupak uspoređivanja sličnih pojava i činjenica odnosno procesa i utvrđivanja jakosti ili intenziteta sličnosti i razlika između njih. Komparacijom se uočavaju sličnosti ili razlike između događaja, pojava i objekata. *U radu će se koristiti kako bi usporedili rezultate prediktivnih tehnika jednu s drugom.*
- **Metoda klasifikacije** - postupak raščlanjivanja općeg pojma na posebne, tj. jednostavnije pojmove. *U radu će se koristiti kako bi se bolje objasnio cijeli proces rudarenja podataka*
- **Metoda dedukcije i indukcije** - dedukcija je definirana kao zaključivanje od općeg prema posebnom, dok je indukcija zaključivanje od posebnog prema općem.
- **Metoda deskripcije** – postupak jednostavnog opisivanja ili očitovanja činjenica, procesa i predmeta. *Tako će se ova metoda u radu koristiti za definiranje osnovnih pojmova poput rudarenja podataka, strojnog učenja,prediktivnih tehnika, prednosti i nedostaci korištenih tehnika i sl.*
- **Metoda kompilacije** je metoda koja je neizostavni dio diplomskih radova, a temelji se na preuzimanju tuđih rezultata znanstvenih radova. Ova metoda nikada se ne koristi sama već se koristi u kombinaciji s drugim metodama

U empirijskom dijelu će se koristiti različite metode kojima će se prikupiti podaci te na tim podacima će biti izvršeno predviđanje klasa kako bi se rezultati mogli analizirati i usporediti ciljem dokazivanja pojedine hipoteze provoditi će se analize te odabrani statistički testovi.

1.6 DOPRINOS ISTRAŽIVANJA

Iako postoje neka istraživanja koja uključuju komparativnu analizu tehnika i algoritama rudarenja podataka ,međutim niti jedno istraživanje nije moglo doći do odgovora koja bi se tehnika mogla koristiti u određenoj domeni tj. koji algoritam bi se trebao koristiti pod specifičnim uvjetima. Teško je određenu tehniku ili algoritam svrstati u neku domenu jer puno toga utječe na krajnji rezultat kao što je varijacija u podatcima i korištenim algoritmima,predprocesni koraci, optimalizacija parametara i sl. Stoga od ovoga rada očekujemo da ćemo uspjeti približite problematiku. Koristiti ćemo 3 različita skupa podatka kako bi evaluirali izvedbu 4 izabrane tehnike predviđanja podataka. Rezultate izvedbi tehnika na pojedinom skupu te izvedbe svake pojedine tehnika na svim pod skupovima će biti uspoređeni i analizirani. Svrha ovoga je identificirati tehniku koja ima najbolju izvedbu za određeni izabrani skup podataka.

Cijeli rad bi trebao približiti problematiku odabira najbolje tehnika u širokom okviru što bi olakšalo i ubrzalo izgradnju modela za predviđanje.

1.7. STRUKTURA DIPLOMSKOG RADA

Rad će biti strukturiran u sedam poglavlja.

U uvodnom, prvom poglavlju diplomskog rada prikazati će se definicija problema, predmet i ciljevi istraživanja te metode istraživanja koje su se u radu koristile. Također će se postaviti hipoteze istraživanja te pružiti pregled strukture rada.

U drugom dijelu rada teorijski će se definirati i objasniti pojmovi vezani uz rudarenja podataka i korištenja tih tehnika.

U trećem poglavlju objasniti će se teorijske značajke tehnika rudarenja podataka. Svaka tehnika će biti pobliže objašnjena kao i prednosti i nedostaci korištenja.

U četvrtom dijelu će se predstaviti skupovi podataka na kojima će se istraživanje izvršiti, priroda tih skupova. Korak pred-procesiranje će biti teoretski obrađen kao i najkorištenije tehnike

Peto poglavlje će predstaviti empirijski dio istraživanja, rezultate i usporedbe tehnika korištenim na svakom skupu podataka.

U 6 dijelu će biti pojašnjena metodologija istraživanja, definiranje istraživanja, rezultati istraživanja te testiranje hipoteza i interpretacija rezultata.

U sedmom dijelu će biti prikazan zaključak do kojeg se došlo na temelju teorijskih spoznaja i promatrane problematike, odnosno rezultata istraživanja.

Na samom kraju dati će se uvid u literaturu koja je korištena prilikom izrade diplomskog rada, popis tablica, grafova i slika.

2.RUDARENJE PODATAKA

2.1 INFORMACIJSKO DOBA

Svijet se danas nalazi u informacijskom dobu. Podaci se generiraju u enormnim količinama čak se 2.5 kvintilijuna(2.5 na osamnaestu) bajtova proizvodi svaki dan te je 90% svjetskih podataka stvoreno samo u posljednje dvije godine. Količina podataka raste eksponencijalno i očekuje se da će rasti 40%godišnje što znači da će se količina podataka otprilike uduplati svako dvije godine . Tijekom 2013-e je proizvedeno 4.4 zeta bajta (10 na dvadeset prvu bajtova) dok se do 2020 godine pretpostavlja da će ta brojka doći do 44 zeta bajta ili 44 trilijuna gigabajta prema gruboj procjeni to je 57 puta više nego broj zrna pijeska na svim plažama. Izvori tih enormnih količina podataka su Internet, transakcijski podaci iz različitih industrija, podaci generirani direktno sa strojeva i razni drugi izvori te zahvaljujući sofisticiranim tehnologijama kao što su kompjuteri,mobiteli,sateliti i dr.. Za eksplozivan rast dostupnih podataka zaslužna je digitalizacija i kompjuterizacija društva te brz razvoj alata i programa za prikupljanje i skladištenje podataka. Poslovni svijet generira ogromne setove podataka koji uključuju prodajne transakcije, trgovanje dionica, opise proizvoda, prodajne promocije, profile tvrtki, učinak tvrtki, korisničke povratne podatke, bankovne transakcije. Znanstvene i inženjerske prakse kontinuirano generiraju veliki broj peta bajta od daljinskih istraživanja,mjerenja,znanstvenih eksperimena,sistemskih perfomanci,injžinjerskih promatratranja , i nadgledanja okoliša- Telekomunikacijske tvrtke također imaju desetke peta bajtova prometa svaki dan. Zdravstvena i medicinska industrija stvara ogromnu količinu podataka od medicinskih zapisa, motrenja pacijenata i slika. Milijune Web pretraga na pretraživačima procesiraju peta bajte podataka svaki dan. Online zajednice i društveni mediji su postali

važan izvor podataka, proizvode digitalne slike i videa, blogove, Web zajednice, i različite vrste društvenih mreža. Lista izvora koja generira ogromne podatke u svijetu je beskrajna.

Ovakav eksplozivni rast, široko dostupan i gigantska masa podataka uistinu čine naše vrijeme dobom podataka. Moćni i svestrani alati su prijeko potrebni da bi se automatski razotkrile vrijedne informacije iz sve te enormne količine podataka i transformirali ih u organizirano znanje. Ta potreba je vodila rođenju rudarenja podataka. Ovo područje je vrlo mlado, dinamično i obećavajuće. Rudarenje podataka ostvaruje veliki napredak u našem putovanju od doba podataka prema nadolazećem informacijskom dobu. Omogućuje nam mnogo stvari koje prije nismo mogli kao što je: primjećivanje poslovnih trendova, sprječavanja katastrofa, borba protiv kriminala, nove izvore ekonomske vrijednosti, kvalitetnije uvide u znanstvene podatke i mnogo drugog.
Rudarenje podataka je pokušaj svladavanja problema koje nam je digitalna informacijska era nametnula, a to je preopterećenje podacima.

2.2. ŠTO JE RUDARENJE PODATAKA

Rudarenje podataka i njegova primjena u otkrivanju znanju su nove tehnike koje predstavljaju neizostavan dio suvremene analize podataka te su još u fazi intenzivnog razvoja ali usprkos tome pokazale su se vrlo praktičnim u raznim industrijama. Ne postoji standardna praksa rudarenja podataka za razliku od recimo primjene statističkih postupaka. Postoje samo pozitivna i manje pozitivna iskustva sa određenim postupcima i njihovom primjenom na konkretnim domenama

Jednostavno rečeno rudarenje podataka je pronalaženje obrazaca i zakonitosti u podacima kako bi dobili informacije koje se mogu iskoristiti u svrhu poslovnog odlučivanja ili stvaranja vrijednosti. Rudarenje podataka je relativno novo i brzorastuće područje računarskih znanosti. Za rudarenje podataka kažemo da je interdisciplinarno jer sjedinjuje mnoštvo znanosti kao što su matematika,statistika,baze podataka, umjetnu inteligenciju i dr.

Postoji mnogo sličnih definicija rudarenja podataka tako prema Pang-Ning Tan(Introduction to Data mining-plava) *Rudarenje podataka je automatski proces otkrivanja korisnih informacija u velikim repozitorijima podataka . Tehnike rudarenje podataka su razvijene kako bi pronašle nove i i korisne obrasce koje bi inače ostale nepoznate*[1]

[1] Pang-Ning Tan(Introduction to Data mining-plava)

Prema David J.Hand i ostali(Princeples of Data Mining-amazon) *Rudarenje podataka je analiza (često velikih) promatranih podataka kako bi pronašli neslućene veze i rezimirali podatke na novi način koji je razumljiv i koristan vlasniku podataka*[2]

Fayyad kaže da je rudarenje podataka netrivijalni proces identificiranja validnih,novih,potencialno korisnih i ultimativno razumljivih obrazaca u podacima[3] (Jerome H.Friedman DM and statistisc)

Za Feruzzu je rudarenje podataka set metoda korištenih u proces otkrivanja znanja kojim izdvaja prethodno nepoznate veze i obrasce unutar podataka[4](Jerome H.Friedman DM and statistisc)

*(*definicije u knjizi Data mining tehniques PUJARI

Pojam rudarenje podataka prvi put koriste istraživačke zajednice u kasnim osamdesetima. U ranim danima nije bilo dogovora što pojam rudarenja podataka obuhvaća što u nekim dijelovima procesa je još slučaj. Općenito se rudarenje podataka može definirati kao set mehanizama i tehnika, realiziranih u softveru, kako bi se izvukle skrivene informacije u podacima. Riječ skrivena u ovoj definiciji je vrlo važna; upitni jezik SQL je sofisticiran Al nije rudarenje podataka. Također termin informacije se treba interpretirati u najširom smisli. Početkom devedesetih rudarenje podataka se obično prepoznavalo kao pod proces u širom proces zvanom Otkrivanje Znanja u Bazama podataka(Knowledge Discovery in Databases – KDD).Najkorištenija definicija „KDD" procesa je pripisana Fayyed et al : „Netrivijalni proces identificiranja važećih,novih,potencijalno korisnih i naposljetku razumljivih obrazaca u podacima[5](Fayed et al.1996). Kao takvo rudarenje podataka bi se trebalo gledati kao pod proces u cjelokupnom „KDD" proces, usredotočeno na otkrivanje „skrivenih informacija". Ostali pod procesi koji formiraju KDD proces su priprema podataka(skladištenje ,čišćenje podataka, pred procesiranje itd.) i analiza/vizualizacija rezultata

Mnogi stručnjaci tretiraju rudarenje podataka kao sinonim za otkrivanje znanja iz podataka (Knowledge discovery from data) dok drugi gledaju na rudarenje podataka samo kao esencijalni korak u procesu otkrivanja znanja. U ovom radu proces otkrivanja znanja[6] (*Knowledge Discovery in Database) (fayyed 1996 et al)*koristi ćemo kao sinonim za rudarenje podataka te se neće ulaziti u analizu razlike između dva izraza Proces otkrivanja znanja ili rudarenja podataka možemo gledati kao na iterativni slijed sljedećih koraka.

[2] David J.Hand i ostali(Princeples of Data Mining-amazon
[3] (Jerome H.Friedman DM and statistisc)
[4] Jerome H.Friedman DM and statistisc
[5] (Fayed et al.1996)
[6] (*Knowledge Discovery in Database) (fayyed 1996 et al)*

2.3 METODOLOGIJA RUDARENJA PODATAKA

Posljednjih godina se događa veliki rast i konsolidacija područja rudarenja podataka. Rudarenje podataka kao proces izvlačenja informacija i prepoznavanja obrazaca se sve više koristi u raznim industrijama za rješavanje poslovnih problem dok je početkom tisućljeća njegova primjena u većini bila u akademske svrhe i znanstvena istraživanja. Rast važnosti područja teži utvrđivanju standarda i metoda provođenja rudarenja podataka. Tako su razvijaju dvije metodologije CRISP-DM i SEMMA. Obje se javljaju kao industrijski standardi i definiraju niz uzastopnih koraka kojim se implementira primjena rudarenja podataka. U ovom radu ćemo koristiti CRISP-DM metodologiju za izvlačenje informacija iz podataka. Važno je naglasiti da su oba metodologije velikim dijelom slične te se neće analizirati njihove razlike.

CRISP-DM

CRISP-DM predstavlja *cross-industry process for data mining*

Razumijevanje problema

Početna faza se odnosi na razumijevanje ciljeva i zahtjeva projekta iz poslovne perspektive

Prva faza crisp-dm-a procesa je razumjeti što korisnik želi postići iz poslovne perspektive. Korisnici često imaju određene ciljeve i ograničenja koja treba izbalansirati. Cilj analitičara je otkriti važne faktore koje mogu utjecati na ishod projekta .Razumijevanja problema je vrlo važno i zanemarivanje ovog koraka može prouzročiti da se mnogo truda uložiti u produciranje pravog odgovora na kriva pitanja.

Razumijevanje podataka

Faza razumijevanja podataka počinje sa inicijalnim skupljanjem podataka i aktivnostima koji nam omogućuju da se upoznamo sa podacima, identificiranje problema kvalitete podataka, napraviti prve uvide u podatke, i detektirati zanimljive pod podatke za formiranje hipoteze odnosno skrivenih informacija.

Priprema podataka

Priprema podataka je faza koja pokriva sve aktivnosti potrebne za izgradnju finalne verzije baze podataka od inicijalne sirove baze podataka. Priprema podataka se obavlja različitim tehnikama više puta bez kakvog propisanog reda. Tehnike uključuju selekciju tablica, polja i atributa kao i transformaciju i čišćenje podataka.

Modeliranje

U ovoj fazi, različite tehnike modeliranja su odabrane i primijenjene, i parametri su kalibrirani na optimalne vrijednosti. Obično, postoji nekoliko tehnika za isti tip problema rudarenja podataka . Neke tehnike imaju specifične zahtjeve s obzirom da oblik podataka zato je često potrebno vraćanje na fazu pripreme podataka.

Evaluacija rezultata

U ovoj fazi projekta, izgradili smo kvalitetni model iz perspektive analitičara. Prije implementiranja modela ,važno je evaluirati i pregledati izvršene korake stvaranja modela, kako bi bili sigurni da model ispravno postiže poslovne ciljeve. Ključni cilja je odrediti postoji li neko važno pitanje koje nismo uzeli u obzir. Na kraju ove faze, se odlučuje o korištenju rezultata rudarenje podataka

Primjena/implementacija rezultata

Stvaranje modela generalno nije kraj projekta. Ako je cilj modela povećanje znanja o podacima. Dobiveno znanje se treba organizirati i prezentirati na način da ga korisnik može iskoristiti.. Često uključuje primjernu „živih" modela u procesu donošenja odluka. Ovisno o zahtjevima, faza implementacije rezultata može biti jednostavna kao neko izvješće ili kompleksno kao implementiranje procesa rudarenja podataka u cijeloj kompaniji ili organizaciji. Često korisnik sam implementira rezultate.



Slika 1. CRISP-DM

2.4 KATEGORIZACIJA

Tehnike rudarenja podataka se oslanjaju na tehnike iz područja kao što su strojno učenje, prepoznavanje obrazaca, umjetna inteligencija i statistika kako bi pronašli obrasce u podacima

U ovom dijelu ćemo kategorizirati zadatke i metode rudarenja podataka s obzirom na ciljeve analize podataka. Kategorije možemo raščlaniti na metode ili zadatke analize podataka . Za svaku metodu postoji niz algoritama koje se koriste u analizi

Nakon analize poslovnog problema ,baze podataka, čišćenja i transformiranja podataka potrebno je odabrati prikladne tehnike rudarenja podataka koje odgovaraju korisničkim ciljevima i tehničkim specifikacijama baze podataka. Potrebno je kategorizirati zadatke rudarenja podataka koje odgovaraju ciljevima osobe koja analizira podatke.

Postoje niz obrazaca koje možemo identificirati u podacima kao što su karakterizacija i diskriminacija klasa u podacima, rudarenje čestih uzoraka,asocijacije,korelacije,kluster analize, analiza anomalija, regresija te klasifikacija koja je fokus ovoga rada

Rudarenje podataka kao faza *KDD* procesa uključuje ponavljajuću iterativnu primjenu određenih metoda rudarenja podataka. Ciljevi procesa otkrivanja znanja iz baza podataka su definirani namjerom korištena sustava. Možemo razlikovati dva tipa ciljeva: 1) Verifikaciju i 2) Otkrivanje. Sa verifikacijom sustav je ograničen na verificiranje korisnikove hipoteze. Otkrivanjem sustav samostalno pronalazi nove obrasce .Dakle, ako nam je cilj otkrivanje možemo dodijeliti zadatke rudarenja podataka koje možemo svrstati u dvije kategorije, predviđanje i opisivanje(Shapiro 1996 članak).

Predviđanjem pokušavamo predvidjeti vrijednost određenog atributa s obzirom na druge atribute. Atribut koji predviđamo ima zajednički naziv ciljna ili zavisna varijabla dok atributi pomoću kojih predviđamo ciljnu varijablu se nazivaju eksplanatorne i nezavisne varijable.Prediktivno modeliranje je zadatak izgradnje modela za ciljnu varijablu kao funkcija eksplanatornih varijabli. Dva su tipa produktivnog modeliranja: klasifikacija koja se koristi za kontinuirane ciljne varijable , i regresija, koja se koristi za diskretne ciljne varijable.

Kad radimo deskriptivni model cilj nam je opisati sve podatke. Primjeri takvih opisa uključuju metode kao što su procjena gustoće, luster analiza, modeliranje odnosa. Deskriptivni modeli jednostavno sumiraju podatke na način koji će nam pomoći pri njihovom razumijevanju

Većina autora dijeli zadatke rudarenja podataka na dvije osnovne kategorije a to su *Predviđanje i Opisivanje* . Predviđanje se često naziva i nadzirano učenje dok opisivanje nazivaju nenadzirano učenje.

Postoji veliki broj tehnika rudarenja podataka koji se poklapaju sa tehnikama strojnog učenja te je moguće ove dvije osnovne kategorije podijeliti na više kategorija i supkategorija. Na primjer David Handet et all je podijelio kategorije rudarenja podataka na 5 vrsta.

Prema David Hand et all je proširio zadatke rudarenja podataka na još tri kategorije : Eksplorativna analiza podataka kojoj je cilj istraživanje podataka bez ikakve jasne ideje o onome što tražimo. EAP tehnike su tipično interaktivne I vizualne I postoji dosta efikasnih grafičkih metoda za male baze podataka. Što broj dimenzija raste postaje sve teže vizualizirati podatke. Primjer primjene EAP tehnika je coxcomb grafikon sličan piti u kojem možemo naprimjer prikazati smrtnost po bolnicama u nekom gradu

Sljedeća kategorija je otkrivanje obrazaca i veza u podacima. 3 navedene kategorije zadataka su više orijentirane prema izgradnji modela. Ovdje je naglašena primjena na otkrivanju obrazaca. Naprimjer otkrivanje prevare detektiranjem podataka koji se značajno razlikuju od ostalih. Drugi primjer bi bio u astronomiji gdje detekcijom neobičnih zvijezda i galaksija može voditi do neotkrivenih fenomena. Zadatak pronalaska kombinacije artikala koji se često skupa kupuju je dosta vremena u fokusu rudarenja podataka i to korištenjem algoritama baziranih na asocijacijskim pravilima.

Petu kategoriju naziva dohvaćanje sadržaja. Ovdje korisnik već ima nekakve obrasce te pokušava pronaći slične. Ovaj zadatak se najviše koristi za tekstualne i slikovne baze podataka. Za tekst, obrazac može biti set ključnih riječi i korisnik želi pronaći relevantne dokumente u velikoj bazi podataka relevantnih dokumenata. Za slike, korisnik može imati primjer slike,skicu slike i opis slike i želi pronaći slične sike iz velike baze slika. U obe definicije sličnost je krucijalna kao i detalji strategije potrage

Svaka od kategorija zadataka se može podijeliti na još subkategorija koje se razlikuju po ciljevima koje analitičar želi ispuniti. Iako vidimo da se kategorije zadataka razlikuju jedna od druge one dijele mnogo sličnosti te granice između kategorija nisu jasno definirane tako da neki prediktivni modeli mogu biti deskriptivni i obratno.Svaki zadatak rudarenja podataka se sastoji od različitih tehnika i algoritama koje ćemo pobliže objasniti u sljedećim poglavljima. U ovom radu ćemo se bazirati na zadatak klasifikacije koju smo svrstali u prediktivnu kategoriju.[7]

[7] David Hand et all. Princeples of data mining

2.5 PREDIKTIVNO MODELIRANJE

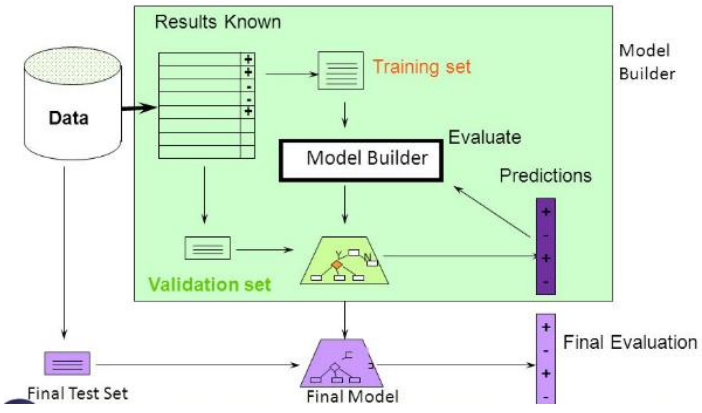
Klasifikacija je proces koji se sastoji od dva koraka ,prvi korak koji možemo nazvati korak učenja gdje se konstruira klasifikacijiski model . Drugi korak je klasifikacijski korak gdje se mode koristi kako bi se predvidile klasne oznake na određenim podacima.(cm 3d edition)

Modeliranje podrazumijeva odabir različitih tehnika modeliranja i njihovu primjenu na ulazni podatkovni set. U tom se procesu također parametri modela kalibriraju na optimalne vrijednosti. Često se određeni problem može riješiti pomoću nekoliko različitih metoda. Određene metode,pak zahtijevaju podatke u određenom obliku tako da je često potrebno vraćanje na fazu priipreme podataka gdje se podaci pripremaju za primjenu određene tehnike. U ovoj fazi rudarenja podataka se odlučuje nad kojim će se podacima raditi model,nad kojima će se kalibrirati parametri modela i nad kojim će se testirati rad modela. U tu se svrhu podatkovni set dijeli na tri ,najčešće odvojena dijela:

1.Trenirajući podatkovni set ili trenig set (eng. Train set,*training* set) - predstavlja one podatka pomoću kojih se model izgrađuje.

2. Validacijski podatkovi set (eng. Validate set) – pomoću tih podataka se optimiziraju parametri modela i poboljšavaju njegove perfomanse.

3. Testni podatkovni set ili test set (eng. Test set) – čine ga podaci koji nisu iskorišteni za izgradnju modea i na njima se ispituje koliko je dobiveni model dobar.



Slika 2. Proces rudarenja podataka (Izvor: http://slideplayer.com/slide/6194398/)

Ako je podatkovni set mali tada postoje posebne metode koje omogućuju maksimalno iskorištavanje svih podataka . To su metode ukrštene validacije (eng. cross validation). Izbor pojedinih jedinica promatranja u određeni podatkovni set,najčešće se čini slučajnim odabirom. Poslije odabira finalnog modela ,model

se validira i procjenjuje se perfomansa. Validacija modela se radi na test setu kako bi se procijenio model koji je izgrađen na na trenig testu, jeli se može generalizirat na neviđene podatke. Pod generalizirati se podrazumijeva uklapa li se model u novi set podataka i evaulira mu se prediktivna moć. Postoji niz tehnika koje se koriste u validaciji modela koje će biti objašnjene u kasnijim poglavljima.

Prediktivno modeliranje definira proces razvijanja modela na način da možemo razumijeti i kvantificirati preciznos t modela na podacima koje ćemo tek vidjeti (Kuhn johnson applied).

Prediktivno modeliranje je bilo koja metoda kojoj je rezultat predviđanje, bez obzira na pristup(galit shumeli to explain or to predict)

Glavni problem kod modeliranja je pronaći model ,koji opisiuje veze između ciljne variable Y i niza drugih varijabli,zvanih eksplanatarone varijable X. Kad nam je zadatak predviđanje ali i u nekim deksriptivnim zadacima ,ciljna vrijednost varijable T[Y] = y nam je dana u podacima. Kod strojnog učenja takve zadatke zovemo *nadzirano učenje (supervised learning)*.

(Introduction to Supervised Learning Erik G. Learned-Miller) .Nadzirano učenje je jednostavno formalizacija ideje učenja kroz primjere. Kod nadziranog učenja ,učeniku (tipično ,računalni program) je dano 2 seta podataka,trenig set i test set. Ideja je da učenik „uči“ iz niza označenih primjera u trenig testu kako bi mogao identificirati neoznačene primjere(ciljnu varijablu) u test setu sa što većom preciznošću. To je cilj učenika ,razviti pravila, program ili proceduru koja će klasificirati nove primjere(u test setu) analizirajući prijmjere koje već imaju oznaku klase. Na primjer, trenig baza podataka se može sastojati od slika različitih vrsta voća (jabuke,breskve i banane), gdje je identite(klasa)t voća već dan učeniku. Test baza podataka će se sastojati od više neidentificiranih dijelova voća iz više klasa. Cilj je razviti pravila koja če identificirati elemente u test bazi podataka.

Nadzirano učenje možemo podijeliti u dvije kategorije: klasifikaciju i regresiju. Kod klasifikacije je cilj dodijeliti klasu (ili oznaku) iz poznatih nizova klasa observaciji. To je kategorična varijabla. Kod regresije je cilj predvidjeti kontinuiranu mjeru za observaciju .

Cijeli set podataka se može promatrati kao heterogena matrica. Redovi matrice se zovu observacije,primjeri, ili instance i svaki sadrži niz mjera za subjekta(voće iz primjera poviše). Redovi matrice se zovu prediktori,atributi, ili značajke i sve su varijable koje predstavljaju mjere izvršene na svakom subjetku(boja voća,oblik,okus itd). Postoji i ciljna varijabla(jabuka,breska banana) koja nam je poznata

Formalno gledajući na nadzirano učenje, *training* set se sastoji od **n** parova (x1,y1),(x2,y2)...(xn,yn), gdje je xi mjera ili niz mjera,a yi je klasa. Na primjer xi također možebiti i grupa(vektor) od 5 mjera (u

medicinskoj bazi podataka to može biti težina,visina,temperatura, šećer i tlak). Odgovarajući yi može biti klasifikacija pacijenta na „zdrave“ i „ ne zdrave“.Kod test podataka u nadziranom učenju je još jedan niz mjera ali bez oznaka (xn+1,xn+2...xn+m) .Kao što je već opisano,cilj je pogoditi oznake (klase) za test podatke („zdrave“ ili „nezdrave“) kao zaključak iz trenig podataka.

Klasifikacija je predviđanje kategoričnih varijabli,ako želimo numeričko predviđanje odnosno predvidjeti vrijednost kontinuirane varijable korističemo metodu zvanu regresija. Klasifikacija i regresija su dvije glavne vrste predviđanja. Na primjer ako marketing menadžer želi predvidjeti koliko će korisnik potrošiti novaca u prodavonici koristit ćemo statističku metodu regresijske analize gdje izgrađeni model predviđa kontinuiranu vrijednost za razliku od klasifikacije i predviđanja klasa.Ovaj rad se fokusira na klasifikacijske tehnike predviđanja

Deskriptivni modeli kao što je već rečeno jednostavno sažimaju podatke na prikladan način koji nam omogućuje što bolje razumijevanje podataka i problema. Usporedno prediktivno modeliranje ima specifičan cilj predviđanja nepoznatih vrijednosti varijabla na temelju drugih varijabla. [8]

2.6 KLASIFIKACIJA

Što Je klasifikacija? Kad banka daje kredit mora analizirati podatke kako bi prepoznala koji zahtjevi za kredit su sigurni a koji riskantni. Marketing menadžeru trebaju podaci za analizu koji će mu pomoći odrediti koji profil kupca će kupiti novi kompjuter. Medicisnski istražitelj želi analizirati podatke pacijenata s rakom dojke kako bi predvidio koju od 3 specifična tretmana bi pacijent trebao primiti. To su neki od primjera klasifikacije.Na primjeru ispod vidimo kako to izgleda u praksi . Skup podataka sa 10 instanci i 4 atributa pomoću kojih predviđamo ciljnu varijablu „Buy widget“ . Na desnoj strani slike je rezultat kojih smo dobili nakon što smo primjenili metodu stabla odlučivanja. Možemo vrlo lako isčitati da muškarci sa plaćom većom od 1800 će kupiti proizvod.

[8] Michelin Kamber et all. Data mining: Concepts and tehniques



Slika 3. Primjer klasifikacije (Izvor: https://fluxicon.com/blog/2012/02/data-requirements-for-process-mining/)

Klasifikacija je zadatak u procesu rudarenja podataka te spada u područje strojnog učenja i inspirirano je prepoznavanju obrazaca gdje je cilj klasificirati objekte u kategorije koje zovemo klase.To je oblik analize podataka u kojem izvlačimo model koji najbolje opisuje klase podataka. Takve modele nazivamo klasifikatorima koji predviđaju oznake klasa. Ovakve analize nam omogućavaju bolje razumijevanje podataka. Mnoge klasifikacijske metode su predložili istražitelji iz disciplina kao što su strojno učenje,prepoznavanje obrazaca i statistike. Mnogi algoritmi se nastanjuju u memoriji što tipično pretpostavlja malu bazu podataka. Nedavna istraživanja u rudarenju podataka se nadograđuju na to i razvijaju skalabilne klasifikacijske i prediktivne tehnike sposobne svladati velike količine podataka. Klasifikacija ima brojne aplikacije u detektiranju prijevara,direktnom marketingu,predviđanju perfomansi ,proizvodnji,medicinski dijagnozama i dr. [9]

Ulazne jedinice u zadatku klasificiranju su kolekcije zapisa. Svaki zapis, može se još nazvati instancom ili primjerom, je karakteriziran n-torkom (x,y), gdje je x set atributa, a y posebni atribut dizajniran kao klasna oznaka(klasna/kategorična ili ciljna varijabla). Oznaka klase kod klasifikacije je uvijek diskretna varijabla a kod regresija kontinuirana.

Klasifikacija je zadatak u rudarenju podataka koji „uči“ ciljnu funkciju f koja raspodijeljuje svaki atributski set x u jedne od predefiniranih klasnih oznaka y. Ciljna funkcija se neformalno zove klasifikacijski model. Klasifikacijski model se koristi za :

1) Deskriptivno modeliranje

[9] Michelin Kamber et all. Data mining: Concepts and tehniques

Klasifikacijiski model se može koristiti kao eksplonatorni alat za razlikovati klase između objekata. Na primjer, bilo bi korisno za biologe i druge da imaju deskriptivni model koji sumira podatke o npr. Životinjama i objašnjava koje značajke definiraju sisavce,reptile,ribe ili ptice.

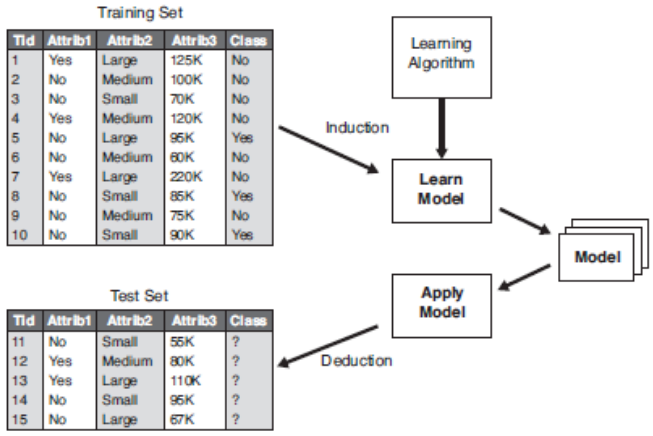
2) Prediktivni model

Klasifikacijski model se također može koristiti za predviđanje klasa nepoznatih zapisa. Klasifikacijski model možemo tretirati kao crnu kutiju koja automatski dodijeljuje klase nepoznatim zapisima.

Klasifikacijske tehnike su prilagođene za predviđanje i opisivanje seta podataka s binarnim ili nominalnim kategorijama. Manje su efikasne za ordinalne kategorije jer ne uzimaju u obzir implicitni poredak među kategorijama.[10]

Generalni pristup rješavanju klasifikacijskog problema-

Klasifikacijske tehnike(ili klasifikator) je sistematski pristup izradi kvalifikacijskog modela od ulaznih jedinica u setu podataka. Primjeri su: stabla odlučivanja,rule-based classifier, neuronske mreže,SVM, naivni Baeyes klasifikator i drugi. Svaka tehnika upotrebava učeći algoritam za identificiranje modela koji najbolje odgovora vezi između atributa i klasa u podatkovnom setu. Algoritmi učenja generiraju model trebali bi odgovarati ulaznim podacima i točno predviđati klase za zapise koje nisu „vidjeli“ . Prema tome ključni cilj algoritama učenja je izgraditi model sa dobrim generalizacijskim sposobnostima tj. Model koji najpreciznije predviđa klase neviđenih podataka. **Slika:**



Slika 4. Izgradnja modela

[10] Kumar et all. Introduction to data mining

Nakon što smo napravili klasifikacijski model,željeli bi smo procijeniti koliko točno klasifikator predviđa vrijednosti podataka koje nismo koristili u trenig setu. Postoji opcija da smo izgradili više klasifikacijskih metoda i sad želimo usporediti njihovu preciznost. Dolazimo do pitanja što znači preciznost i kako ćemo je izmjeriti te kako ćemo procijeniti koji model je najbolji.

2.7 EVALUACIJA MODELA

Osnovni pojam pri evaluaciji klasifikacijskog modela ,jest pojam greške. Jednostavno rečeno,greška u slučaju klasifikacijskih problema je pogrešna klasifikacija: ako se model(klasifikator) primjeni na određeni primjer iz skupa podataka ,on ga klasificira u pogrešnu klasu(kategoriju) ciljnog atributa. Ukoliko su sve greške iste težine,tada omjer broja grešaka prema ukupnom broju klasificiranih primjera predstavlja dobru mjeru rada klasifikatora(modela). No, u mnogim primjenama,razlike u određenim tipovima grešaka su od velike važnosti.

U ovom djelu ćemo predstaviti različite mjere za procjenu koliko klasifikator dobro predviđa klasne oznake. Predstaviti holdout,random subsampling,cross-validation i bootstrap metode koje su česte tehnike procijenjivanja preciznosti na osnovi nasumičnog uzimanja uzoraka iz podataka te selekciju modela odnosno kako izabrati jedan klasifikator nad drugim

Klasifikacijski model ćemo evaluirati prema broju zapisa koji je model točno ili netočno predvidio. Zapise ćemo tablično prikazati pomoću konfjukcijske matrice.

Evaulacija modela se može oblikovati na 3 razine

a) Mjere za evaulaciju perfomansi klasifikatora- procjena koliko dobro klasifikator predviđa klase
b) Procjena preciznosti na osnovi nasumičnog uzimanja zoraka iz podataka
c) Procjena između klasifikatora odnosno odabir jednog klasifikatora nad drugim

Mjere za evaluciju perfomansi klasifikatora

Postoje mjere koje procjenjuju koliko je točan klasifikator odnosno model koj smo izgradili čiji je zadatak predviditi klase za svaku pojedinačnu observaciju, Pri objašnjavanju postaviti ćemo pretpostavku da su klase podjedanko distribuirane. Mjere evaluacije su točnost(stopa raspoznavanja),senzitivnostt,specifičnost,preciznost,odziv,F-mjere.Točnost je specifična mjera ali i generalni termin koji koristimo za mjerenje sposobnosti predviđanja klasifikatora.

Mjerenje točnosti klasifikatora nad trenig podacima može rezultirati varljivim i preoptimističnim rezultatima zbog prespecijalizacije algoritma podacima. Zato je bolje mjeriti točnost nad test podacima koji se sastoji od observacija koje nisu korištene u trenig setu.

Na jednome primjeru ćemo se upoznati sa terminologijom. Za svaki red u tablici možemo reći da je jedan primjer,observacija ili zapis tako da kad govorimo u terminu negativnih i pozitivnih observacija,pozitivna se odnosi bazu korisnika koji će kupiti kompjuter a negativna na primjere koji neče kupiti kompjuter. P ćemo označati kao broj pozitivnih primjera a N je broj negativnih primjera. Za svaki primjer ćemo usporediti klasu(P ili N) koju je klasifikator predvidio s stvarnom klasom. Postoje još 4 termina s kojima se trebamo upoznati i koji su građevne jedinice mnogih mjera evaulacije.

	Klasa pozitivnih primjera	**Klasa negativnih primjera**
Pozitivna predikcija	Stvarno pozitivni (SP)	Lažno negativni(LN)
Negativna predikcija	Lažno negativni(LN)	Stvarno negativni(SN)

Tablica 2: Matrica grešaka

STVARNO POZITIVNI (SP) : se odnosi na pozitivne primjere koji su točno označeni od strane klasifikatora.

STVARNO NEGATIVNI (SN): su negativni primjeri koji kojima je točno dodijeljena klasa od strane

LAŽNO POZITIVNI (LP)- su negativni primjeri koji su krivo klasificirani kao pozitivni

LAŽNO NEGATIVNI (LN)- su pozitivni primjeri koji su krivo klasificirani kao negativni

Izrazi su sumirani u tablici:

	Predviđena klasa			
		Da	**Ne**	**Ukupno**
Stvarna klasa	**Da**	SP	LN	P
	Ne	LP	SN	N
	Ukupno	p'	n'	P + N

Tablica 3.Matrica grešaka 2

Matrica grešaka se sastoji od primjera koji su točno klasificirani i oni se nalaze uzduž dijagonale matrice te broju netočno klasificiranih primjera kojima je predviđene kriva klasa. Matrica grešaka je dobra metoda za analizirati koliko nam dobro klasifikator prepoznaje primjere različitih klasa. TP i TN nam govori kad su predviđanja točna dok FP i FN kada klasifikato griješi.

Prva mjera evaulacije koliko dobro klasifikator ili model predviđa je točnost. Točnost klasifikatora je postotak točno klasificiranih primjera na datom test setu.

Točnost = $\frac{SP+SN}{P+N}$

U nekim izvorima podataka točnost se naziva i stopa prepoznatljivosti(eng. recognition rate) klasifikatora koja pokazuje koliko dobro klasifikator prepoznaje primjere s različitim klasama.

Stopa greške ili greška klasifikacije klasifikatora koja se računa kao 1 – točnost(M) gdje M predstavlja model. Može se izračunati i kao:

Stopa greške = $\frac{LP+LN}{P+N}$

Ako ćemo koristiti trenig set podataka umjesto test seta za procjenu stope pogreške modela , ova mjera je poznata kao *resubstition error.* Ova procjena greške je optimistični prikaz stvarne stope pogreške jer je model testiran samo na primjerima kojima već zna klasu.

Pod pretpostavkom da postoji problem nebalansiranosti klasa. Na primjer podatkovni set je distrubuiran većim dijelom ka negativnim klasama a malim djelom pozitivnim klasama. Na primjeru,otkrivanja prevare, klasa koja nas zanima je postoji li „prevara“ (pozitivna klasa), koja se pojavljuje puno rjeđe nego negativna „nije prevara“ klasa. Kod medicinskih podataka,postoji isto tako klasa koja se pojavljuje rijetko npr. „ rak“. Pretpostavimo da smo istrenirali klasifikator da klasificira medicinske observacije,gdje je klasni atribut „ rak“ a moguće klasne vrijednosti su „da“ i „ne“. Stopa točnosti od 97% se može činiti prilično točna,ali ako pogledamo da 3% primjera iz trenig seta su su stvarno rak. Očito onda da stopa točnosti od 97% može biti neprihvatljiva jer klasifikator možda precizno označava samo primjere koje nemaju rak,i dekvalificira sve primjere koje su rak. Zato koristimo druge mjere, koje mjere kako dobro klasifikator prepoznaje pozitivne primjere (rak = da) i negativne primjere (rak = ne).

Mjere osjetljivost i specifičnost se mogu koristiti u svrhu rješavanja ovoga problema. Osjetljivost se referira kao *stvarno pozitivna stopa* (razmjer pozitivnih primjera koji su točno identificirani), dok je stopa *stvarno negativni* mjera specifičnosti koja se odnosi razmjer negativnih primjera koji su točno identificirani. Ove mjere se definiraju kao :

Senzitivnost = $\frac{SP}{P}$

Specifičnost = $\frac{TN}{N}$

Možemo prikazati točnost kao funkciju osjetljivosti i specifičnosti:

$$\text{Točnost} = \text{osjetljivost}\ \frac{P}{(P+N)} + \text{specifičnost}\ \frac{N}{(P+N)}$$

Preciznost i odziv su mjere koje se često koriste kod klasifikacije. Preciznost možemo gledati kao mjeru korektnosti (koji postotak primjera koji su označeni kao pozitivni i to stvarno jesu), dok je odziv mjera potpunosti(koji postotak pozitivnih primjera je označen takvim). Možemo vidjeti da je odziv isto kao i osjetljivost (*true positive rate).* Mjere možemo izračunati na način :

$$\text{Preciznost} = \frac{SP}{SP + LP}$$

$$\text{Odziv} = \frac{SP}{SP+LN} = \frac{SP}{p}$$

Savršena preciznost se postiže kao 1.0 za neku klasu što znači svaki primjer koji je klasifikator označio da pripada određenoj klasi stvarno njoj i pripada. Međutim ne kaže nam nista o broju klasa koje je klasifikator krivo označio. Postoji inverzna veza između preciznosti i odziva, gdje je moguće povećati jednu mjeru po cijeni smanjenja druge. Na primjer, medicinski klasifikator može postići visoku preciznost označavajući sve primjere s rakom u određenu klasu ali može imati mali odziv ako netočno klasificira ostale primjera. Preciznost i odziv se tipično skupa koriste. Alternativni način kako možemo koristiti ove dvije mjere je da ih kombiniramo u jednu mjeru . Ovaj pristup poznat kao F mjere(F1 ili *F-score)* i Fb mjere. Definirane su kao :

$$\text{F} = \frac{2\ X\ preciznost\ x\ odziv}{preciznost + odziv}$$

$$F_\beta = \frac{(1+\beta^2)\ x\ preciznost\ x\ odziv}{\beta^2\ x\ preciznost\ x\ odziv}$$

Gdje je β nenegativni realni broj. F mjera je harmonijska sredina preciznosti i odziva . Daje jednaku težinu preciznosti i odzivu. F_β mjera je ponderirana mjera preciznosti i odziva. Dodjeljuje jednaku važnost koliko odzivu toliko i preciznosti.

Pored mjera koje su na bazi točnosti,klasifikatori se također mogu usporediti prema sljedećim aspekima

- Brzina : Ovo se odnosi na računalne troškove uključene pri generiranju i korištenju klasifikatora
- Robusnost : Ovo je sposobnost klasifikatora da napravi točne predikcije uz podatkovni set koji se sastoji od primjera s nedostajućim vrijednostima i šumovima. Robusnost procjenjujemo na način da imamo niz podatkovnih setova koji se razliku po količini podataka sa šumovima i nedostajućim vrijednostima.
- Skalabilnost: Ovo se odnosi na sposobnost konstruiranja efikasnog klassifikatora za velike količine podataka. Skalabilnost se tipično procjenjuje sa nizom podatkovnih setova kojima povećavamo veličinu
- Interpretabilnost : Ovo se odnosti na nivo razumijevanja i uvid koji nam omogućava klasifikator ili prediktor. Interpretabilnost je subjektivna i dosta teža za procijeniti. Stabla odlučivanja i klasifikacijska stabla je lako interpretirati, a ipak njihova interpretabilnost može ovisiti o njihovoj kompleksnosti

Prezentirano je nekoliko evaluacijskih mjera. Točnost je najbolje koristiti kada su klase u podacima jednako distribuirane. Druge mjere, kao senzitivnost (ili odziv) ,specifičnost, preciznost I F mjere je bolje koristiti kada se pojavljuje probem nebalansiranih klasa

Tehnike višestrukog particioniranja za generiranje metrike klasifikacijskog modela

1. **Holdout metoda**

U ovoj metodi podatkovni set se nasumično dijeli na dva nezavisna seta, trenig set i test set. Uobičajeno, dvi trećine seta je alocirano u podatkovni set za treniranje kojeg zovemo trenig set te ostala jedna trećina u podatkovni set za testiranje, test set. Trenig test se koristi za stvaranje modela. Točnost modela se procijenjuje na test setu.

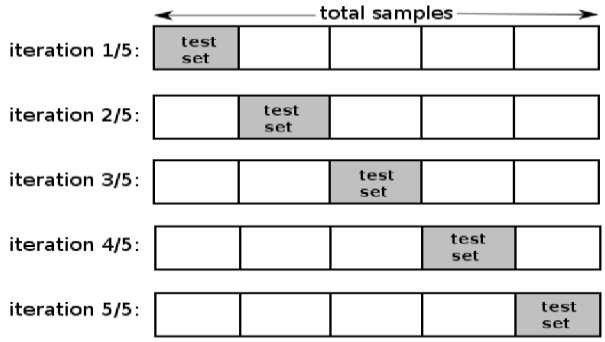
Random subsampling je varijacija holdout metode u kojoj se holdout metoda ponavlja *k* puta. Ukupna procjenjena točnost se uzima kao prosjek svih točnosti iz svake iteracije.

2. **Unakrsna validacija**

Ova metoda daje bolju procjenu stvarne greške modela od jednostruke podjele skupa podataka na skup za učenje modela i na skup za testiranje,pogotovo na manjim skupovima podataka.

U *k*-fold cross validaciji inicijalni podaci su nasumično podijeljeni u *k* uzajamno isključivih podsetova D1,D2,....Dk gdje je svaki otprilike iste veličine. Trenig i test nad podacima se radi *k* puta. U iteraciji *i,*

dio Di je rezerviran kao test set,a dok su ostali dijelovi kolektivno korišteni za treniranje modela. Znači u prvoj iteraciji ,podsetovi D2,....Dk se koriste kao trenig set podaci na kojem će se razviti model koji će se testirati na podsetu D1; druga iteracija se trenira na podsetovima D1,D3....,Dk i testira na podsetu D2; i tado dallje.U ovoj se metodi se svaki uzorak koristi jednak broj puta za treniranje i jednom za testiranje. Procjena točnosti kod klasifikacije je ukupan broj točnih klasifikacija iz *k* iteracija,podijeljen sa ukupnim brojem instanci iz inicijalnih podataka.



Slika 5.Unakrsna validacija.(Izvor: https://www.researchgate.net/figure/266617511_fig10_Figure-211-5-Fold-Cross-Validation)

Leave-one-out je poseban slučaj k-fold unakrsne validacije gdje se privremeno izdvoji primjer iz skupa primjera za učenje te nauči model na preostalin primjerima i testira na tom izdvojenom i tako N puta.

Odabir modela

Pretpostavimo da imamo 2 klasifikacijska modela, M1 i M2. Izveli smo 10-fold cross-validation kako bi dobili stopu greške za svaki model. Intuitavno se može činiti da ćemo izabrati model s najmanjom stopom greške; međutim,stopa pogreške je samo procjena greške stvarne populacije na budućim slučajevima. Može postojati značajna varijanca između stopa pogreške u bilo kojem 10-fold cross validation ekspirimentu. Iako dobivena stopa pogreške iz M1 i M2 modela može biti različita ona ne mora biti statistički značajna.

Kako bi se odlučilo postoji li prava razlika u stopama pogreške dva modela,mora ćemo iskoristiti statistički test značajnosti. Za to možemo koristiti testiranje hipoteze t-test ili Student s t-test. Hipoteza je da su oba modela jednaka,drugim riječima,razlika između stopa rata između njih je nula.Ako možemo odbiti ovu hipotezu onda možemo zaključiti da razlika među modelima je statistički znaćajna,u tom slučaju možemo izabrati model s manjom stopom pogreške.

SP, SN, LP i LN su koristi kod vrednovanja troškova i koristi (rizika ili prednosti) asociranih sa klasifikacijom modela . Troškovi asocirani sa lažnim negativima(kao netočno predviđanje da pacijent sa rakon nema rak) su daleko važniji nego oni kod lažnih pozitiva(netočno označavanje pacijenata kao da imaju rak). U takvim slučajevima,trebamo razlikovati tipove grešaka po važnosti pridavajući im različite troškove. Ovi troškovi podrazumijevaju opasnost za pacijenta,financijski trošak terapije ili drugi bolnički trošak.

Receiver operating characteristic curves (ROC) je koristan alat za organiziranje klasifikatora i vizualiziranje njihovih perfomansi. ROC grafovi su često korišteni u medicini ali posljednih godina raste njihova upotreba u područjima kao što su strojno učenje i rudarenje podataka. Jedan od prvih usvojitelja ROC grafova u strojnom učenje je bio **Spackman(1989)**,koji demonstrira vrijednosti ROC krivulja u evaulaciji i usporedbi algoritama. Usto što je generalno dobra metoda za vizualizaciju izvedbi,postoje određena svojsta koje ih čine posebno korisnim u domeni sa asimetričnom klasnom distribucijom i nejednakim klasifikacijskim troškovima greške.

Klasni probem započinje uzimajući u obzir recimo dvije klase. Formalno,svaka instanca I je mapirana do jednog elementa iz skupa {p,n} pozitivnih i negativnih klasnih oznaka. Klasifikacijski model(klasifikator) mapira instance u predviđene klase. Neki klasifikacijski modeli proizvode kontiunirane izlazne jedinice na koje mogu biti primjenjeni različiti pragovi za predviđanje klasnog člasnstva. Drugi modeli proizvode diskretne klasne oznake koje indiciraju samo predviđenu klasu za instancu. Kako bi razlikovali stvarnu klasu i predviđenu klasu koristimo oznake {P,N} za predviđenu klasu koju je stvorio model.

S obzirom na klasifikator i instancu,postoje 4 moguća ishoda. Ako je instanca pozitivna i klasificirana je kao pozitivna ,broji se kao STVARNO POZITIVNA; ako je klasificirana kao negativna označuje se kao LAŽNO NEGATIVNA. Ako je instanca negativna i klasificira se kao negativna,označavamo je kao STVARNO NEGATIVNA; ako je klasificirana kao pozitivna,označavamo je kao LAŽNO POZITIVNA. Dobivenim klasifikator om i skupom instanci(test skup) možemo izgraditi KONFUKSIJCKU MATRICU. Matrica predstavlja osnovu za mnoge metrike.

ROC graf je dvodimenzionalni graf gdje se *tp stopa* nalazi na osi y a *fp stopa* na osi x. ROC graf prikazuje relativan kompromis između prednosti(true positive) i troškova(false positive)

4.METODE KLASIFIKACIJE

4.1 STABLO ODLUČIVANJA

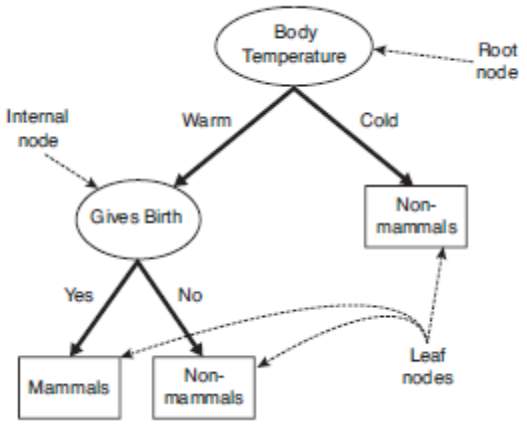
Kako bi ilustrirali kako stabla odlučivanja FUKCIONIRAJU,uzećemo za primjer bazU podataka za kičmenjake.Kičmenjake ćemo klasificirati prema dvije kategorije: sisavci i ne sisavci. Pretpostavimo da je nova vrsta otkrivena. Koliko u tom slučaju možemo dobro procijeniti je li sisavac ili nije.Jedan pristup je postaviti seriju pitanja o karakteristikama vrste. Prvo pitanje koje bi mogli pitati je li vrsta toplokrvna ili hladnokrvna.Ako je hladnokrvna onda defitivno nije sisavac. Inače je ili ptica ili sisavac. Pitanje u kasnijoj fazi može npr. Biti: Da li ženke rađaju mlade? One koje rađaju su definitivno sisavci,dok one koje ne rađaju uglavnom nisu sisavci.

Ovaj primjer nam pokazajue kako možemo riješiti klasifikacijski problem s nizom pažljivo izrađenih pitanja o atributima test seta. Svaki put kad primimo odgovor, nadolazi prateće pitanje sve dok ne dosegnemo zaključak o klasnoj oznaci instance. Niz pitanja i mogućih odgovora se može organizirati u formi stabla odlučivanja,što je hijerarhijska struktura koja se sastoji od čvorova i rubova. Stabloima tri tipa čvorova.

- početni čvor- nema ulaznih grana i nula ili više izlaznih grana
- unutrašnji čvor,svaki ima točno jednu ulaznu granu i 2 ili više izlaznih grana
- krajnji čvor - svaki ima točno jednu ulaznu granu bez izlaznih grana

U stablu odlučivanja, svakom krajnom čvoru je dodijeljena klasna oznaka. Čvorovi koji nisu krajnji, znači početni i unutrašnji čvorovi,sadrže atributne uvjete koje dijele instance s različitim karakteristikama.

Elementi znanja kod stabla odlučivanja su čvorovi i grane.Grane povezuju „ roditeljske čvorove“ s „dječjim čvorovima“. Čvor bez rodteilja naziva se „korijenski čvor“, a čvorovi bez djece su „ listovi“ - Listovi se još nazivaju „čvorovi odgovra“ jer oni predstavljaju sva moguža rješenja zadanog problema. Svi ostali čvorovi su „čvorovi odluke“



Slika 6. Primjer stabla odlučivanja

ALGORITMI

ID3,C4.5 i CART su najpoznatiji algoritmi koji na rekurzivan način od gore prema dolje u maniri podijeli pa vladaj konstruiraju stablo odlučivanja. Većina algoritama za stabla odlučivanja koriste ovaj pristup,koji počinje trenig setom observacija s nekom klasnom oznakom. Trenig set se rekurzivno particira na manje podsetove i tako se stablo gradi. Osnovni algoritam stabla odlučivanja možemo sumirati na sljedeći način:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C, then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

Slika 7. Indukcija stabla odlučivanja

- Algoritam se poziva s 3 parametra : D, atribut_lista i Atribut_metoda_selekcije. Na D se odnosi djeljenje podataka. Parametar atribut_lista je lista atributa koji opisuju observaije. Atribut_metoda_selekcije je parametar koji specifira istraživačku proceduru za izabrati atribut koji najbolje dijeli observacije po klasama.Ova procedura sadži mjeru selekciju infomacijski dobitak ili Gini index.Neke mjere za selekciju atributa kao Gini indeks primjenjuju binarni izgled stabla. Druge,kao informacijska dobit,omogućuju višestruku podijelu.
- Stablo započinje kao jedan čvor,N i predstavlja trenig observacije u skupu D
- Ako su observacije u skupu D iste klase,onda čvor N postalje krajnji čvor (leaf) i označen je u tom razredu(korak 2 i 3). Koraci 4 i 5 su uvjeti otkazivanja.Svi uvjeti otkazivanja su objašnjeni na kraju algoritma
- Inače,algoritam poziva parametar *Attribute_selection_method* koji određuje kriterije podjele. Kriterij podjele nam govori koji atribut treba testirati u čvoru N te određuje najbolji način za podijeliti observacije iz seta D u individualne klase (korak 6) .Kriteriji podjele također odlučuju koje grane rastu iz čvora N . Ako želimo biti specifični ,kriterij podjele zapravo odlučuje o atributu podjele i naznačuje točku dijeljenja ili podijeljeni podset.

  Kriterij podjele određuje u idealnom slučaju da podijele od svih grana budu što„čišće“. Particije ili podjele su čiste ako sve observacije u njima pripadaju istom razredu. Drugim

riječima ako podijelom observacije u setu D prema međusobno isključivim ishodima kriterija podjele,nadamo se da će nastale podjele biti što čišće

- Čvor N je označen s kriterijom podjele,koji služi kao test za čvor(korak 7). Za svaki ishod kriterija podjele izrasta grana iz čvora N. Observacije u setu D su particirane prema (korak 10 i 11). Postoje 3 moguća scenarija. Na primjer neka A bude atribut podjele . A ima v posebnih vrijednosti {a1,a2,....,av),bazirane na *training* podacima.

  1. A je diskretna vrijednost: U ovom slučaju,ishodi testa čvora N direktno odgovaraju poznatim vrijednostima A. A grana je kreirana za svaku poznatu vrijednost,aj,od A i označeno s tom vrijednosti. Particija Dj je podset klasnih oznaka observacija u D podsetu s vrijednostima aj atributa A. Zato što sve observacije u pojedinim particijama imaju istu vrijednost za atribut A te nema potrebe da se A uzima u obzir u budućim particijama observacija.
  2. A je kontuinirana vrijednost. U ovom slučaju čvor N ima 2 moguća ishoda,koja odgovaraju uvjetima A< *split-point* i A>split_point. Observacije su podijeljene tako da D1 sadrži podset klasnih oznaka observacija u D gdje je A < *split_point,* dok D2 sadrži ostale.
  3. A je diskretna vrijednost kojom se kreira binarno stablo. Test na čvoru N je u formi „ A ϵ Sa“ gdje je Sa podijeljeni podset atributa A. To je podset poznatih vrijednosti atributa A. Ako određena observacija ima vrijednost aj atributa A i ako je u skupu Sa,tada je test na čvoru N zadovoljen.Dvije grane su izrađene iz čvora N. Po konveciji,lijeva grana iz čvora N je označena sa *da* tako da D1 odgovora podsetu podataka u setu D koji odgovoaraju testu,a desna je označena sa *ne* analogno tome i podskup D2 odgovora podsetu podataka koji sad sadrži klasno označene instance iz skupa D koji zadovoljavaju test
- Algoritam koristi isti proces rekursivno kako bi formirao stablo odlučivanja za observacije za svaku nastalu particiju Dj od skupa D
- Rekursivno particiranje se stopira kad je jedan od uvjeta otkazivanja zadovoljen:

  1.Sve instance u partitivnom djelu D (zastupljene u čvoru N) pripadaju istoj klasi(korak 2 i korak 3)

2.Nema više atributa po kojima bi se instance mogle dijeliti(korak 4). U ovom slučaju većinsko glasovanje je upotrebljeno(korak 5) . Ovo uključuje konverziju čvora N u list i označavanje s najčešćom klasom u skupu D.

3. Nema observacija za pojedinu granu,što znači da je podskup Dj prazan (korak 12). U ovom slučaju,list je stvoren s većinskom klasom iz skupa D.

Nastalo stablo odlučivanja je vraćeno (korak 15) [11]

MJERA ODABIRA ATRIBUTA

Glavni cilj provedbe selekcije atributa je izabrati podskup ulaznih atributa kako bi se eliminirali atributi koji nisu relevantni i koji ne daju prediktivnu informaciju te konačno, postizanje visoke točnosti klasifikacije (**Ramaswami i Bhaskaran, 2009.).**

Mjera po kojoj se odabiru atributi je heuristična po tome što odabire kriterije podjele koji na „najbolji" način dijele podatke. Ako želimo podijeliti skup D u manje dijelove ili particije prema ishodu kriterija podjela,idealno bi bilo da svaka particija bude čista (sve instance u jednoj particiji pripadaju istoj klasi). Konceptualno , „najbolji" kriterij podijele je onaj koji daje takve rezultate ili bar približno. Mjere odabira atributa se zovu i pravila podjele jer određuju kako se instance na određenim čvorima dijele. Atribut koji je izmjeren kao najbolji je odabran kao atribut podjele za postojeće instance. Ako je atribut podjele kontinuirana vrijednost onda se točka podjele ili podskup podjele određuje kao dio kriterija za podijelu. Čvor na stablu kreiran za određen dio je označen kriterijom podjele,grane izrastaju za svaki ishod kriterija i prema njima dodjeljuju instance. U sljedećom odjeljku ćemo objasniti 3 popularne mjere izbora atributa- informacijska koristi, gain ratio i Gini index

[11] Kumar et all. Introduction to data mining

Informacijska dobit

Informacijska dobit se koristi kad želimo odrediti koji atribut u danom trenig vektoru značajki ćemo koristiti za diskriminaciju među klasama. Informacijska dobit nam kaže koliko je važan dani atribut u istaknutom vektoru.Koristi se za odlučivanje poretka atributa u čvorima stabla odlučivanja.

Informacijka korist = entropija(roditelj čvor) - [prosjek entropije(djeca čvor)]

Prilikom izgradnje stabla odlučivanja bira se atribut s najvećom informacijskom dobiti iz trenig seta i i postavlja se kao korijen čvor stabla zatim se stvaraju podređeni čvorovi zvani djeca čvorovi . U svakom čvoru postoji podskup vektora u kojem atribut ima određenu vrijednost. Ovaj postupak se se nastavlja ponavljati rekursivno.(slajd-glava)

Informacijska dobit predstavlja razliku entropije prije grananja i entropije nakon grananja nad atributom A.

ID3 koristi informacijsku dobit kao mjeru selekcije atributa. Recimo da čvor N predstavlja ili sadržava instance skupa D. Atribut sa najvećom informacijskom dobiti je izabran kao atribut podjele za čvor N. Ovaj atribut minimalizira informacije potrebne za klasificiranje instanci u proizašle particije i reflektira njihovu „čistoću“ . Ovakav pristup minimalizira očekivani broj testova potrebnih za klasificiranje instanci i garantira pronalazak jednostavnog stabla.

OMJER DOBITI

Mjera informacijska dobit je pristrana prema testovima s mnogo ishoda. Na primjer ako imamo atribut koji ima ulogu jedinstvenog identifikatora kao npr proizvod_ID ili radnik_ID. Podjela prema atributu proizvod_ID će rezultirati sa velikim brojem podjela(onoliko koliko ima vrijednosti),svaka sadržavajući samo jednu instancu. Tada je svaka particija čista te je informacijska dobit na ovome atributu maksimalna. Očito,da je takvo particiranje beskorisno za klasificiranje. C4.5 nasljednik ID3,koristi ekstenziju poznatu kao omjer dobiti,koja pokušava prevladavati ovu pristranost.

GINI INDEX

Gini index koristi algoritam CART. Gini index mjeri nečistoću skupa D,dijela skuipa ili trenig seta

Gini index stvara binarni prelom za svaki atribut. Na primjer ako nam je A atribut diskretne vrijednosti s v različitih vrijednosti {a1,a2,....,av} u skupu D. Kako bi odredili najbolji binarni prelom atributa A, ispitujemo sve moguće podskupove koje se mogu formirati koristeći vrijednost A . Na primjer,ako

dohadak ima 3 moguće vrijednosti { *low,medium,high* } onda mogući podskupovi su *{low,medium,high}*, {*low,medium}, {low,high},{medium,high},{low},{medium},{high}, i {}. Isključit ćemo podskup {low,medium,high}* i prazni set iz razmatranja jer konceptualno oni ne predstavljaju podijelu. Prema tome, postoji 2^v2 mogućih načina formiranja dvije particije podataka skupa D,bazirane na binarnom splitu atributa A.

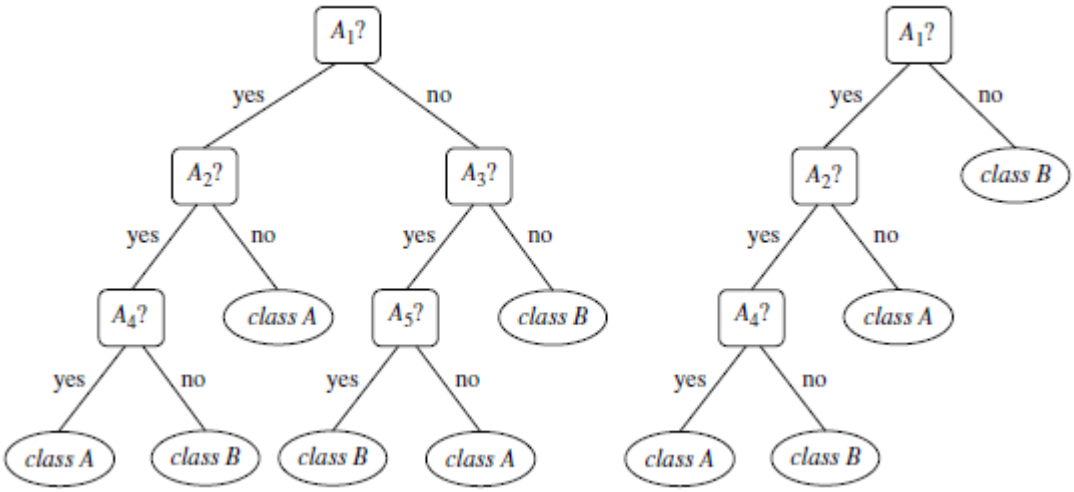
Gini index često se opisuje kao mjera „čistoće“ čvora.Čistoća čvora predstavlja one čvorove u kojima je visok postotak instanci koji pripadaju istoj klasi. Mala vrijednost gini indeksa ukazuje na „čiste“ čvorove

Prikazane su 3 mjere koje se često koriste u izgradnji stabla odlučivanja. Ove mjere imaju i svoje slabosti. Informacijska dobit je pristrana prema atributima s više vrijednosti. Omjer dobiti se prilagođava ovoj pristranosti,ali preferira nebalansirane podijele gdje je jedna particija puno manja od druge. Gini index je pristran prema atributima s više vrijednosti i ima poteškoće ako je broj klasa velik. Također preferira testove koji rezultiraju jednako velikim particijama i čistoći u obe particije. Iako su mjere pristrane,daju razumno dobre rezultate u praksi.

Postoje i druge mjere odabira atributa. Na primjer CHAID,algoritam stabla odlučivanja koji je popularan u marketingu,koristi mjeru odabira atributa koja je bazirana na statističkom X^2testu za neovisnost. Postoje i druge mejere kao što su C-SEP ,G-statistic i MDL koji ima najmanju pristranost prema atributima s više vrijednosti.

Skraćivanje stabla odlučivanja

Kad se stablo odlučivanja izgradi ,neke grane će reflektirati anomalije u *training* podacima zbog šumova i outliera. Metoda skraćivanja rješava problem preuklapanja podataka. Ovakve metode tipično koriste statističke mjere za uklanjanje najmanje pouzdanih grana. Ne podrezano stablo i podrezana verzija je prikazana na slici.



Slika 8.Podrezivanje stabla odlučivanja

Podrezana verzija je manja i manje kompleksna te lakša za shvatiti. Obično su brža i bolja pri klasificiranju nezavisnih test podataka nego nepodrezana stabla. Postoje dva pristupa za skraćivanje stabala:*podrezivanje i nadrezivanje.*

Pri pristupu podrezivanja stabala ,stablo se skraćuje na način da se odlučuje da se dijeljenje neće nastaviti na određenom čvoru. Nakon zaustavljanja,čvor postaje list. List može sadržavati najfrekfentniju klasu među podskupovima . Kod izrade stabla,mjere kao što su statistička značajnost,informacijska dobit,gini index i druge mogu se koristiti za procijeniti kvalitetu podijele ili prijeloma. Postoje poteškoće oko odabira praga za skraćivanje. Visoki prag rezultira prepojednostavljenom stablu,dok visok prag rezultira niskom jednostavnošću.

Drugi i više korišteni pristup skraćivanja stabla je nadrezivanje, koji uklanja podstabla iz potpuno izgrađenog stabla. Podstablo do nekog čvora je skračeno uklanjanjem grana i njihovom zamjenom sa listom. List je označen sa najfrekfentnijom klasom među podstablom koje je bilo uklonjeno. Koristi se posebni algoritam skraćivanja poznat kao trošak kompleksnosti koji je funkcija broja listova i stope pogreške stabla(postotak krivo klasificiranih instanci). Algoritam na taj način određuje koja će podstabla se zamijeniti a koja ostati. Osim ove postoje druge metode skraćivanja stabla zasnovane na različitim funkcijama.

Alternativno,podrezivanje i nadrezivanje se može kombinirati. Nadrezivanje zahtijeva veću računalnu moć ali vodi do pouzdanijeg stabla. Niti jedna metoda skraćivanja nije superiornija nad ostalima. Neke metode

ovise o dostupnosti podataka za skraćivanje,iako ovo nije problem kad se radi sa velikim bazama podataka.

Iako stabla koja su skraćena imaju tendeciju da su više kompaktna nego neskraćena stabla,svejedno mogu biti velika i kompleksa.Kod stabla odlučivanja može doći do problema **ponavljanja i replikacije.** Problem ponavljanja se događa kad se atribut uzastopno testira niz više grana u stablu odlučivanja. Problem replikacije kad postoji više istih podstabala u jednom stablu odlučivanja

Skalabilnost stabla odlučivanja

Koliko je skalabilna indukcija stabla odlučivanja,tj. ako podaci ne stanu u memoriju?

Efikasnost postojećih algoritama,kao što su ID3,C4.5 I CART, je dobro uspostavljena za male skupovne podatke. Efikasnost postaje problematična kad se ovi algoritmi primjenjuju na velikim stvarnim bazama podataka. Spomenuti algoritmi stabla odlučivanja imaju ograničenje, *training* instance moraju se nalaziti u memoriji.

U rudarenju podataka,veliki trenig skupovi sa miljonima instanci su česti. Većinu puta,trenig podaci će biti toliko veliki da neće moć stat u memoriju računala. Dakle, izgradnja stabla odlučivanja može postati neefikasna. Pristupi koj su skalabilniji,sposobni upravljati *training* podacima koji su preveliki, su potrebni. Rane strategije za čuvanje mjesta u memoriji uključuju diskretizaciju kontiuniranih vrijednosti atributa i uzorkovanje podataka na svakom čvoru. Ove tehnike, međutim,još predstavljaju da *training* set može stati u memoriju.

Nekoliko skalabilnih metoda indukcije stabla odlučivanja postoje,a jedno od njih je *RainForest* koji se prilagođava količini dostupne glavne memorije i primjenjuje se na bilo koji algoritam indukcije stabla odlučivanja**.**

U teoriji,postoji eksponecialno mnogo stabla odlučivanja koje se mogu izgraditi od danog niza atributa. Dok su neka stabla odlučivanja mnogo preciznija od drugih, naći optimalno stablo je računalno neisplativo zbog eksponencijalnoe veličine prostora za pretraživanje. Štoviše, efikasni algoritmi su razvijeni koji mogu inducirati dovoljno precizno stablo odlučivanja u razumnom vremenu.

Stabla odlučivanja su jedna od najčešće korištenih metoda induktivnog zaključivanja. To je metoda za procjenu diskretnih funkcija koje su robusne na šum u podacima i sposoben „učiti“ disjuktivne izraze. „Naučena funkcija"

4.2 NEURONSKE MREŽE

Neuronske mreže su simultano jedno od najstarijih i najnovijih područja strojnog učenja. Počeli su se razvijat već u četrdesetima prošlog stoljeća kada su ljudi počeli izgrađivati modele na osnovi funkcioniranja mozga. Prva neuronska mreža,percepton se razvio u pedesetima i bio je dosta popularan te se primjenjivao na velikom broju problema.Percepton je najjednostavnija neuronska mreža koja se sastoji od jednog neurona i kao takva predstavlja osnovnu građevnu jedinicu višeslojnih neuronskih mreža Međutim ti problemi na kojima se primjenjiva s u bili s dosta mana te se percepton pokazao limitiran. Poslije toga dvoje ljudi je objavilo dosta utjecajan članak o neuronskim mrežama te njihovo istraživanje i korištenje praktički zamrlo jer je bilo toliko priče o neuronskim mrežama a nisu mogli riješiti jednostavne procese. Njihovo ponovno „rođenje“ se dogodilo u osamdesetima prošloga stoljeća kad su ljudi shvatili kako koristiti višestruke perceptone i spojiti ih u zajedničku mrežu. Došlo je do dva velika postignuća,backprogaation algoritam i simetrična mreža koja su mogla riješiti raznolike probleme učenja,međutim došlo je do bizarne situacije da su samo dva stručnjaka znali izgraditi mreži i koristitije za rješavanje problema u cijelome svijetu. U to vrijeme su se pojavili moćni linearni klasifikatori kao *support vector machines* te su neuronske mreže opet ispale iz mode i zamrle . Nakon dvije godine su se opet pojavile kad su ljudi shvatili kako ih jednostavno i masovno koristiti, velike zasluge za to imaju hardwerske promjene. Trenutno je jedno od najistraživanijih područja i najaktivinih algoritama.

Razvoj neuronskih mreža

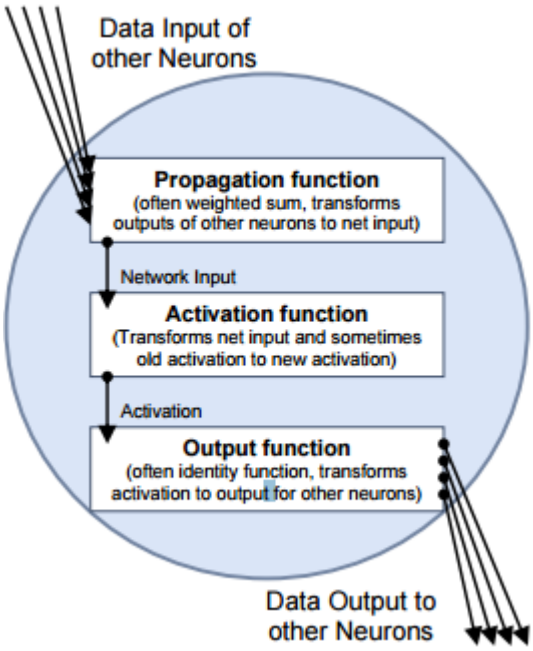
Već u početcima razvoja neuronskih mreža,neuronsko računalstvo se profiliralo kao jedna od grana umjetne inteligencije, točnije pedesetih godina prošlog stoljeća na konferenciji Dartmouth Summer Rreearch Project on Artifical Intelligence na kojoj se prezentirala vizija računalnog modela koji oponaša ljudski mozak. Kao što je već spomenuto neuronsko računalstvo nastji simulirati ili ostvariti paralelnu obradu informacija koju koristi ljudski mozak dok razmišlja,sjeća se i rješava probleme. Za razvoj neuronskih mreža od predsudnog značaja je nekoliko događaja

- 1943 - McCullock i Pitts postavljaju temelje za razvoj neuronskih mreža tako što prvi dokazuju da neuroni mogu imati dva stanja (pobuđujuće i umirujuće) i da njihova aktivnost ovisi o nekom pragu vrijednosti.
- 1949 - Hebb prvi predložio pravilo kojim se opisuje proces učenja (Hebb-ovo pravilo)

- 1956 - Dartmouth Summer Conference na kojoj su Rochester i skupina autora predstavili prvu simulaciju Hebb-ovog modela koja je preteča modela neuronskih mreža
- 1958 - Rosenblatt razvio prvu neuronsku mrežu perceptron, koja je dvoslojna i nije mogla rješavati probleme klasifikacije koji nisu linearno djeljivi (npr. XOR problem)
- 1974 - razvijena višeslojna perceptron mreža - MLP (Paul Werbos), kao preteča Backpropagation mreže, koja prevladava nedostatak perceptrona uvođenjem učenja u skrivenom sloju
- 1986 - Backpropagation mrežu usavršuju Rumelhart, Hinton i Williams, ona vraća ugled neuronskim mrežama, jer omogućuje aproksimiranje gotovo svih funkcija i rješavanje praktičnih problem ate ima najveću komercijalnu upotrebu danas.

Dakle neuronske mreže vuču inspiraciju iz ljudskog mozga I imaju cilj spojiti sposobnosti ljui da da dobro prepoznaju oblike,likce I glasove I sposobnost računala da izvršava numeričke proračune I radi s velikom količinom podataka

.**Obrada informacija u neuronskoj mreži**



Slika 9.Obrada informacija u neuronskoj mreži izvor: (http://eris.foi.hr/11neuronske/nn-predavanje4.html)

Definicija neuronski mreža

Neuronska mreža je međusobno povezana nakupina jednostavnih elemenata obrade, jedinica ili čvorova, čiji se načini djelovanja otprilike temelji na neuronima kod životinja. Sposobnost obrade mreže je posljedica jačine veza među tim jedinicama, a postiže se kroz proces adaptacije ili učenjem iz skupa primjera za uvježbavanje. (SAJT)Pokazat ćemo kako se obrađuju informacije na neuronskoj mreži koja se naziva širenje unatrag (engl. Backpropagation). Ta mreža se intenzivno upotrebljava za različite klase problema te imajednostavan modelkojisemože lako opisati i naučiti. [12]

Učenje

Učenje je proces mijenjanja težina u mreži, a odvija se kao odgovor na podatke izvana koji su predstavljeni ulaznom sloju i u nekim mrežama izlaznom sloju. Podaci koji se predstavljaju izlaznom sloju su željene vrijednosti izlaznih varijabli. Ukoliko su one poznate, radi se o tzv. nadgledanom učenju. Na primjer, nadgledani algoritmi su: mreža širenje unatrag, mreža s radijalno zasnovanom funkcijom, modularna mreža, vjerojatnosna mreža, LVQ (mreža učeće vektorske kvantizacije), i drugi. Ukoliko je ulazni vektor jednak izlaznom vektoru, radi se o autoasocijativnim mrežama, a ukoliko je različit, radi se o heteroasocijativnim mrežama. Kod nekih mreža željeni izlaz ne mora biti predstavljen mreži. U tom slučaju radi se o tzv. nenadgledanom učenju. Najčešći nenadgledani algoritmi su Kohonenova mreža, mreža konkurentskog učenja, te ART (mreža adaptivne rezonantne teorije).

Prije samog učenja potrebno je definirati model (ulazne i izlazne varijable), te prikupiti podatke iz prošlosti na kojima će se primijeniti mreža. Prikupljene podatke treba podijeliti u dva poduzorka (uzorak za treniranje i uzorak za testiranje), a ukoliko se za vrijeme učenja planiraju koristiti optimizacijske tehnike za optimiranje duljine učenja i strukture mreže, potrebno je ukupan uzorak podijeliti na tri poduzorka (za treniranje, testiranje i konačnu validaciju). Pravila za ovu podjelu nema, osim što se preporuča najveći dio podataka ostaviti za treniranje mreže, a manji dio podataka za testiranje i validaciju (npr. 70% za treniranje, 10% za testiranje i 20% za validaciju). Podaci se raspoređuju u poduzorke slučajno, osim kod vremenskih serija gdje treba poštovati vremenski slijed nastajanja promatranja, tj. trenirati mrežu na starijim, a testirati na novijim podacima.

Nakon što je definiran model, pripremljeni ulazni podaci i izabran NM algoritam, te pravilo učenja i potrebne funkcije, mrežu treba učiti ili trenirati na pripremljenim podacima iz prošlosti, kako bi ona prepoznala vezu između podataka i bila u mogućnosti na osnovu ulaznih vrijednosti predviđati izlaze.

[12] (http://eris.foi.hr/11neuronske/nn-predavanje4.html

Sama faza učenja je proces podešavanja težina u mreži, koje se odvija u više iteracija ili prolaza kroz mrežu. Jedna iteracija predstavlja učitavanje jednog promatranja iz podataka (jednog ulaznog i izlaznog vektora), ali se zbog povećanja brzine učenja ponekad preporuča učitati više promatranja odjednom, pri čemu se broj promatranja koji se obrađuju u jednoj iteraciji zove epoha. U svakoj iteraciji računaju se nove težine, a kod nadgledanih algoritama i nova greška. Obično se mreža trenira u nekoliko tisuća iteracija.

Najvažnije pitanje u ovoj fazi je koliko dugo trenirati mrežu kako bi ona dala što bolji rezultat, odnosno najmanju grešku. Ne postoje egzaktna pravila za dužinu treniranja, te odgovor na ovo pitanje treba potražiti vlastitim eksperimentiranjem ili primjenom optimizacijskih tehnika kao npr. tehnika unakrsnog testiranja. Ova se tehnika može opisati u nekoliko koraka:

- mreža se najprije trenira na određenom broju iteracija (npr. 10000),
- tako naučena mreža se testira na uzorku za testiranje, i pohrani dobiveni rezultat i mreža.
- mreža se zatim nastavlja trenirati na još tolikom broju iteracija (npr. još 10000), te se dobiveni rezultat uspoređuje s prethodno pohranjenim. Ukoliko je u ponovnom učenju dobiven bolji rezultat, pohranjuje se novi rezultat i nova mreža.
- postupak se ponavlja sve dok se rezultat prestane poboljšavati, a najbolja pohranjena mreža ulazi u daljni postupak validacije.

Rezultat (npr. RMS greška) dobiven u fazi učenja nije mjerodavan za ocjenjivanje mreže, jer ne pokazuje ponašanje mreže na novim podacima.

2)Testiranjemreže

Testiranje mreže je druga faza rada neuronske mreže, i ona je odlučujuća za ocjenjivanje mreže. Razlika između faze učenja i faze testiranja je u tome što u ovoj drugoj fazi mreža više ne uči, a to znači da su težine fiksne na vrijednostima koje su dobivene kao rezultat prethodne faze učenja. Takvoj mreži se predstavljaju novi ulazni vektori koji nisu sudjelovali u procesu učenja, a od mreže se očekuje da za predstavljen novi ulazni vektor proizvede izlaz. Ocjenjivanje mreže obavlja se izračunavanjem greške ili nekog drugog mjerila točnosti (npr. stope točnosti), na način da se izlaz mreže uspoređuje sa stvarnim izlazima.

Dobivena greška mreže na uzorku za validaciju je rezultat kojim se tumači uspješnost ili neuspješnost neuronske mreže i njezina korisnost u primjeni za predviđanje na budućim podacima.

Kako se dizajnira umjetna neuronska mreža

Ovisno o temeljnim formulama koje se koriste za učenje, ulazne i izlazne funkcije, postoje različiti algoritmi NM, a unutar svakog algoritma moguće su intervencije u strukturi mreže (topologiji) i izboru parametara učenja, te tako postoji široki spektar NM arhitektura. One se međusobno razlikuju prema kriterijima:

- broju slojeva (dvoslojne i višeslojne),
- tipu veze između neurona (inter-slojne veze i intra-slojne veze),
- vezi između ulaznih i izlaznih podataka (autoasocijativne i heteroasocijativne),
- ulaznim i izlaznim (prijenosnim) funkcijama,
- pravilu učenja,
- ostalim parametrima (sigurnosti ispaljivanja, vremenskim karakteristikama, i dr.).

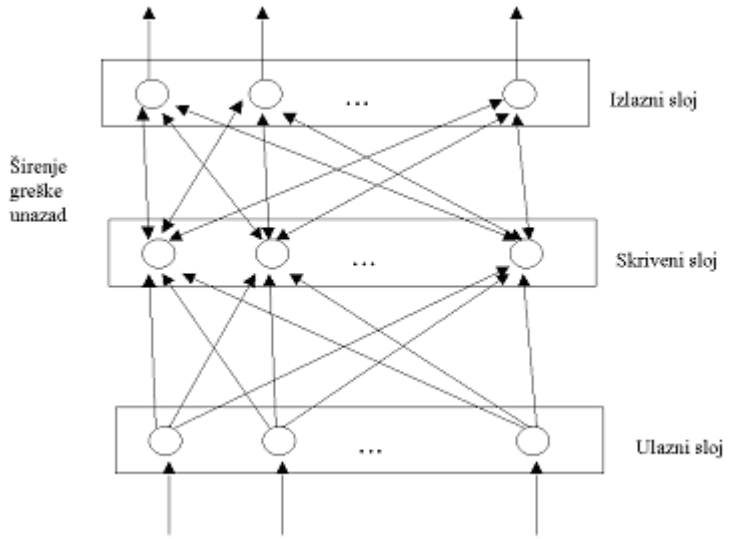
Od gore navedenih karakteristika, za razlikovanje algoritama od osobitog je značenja pravilo učenja. Pravilo učenja specificira način na koji se podešavaju težine u mreži.

Mreža "širenja unatrag"

Algoritam mreže "širenje unatrag" bio je presudan za široku komercijalnu upotrebu ove metodologije, te je neuronske mreže učinio široko upotrebljavanom i popularnom metodom u različitim područjima. Njezin prvi kreator bio je Paul Werbos 1974., a proširena je od strane Rumelhart-a, Hinton-a i Williams-a 1986. Bila je to prva neuronska mreža s jednim ili više skrivenih slojeva. U osnovi, ova mreža propagira input kroz mrežu od ulaznog do izlaznog sloja, a zatim određuje grešku i tu grešku propagira unazad sve do ulaznog sloja ugrađujući je u formulu za učenje. Standardni algoritam mreže "širenje unatrag" uključuje optimizaciju greške koristeći deterministički algoritam gradijentnog opadanja (eng. gradient descent). Glavni nedostatak ovog algoritma je problem čestog pronalaženja lokalnog umjesto globalnog minimuma greške, stoga novija istraživanja uključuju njegovo unapređivanje nekim drugim determinističkim (npr. metode drugoga reda) ili stohastičkim metodama (npr. simulirano kaljenje).

Strukturu mreže čine ulazni sloj, izlazni sloja i najmanje jedan skriveni sloj, s vezom unaprijed. Tipična arhitektura "širenje unatrag" prikazana je na donjoj slici (zbog jasnoće je prikazan samo jedan skriveni sloj):

Slika10.
Arhitektura mreže "širenje unatrag"



Tok podataka kroz mrežu može se ukratko opisati u nekoliko koraka:

1. od ulaznog sloja prema skrivenom sloju: ulazni sloj učitava podatke iz ulaznog vektora X, i šalje ih u prvi skriveni sloj,
2. u skrivenom sloju: jedinice u skrivenom sloju primaju vagani ulaz i prenose ga u naredni skriveni ili u izlazni sloj koristeći prijenosnu funkciju,
3. kako informacije putuju kroz mrežu, računaju se sumirani ulazi i izlazi za svaku jedinicu obrade,
4. u izlaznom sloju: za svaku jedinicu obrade, računa se skalirana lokalna greška koja se upotrebljava u određivanju povećanja ili smanjenja težina,
5. propagiranje unazad od izlaznog sloja do skrivenih slojeva: skalirana lokalna greška, te povećanje ili smanjenje težina računa se za svaki sloj unazad, počevši od sloja neposredno ispod izlaznog sve do prvog skrivenog sloja, i težine se podešavaju.

Mreža "širenje unatrag" je univerzalni algoritam primjenjiv na probleme previđanja, gdje je potrebno predvidjeti vrijednost jedne ili više izlaznih varijabli, no moguće ga je koristiti i za probleme klasifikacije, gdje se ulazni vektor raspoređuje u jednu od klasa zadanih na izlazu, npr. određivanje da li neka dionica pripada u grupu rastućih ili padajućih dionica u nekom razdoblju. U tu svrhu, standardni algoritam ove mreže potrebno je proširiti Softmax aktivacijskom funkcijom, koja osim što podstiče mrežu da što jasnije rasporedi ulazni vektor u jednu od klasa, omogućava i usporedbu rezultata sa statističkim metodama koje kod klasifikacije daju vjerojatnosti da će ulazni vektor pripadati u neku od klasa. Mreža "širenje unatrag" ne preporuča se za upotrebu na nestacionarnim podacima, ili za slučajeve kada podaci u sebi skrivaju više, u osnovi različitih, problema. Rješenje za takve probleme može se pronaći u

upotrebi nekoliko neuronskih mreža od kojih će svaka rješavati pojedini problem zasebno, ili u izboru nekog drugog algoritma.

Model neurona

Neuron je osnovni procesni element neuronske mreže, zamišljen kao matematička funkcija koja imitira primitivan model biološkog neurona. Model neurona prikazan je na slici 1

Elementi modela neurona su:

- skup sinapsi, tj. ulaza (x1...xp) od kojih svaki ima svoju težinu (wk1...wkp) (signal xj na ulazu j neurona k ima težinu wkj),
- sumator za zbrajanje otežanih ulaza, tj. računanje linearne kombinacije ulaza,
- aktivacijska funkcija koja ograničava izlaz neurona na interval [0,1].

Aktivacijska funkcija može biti linearna ili nelinearna. Neke od najčešćih aktivacijskih funkcija prikazane su na slici 2.

Višeslojni perceptron karakterizira nelinearna aktivacijska funkcija, najčešće sigmoidna funkcija.[13]

Tehnički neuronske mreže se sastoje od jednostavnih procesnih jedinica,neuroona, I usmjerenih ,ponderiranih veza među neuronima. Snaga veze između dva neurona *i* I *j* referera se kao Wij. Neuronska mreža je sortirana u 3 dijela (N,V,w) sa dva seta N,V I funkcijom w, gdje je N set neurona I V set {(i,j)|I,j e N} čiji elementi se zovu veze ili konekcije između neurona i i neurona j. Funkcija w :V → Rdefiniria pondere,gdje ((i,j)), , ponder veza među neuronom I I neuronom j, je skraćeni izraz Wi,j.

Ponderi se mogu implementirati u kvadratnoj ponderiranoj matrici W ili, po želji, u ponderiranin vector W, gdje red matrice ukazuje gdje veza/konekcija počinje,a broj stupca matrice ukazuje,koji je neuron ciljan

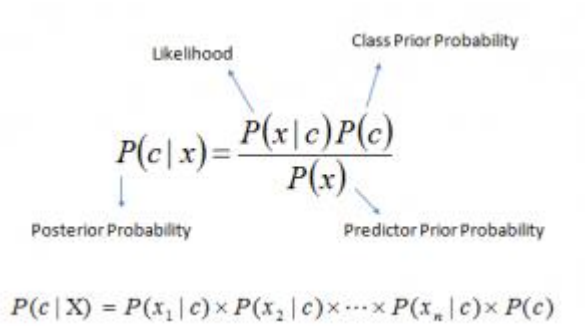
[13] http://www.zemris.fer.hr/predmeti/kdisc/Sem2.pdf)

(target). U ovom slučaju sa 0 se označuju ne postojeće veze. Ovakva prezentacija matrice se zove Hintonov dijagram.Podaci se prenose među neuronima kroz veze sa povezanim ponderima koji su ekscitatorski i inhibitorski. Veze među neuronima su usmjerene da bi znali kojim putem informacije idu. Veze također imaju razliite vrijednosti jer neke veze su važnije od drugih, vrijednost veze nazivamo ponderom ili težinom veze.[14]

4.4 NAIVNI BAYESOV ALGORITAM

Naivni bayesov klasifikator je klasifikacijska tehnika temeljena na Bayesovom teoremu sa pretpostavkom nezavisnosti među prediktorima. Jednostavnije rečeno, utjecaj atributa na klasu je nezavisan o vrijednostima drugih atributa. Na primjer , za neko voće ćemo reći da je jabuka ako je crveno I okruglo. Iako ove značajke ovise jedna o drugoj ili o postajanju drugih značajki , sva ova svojsta nezavisno pridonose vjerojatnosti da to neko voće je jabuka I zato je algoritam dobio naziv naivni. Model naivni bayes je lagan za izgradit I posebno koristan za velike baze podataka. Vrlo je jednostavan i algoritam koji je poznat da može nadigradti visoko sofisticirane klasifikacijske metode

Bayesov teorem pruža način za izračunavanje aposteriorne vjerojatnosti P(c|x) iz P(c) ,P(x) I P(x|c)



Slika 11. Bayesov teorem

- P(c|x) je posteriorna vjerojatnost klase(c,ciljna vrijednost) s obzirom na predictor (x,atribut)
- P(c) je apriorna vjerojatnost klase
- P(x|c) je vjerojatnost prediktora s obzirom na klasu
- P(x) je apriorna vjerojatnost prediktora

Bayesov teorem

[14] Kriesel Neural networks

Bayesov teorem je dobio ime po Thomas Bayesu, nekonformistički engleski svećenik koji je radio s vjerojatnosima I teorijema odlučivanja vrlo rano tijekom 18 stoljeće.

Neka X bude podatak tj. jedna instance. U Bayesovim izrazima, X se smatra "dokazom". Kao obično, opisano je mjerama napravljenima na setovima od n atributa. Neka H bude neka hipoteza kao naprimjer da instance X pripada nekoj klasi C. Za klasifikacijski problem, želimo odrediti P(H|X), vjerojatno da hipoteza H uz uvjet X (promatrana instance) ili "dokaz". Drugim riječima, gledamo vjerojatnost da instance X pripada klasi X, s obzirom da znamo atributni opis istance X.

P(H|X) je posteriorna vjerojatnost, da je H uvjetovan X. Na primjer, pretpostavimo da naša baza podataka korisnika je ograničena atributima *godine I prihod,* znači x je 35 godišnji korisnik sa prihodom od 40 000 HRK. Pretpostavimo da je H hipoteza koja kaže da će naš korisnik kupiti računalo. Onda P(H|X) reflektira vjerojatnost da korisnik X će kupiti kompjuter uz uvjet da znamo korisnikove godine I prihod.

Za razliku, P(H) je apriorirna vjerojatnost od H. Na primjer, ovo je vjerojatnost da će bilo koji korisnik kupiti kompjuter, bez obzira na godine,prihod I bilo koju drugu informaciju. Posteriorna vjerojatnost, P(H|X) je bazirana na informacijama(u ovom primjer korisničkim informacijama) dok je apriorna vjerojatnost P(H) nezavisna od X.

Slično tome, P(H|X) je posteriorna vjerojatnost X uvjetovana sa H. To je vjerojatnost da će korisnik,X,35 godišnjak koji zarađuje 40 000HRk, uz uvjet da znamo da taj korisnik kupuje računalo.

P(X) je apriori vjerojatnost od X. Koristeći isti primjer, to je vjerojatnost da osoba iz naše baze korisnika stara 35 godina I zarađuje 40 000HRK

"Kako se ove vjerojatnosti procjenjuju?" P(H), P(X|H), I P(X) se može procijeniti iz danih podataka. Bayesov teorem je koristan jer pruža način računanja posteriorne vjerojatnosti P(H|X) iz P(H),P(X|H) I P(X). Kao što je gore napisano Bayesov teorem je:

$$P(H|X)= P(X|H)* P(H)/P(X)$$

U sljedećem odjeljku ćemo vidjeti kako se Bayesov teorem koristi u Naivnom Bayesov algoritmu

Naivni Bayesov klasifikator - primjer

Rad klasifikatora se može pokazati na primjeru scenarija „Dan za građevinske radove". Radi se o problemu određivanja da li je pojedini dan pogodan za rad vani – moguće klasifikacije su „Da" i „Ne".

Svaki dan se prikazuje nizom atributa: „Vrijeme“, „Temperatura“, „Vlažnost“ i „Vjetar“. Primjeri iz skupa za učenje:

Dan	Vrijeme	Temperatura	Vlažnost	Vjetar	Dan za građevinske radove
D1	Sunčano	Vruće	Visoka	Slab	Ne
D2	Sunčano	Vruće	Visoka	Jak	Ne
D3	Oblačno	Vruće	Visoka	Slab	Da
D4	Kišno	Ugodno	Visoka	Slab	Da
D5	Kišno	Hladno	Normalna	Slab	Da
D6	Kišno	Hladno	Normalna	Jak	Ne
D7	Oblačno	Hladno	Normalna	Jak	Da
D8	Sunčano	Ugodno	Visoka	Slab	Ne
D9	Sunčano	Hladno	Normalna	Slab	Da
D10	Kišno	Ugodno	Normalna	Slab	Da
D11	Sunčano	Ugodno	Normalna	Jak	Da
D12	Oblačno	Ugodno	Visoka	Jak	Da
D13	Oblačno	Vruće	Normalna	Slab	Da
D14	Kišno	Ugodno	Visoka	Jak	Ne

Slika 12. Skup podataka

Vrijednosti parametara koji obilježavaju događaje su:

- Vrijeme: Oblačno, Kišno, Sunčano
- Temperatura: Ugodno, Vruće, Hladno
- Vlažnost: Visoka, Normalna
- Vjetar: Jak, Slab

Pretpostavka je da je Bayesov naivni klasifikator naučio skup primjera za učenje i da mu se sada za klasifikaciju predstavlja novi primjer:

(Vrijeme = sunčano. Temperatura = hladno, Vlažnost = visoka. Vjetar = jak)

Zadatak klasifikatora jest predvidjeti klasifikaciju primjera za naučeni ciljni koncept „Dan za građevinske radove“. Jednadžba naivnog Bayesovog klasifikatora za ovaj primjer:

$vNB = argmax\ p(vj) \prod p(ai\ |vj) = argmax\ p(vj)P(vrijeme = sunčano|\ vj{\in}[Da,Ne]\ vj)$ $p(temperatura = hladno|vj)p(vlažnost = visoka|vj)p(vjetar = jak|vj)$ (29)

U drugom je redu varijabla a nadomještena s konkretnim vrijednostima. Za izračun je potrebno odrediti 10 različitih vrijednosti (po 4 vjerojatnosti za navedene vrijednosti atributa za svaku kategoriju, plus apriori vjerojatnosti za svaku od kategorija). Prvo se određuju apriori vjerojatnosti po kategorijama:

$(dan\ za\ radove = Da)$ = 9 14 = 0,64 p

$(dan\ za\ radove = Ne)$ = 5 14 = 0,36

Na sličan način se računaju i ostale potrebne vrijednosti, npr. za vjetar:

$(vjetar = jak|dan\ za\ radove = Da)$ = 3 9 = 0,33

$(vjetar = slab|dan\ za\ radove = Ne)$ = 3 5 = 0,6

Nakon što su izračunate sve potrebne vrijednosti, računaju se vjerojatnosti potrebne za klasifikaciju:

$p(Da)p(sunčano|Da)p(hladno|Da)p(visoka|Da)p(jak|Da)$ = 0,0053
$p(Ne)p(sunčano|Ne)p(hladno|Ne)p(visoka|Ne)p(jak|Ne)$ = 0,0206

Na temelju dobivenih vjerojatnosti, Bayesov naivni klasifikator daje odgovor „Dan za građevinske radove“ = „Ne“. Kao i u primjeru za Bayesov teorem, kategorije u ovom primjeru su neovisne i međusobno isključive, pa se njihove aposteriori vjerojatnosti mogu normalizirati tako da njihov zbroj bude jednak jedan:

$p(Da|sunčano,\ hladno,\ visoka,jak)$ = 0,0053 0,0053 + 0,0206 = 0,2 $p(Ne|sunčano,\ hladno,\ visoka,jak)$ = 0,0206 0,0053 + 0,0206 = 0,8

5.SKUPOVI PODATAKA

Predstavljena su 3 skupa podataka. Skupovi podataka su iz 3 različita područja preuzeta su sa UCI Machine Learning repozitorija. Skupovi se razlikuju po svojim karakteristikama, a karakterizirani su s : brojem atributa,brojem instance i brojem klasa.

5.1 'Bank marketing' skup podataka

Bank marketing skup podataka je javan i dostupan za istraživanja (moro et all.,2011). Skup podataka je preuzet sa UCI repizotorij. Sastoji se od 17 varijabli i 45211 instanci. 17 varijabli je podijeljeno na dva dijela 16 nezavisnih varijabli I jedno zavisnu (ciljna varijabla). Cij klasifikacije je predvidjeti hoće li klijent potpisati ugovor. Ove varijable prema broju stupaca u bazi podataka izgledaju :

1. age (brojčano)
2. job : vrsta posla (kategorija: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3. marital: bračni status (kategorija : "married","divorced","single"; note: "divorced" means divorced or widowed)
4. aducation (kategorija: "unknown","secondary","primary","tertiary")
5. default : ima kredit po defultu (binarno:" yes","no")
6. balance : prosječna godišnja bilansa,u eurima (numerički)
7. housing osoba ima zajam za nekretnine? (binarno;"da","ne")
8. loan: osobni zajam? (binarno"da","ne")
9. contact: način uspostave kontakta (kategorija: "unknown","telephone","cellular")
10. day: posljedni kontakt,dan u mjesecu(brojčano)
11. month: zadnji kontakt,mjesec u godini(kategorija: "jan", "feb", "mar", ..., "nov", "dec")
12. duration: vrijemecap zadnjeg kontakta, u sekundama (brojčano)
13. campaign: broj kontakata sa klijentom tijekom ove kampanje(brojčano,uključuje posljedni kontakt)
14. pdays(broj dana koje je prošlo otkad je klijent kontaktiran u prošloj kampanji(brojčano, -1 znači da klijent nije prethodno kontaktiran)
15. previous: broj kontakata prije ove kompanje za tog klijenta(brojčano=
16. poutcome: ishod prethodne marketinške kampanje(kategorija: "unknown","other","failure","success")
17. y: je li se klijent pretplatio za deposit? (binarno:"da","ne")

Proporcija klasa: “no”39922 “yes”5289

16 atributa(7 numeric + 9 nominalnih) + jedna ciljna (nominalna)

	Broj atributa	Broj instanci	Broj klasa
Skup podataka ‘bank’	17	45211	2

Tablica 4. Skup podataka ‘bank’

5.2 ‘Squash-stored’ skup podataka

Squash Harvest stored je javan I dostupan skup podataka za istraživanje,preuzet sa tunedit repizotorija koji sadrži niz skupova za eksperimentiranje. Cilj istraživanja je bio utvrditi promjene koje se događaju zgnječenom voća tijekom zrenja I dozrijevanjakako bi se odredilo najbolje vrijeme koje daje najbolju kvalitetu na tržištu. Originalan cilj je bilo utvrditi varijable utjecu na kvalitetu zgnječenog voća poslije različitih vremenskih perioda. Ovo je određeno varijablom prihvatljivost koja je ujedno I ciljna varijabla koju predviđamo s klasama “neprihvatljivo”,”prihvatljivo” I “odličan”. Skup podataka se sastoji od 52 instance I 25 atributa od kojih su 4 nominalne I 21 numerička varijabla.

Atributi:

1. site - lokacija voća
2. daf - broj dana nakon cvjetanja
3. fruit - individualni broj voća
4. weight - težina u gramima
5. storewt - težina voća nakon skladištenja
6. pene - penetrometer pokazuje zrelost voća nakon žetve
7. solids_% - test suhoće
8. brix - a refektometer se koristi za mjerenje slatkoće voća

9. a* - notacija za mjeru boja

10. egdd - akumulacija grijanja poviše 8c od pojave biljke do žetve

11. fgdd - akumulacija grijanja povise 8c od cvjetanja do žetve

12. groundspot_a* - broj koji pokazuje boju kože na kojoj je bilo voće

13. glucose - mjereno z mg/100g

14. fructose - mjereno u mg/100g

15. sucrose - mjereno u mg/100g

16. total - mjereno u mg/100g

17. glucose+fructos- mjereno mg/100g

18. starch - mjereno u mg/100g

19. sweetness - srednja vrijednost od 8 okusa od 1500

20. flavour srednja vrijednost od 8 okusa od 1500

21. dry/moist - srednja vrijednost od 8 okusa od 1500

22. fibre srednja vrijednost od 8 okusa od 1500

23. heat_input_emerg - količina grijanja tijekom rasta

24. heat_input_flower - količina grijanja prije cvjetanja - real

Klasa:

25. Acceptability - prihvatljivost voća

Distribucija klasa:

excellent - 23

ok - 21

not_acceptable - 8

	Broj atributa	Broj instanci	Broj klasa
Skup podataka ‘bank’	25	52	3

Tablica 5. Skup podataka ‘squash’

5.3 ‘Nursery’ skup podataka

Nursery skup podataka je javan i dostupan za istraživanje te preuzet sa UCI repizotorija.Cilj klasifikacije je rankiranje aplikacija za dječje vrtiće. Baza podataka je izgrađena osamdesetih godina prošlog stoljeća kako bi se preciznije objasnile odbijene prijave. Odluka se procjenjivala na temelju tri pod problema zanimanju roditelja,obiteljskoj strukturi i financijama te su za kvalitetnu procjenu prikupljene navedene varijable:

1. parents-zanimanje roditelja
2. has_nurse- ide u dječji vrtić
3. form- struktura obitelji
4. children-broj djece
5. housing-uvjeti stanovanja
6. finance-financijsko stanje obitelji
7. social- socijalno stanje obitelji
8. health-zdrastveno stanje obitelji
9. class- evaluacija aplikacije za medicinsku školu

Distribucija klasa

not_recom 4320 (33.333 %)
recommend 2 (0.015 %)
very_recom 328 (2.531 %)
priority 4266 (32.917 %)
spec_prior 4044 (31.204 %)

9 variabli, 8 Nominalnih i 1 klasna nominalna varijabla

	Broj atributa	Broj instanci	Broj klasa
Skup podataka ‘squash’	9	12960	5

Tablica 6. Skup podataka ‘nursery’

6.KOMPARATIVNA ANALIZA I REZULTATI

U poglavlju 5, različiti skupovi podataka su predstavljeni i analizirani kako bi se dobio početni uvid u njihovu prirodu. Na svakom skupu podataka je iskorišteno više tehnika rudarenja podataka kako bi se izgradili modeli kojima će se usporediti njihove prediktivne sposobnosti. Rezultati preciznosti predikcija ciljnih varijabla se uspoređuju i razlika će se izmjeriti komparativnim pristupom.Odabir najprihvatljivije klasifikacijeske metode za određeni zadatak je problem i ne postoji jednostavan odgovor. Za izgradnju modela i generiranje metrike krištena je unakrsnu validaciju. Podaci su testirani na način da je skup podataka podijeljen na 10 izdvojenih skupova od kojih je 9 bilo za treniranje i 1 za testiranje modela.

Dobiveni rezultati na svakom pojedinom skupu podataka su uspoređeni prema sposobnosti predviđanja klasne varijable. Metrika kao što je točnost,preciznost,osjetljivost,odziv, F-mjere,ROC područje i matricia grešaka pomaže dobiti uvid u moć predikcije određenog algoritma na skupu podataka. Komparativnim pristupom ćemo odabrati najbolji algoritam za pojedine skupove podataka

Prethodna empirijska istraživanja su pokazala da izbor optimalnog klasifikatora ovisi o korištenim podacima (Michie et. al, 1994.). Van der Walt ispituje koje karakteristike podataka utječu na učinkovitost klasifikacije i razvija mjere za mjerenje tih karakteristika podataka. Ove mjere omogućuju definiciju odnosa između karakteristika podataka i učinkovitosti klasifikatora.

Utjecaj karakteristika podataka na učinkovitost klasifikacije je ispitivao Van der Walt u prijašnjim istraživanjima (michie et.al,1994). Razvio je mjere koje su omogućavale mjerenje tih karakterisitika podataka. Na taj način je pokušao definirati odnos između učinkovitosti klasifikatora I karakteristika podataka. Mjere su grupirane u sljedeće kategorije: standardne mjere,mjere oskudnosti podataka,statističke mjere,mjere teorije informacija,mjeregranica odluka,topološke mjere I mjere šuma.

Karakteristike	Mjera
Standardne mjere	
Dimenzionalnost	d
Broj instance	N
Broj klasa	C

Tablica 7. Standardne mjere

U ovom poglavlju se istražuje odnos između standardnih mjera karakteristika skupa podataka i učinka klasifikatora. 3različita skupa podataka su predstavljena i utjecaj klasifikatora na predviđanje njihovih klasa. Skupovi podataka sadrže različiti broj atributa ,klasa i instanci. U provedenom eksperimentu se može vidjeti da veličina uzorka ima veliki utjecaj na učinkovitost klasifikatora kao i broj klasa. Također i broj atributa po klasi utječe na učinkovitost klasifikacije,jer određuje količinu informacija dostupne za treniranje modela.

U provedenom eksperimentu možemo vidjeti da 3 skupa podataka sa različitim veličinom uzoraka,broja klasa i broja instance reagiraju drukčije te da osim točnosti algoritama imaju raznoliku metriku. Kako bi se testirale hipoteze rada provedeno je istraživanje napravljene su usporedbe na 3 javno dostupna skupa podataka. Za svaki skup analizirane su standardne mjere karakteristike skupa. Izabrani skup podataka se testira koliko je točan uporabom 4 klasifikatora. Svaki klasifikator se testira na testnom skupu procesom unakrsne validacije kako bi se evaluirale perfomanse klasifikacijskih algoritama. Točnost klasifikatora na testnim podacima ćese međusobno uspoređivati

U prvom koraku,provjeravaju se standardne karakteristike skupa podataka .Izmjerene su sve vrijednosti za svaki skup. Svaka vrijednost je klasificirana u jednu od dvije kategorija.

Kategorije za svaku od karakteristika skupa podataka su slijedeće:

broj atributa: mikro,mali, veliki,

broj instanci: mikro,mali, veliki,

broj klasa: binomna,multiklasna

	Broj atributa	Broj instanci	Broj klasa	Opis
Dataset bank	17	45211	2	mali,veliki,binomna
Dataset squash	25	52	3	Mali,mali,multiklasna
Dataset nurses	9	12960	5	Mali,veliki,multiklasna

Tablica 8. Opis skupova

.

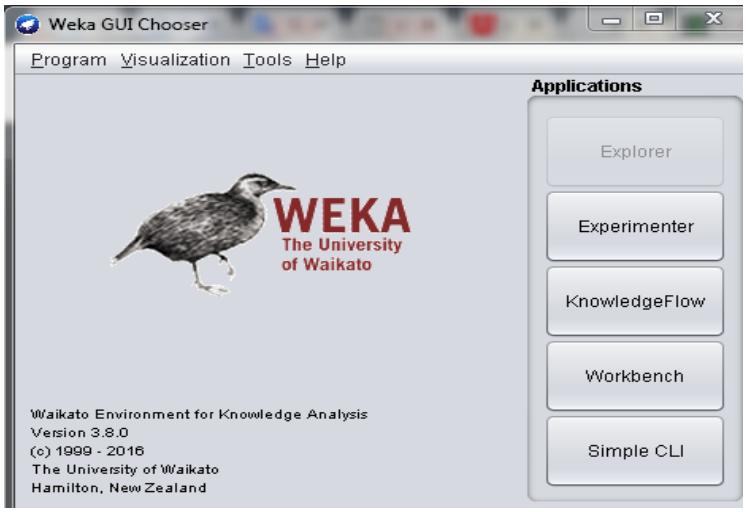
6.1 WEKA-alat za rudarenje podataka

U ovom istraživanju korišten je popularni softwer za strojno učenje i rudarenje podataka napisan u Javi, razvijen na sveučilištu Waikato,Novi Zeland. Weka je „open-source“ alat besplatan i dostupan svima za korištenje. Weka podržava više faza rudarenja podataka kao što je pretprocesiranje podataka,klusterizacija,klasifikacija,regresija,vizualizacija i odabir atributa. Sadrži brojne algoritme strojnog učenja ,npr. J48,naivni Bayes, Multilayer Percepon i druge te uključuje cijeli proces rudarenja podataka od pretprocesiranja do vizualizacije podataka.

Weka koristi vlastiti formaliziran prikaz ulaznih podataka poznat kao .arff format. Sastoji se od zaglavlja u kojima su opisani atributi te su podaci odijeljeni zarezom. Atributi u alatu Weka mogu biti numerički ili norminalni. Pod numeričke se podrazumijeva kontiuniran atributi koji mogu biti realne ili cijelobrojne vrijednosti. Nominalni atributi sadrže određeni skup vrijednosti.

Okvir koji je korišten u ovom istraživanju od pozivanja podataka do evaluacije je opisan u nastavku rada. Nakon toga će prikazani biti stvarni rezultati iz Weka alata za svaki od algoritama koji će se uspoređivati te objašnjen proces koji je korišten kako bi se došlo do tih rezultata

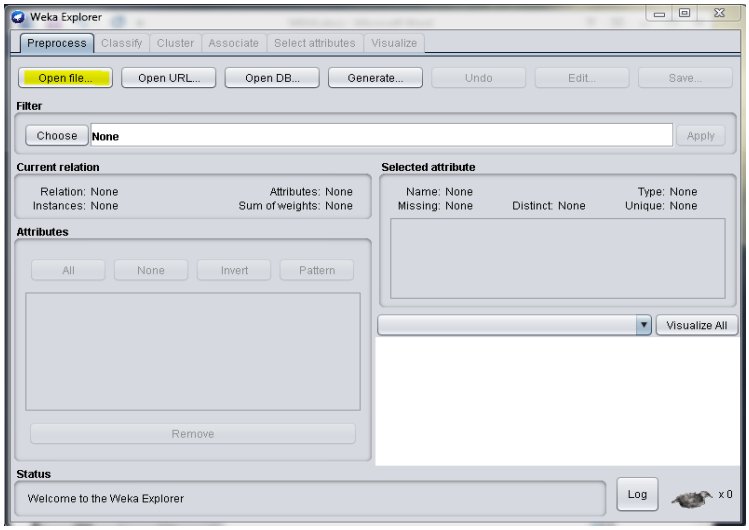
U početnom prozoru sustava (slika) možemo izabrati 5 načina na koji ćemo koristiti programski paket Weka:



Slika 13. Početni prozor Weke

1. **Simple CLI** pruča jednostavno sučelje koje omogućava izravno izvršenje Weka naredbi
2. **Workbench**
3. **Explorer** je okolina za istraživanje podataka
4. **Experimenter** je okolina za izvršavanje eksperimenata i vođenje statističkih testova među modelima
5. **KnowledgeFlow** je „Java-Beans-based“ sučelje za postavljanje i pokretanje eksperimenata strojnog učenja

U ovom radu će se koristiti opcija *Explorer* za analizu podataka i izgradnju prediktivnog modela koji može generirati evaluacijsku metriku koju možemo uspoređivati među modelima. Klikom na Explorer otvara se prozor *Weka Explorer* i prvi tab *Pretprocess*



Slika 14. Prozor 'preprocess' u Weki

Klikom na *Open File* pozivamo podatke na kojima će se raditi analiza u weka sustavu. To je prvi korak koji nam omogućava daljne analiziranje i istraživanje podataka. Kad su podaci pozvani ,otvara nam se niz funkcija koji se mogu koristiti za razvijanje modela koji će imati najbolje prediktivne sposobnosti. Podacise mogu unijeti sa datoteka sa različitim formatima : ARFF,CSV,C4.5,binarni, mogu se također učitati sa URL ili SQL baze podataka(koristeći JDBC). Način koji je korišten u ovom radu je učitavanja podataka u WEKU formatom „Attribute-Relation File Format“ (ARFF) .

Kad je sustav napunjen podacima,WEKA prepoznaju atribute koji su prikazani u prozoru „Attributes“ te nam pokazuje listu atributa.

No. je broj koji identificira red atributa onako kako je u datotec

Polja nam dopuštaju da klikom izaberemo atribute s kojima će se raditi

Name je ime atributa koje ima u izvornoj datoteci

„Current relation“ dio poviše prozora „Attributes“ prikazuje ime relacijske tablice,broj atributa ,broj instanci i sumu pondera.

Tijekom skeniranja podataka, WEKA izračunava osnovne statističke mjere za svaki atribut. Statistika je prikazana na desnoj strani panela 'Preprocess' u kutiji pod imenom 'Selected Attributes':

'Name' je ime atributa

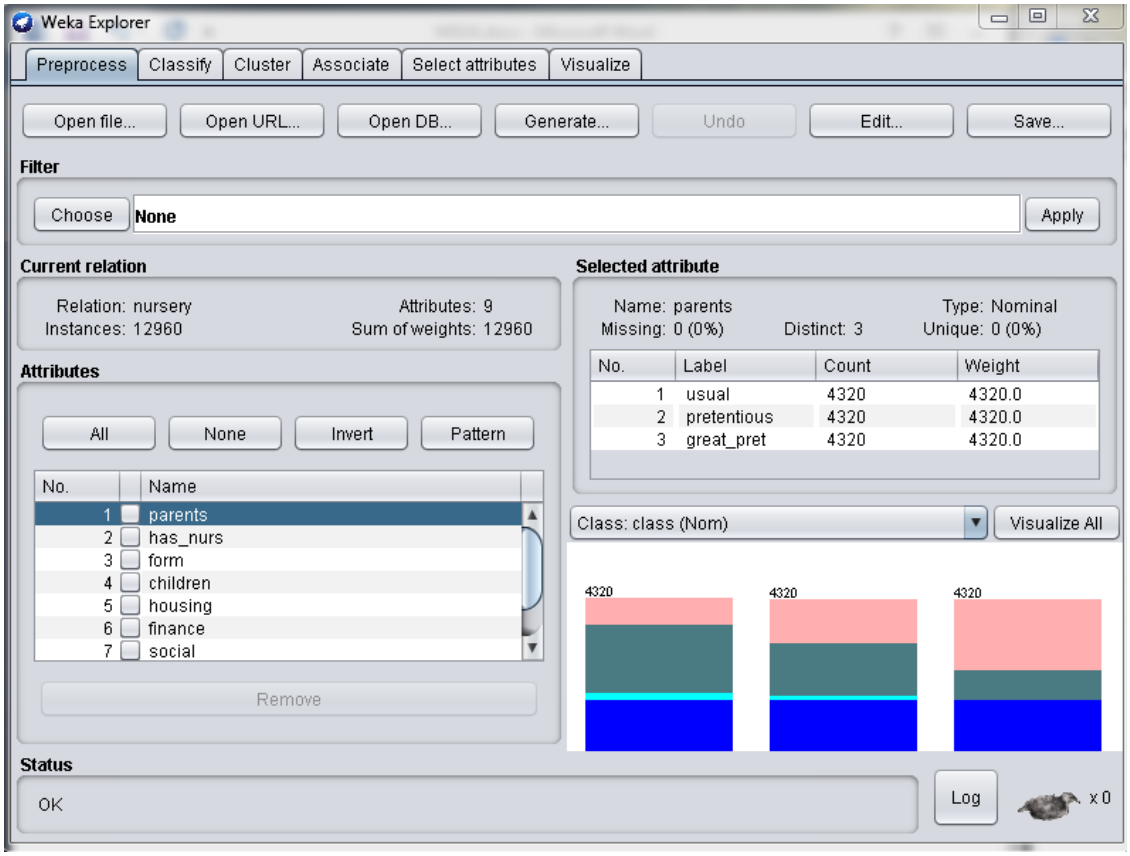
'Type' je tip atributa obično Nominalni ili Numerički

'Missing' je broj (postotak) instanci koji nisu specifirani u podacima

'Distinct' je broj različitih vrijednosti koji se nalaze u podacima za određeni atribut

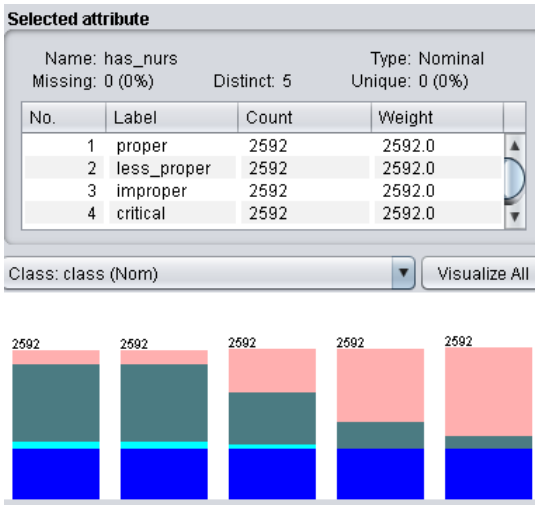
'Unique' je broj(postotak) instanci u podacima koji ima vrijednost koja se pojavljuje samo jednom u podacima

Atribut se može izbrisati iz prozora 'Attributes'. Klikom na atribut se može vidijeti osnovna statistika toga atributa.



Slika 15. Prozor 'preprocess' u Weki

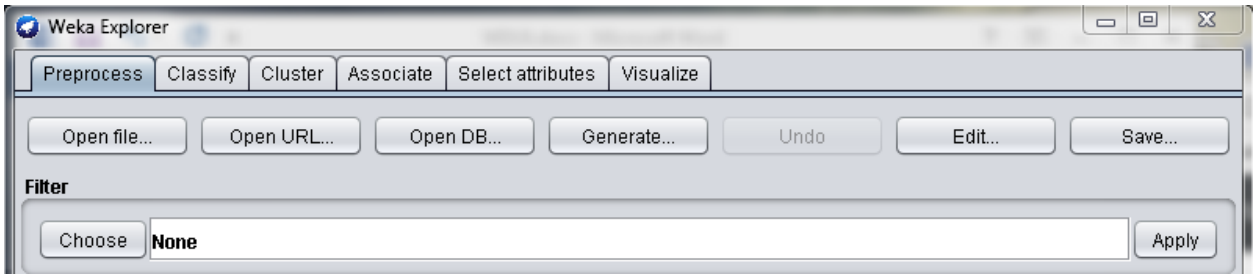
Posljedni atribut u prozoru 'Attributes' je zadani ciljni atribut koji se može promijeniti u prozoru 'Class(slika)'. Atribute možemo i vizualirati s obzirom na klasu klikom na taj atribut ili pod tipkom 'Visualize All'



Slika 16. dio'Selected attributes' u prozoru 'preprocess'

Postavljanje filtera u WEKI

Alati za predprocesiranje u WEKI zovu se 'filteri'. WEKA sadrži filtere za diskretizaciju,normalizaciju,uzorkovanje,selekciju atributa,transforomaciju i filteri za kombiniranje atributa. U ovom radu korišten je filter selektiranje atribura koji je pobliže objašnjen u sljedećim dijelovima.Klik na botun 'Choose' otvara se padajući menu gdje su prikazani mogući filteri raspoređeni u dvije podgrupe 'Supervised ' i 'Unsupervised'. Odabir filtera ovisi o više varijabli ali cilj je uvijek isti a to je poboljšanje modela. Tipkom 'Apply' procesuiramo filter i ishod se može odmah vidjeti u prozoru 'Attributes'



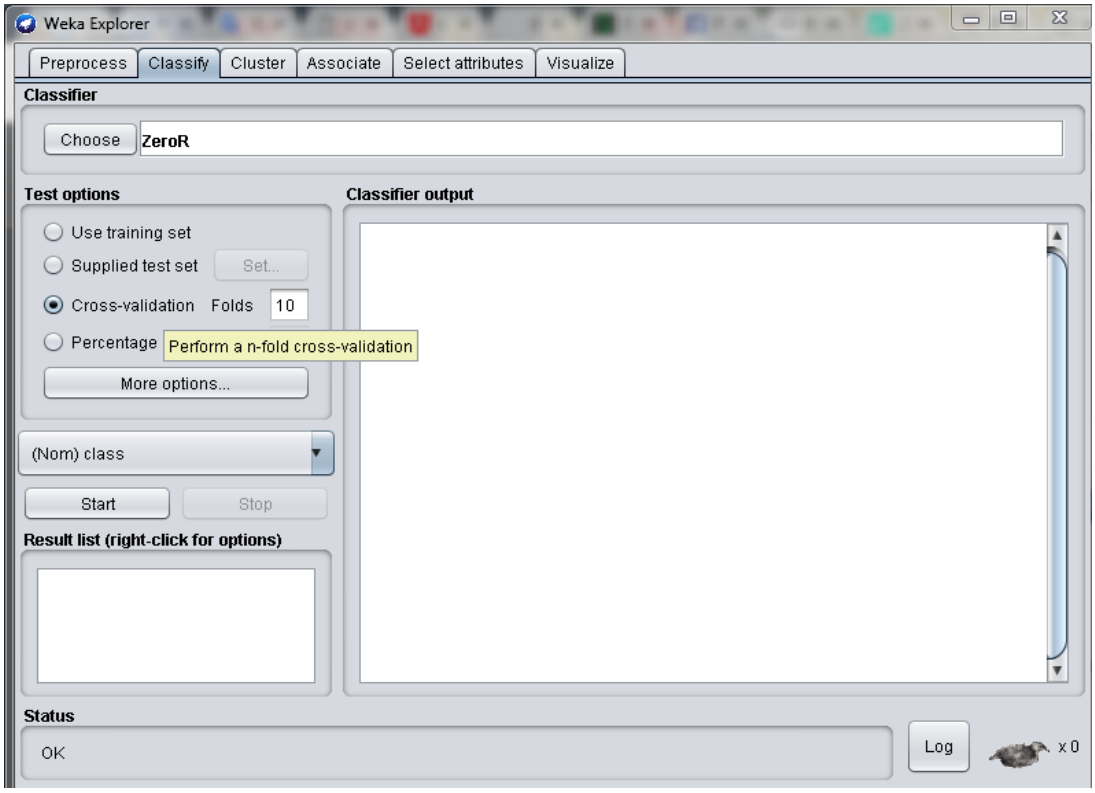
Slika 17. dio'Filter'u prozoru 'preprocess'

Klasifikatori u WEKI su modeli za predviđanje nominalnih ili numeričkih vrijednosti. Dostupan je veliki broj algoritama u WEKI koji uključuju i 4 metode i algoritma koji su prestavljeni u ovom radu : naivni Bayes,stablo odlučivanja,neuronske mreže i logistička regresija. Nakon što smo napunili sustav podacima i napravili pretprocesiranje sukladno ciljevima analize dolazi se do sljedeće faze izgradnje modela.Klikom na tipku 'Classify' otvara se prozor u kojom ćemo postaviti parametre za izgradnju klasifikatora

Klikom na tipku 'Choose' otvara se padajući izbornik s nizom algoritama za predviđanje. Podgrupe su podijeljene po karakteristikama algoritama tj. Po metodama. Prije nego što pokrenemo klasifikacijski algoritam postavljaju se test opcije. Opcije za testiranje su dostupne u dijelu 'Test Options' a to su :

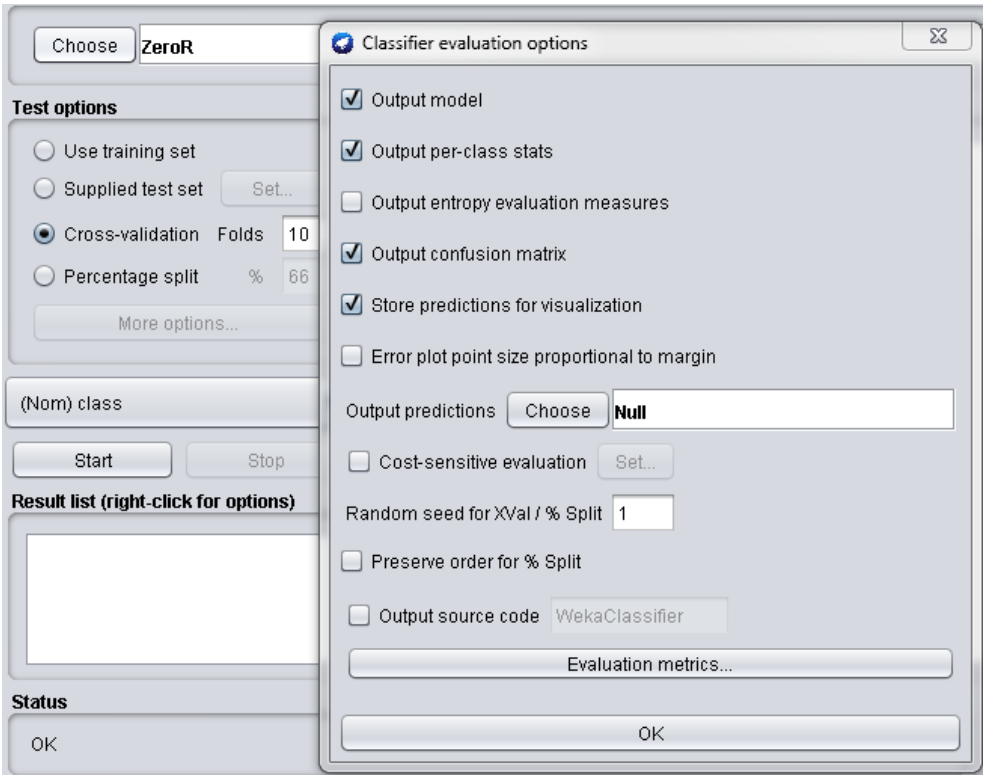
1. **'Use training set'**. Evaulira klasifikator s obzirom na instance koje su korištenje za treniranje podataka
2. **'Supplied test set'**. Evaulira klasifikator s obzirom na instance koje su dodatno dodane za testiranje

3. **'Cross-validation**'. Evaulira klasifikator unakrsnom validacijom koristeći podskupove ,njihov broj je moguće odrediti prije procesiranja. U ovom radu korišten je ovaj pristup testiranja podataka za sve modele
4. **'Percantage split'.** Evaulira klasifikator na način da određeni postotak skupa podataka koristi za treniranje a jedan za testiranje. Postotke je moguće manualno odrediti.



Slika 18. prozor 'Classify'

Opcija 'More Options' nudi mogućnost odabira ishoda koje će generirati model. U ovom radu označeni ishodi modela su vidljivi na slici ispod.



Slika 19. Evaluacijske opcije

1.'Output model'. Ishod je klasifikacijski model na punom trenig setu dostupan za promatranje ,vizualizacije i dr.

2.'Output per-class stats'. Preciznost/Odziv i stvarna/lažna statistika za svaku klasu ishoda

3.'Output confusion matrix'. U ishod je uključene matrica grešaka

4.'Store prediction for visualization'. Predviđanje klasifikatora su zapamćena kako bi se mogla vizualizirati

5. Postavi '**Random seed for Xval / % Split'** na 1. Ovo znači da se podaci slučajno dijele prije evaluacijskog procesa

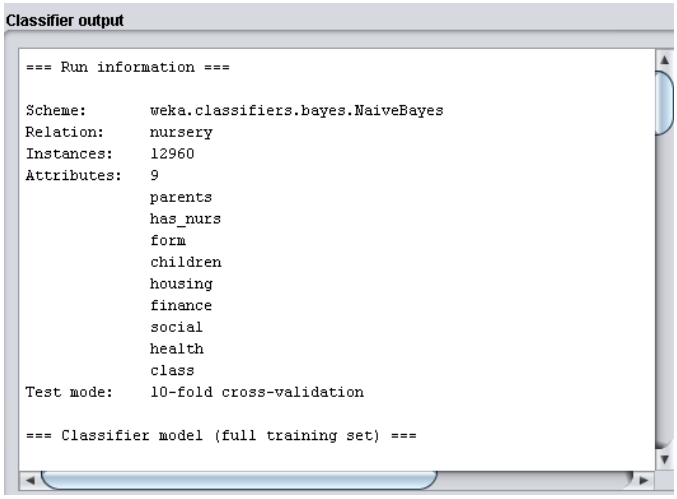
Nakon što smo specifirali opcije ,algoritam se može pokrenuti klikom na tipku 'Start'. Desni dio panela je područje na kojom promatramo rezultate trenig i test skupa.

Primjer pogleda na rezultate(slika).

Informacije pri pokretanju:

- Korišteni algoritam
- Ime relacije
- Broj instanci

- Broj atributa u relaciji
- Opcija testiranja



Slika 20. Rezultat

U ovom dijelu reulzata je dio gdje su se podaci trenirali na cijelom trenig skupu podataka. Ovdje se može vidjeti struktura algoritama(broj čvorova,veličinu stabla itd. u primjeru stabla odlučivanja)

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                      Class
Attribute          not_recom  recommend very_recom   priority spec_prior
                      (0.33)         (0)     (0.03)     (0.33)     (0.31)
============================================================================
parents
  usual               1441.0         3.0      197.0     1925.0      759.0
  pretentious         1441.0         1.0      133.0     1485.0     1265.0
  great_pret          1441.0         1.0        1.0      859.0     2023.0
  [total]             4323.0         5.0      331.0     4269.0     4047.0

has_nurs
  proper               865.0         3.0      131.0     1345.0      253.0
  less_proper          865.0         1.0      133.0     1345.0      253.0
  improper             865.0         1.0       67.0      905.0      759.0
  critical             865.0         1.0        1.0      465.0     1265.0
  very_crit            865.0         1.0        1.0      211.0     1519.0
  [total]             4325.0         7.0      333.0     4271.0     4049.0

form
  complete            1081.0         3.0      119.0     1153.0      889.0
  completed           1081.0         1.0      101.0     1093.0      969.0
  incomplete          1081.0         1.0       71.0     1039.0     1053.0
  foster              1081.0         1.0       41.0      985.0     1137.0
```

Slika 21. Rezultat

Evaluacija se radi na testnom skupu . Ovaj dio rezultata daje procjenu perfomanse algoritma. Ispisana je lista metrika koje sumiraju točnost klasifikatora. U ovom radu će se koristiti metrika kao što je točnost,preciznos,senzitivnost i odziv. Uz navedene možemo isčitati i niz statističkih mjera grešaka klasifikatora te matricu greške

6.2 Kriteriji korišteni u ovoj komparativnoj analizi

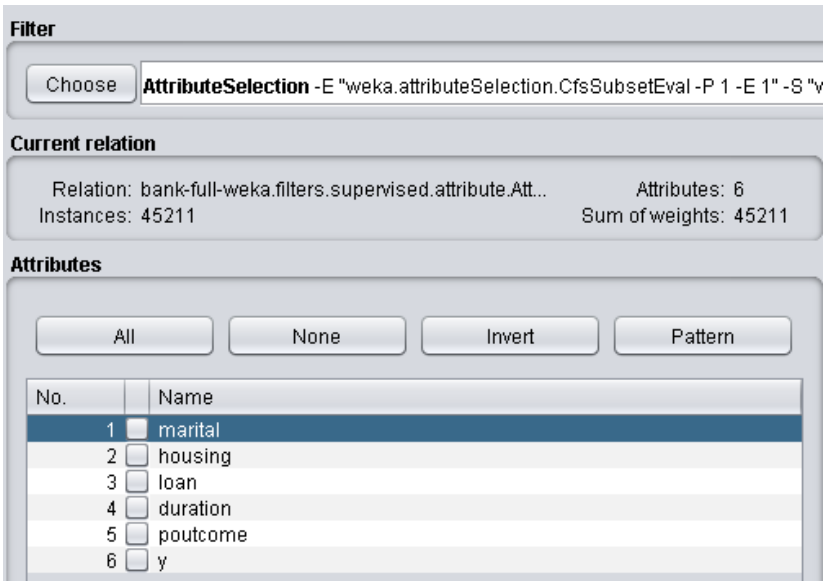
U ovom djelu kriterija je predstavljeno koji su se iskoristili za komparaciju različitih tehnika rudarenja podataka.

1. Točnost(eng. accuracy)- udio točno klasificiranih primjera na datom test skupu
2. Preciznost (eng. precision)-omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa
3. Odziv/osjetljivost (eng. recall) -omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa
4. Specifičnost- razmjer negativnih primjera koji su točno identificirani

Generiranje navedene metrike se izvodi metodom unakrsne validacije. Postavljeno je 10 iteracija na kojema će se evaulirati model svaka iteracija nasumično dijeli skup podataka na trenig i test podskup. Sposobnost da se predvidi testni ishod nam kaže koliko je dobar model i to je mjera perfomanse modela. Rezultati predviđanja trening skupa podataka nisu prezentirani u ovoj analizi jer perfomanse na testnom skupu podataka su važnije za ovo istraživanje

6.3 Analiza 'Bank' skupa podataka

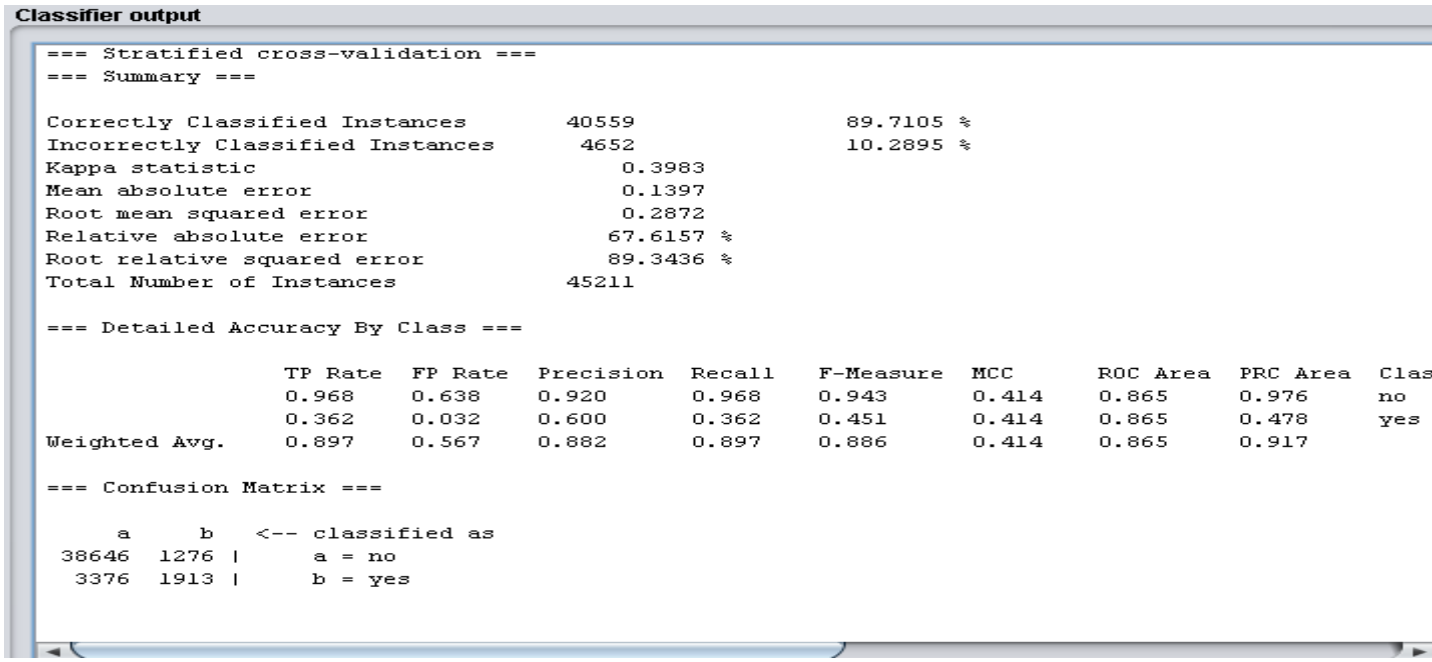
'Bank' skup podataka ima 16 prediktora i jednu ciljnu,zavisnu varijablu. 4 tehnike rudarenja podataka su primjenje za predviđanje zavisne varijable. Za poboljšljavanje točnosti modela kako bi se povećala prediktivna moć modela nisu korišteni svi atributi.Selekcijom atributa je smanjen njihob broj na njih pet prediktivnih atributa i jednu klasnu koja se predviđa. Evaluator atributa koji je korišten je CFSSubsetEva kojie valulira podskup atributa koji najsnažnije koreliraju sa klasnom varijablom a međusobno imaju nisku korelaciju . Na sljedećim slikama će biti prikazani rezultati



Slika 22. Predprocesiranje 'bank' skupa podataka

Izgradnja modela na 'Bank' skupo podataka metodom naivnog Bayesa

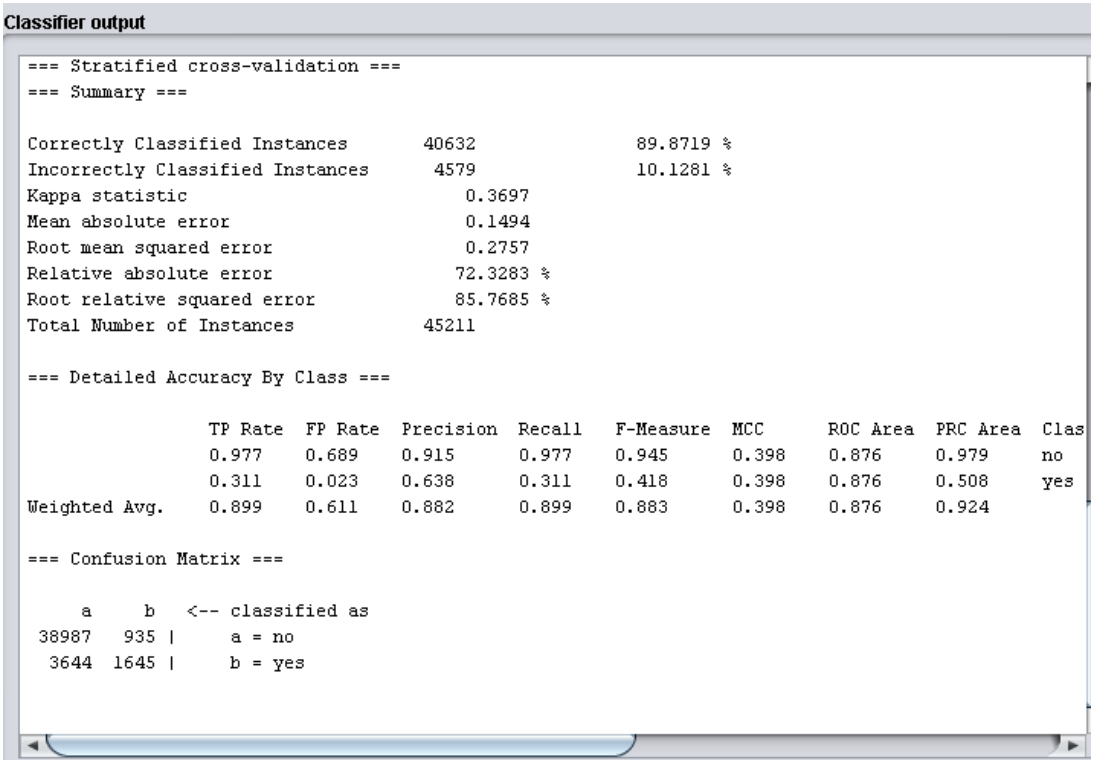
Naivni Bayes ima točnost od 89.7105% točno klasificiranih primjera tj. primjera kojima je očno pogođena klasa ima 40 559 dok je netočnih 4652 (10,2895%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.882 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.897 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,567 (specifičnost).



Slika 23. Rezultati klasifikacije 'bank' skupa podataka metodom naivnog bayesa

Izgradnja modela na 'Bank' skupo podataka metodom logističke regresije

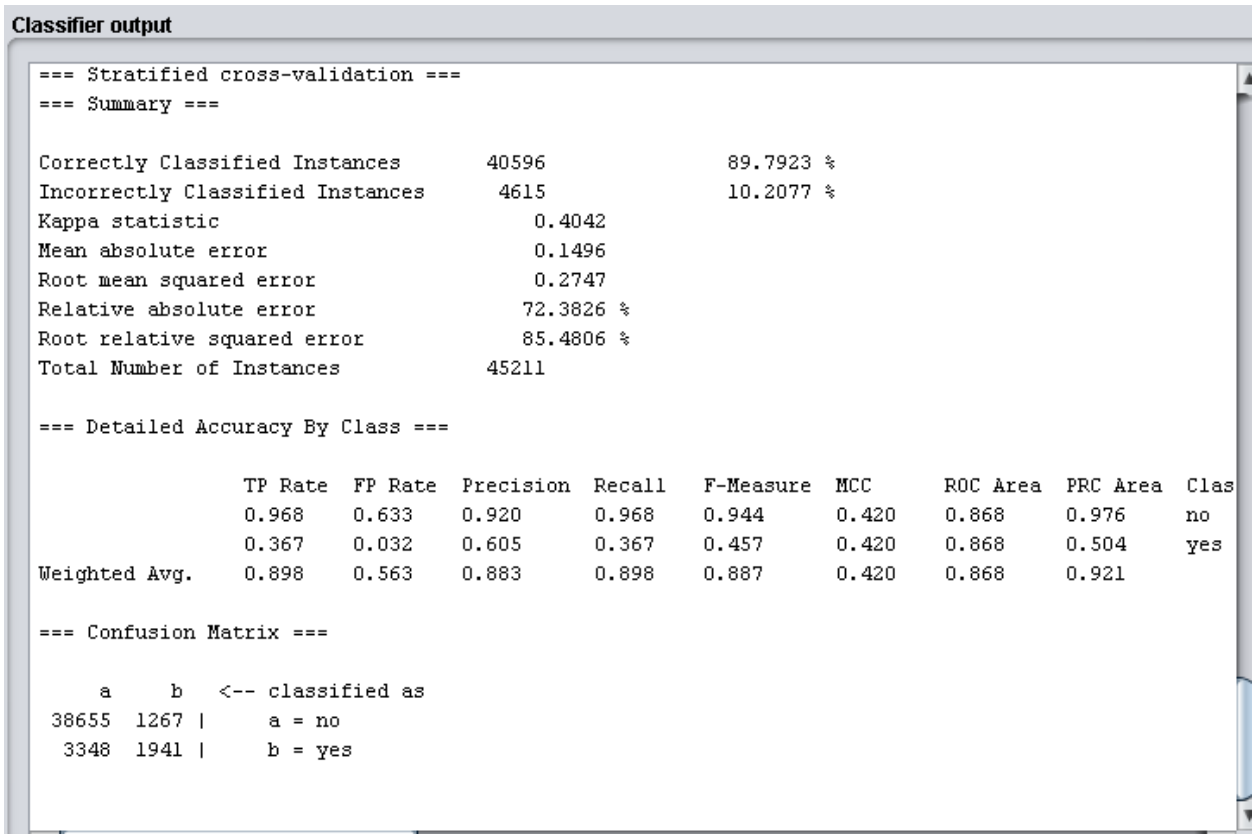
Logistička regresija ima točnost od 89.8719% točno klasificiranih primjera tj. primjera kojima je očno pogođena klasa ima 40 632 dok je netočnih 4579 (10,1281%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.882 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.899 (odziv). Razmjer negativnih primjera koji su točno identificirani je 0,611 (specifičnost).



Slika 24. Rezultati klasifikacije 'bank' skupa podataka metodom logističke regresije

Izgradnja modela na 'Bank' skupo podataka metodom neuronskih mreža

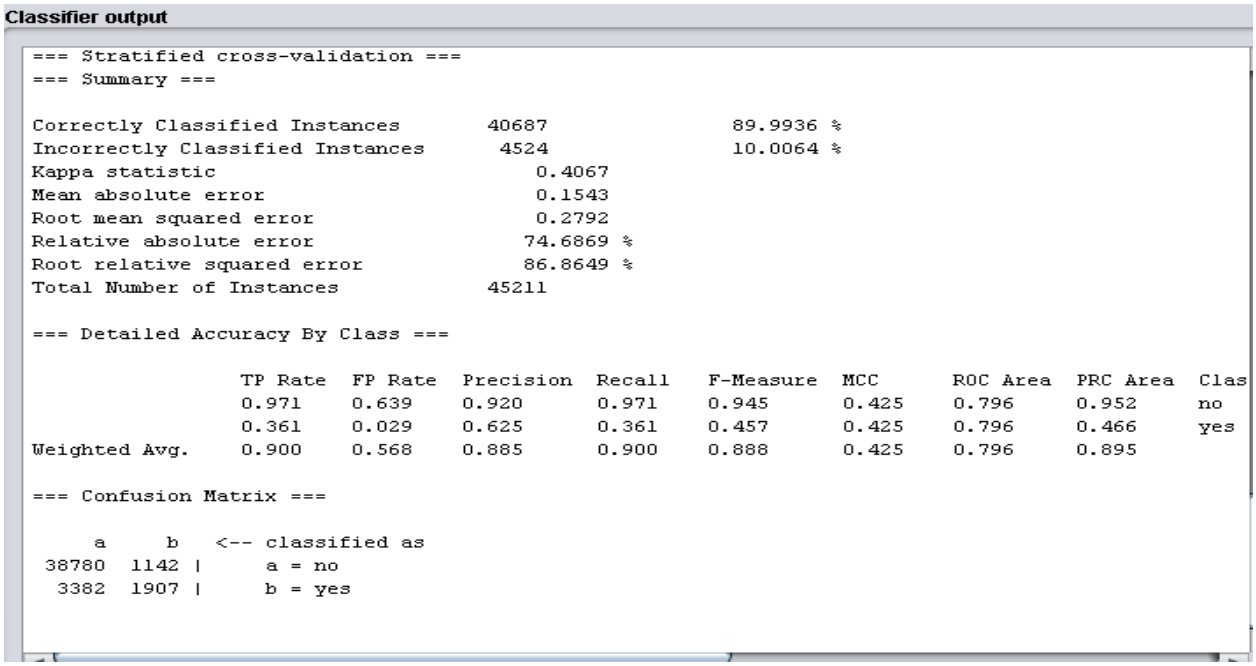
Neuronske mreže imaju točnost od 89.7923% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 40 596 dok je netočnih 4615 (10,2077%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.883 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.898 (odziv)..Razmjer negativnih primjera koji su točno identificirani je 0,563 (specifičnost).



Slika 25. Rezultati klasifikacije 'bank' skupa podataka metodom neuronskih mreža

Izgradnja modela na 'Bank' skupo podataka metodom stabla odlučivanja

Stablo odlučivanja ima točnost od 89.99365% točno klasificiranih primjera tj. primjera kojima je očno pogođena klasa ima 40 687 dok je netočnih 4524 (10,2895%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.885 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.900 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,568 (specifičnost).



Slika 26. Rezultati klasifikacije 'bank' skupa podataka metodom stabla odlučivanja

Odabir najboljeg algoritma na 'Bank' setu podataka

Pogledom na rezultate može se uočiti da su ishodi algoritama imaju vrlo visoku generalnu točnost. Velik broj primjera je sigurno pridonio tome kao i dobra povezanost prediktora i prediktanta. Pomnijim pogledom na rezultate može se primjetiti da je dosta velik broj krivo klasificiranih primjera pozitivne klase 'yes' što vidimo metrikom specifičnosti (*FP Rate)* koja je vrlo visoka na svim ishodima stoga s obzirom na izjednačenost generalne točnosti algoritama .

Metoda	Točnost	Specifičnost	Preciznost	Odziv
Naivni Bayes	**89.7105%**	**0,567**	**0.882**	**0.897**
Logistička regresija	**89.8719%**	**0,611**	**0.882**	**0.899**
Neuronske mreže	**89.7923%**	**0,563**	**0.883**	**0.898**
Stablo odlučivanja	**89,9936%**	**0,568**	**0,885**	**0,900**

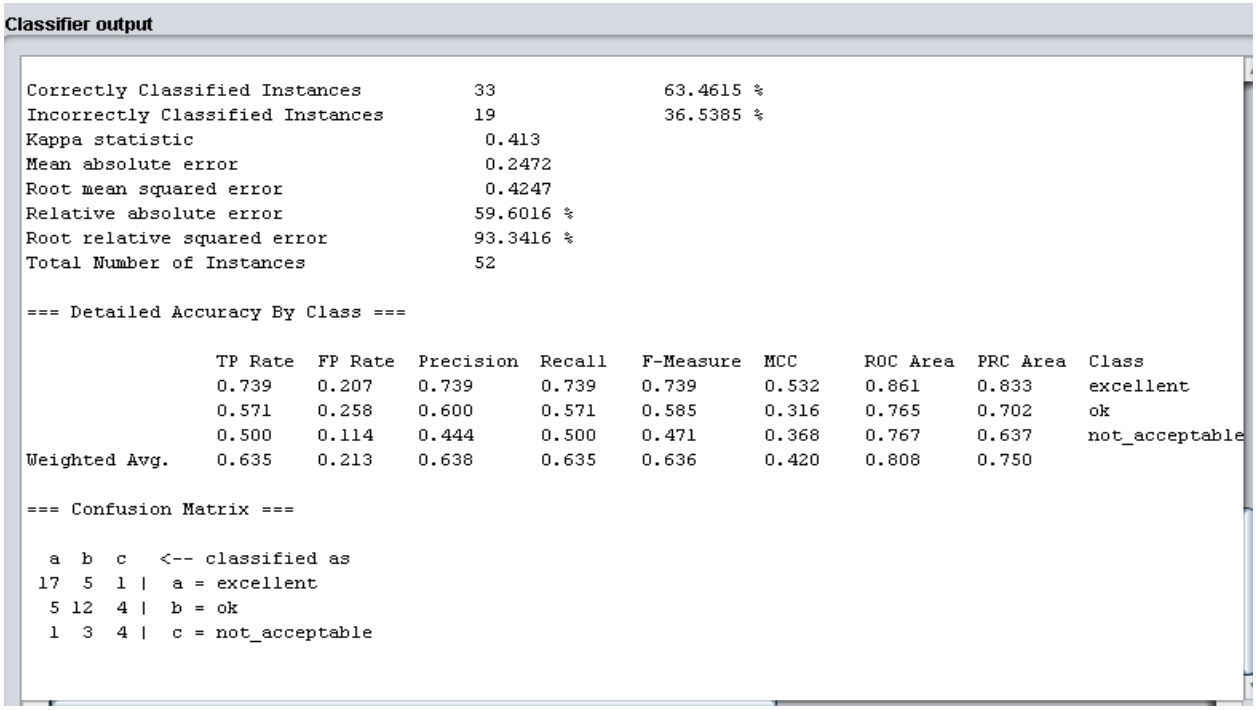
Tablica 9. Komparativna analiza algoritama na 'bank' skupu podataka

6.4.Analiza 'Squash-stored' skupa podataka

'Squash-stored' skup podataka ima 24 prediktora i jednu ciljnu,zavisnu varijablu. 4 tehnike rudarenja podataka su primjenje za predviđanje zavisne varijable. Za poboljšljavanje točnosti modela kako bi se povećala prediktivna moć modela nisu korišteni svi atributi.Selekcijom atributa je smanjen njihob broj na njih 6 prediktivnih atributa(pene,a*,fgdd,total,sweetness,heat_input_flower) i jednu klasnu varijablu(Acceptability) koja se predviđa. Evaluator atributa koji je korišten je CFSSubsetEva kojie valulira podskup atributa koji najsnažnije koreliraju sa klasnom varijablom a međusobno imaju nisku korelaciju . Na sljedećim slikama će biti prikazani rezultati

Izgradnja modela na 'Squash-stored' skupo podataka metodom naivnog Bayesa

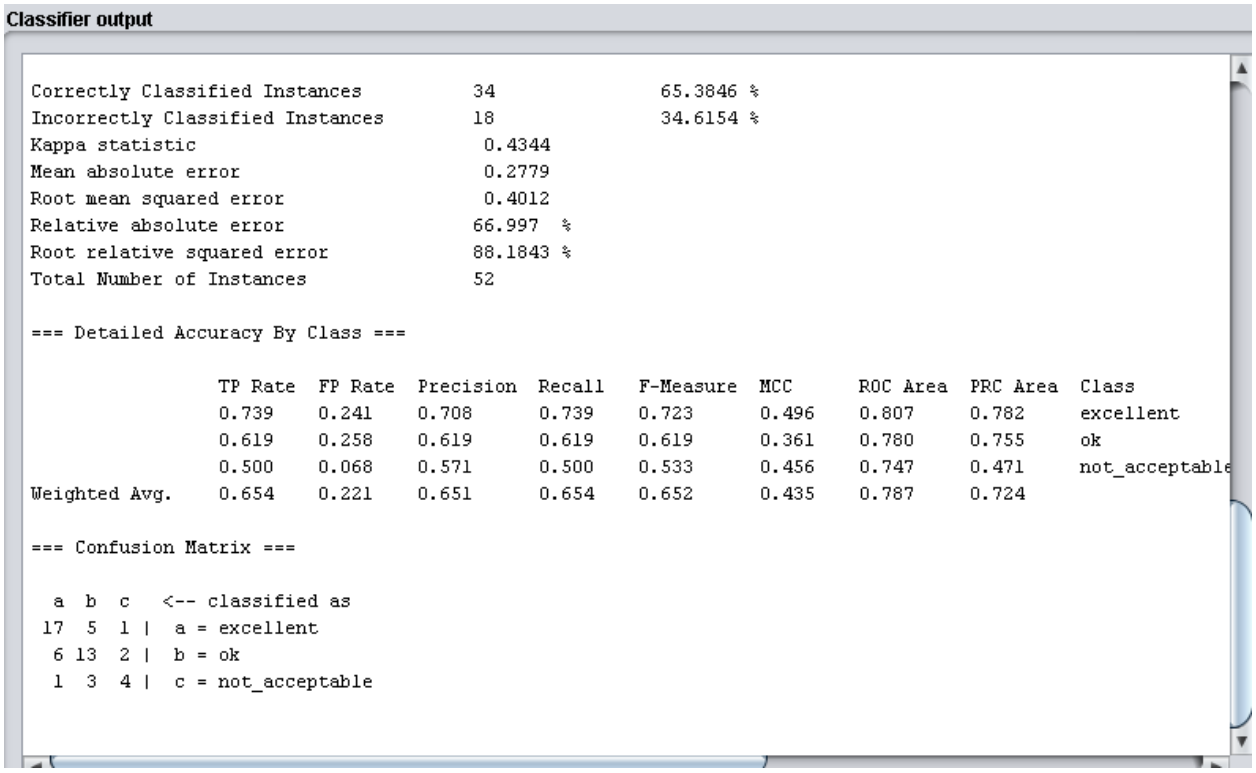
Naivni Bayes ima točnost od 63,4615% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 33 dok je netočnih 19 (36,5385%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.638 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.635 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,213 (specifičnost).



Slika 27. Rezultati klasifikacije 'squash' skupa podataka metodom naivnog Bayesa

Izgradnja modela na 'Squash-stored' skupu podataka metodom logističke regresije

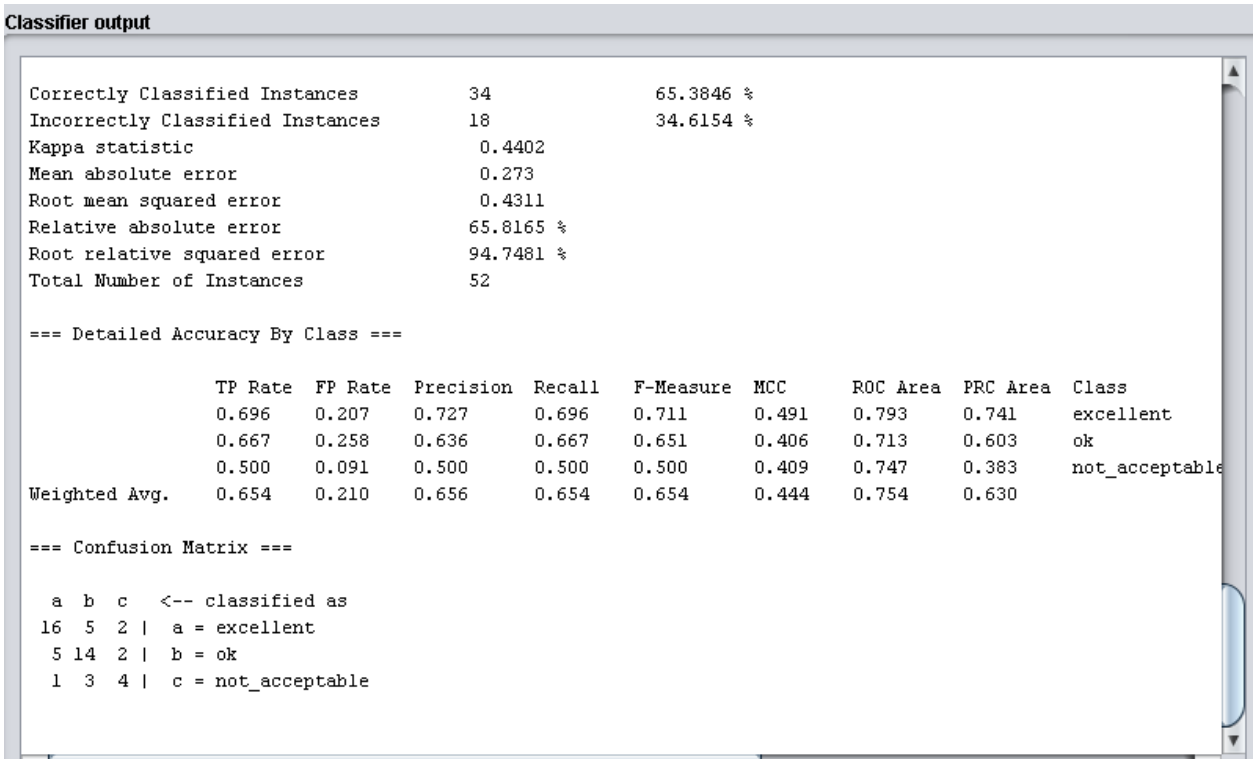
Logistička regresija ima točnost od 65,3486% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 34 dok je netočnih 18 (36,5385%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.651 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.654 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,221 (specifičnost).



Slika 28. Rezultati klasifikacije 'squash' skupa podataka metodom logističke regresije

Izgradnja modela na 'Squash-stored' skupo podataka metodom neuronskih mreža

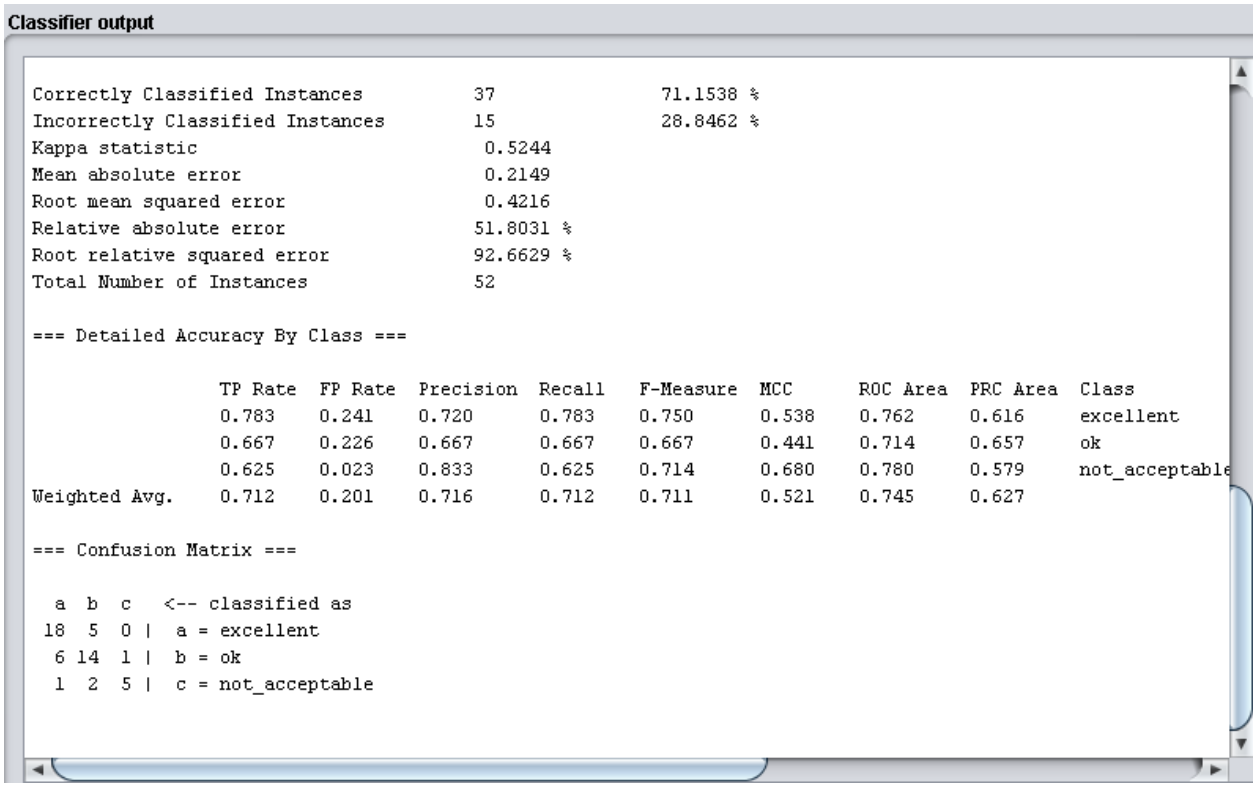
Neuronske mreže imaju točnost od 65,3846% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 34 dok je netočnih 1988 (36,5385%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.638 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.654 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,210 (specifičnost).



Slika 29. Rezultati klasifikacije 'squash' skupa podataka metodom neuronskih mreža

Izgradnja modela na 'Squash-stored' skupo podataka metodom stabla odlučivanja

Stabla odlučivanja imaju točnost od 71,1538% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 37 dok je netočnih 15 (36,5385%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0.716 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.711 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,201 (specifičnost).



Slika 30. Rezultati klasifikacije 'squash' skupa podataka metodom stabla odlučivanja

Odabir najboljeg algoritma na 'Squash-stored' setu podataka

Pogledom na rezultate može se uočiti da metoda stablo odlučivanja ima majveću generalnu točnost. Specifičan skup podataka s mnogo atributa s velikom dimenzionalnošću te malim brojem instanci je utjecao na lošije rezultate. Stablo odlučivanja je očigledno najbolja tehnika za ovajskup podataka .Inače stablo odlučivana je robusna metoda analize podataka koja se vrlo dobro nosi s velikom dimenzionalnošću te se toi ovdje se tako pokazalo.

Metoda	**Točnost**	**Specifičnost**	**Preciznost**	**Odziv**
Naivni Bayes	**63,4615%**	**0,213**	**0.638**	**0.635**
Logistička regresija	**65,3486%**	**0,221**	**0.651**	**0.654**
Neuronske mreže	**65,3846%**	**0,210**	**0.638**	**0.654**
Stablo odlučivanja*	**71,1538%**	**0,201**	**0.716**	**0.711**

Tablica 10. Komparativna analiza algoritama na 'squash' skupu podataka

6.5.Analiza 'Nurses' skupa podataka

'Nurses' skup podataka ima 8 prediktora i jednu ciljnu,zavisnu varijablu. 4 tehnike rudarenja podataka su primjenje za predviđanje zavisne varijable. S obzirom na čistoću podataka 'nurses' nije bilo potrebno koristiti nikakvu selekciju atributa za poboljšavanje točnosti modela. Na sljedećim slikama će biti prikazani rezultati

Izgradnja modela na 'Nursery' skupa podataka metodom naivnog Bayesa

Naivni Bayes ima točnost od 90,3241% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 11706 dok je netočnih 1254 (9,6759%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0,906 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.903 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,046 (specifičnost).

Classifier output

```
Correctly Classified Instances       11706               90.3241 %
Incorrectly Classified Instances      1254                9.6759 %
Kappa statistic                          0.8567
Mean absolute error                      0.0765
Root mean squared error                  0.1767
Relative absolute error                 28.0234 %
Root relative squared error             47.8152 %
Total Number of Instances            12960

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     not_recom
                 0.000    0.000    0.000      0.000    0.000      0.000    0.906     0.001     recommend
                 0.058    0.000    0.905      0.058    0.109      0.226    0.995     0.827     very_recom
                 0.903    0.096    0.821      0.903    0.860      0.789    0.966     0.937     priority
                 0.869    0.047    0.894      0.869    0.882      0.829    0.978     0.943     spec_prior
Weighted Avg.    0.903    0.046    0.906      0.903    0.894      0.857    0.982     0.957

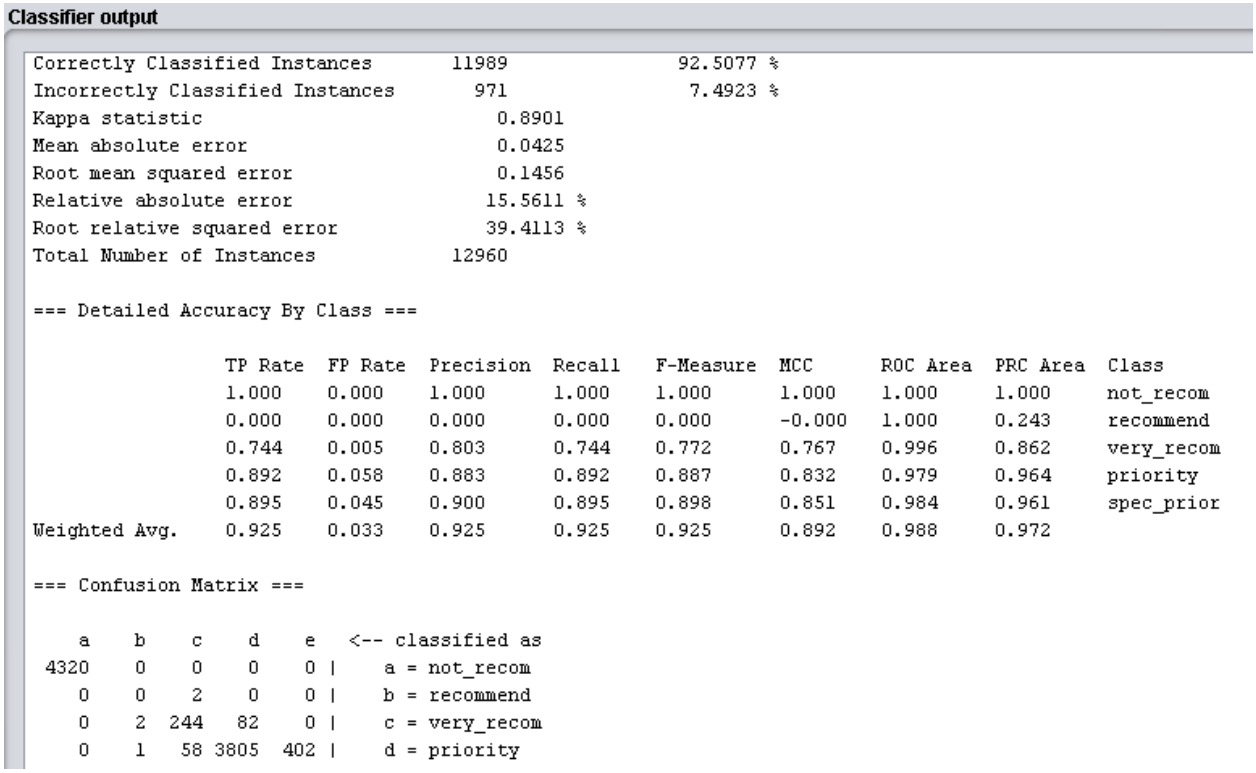
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4320    0    0    0    0 |    a = not_recom
    0    0    2    0    0 |    b = recommend
    0    0   19  309    0 |    c = very_recom
    0    0    0 3851  415 |    d = priority
```

Slika 31. Rezultati klasifikacije 'nursery' skupa podataka metodom naivnog Bayesa

Izgradnja modela na 'Nurses' skupa podataka metodom logističke regresije

Logistička regresija ima točnost od 92,5077% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 11989 dok je netočnih 971 (7,4293%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0,925 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.925 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,033 (specifičnost).



Classifier output

```
Correctly Classified Instances       11989               92.5077 %
Incorrectly Classified Instances       971                7.4923 %
Kappa statistic                          0.8901
Mean absolute error                      0.0425
Root mean squared error                  0.1456
Relative absolute error                 15.5611 %
Root relative squared error             39.4113 %
Total Number of Instances            12960

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     not_recom
                 0.000    0.000    0.000      0.000    0.000      -0.000   1.000     0.243     recommend
                 0.744    0.005    0.803      0.744    0.772      0.767    0.996     0.862     very_recom
                 0.892    0.058    0.883      0.892    0.887      0.832    0.979     0.964     priority
                 0.895    0.045    0.900      0.895    0.898      0.851    0.984     0.961     spec_prior
Weighted Avg.    0.925    0.033    0.925      0.925    0.925      0.892    0.988     0.972

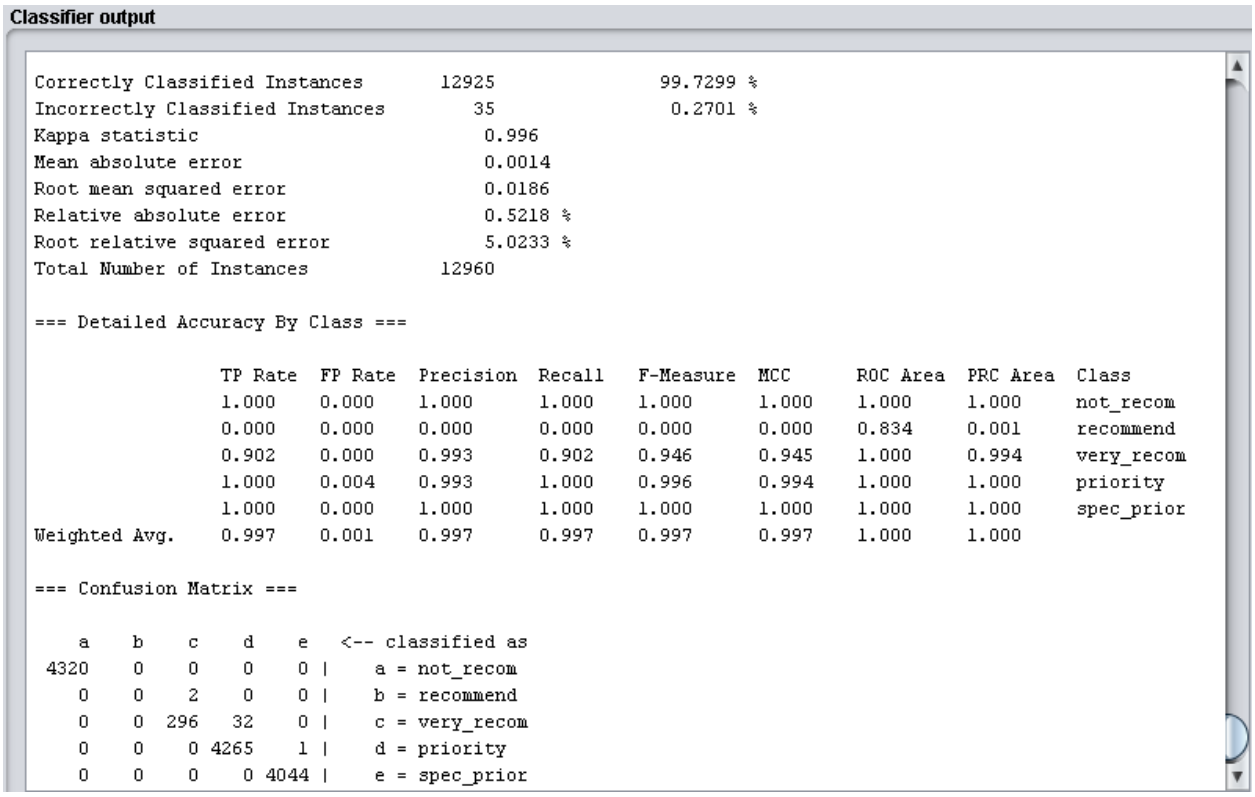
=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
 4320    0    0    0    0 |    a = not_recom
    0    0    2    0    0 |    b = recommend
    0    2  244   82    0 |    c = very_recom
    0    1   58 3805  402 |    d = priority
```

Slika 32. Rezultati klasifikacije 'nursery' skupa podataka metodom logističke regresije

Izgradnja modela na 'Nurses' skupa podataka metodom neuronskih mreža

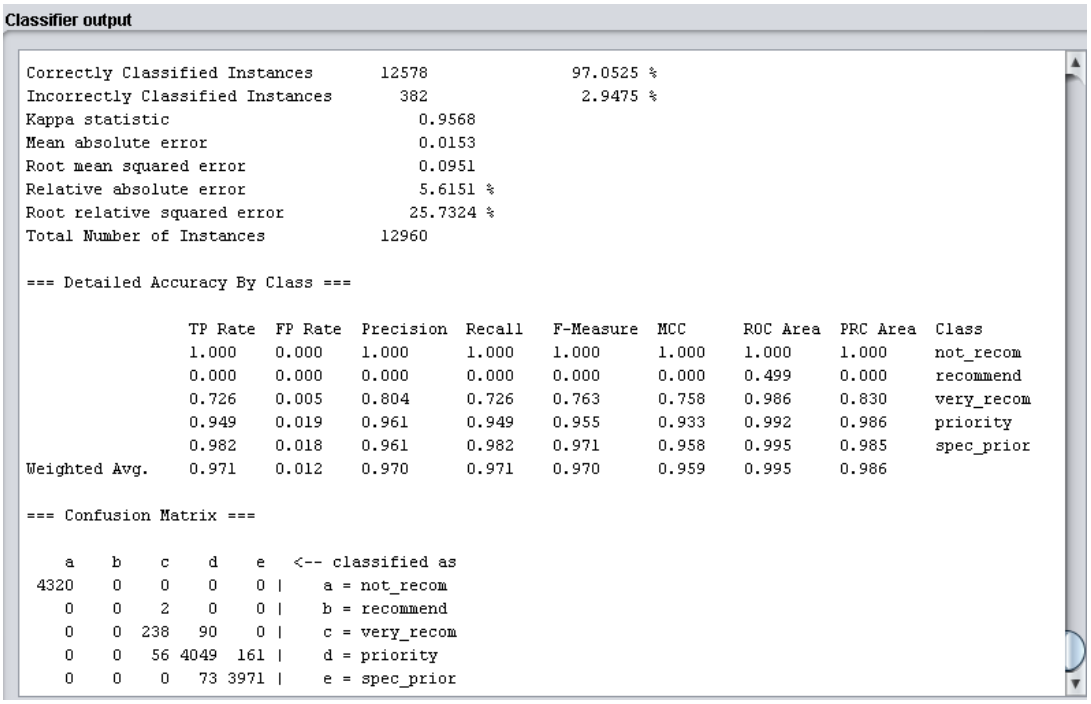
Neuronske mreže imaju točnost od 99,7299% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 12925 dok je netočnih 35 (0,2701%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0,957 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.997 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,001 (specifičnost).



Slika 33. Rezultati klasifikacije 'nursery' skupa podataka metodom neuronskih mreža

Izgradnja modela na 'Nurses' skupa podataka metodom stabla odlučivanja

Stablo odlučivanja ima točnost od 97,0525% točno klasificiranih primjera tj. primjera kojima je točno pogođena klasa ima 12578 dok je netočnih 382 (2,9475%). Omjer pozitivnih primjera koji su točno identificirani u broju pozitivnih predviđenih klasa je 0,970 (preciznost). Omjer pozitivnih primjera koji su točno identificirani u broju stvarnih pozitivnih klasa je 0.971 (odziv).Razmjer negativnih primjera koji su točno identificirani je 0,012 (specifičnost).



Slika 34. Rezultati klasifikacije 'nursery' skupa podataka metodom stabla odlučivanja

Odabir najboljeg algoritma na Nursery setu podataka

Pogledom na rezultate može se lako uočiti da metoda neuronskih mreža ima skoro stopostotnu točnost. Dobro balansirani.distrubuiran i čist skup podataka je utjecao na ovako dobru prediktivnu sposobnost

Metoda	**Točnost**	**Specifičnost**	**Preciznost**	**Odziv**
Naivni Bayes	**90,3241%**	**0,046**	**0,906**	**0.903**
Logistička regresija	**92,5077%**	**0,033**	**0,925**	**0.925**
Neuronske mreže*	**99,7299%**	**0,001**	**0,957**	**0.997**
Stablo odlučivanja	**97,0525%**	**0,012**	**0.970**	**0.971**

Tablica 11. Komparativna analiza algoritama na 'nursery' skupu podataka

6.6.OSVRT NA TEZE

Tijekom rada 4 najpopularnije klasifikacijske tehnike su korištene kako bi se analizirala 3 skupa podataka. Točnije,mjerila se prediktivna moć pojedinog algoritma i međusobno uspoređivala po 4 kriterija perfomansi. U ovom istraživanju uzele se u obzir,veličina podataka,veličina atributa te klasna distribucija skupa podataka te se dovelo u odnos s odabirom tehnike za rudarenje podataka

Rezultati koji smo dobili su sumirani u sljedećoj tablici. Može se isčitati da sva 3 skupa podataka se razlikuju te da je ta razlika direktno vezana za odabir najboljeg klasifikatora. Hipoteza koju smo postavili na početku istraživanja da različiti skupovi podataka utječu na izbor tehnike za predviđanje je zadovoljena

	Broj atributa	Broj instanci	Broj nominalnih varijabli	Broj numeričkih varijabli	Broj klasa		Najbolja metoda
Skup podataka'bank'	17	45211	10	7	2		Stablo odlučivanja,Neuronske mreže,log regresija,naivni Bayes
Skup podataka 'Squash-stored'	25	52	4	21	5		Stablo odlučivanja
Skup podataka 'nuresery'	9	12960	8	1	3		Neuronske mreže

Tablica 12. Komparativna analiza najboljih tehnika rudarenja podataka

Ako promotrimo i drugu hipotezu da se preciznost predviđanja tehnika rudarenja podataka značajno razlikuje na istim skupovima podataka se može isčitati u 6 dijelu rada kad su sva 4 algoritma procesirana na 3 različita skupa podataka te za svaki skup podataka algoritmi su davali različite rezultate

Iz svega je lako isčitati da ne postoji dominatna generalna tehnika za rudarenje podataka te ona ovisi i o veličini skupa,broju atributa ,klasnoj distribuciji te o mnogim drugim parametrima.

7.ZAKLJUČAK

Razvijanje moćnih i svestranih alata je prijeko potrebno da bi se automatski razotkrile vrijedne informacije iz sve te enormne količine podataka i transformirali ih u organizirano znanje. Ogromne količine podataka koje svijet genereira prati i razvoj alata za njihovu skladištenje. Dolazi do problema kako analizirati sve te podatke . Sve to vodi do razvoja novog područja u u računalnoj znanosti koje je vrlo obećavajuće i inovativno.

Rudarenje podataka, često se naziva i otkrivanje znanja iz podataka , je automatsko izvlačenje obrazaca iz velikih podataka pohranjenih u masivnim bazama podataka,skladištima ili repozitorijima

Razvijene su nove metodologije,sustavi,aplikacije za savladavanje svih vrsta podataka. Vrlo je važno znati ciljeve analize podataka kako bi odredili koje zadatke i metode je najbolje koristiti za dobivanje informacija. Klasifikacija je metoda kojoj je cij dodijeliti klasu iz poznatih primjera sa već određenom klasom te se koristi kad seželi predvidjeti kategorične varijable.

U 6. poglavlju korištene su 4 metode rudarenja podataka odnosno 4 algoritma. Struktura i način kako algoritam 'uči' podatke se vrlo razlikuju jedan od drugoga te se može pretpostaviti i da nemaju iste rezultate bez obzira na jednaku svrhu da predvide podatke. Promatrajući rezultate svakog modela na 3 različita skupa podataka očigledna je razlika prediktivnih sposobnosti algoritama na koje najviše utječe priroda podataka,veličina skupa podataka,broj atributa te broj klasa u podacima.

Vrlo je teško izabrati algoritam kojim će se dobiti najbolja točnost predviđanja te ovisi i o mnogim drugim parametrima. Nakon istraživanja tehnika rudarenja podataka neuronskih mreža,stabla odlučivanja,logističke regresije i naivnog Bayesa očita je razlika u načina fukcioniranja svakog algoritma te njihov odgovor na skupove podataka koji se mogu razlikovati po velikom broju parametara te je teško odrediti koji su to sve parametri koji utječu na odabir najbolje tehnike i koji su najvažniji parametri za olakšavanje izbora . Jedan od načina rješavanja ovog problema je korištenje više tehnika rudarenja podataka te zatim evaluacija i usporedba klasifikatora s ciljem odabira najtočnijeg. Evauluacija modela je vrlo bitna te generira metriku temeljem koje se mogu uspoređivati modeli.Korištena metrika u ovom radu je

točnost,preciznost,specifičnost i odziv. Kao što se vidi da korištena tehnika ovisi o skupu podataka tako i evaluacijska metrika ovisi o različitim parametrima kao što je balansiranost klasa u skupovima podataka i dr te ne postoji jedna mjera po kojoj se uspoređuju podaci već ovisi o cilju i zadatku analize.

Ovim radom predstavio se generalni koncept rudarenja podatak te je opisana najkorištenija metoda klasifikacija. 4 klasifikacijske tehnike su analizirane i uočena je veza između algoritama i skupova podataka te napomenuta važnost koji skupovi podataka imaju na odabir algoritma.

8. POPIS SLIKA I TABLICA

Tablica 1. Generalna kategorizacija tehnika rudarenja podataka

Tablica 2: Matrica grešaka

Tablica 3.Matrica grešaka 2

Tablica 4. Skup podataka ‘bank’

Tablica 5. Skup podataka ‘squash’

Tablica 6. Skup podataka ‘nursery’

Tablica 7. Standardne mjere

Tablica 8. Opis skupova

Tablica 9. Komparativna analiza algoritama na 'bank' skupu podataka

Tablica 10. Komparativna analiza algoritama na 'squash' skupu podataka

Tablica 11. Komparativna analiza algoritama na 'nursery' skupu podataka

Tablica 12. Komparativna analiza najboljih tehnika rudarenja podataka

Slika 1. CRISP-DM

Slika 2. Proces rudarenja podataka

Slika 3. Primjer klasifikacije

Slika 4. Izgradnja modela

Slika 5.Unakrsna validacija

Slika 6. Primjer stabla odlučivanja

Slika 7. Indukcija stabla odlučivanja

Slika 8.Podrezivanje stabla odlučivanja

Slika 9.Obrada informacija u neuronskoj mreži

Slika10. Arhitektura mreže "širenje unatrag"

Slika 11. Bayesov teorem

Slika 12. Skup podataka

Slika 13. Početni prozor Weke

Slika 14. Prozor 'preprocess' u Weki

Slika 15. Prozor 'preprocess' u Weki

Slika 16. dio'Selected attributes' u prozoru 'preprocess'

Slika 17. dio'Filter'u prozoru 'preprocess'

Slika 18. prozor 'Classify'

Slika 19. Evaluacijske opcije

Slika 20. Rezultat

Slika 21. Rezultat

Slika 22. Predprocesiranje 'bank' skupa podataka

Slika 23. Rezultati klasifikacije 'bank' skupa podataka metodom naivnog bayesa

Slika 24. Rezultati klasifikacije 'bank' skupa podataka metodom logističke regresije

Slika 25. Rezultati klasifikacije 'bank' skupa podataka metodom neuronskih mreža

Slika 26. Rezultati klasifikacije 'bank' skupa podataka metodom stabla odlučivanja

Slika 27. Rezultati klasifikacije 'squash' skupa podataka metodom naivnog Bayesa

Slika 28. Rezultati klasifikacije 'squash' skupa podataka metodom logističke regresije

Slika 29. Rezultati klasifikacije 'squash' skupa podataka metodom neuronski mreža

Slika 30. Rezultati klasifikacije 'squash' skupa podataka metodom stabla odlučivanja

Slika 31. Rezultati klasifikacije 'nursery' skupa podataka metodom naivnog Bayesa

Slika 32. Rezultati klasifikacije 'nursery' skupa podataka metodom logističke regresije

Slika 33. Rezultati klasifikacije 'nursery' skupa podataka metodom neuronski mreža

Slika 34. Rezultati klasifikacije 'nursery' skupa podataka metodom stabla odlučivanja

9.LITERATURA

1.Han,J.,Kamber,M.,Pei,J.(2012):Data Mining Concepts and Tehniques,Waltham

2.Tan,P.,Steinbach,M.,Kumar,V. (2006): Introduction to data mining,Boston

3.Osmar R.Zaine(1999):Principles of Knowledge Discovery in Database, Introduction to Data mining,str. 1-15

4. Smola,A.,Vishwanathan,S.V.N.(2008):Introduction to machine learning,Cambridge

5.Jackson,J.(2002):Data mining. A conceptual overview,Communications of the Association for Information Systems(Volume 8,) str.267 - 296

6.Demšar,J.,(2006): Statistical Comparisons of Classifiers over Multiple Data Sets,Journal of Machine Learning Reaserch 7 ,str.1-30

7. Friedman,J.H.,(): Data Mining and Statistics:What's the connection?,Stanford

8. Two Crows Corporation(2005):Introduction to Data Mining and Knowledge Discovery,Third Edition,Potomac

9. Niculescu-Mizil,A.,Caruana,R.():An Emperical Comparison of Supervised Learning Algorithms,NY

10 Villacampa,O.(2015): Feature Selection and Classification Methods for Decision Making: A Comparative Analysis,Nova Southeastern

11.Institut Ruđer Bošković:Otkrivanje znanja dubinskom analizom podataka,priručnik za istraživače i student

12. Bravo,H.,C.; Irizarry,A.,R.(2010):Model selection and assessment

13.Jović,A.,() : Postupci dubinske analize podataka,FOI

14. Aksenova,S.,S.(2004): Machine learning with WEKA WEKA EXPLORER TUTORIAL,Sacramento

15. Kusonmano et all.(2009): Evalution of the Imapct of Dataset Characteristivs for Classification Problems in Biological Applications, International Journal of Medical,Health,Bioengineering and Pharmaceutical EngineeringVol.3,No:10

16. Usama,M.Fayyad(1996); Data-Mining and Knowledge Discovery: Making sense out of Data

17. Weiss, Sholom M. et al.,(1998): Predictive Data-Mining: A Practical Guide, San Francisco

18. Salzberg,L.,S.(1997): Data Mining and Knowledge Discovery,On Comparing Classifiers:Pitfalls to Avoid and a Recommended Approach1,str.317-328,Boston

19, Moro,S.,Lauereano.,M.,S.,Cortez,P,.(): Using data mining for bank direct marketing: An application of the Crisp-DM methodology,Lisbou

20.Garcia-Salz,D.,(2011):Workshop and conference proceedings, Comparing classification methods for predicting distance student'sperfomance,str.25-32

21. Holte,R.,C.: Very simple Classification Rules Perform Well in Most Commonly Used Datasets,Otawa

22. Niculescu-Mizil,A.,Caruana,R.(2006):An Emperical Comparison of Supervised Learning Algorithms Using Different PerfmanceMetrics,NY

23.Entezari-Maleki et all: Comparison of Classification Methods Based on the type of Attributes and Sample Size,Tehran

24. Nsfor,C.,G.(2006):A comparative analysis of predictive data.mining teqniques,Knoxville

25. Fayyad et all(1996): Al Magazine Vol.17,No.3:From data Miningto Knowledge Discovery in Database

26. Bellaachia,A.: Classification and Prediction

27.Larose,D.,T.(2006): Data mining methods and models,New Jersey

28.Danso,S.,O.(2006):An Exploration of Classification Prediction Tehniques in Data Mining: The insurance domain,Boumemouth University

29. Ramagari,M.,B.(): Indian Journal of Computer Science and Engineering,vol.1,no.4, Data mining tehniques and applications

30, Johnson,K.,Kuhn,M.(): Applied Predictive Modeling

31. Hamalainen,W.(2006): Descriptive and Predictive Modeling Tehniques for Educuational Technology,University of Joensuu

32. Miller.,Thomas.,W.(2014):Modeling Techniques in Predictive Analytics,New Jersey

33. Chapman, P. et al.,(200) "CRISP-DM 1.0: Step by Step Data Mining Guide," *CRISPDM*

34.Singh,Y.,Chauhan,A.,S(2006-2009):Journal of Theoretical and Applied Information Technology, Neural Networksin Data Mining,Allahabad

35, Shaley-Shwartz,S.,Ben-David,S.(2014): Understanding Machine Learning: From Theory to Algorithms,Cambridge University Press

36,Pyle.D.(1999):Data Preparation for data Mining,Morgan Kaufman