Primjena tehnika rudarenja podacima u telekomunikacijama

Kundek, Zoran

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Sveučilišni studij

PRIMJENA TEHNIKA RUDARENJA PODACIMA U TELEKOMUNIKACIJAMA

Diplomski rad

Zoran Kundek

Osijek, 2017.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

IZJAVA O ORIGINALNOSTI RADA

Osijek, Rujan 2017.

Ime i prezime studenta:	Zoran Kundek
Studij :	Diplomski studij elektrotehnike
Mat. br. studenta, godina upisa:	D-451, 2011./2012.

Ovom izjavom izjavljujem da je rad pod nazivom**:**
Primjena tehnika rudarenja podataka u telekomunikacijama

izrađen pod vodstvom mentora

Prof. dr. sc. Drago Žagar, dipl. ing.

i su-mentora

Dr. sc. Višnja Križanović, dipl. ing.

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.

Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, su-mentora i drugih osoba, a što je izričito navedeno u radu.

Potpis studenta:

SADRŽAJ

1. UVOD

Rudarenje podataka (eng. *Data mining*) je postupak dobivanja određenih relevantnih i korisnih informacija nakon izvršenog sortiranja, organiziranja te grupiranja velikog broja podataka i njihovih korelacija, te prepoznavanja uzoraka koje oni tvore. Ono uključuje njihovo daljnje analiziranje iz različitih perspektiva te oblikovanje u razumljive i korisne informacije [1].

Podaci kojima raspolažemo mogu biti organizirani u baze podataka, ali isto tako oni mogu biti i neorganizirani i nestrukturirani, dobiveni iz drugih izvora poput prikupljenih subjektivnih dojmova i iskustava korisnika s web-a ili iz provedenih anketa ne ciljanog tipa.

Odabir metoda rudarenja podataka ovisi uglavnom o području primjene. Tako one mogu biti ili odabrane, ili preuzete iz drugih disciplina i područja znanosti, kao što su primjerice matematika, statistika i slično, ili modelirane iz prethodno utvrđenih poveznica i korelacija unutar same baze podataka koja se obrađuje.

Rudarenje podataka je neizostavni dio menadžmenta znanja (eng. *knowledge managementa*) procesa prikupljanja i distribuiranja, te efektivnog korištenja znanja u svrhu izvlačenja konkretnih zaključka. U slučaju kada se radi o velikoj količini prikupljenih podataka, dobivenih nakon izvršenog sortiranja, organiziranja te grupiranja istih uz pomoć metoda i tehnika rudarenja podataka dolazi se do zaključka kako je rudarenje podataka vrlo bitan postupak unutar procesa otkrivanja znanja (eng. *knowle dgediscovery*). Primjena inteligentnih tehnika i naprednih znanja omogućuje pretraživanje podataka uz uočavanje traženih podataka kroz prikazane uzorke, stvaranje novih pravila, primjenu novih ideja, postavljanje bitnih pitanja te predviđanjem budućih ishoda i lakše poduzimanje ispravnih koraka. Pri tome je primjena rudarenja podataka nužna. Iz navedenog proizlazi kako je otkrivanje znanja postupak koji nastaje međusobnom integracijom ponavljajućih procesa rudarenja podataka kombiniranih s provedbom ekspertiznih analizi dobivenih rezultata od strane stručnjaka, a s ciljem izvođenja novih zaključaka koji se mogu formirati u novo znanje o podacima.

Cilj ovog rada je prikazati i opisati praktičnu primjenu tehnika i metoda rudarenja podataka strukturiranog u naredna četiri poglavlja.

Nakon kraćeg uvoda o rudarenju podataka danog u uvodnom poglavlju, u drugom pogavlju je opisan je teorijski postupak rudarenja podataka podijeljenog na primjenu rudarenja podataka, postupke otkrivanja znanja kao i tehnike te metode rudarenja podataka. Kroz treće poglavlje rad se bavi s pregledom mogućnosti *Orange* alata, kao što je razvoj tijeka rada, priprema dokumenata za učitavanje u programski alat, kao i njegovih funkcionalnosti koje je moguće primijeniti na učitane podatke. U četvrtom poglavlju su dani primjeri primjene tehnika i metoda rudarenja podataka u telekomunikacijama. Dok četvrtom poglavlju je obrađen praktični primjer rudarenja podataka kroz klasifikaciju i grupiranje korisnika na osnovu usluga koje koriste.

2. TEORIJSKI UVOD U POSTUPAK RUDARENJA PODATAKA

2.1. Primjena rudarenja podataka

Rudarenje podataka (eng. *Data mining*) je postupak dobivanja određenih relevantnih i korisnih informacija nakon izvršenog sortiranja, organiziranja te grupiranja velikog broja podataka i njihovih korelacija, te prepoznavanja uzoraka koje oni tvore. Ono uključuje njihovo daljnje analiziranje iz različitih perspektiva te oblikovanje u razumljive i korisne informacije [1].

Bitno je napomenuti kako je rudarenje podataka postupak koji se sastoji od više koraka, od čega se niti jedan od njih ne bi smio u potpunosti preskočiti, dok je neke od njih potrebno i više puta ponavljati. Pri razmatranju rudarenja podataka kao tehničkog procesa, potrebno je omogućiti prenošenje analizirane predmetne problematike u tehničku domenu, čime se omogućuje primjena postupka rudarenja podataka, poput istraživanja, uzorkovanja, modeliranja, ocjene modela, implementacije na produkcijsku okolinu, te ocjene dobivenih rezultata. Kod prenošenja predmetne problematike, kompleksnost se proporcionalno povećava s kompleksnošću analize koju je potrebno provesti u sklopu procesa rudarenja podataka. Stoga je vrlo velika vjerojatnost da bi manji neoprez mogao uzrokovati nerješive prepreke, poput izvlačenja pogrešnih zaključaka i informacija. Takvu vrstu scenarija moguće je izbjeći korištenjem preporučenih metoda temeljenih na iskustvu, praksi i dizajnu procesa, kako bi se izbjegli neželjeni rezultati procesa učenja kao što su učenje neistinitih činjenica i učenje istinitih činjenica koje nisu korisne.

Početni korak u postupku rudarenja podataka koji nameće je dobro razumijevanje zadatka poslovnog zahtjeva koji je postavljen kao početni uvjet, nakon čega je tek moguće njegovo prevođenje i prilagodba u zadatak postupka rudarenja podataka. Nakon provedenog zadatka, potreban je ispravan odabir jednog od postupaka kojim će se raditi postupak rudarenja podataka. To može biti usmjereni, odnosno nadgledani (eng. *supervised*) postupak rudarenja podataka, ili neusmjereni odnosno nenadgledani (eng. *unsupervised*) postupak rudarenja podataka.

Kod usmjerenog ili nadgledanog postupka rudarenja podataka ciljani rezultat je poznat. Dakle cilj građenja modela ove vrste bi bio prikazivanje odnosa između izlazne varijable i ulaznog skupa podataka kojeg bi mogli prikazati njegovim najbitnijim primjenama kao što su:

- klasifikacije
- procjene
- predviđanja

Klasifikacija je raspodjela nestrukturiranih podataka u čvrsto definirane klase koje su definirane na osnovu tih nestrukturiranih podataka. Nameće se kao vrlo čest i koristan postupak u rudarenju podataka zbog toga što su izlazne varijable prikazane u kategoričkom obliku (npr.: 1/0). Ovaj postupak se najčešće koristi pri modeliranju uz primjenu tehnika rudarenja podataka kao što su kreiranje stabla odluke, tehnika grupiranja, najmanje udaljenosti ili najbližih susjeda (eng. *neares neighbor technique*), tehnika analize mreža, a naravno primjenjiv je i u tehnici neuronskih mreža.

Procjena se vrlo malo razlikuje u izlaznoj varijabli od klasifikacije jer daje odgovor u kontinuiranom numeričkom obliku, tj. kao niz brojeva u određenom rasponu (npr.: 1-10). Korištenjem takve vrste kontinuiranih varijabli dobivaju se varijable kreditne sposobnosti, računi i sl. Kao i kod klasifikacija, procjene je postupak čija je primjena prikladna kod neuronskih mreža i regresije.

Predikcija se zasniva na primjeni niza ulaznih podataka zabilježenih unutar određenog vremenskog perioda koji prethodio nastanku vrijednosti izlazne varijable. Jednostavnije definirano, postupak gotovo jednak kao kod klasifikacije i procjene, ali izlaznu varijablu određuje na osnovu predviđanja nekog budućeg ishoda ili vrijednosti.

Kod neusmjerenog ili nenadgledanog postupka rudarenja podataka rezultat ne postoji. Ova vrsta postupka ne nudi nikakvu izlaznu varijablu, već se bavi unutarnjom korelacijom i strukturalnim vezama unutar ulaznog skupa podataka kako bi ih prikazao u obliku nakupljanja, odnosno definirao ih kako bi bile jasnije vidljive.

- hijerarhijsko grupiranje u logičke skupove (eng. *hierarhical clustering*)
- profiliranje (eng. *profiling*)
- grupiranje prema zajedničkim obilježjima (eng. *affinity grouping*)

Hijerarhijsko grupiranje u logičke skupove je postupak grupiranja podataka koji dijele zajedničke karakteristike. Stručnije opisana korelacija ovakvih podataka bila bi segmentiranje heterogene populacije u više manjih homogenih skupina unutar kojih je

sličnost zapisa maksimizirana, dok je sličnost između samih skupova minimizirana. Hijerarhijsko grupiranje je specifično po tome što njegova izlazna varijabla nije prikazana u obliku vrijednosti koja bi se dobila iz neke klasifikacije, procjene ili nekog predviđanja, već je prikazana samo kao parametar u postupku rudarenja podataka. Pri tome gdje otkrivene hijerarhije koriste kao ulazni parametri za različite tehnike modeliranja kako bi pojedini zasebni modeli u konačnici dali bolje rezultate u konačnici.

Profiliranje je bitan alat analitičara pri opisu uzoraka unutar ulaznog skupa podataka kojim se poboljšava razumijevanje izvora podataka (npr. profila ljudi, načina prikupljanja podataka, i sl.) te dolazi do njihovog objašnjenja. Prigodna tehnika za profiliranje podataka bi bila stablo odluke.

Grupiranje prema zajedničkim obilježjima ili pravila pridruživanja (eng. *Association rules*) kako i sama riječ kaže povezuje zajedničke atribute. Ukoliko dva ili više produkata imaju zajednički ili sličan atribut bit će grupirani zajedno.

Podaci kojima raspolažemo mogu biti organizirani u baze podataka, ali isto tako oni mogu biti i neorganizirani i nestrukturirani, dobiveni iz drugih izvora poput prikupljenih subjektivnih dojmova i iskustava korisnika s web-a ili iz provedenih anketa ne ciljanog tipa. Nastavno na početni korak te odabir prikladnog postupka za obradu inicijalnog skupa podataka, kao njihove obrade, potrebno je napraviti selekciju, te transformaciju obrađenih podataka u konkretnu korisnu informaciju. Formiranje korisnih informacija se provodi na način da se razjašnjavaju pojedina značenja podatkovnih faktora te otkrivaju potencijalni problemi unutar podataka, što zatim dovodi i do boljeg razumijevanja tih podataka.

U narednim koracima podaci se grupiraju u prikladan skup podataka podijeljen na dva do tri podskupa koji se koriste za učenje (odnosno treniranje), validaciju te testiranje modela. Koraci koji se bave identifikacijom i pripremom odnose najveći udio resursa u cijelom procesu rudarenja podataka (ETL (eng. *Extract Transform Load*) proces rudarenja podataka). Tako se kreiranjem novih varijabli prilikom transformacije podataka omogućuje otkrivanje novih aspekata, omjera i kombinacija varijabli, nakon čega slijedi konačno kreiranje samog modela. Svaki model je moguće ocijeniti ili omogućiti ocjenu alatima neovisnima o modelu, kao što su matrice podudarnosti (eng. *coincidence matrix*) ili matrice pogreške (eng. *confusion matrix*), koje mogu prikazati neispravne ili pogreške klasifikacije izlaznih klasa varijabli. Sami postupak rudarenja podataka moguće je prikazati kroz deset koraka:

1. prevođenje poslovnog problema u problem dubinske analize podataka

2. odabir podataka

3. upoznavanje s podacima

4. kreiranje skupa podataka za modeliranje (eng. *model set*)

5. korekciju podataka

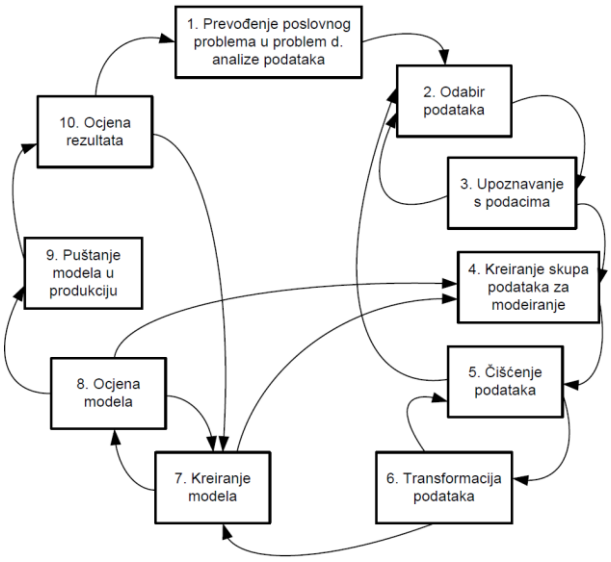
6. transformaciju podataka

7. kreiranje modela

8. ocjenu modela

9. puštanje modela u produkciju

10. ocjenu rezultata modela



Slika 2.1. Postupak rudarenja podataka

Kako je prikazano na slici 2.1., postupak rudarenja podataka je transparentnije i najbolje prikazati kao skup petlji koje su međusobno povezane, a čiji postupak izvođenja ima svoj redoslijed. Iako postoji redoslijed, sami koraci započinju prije nego je prethodni korak u potpunosti završen. Takvom izvedbom se dobiva revidirajući stanje koraka postupka koji činjenice naučene u kasnijim koracima primjenjuje na prethodne korake i tako ih revidira.

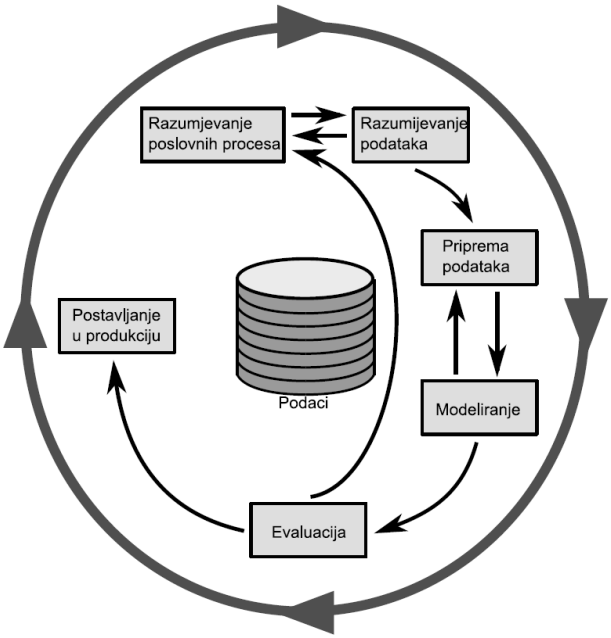
Uzorkovanje je postupak identifikacije ulaznih podataka, uzorkovanja velikih skupova podataka te podjele podataka u skupove podataka za treniranje, validaciju i testiranje.

Istraživanje je postupak koji obuhvaća statističko i grafičko istraživanje podataka, odnosno statističko proučavanje podataka, identifikacija ključnih varijabli, te njihov grafički prikaz.

Prilagodba ili modifikacija je postupak koji obuhvaća pripremu podataka za modeliranje, kreiranje dodatnih potrebnih varijabli, transformaciju postojećih varijabli zbog lakše uporabe i čitanja, identifikaciju glavnih čimbenika (eng. *outlier*), zamjenu nepostojećih vrijednosti, te analiza grupiranja (eng. *Clustering analysis*).

Modeliranje je postupak kreiranja prediktivnog modela uz primjenu regresijskog modela za točniju predikciju, stabla odluke, neuronskih mreža ili predefiniranog korisničkog modela.

Ocjena je postupak usporedbe modela korištenjem grafova koji prikazuju postotak točnosti, razdiobu i frekvenciju ponavljanja, te graf grupiranja podataka.



Slika 2.2. Koraci referentnog model

2.2. Postupci otkrivanja znanja

Prema [2] rudarenje podataka je proces višestrukog ponavljanja postupka poznatijeg kao otkrivanje znanja prema kojem postoji niz koraka koje je potrebno poštovati.

Definiranjem problematike određuje se cilj zadatka projekta zbog kojeg će se provoditi postupak rudarenja i otkrivanja znanja. Pri tome je bitna provjera opravdanosti i korisnosti cilja. Ukoliko cilj nije koristan ili opravdan postupak otkrivanja znanja nije vrijedno raditi, a ukoliko je cilj opravdan i koristan onda je otkriveno znanje moguće i primijeniti. Cijelom postupku prethodi identificiranje korisnih ulaznih podataka, te njihovog grupiranja u korisne skupove podataka. Nakon završenog postupka definiranja problematike može se tvrditi kako postoji odabrani skup podataka na kojem će se raditi otkrivanje znanja.

Postupci selekcije i provjere prikupljenih podataka odnose se na pripremu prilikom njihovog dobivanja iz mnogobrojnih izvora, kao i filtriranje i odbacivanje nepotrebnih ili manje korisnih informacija kako bi se izbjegla međusobna isključivost ili neki od drugih sukoba unutar prikeljenog skupa podataka. Ujedno, provodi se i postupak udruživanja podataka iz različitih izvornih tablica kako bi se dobila homogenost izvornog skupa. Ovaj korak zahtjeva vrlo veliku preciznost, te iziskuje velike napore s gledišta utrošenog vremena i sredstava.

Postupak sažimanja ili kompresije podataka se oslanja na pravilno shvaćanje cilja zadatka projekta definiranog problematikom. Iz jasno definiranog cilja proizlazi pronalaženje korisnih informacija koje će predstavljati podatak koji je konkretan i koristan. Metodama dimenzioniranja i transformacije veliki broj varijabli će zatim biti moguće ograničiti i razjasniti prilikom razmatranja.

Postupci analize i postavljanje hipoteze se odnose isključivo na odabir algoritma za rudarenje podataka kojim će biti moguće prikazati uzorke unutar skupova podataka.

Rudarenje podataka se svodi na pronalazak uzoraka unutar informacija grupiranih u podatke dobivenih sažimanjem. Kvaliteta postupka rudarenja podataka je proporcionalna s kvalitetom pripremljenih podataka, odnosno ovisi o tome koliko kvalitetno je odrađeno definiranje problema, napravljena dobra selekcija i pročišćavanje podataka te sažimanje stvarne suštine cilja. Ukoliko su podaci kvalitetno pripremljeni, zadatak rudarenje podataka postaje relativno trivijalan postupak. Rudarenje podataka se svodi na korištenje raznovrsnih metoda rudarenja, a neke od njih su regresija, stabla odluke, klasifikacija, pojedini oblici grupiranja, kao što su

hijerarhijsko grupiranje i grupiranje prema zajedničkim obilježjima procjene, profiliranje i na posljetku predikcija.

Provjera ispravnosti modela i točnosti hipoteze je postupak testiranja samog modela koji je stvoren na neovisnom transparentnom skupu podataka koji će prouzrokovati jasno vidljiv propust, krivi rezultat izazvan krivo primijenjenim modelom ili lošijom solucijom, te krivo i nepotpuno postavljenom hipotezom. Ispravnost modela ne odnosi se samo njegovu točnost već i brzinu. Brzina je vrlo često važniji faktor jer točnost u dovoljno velikom postotku može biti zadovoljavajuća ukoliko sustav brzo i efikasno radi.

Promatranje modela je konstantni radni proces provjere zadužen za samo-održivost modela. Predstavljanjem novih informacija modelu uslijed bilo kojeg oblika promjene te provjera njegove reakcije, odnosno stabilnosti i ispravnosti izlaznih varijabli određuje mogućnosti njegove primjene u realnom vremenu, nadogradnje ili odbacivanja na osnovu zastarjelosti i neispravnosti.

2.3. Postojeće metode rudarenja podatka

Alate ili metode je moguće podijeliti u osnovne skupine primjene kao što su alati za učitavanje i sadržavanje skupa podataka [3].

Kod analitičkih istraživanja podataka se još uvijek koriste tradicionalni ili klasični pristupi za otkrivanje ključnih čimbenika unutar skupova podataka postavljanjem pretpostavke na osnovu koje se gradi model, koji tada potvrđuje ili pobija predloženu teoriju modela. Takav način rada je dugotrajan proces koji iziskuje prvenstveno temeljito preispitivanje ispravnosti hipoteze, njezinu potvrdu, te prilagodbu ukoliko ne udovoljava početnim uvjetima.

Alati kod rudarenja podataka uvelike olakšaju odabir odgovarajućeg modela, te je na taj način veliku većinu ponavljajućih procesa koje je potrebno napraviti u ranije spomenutom tradicionalnom pristupu, moguće izbjeći odrađivanjem istog postupka putem računalnih simulacija uz primjenu opisnih i prediktivnih analitičkih modela koje čine međusobno povezani skupovi istih alata. Zbog toga je alate moguće podijeliti u osnovne skupine primjene kao što su alati za učitavanje i sadržavanje skupa podataka (npr. razne vrste tablica s predefiniranim primjenama, osnovnim i dodatnim ključevima), klasifikacijski alati za

grupiranje podataka, regresijski alati za procjenu i oblik povezanosti podataka, te alati za vizualni prikaz u obliku grafova i dijagrama.

2.4. Postojeće tehnike rudarenja podatka

Modeliranje se postavlja kao jedna od najvažnijih tehnika rudarenja podataka, te omogućava dobivanje bitnih, dotad nepoznatih podataka koji omogućuju predviđanja i dobivanja informacija o budućim utjecajima i događajima. Modeliranje je u osnovi postupak kreiranja modela kao skupa matematičkih relacija uz primjenu podataka o događaju čiji je ishod od ranije poznat, te primjene kreiranog modela na događaj nepoznatog ishoda. Postoje dva osnovna postupka rudarenja podataka, usmjereni, odnosno nadgledani (eng. *supervised*) postupak rudarenja podataka, i neusmjereni odnosno nenadgledani (eng. *unsupervised*) postupak rudarenja podataka.

Kod usmjerenog ili nadgledanog postupka rudarenja podataka ciljani rezultat je poznat. Dakle cilj građenja modela ove vrste bi bio prikazivanje odnosa između izlazne varijable i ulaznog skupa podataka kojeg bi mogli prikazati njegovim najvažnijim primjenama. Neke od najčešćih tehnika pri procesu modeliranja, tj. tehnika rudarenja podataka jesu regresija, stablo odluke i neuronske mreže.

Kod neusmjerenog ili nenadgledanog postupka rudarenja podataka rezultat ne postoji. Ova vrsta postupka ne prikazuje nikakvu izlaznu varijablu, već se bavi unutarnjom korelacijom i strukturalnim vezama unutar ulaznog skupa podataka kako bi ih prikazao u obliku grupiranja, odnosno definirao ih kako bi bile jasnije vidljive [4].

2.4.1. Regresija

Regresija je zadužena za predviđanje pripadnosti varijable određenoj klasi. U neke tehnike klasifikacije i regresije uključena su i stabla odluke, neuronske mreže.

Redoslijed i uparivanje alata određuje opisne modele pomoću kojih se tvore pravila po kojim se buduće varijable uparuju.

Dvije osnovne vrste regresije koje se koriste su linearna i logistička regresija. Linearna regresija se može podijeliti na osnovnu jednostavnu linearnu regresiju kod koje se ravnim pravcem aproksimira odnos između jedinstvene ulazne i jedinstvene izlazne kontinuirane varijable. Višestruka regresija (eng. *multiple regression*) koristi linearnu površinu ili višedimenzionalnu ravninu za prikazivanje višestrukih ulaznih varijabli i jedne kontinuirane izlazne varijable. Obilježja logističke regresije su, za razliku od linearne regresije, diskretne vrijednosti na izlazu, odnosno binarni izlaz 1/0, odnosno DA/NE.

Metoda linearne regresije je zasnovana na formuli:

$$y_i = f(x_i) + \varepsilon_i \tag{2-1}$$

gdje je y_i vrijednost izlazne varijable i-tog slučaja, x_i je vektor vrijednosti ulaznih varijabli *i*-tog slučaja, dok je ε_i vekror greške, odnosno nekorelirane slučajne varible sa srednjim vrijednosti jednakom nuli [5].

Jednostavna linearna regresija se može zapisati u obliku:

$$y = \beta_0 + \beta_i x + \varepsilon \tag{2-2}$$

koja predstavlja linearan odnos za sve slučajeve i naziva se jednadžba regresije. Član greške ε je potreban kako bi označio neodređenost modela.

Višestruka regresija je nadogradnja jednostavne linearne regresije uz pretpostavku ili analognu komponentu vektora greške, te u slučaju n varijabli može se zapisati u obliku:

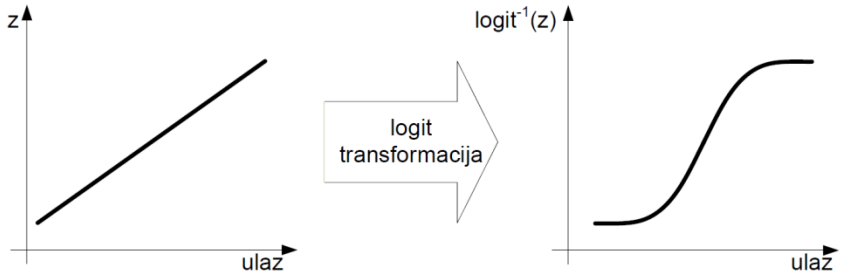
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n \tag{2-3}$$

Logističku regresiju je moguće gledati kao specijalni slučaj linearne regresije. Kao spojnu funkciju koristi log *it* funkciju:

$$g(p_i) = \log it(p_i) = log \frac{p_i}{1 - p_i}$$

(2-4)

gdje je pi vjerojatnost da izlaz ima vrijednost koja se određuje modelom uz uvjet da ulazne varijable imaju vrijednost *i*-tog uzorka [6]. Korištenjem log *it* funkcije (Sl. 2.3.) izlazne vrijednosti varijabli su u rasponu od 0 do 1 za razliku od linearne regresije kojemu je vrijednost izlaznih varijabli u rasponu od -∞ do ∞.

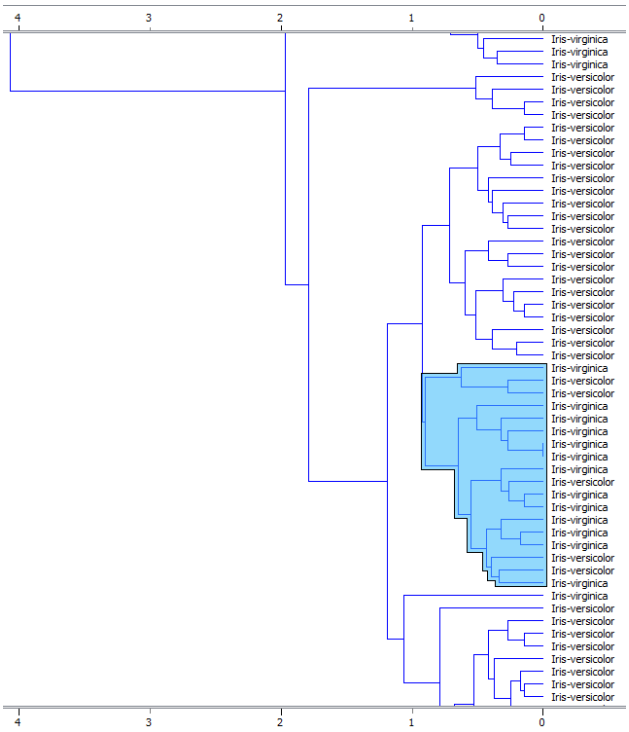


Slika 2.3. Prikaz logističke regresije.

2.4.2. Grupiranje

Grupiranje se bavi razlikovanjem entitetnih vrijednosti i grupira ih u skupine po sličnosti prema obilježjima, odnosno odvaja ih u druge skupine po različitosti. Grupiranje se vrši mjerenjem udaljenosti između vrijednosti. Na temelju njih se provodi njihovo grupiranje u pripadajuće skupine. Ukoliko su rezultati proizašli iz dva različita modela ili algoritma isti, takva provjera potvrđuje točnost grupiranja i zaključaka proizašlih iz njih. Grupiranje je moguće definirati kao računalno učenje bez potrebnog nadzora [7].

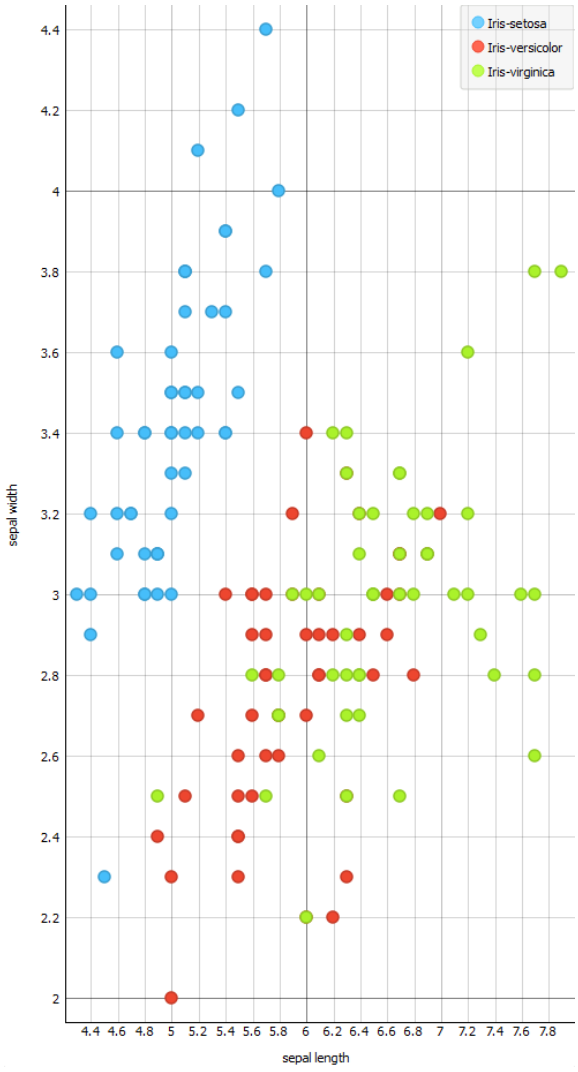
Grupiranje je moguće podijeliti u dvije skupine, i to na hijerarhijsko grupiranje i grupiranje s podjelom prema obilježjima. Kod hijerarhijskog grupiranja izlazna varijabla je prikaz sličan stablu koji se naziva dendogram, prikazan na slici 2.4., Kod ove vrste grupiranja ne postoji jasno preklapanje grupa, dok kod grupiranja s podjelom prema obilježjima postoji jasno preklapanje. Dendogram je dijagramsko stablo u kojem je moguće zabilježiti udaljenosti pridruživanja ili razdvajanja pomoću kojih je moguće jasno definirati broj skupova.



Slika 2.4. Dendogram, prikaz hijerarhijskog grupiranja.

Hijerarhijsko grupiranje se zasniva na dva osnovna tipa algoritma. Kod prvog algoritma značajka je grupiranja podataka u okruglim grupama, a razlog tomu je to što se obrada izvršava od dna prema gore, tijekom čega se provodi povezivanje susjednih podataka. Postupak se ponavlja skroz dok svi podaci ne zadovoljavaju značajke iste grupe. Drugi algoritam radi razdvajanje podataka u grupe uz početnu pretpostavku kako svi podaci već pripadaju istoj grupi [8].

Algoritmi u velikom broju slučajeva koriste metodu stvaranja okruglih grupa prilikom čega obilježja grupe pohranjuju u matricu udaljenosti. Svaka grupa predstavlja jesnu grupu, te logično se da zaključiti kako će broj inicijalnih grupa biti jednak broju grupa. Postupak se nastavlja s ponavljanjem skroz dok se ne formira jedna jedinstvena grupa iz koje se tada nadograđuje matrica udaljenosti. Pretpostavkom kako je svaki podatak definiran mjerama q_i i p_i, otvara se mogućnost njihovog prikaza u dvodimenzionalnom koordinatnom sustavu.



Slika 2.5. Prikaz hijerarhijskog grupiranja podataka
u dvodimenzionalnom koordinatnom sustavu

Udaljenost između točaka je moguće izračunati pomoću Euklidove razdiobe.

$$d_{qp} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2-5}$$

Prilikom izračuna udaljenosti Euklidovom razdiobom stvara se simetrična matrica udaljenosti prikazana u tablici (Tab 2.1.). Simetričnost tablice označava da se radi o istoj udaljenosti između dva podatka neovisno mjeri li se udaljenost podatka A od podatka B ili obratno, te time se može reći kako se radi o dijagonalnoj matrici pri čemu dijagonalne vrijednosti su jednake nuli jer predstavljaju udaljenost same od sebe.

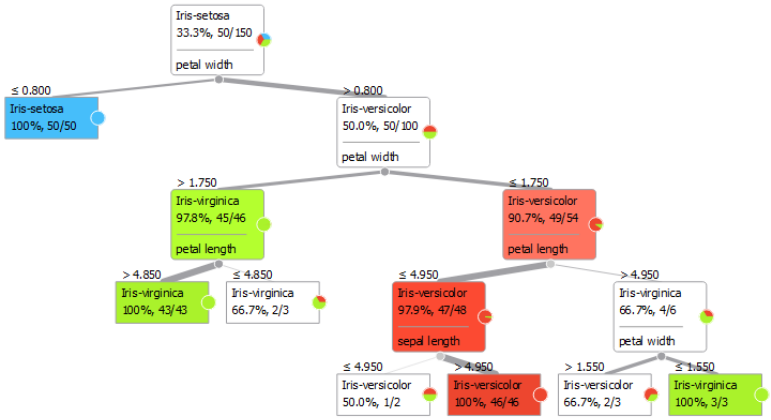
Tablica 2.1. Prikaz primjera tablice matrice udaljenosti.

d_{q0p0}	d_{q1p1}	d_{q1p2}	d_{q1p3}	d_{q1p4}	d_{q1p5}
d_{q1p1}	d_{q0p0}	d_{q1p2}	d_{q1p3}	D_{q1p4}	d_{q1p5}
d_{q2p1}	d_{q2p1}	d_{q0p0}	d_{q2p3}	d_{q2p4}	d_{q2p5}
d_{q3p1}	d_{q3p1}	d_{q3p2}	d_{q0p0}	d_{q3p4}	d_{q3p5}
d_{q4p1}	d_{q4p1}	d_{q4p2}	d_{q4p3}	d_{q0p0}	d_{q4p5}
d_{q5p1}	d_{q5p1}	d_{q5p2}	d_{q5p3}	d_{q5p4}	d_{q0p0}

2.4.3. Stablo odluke

Stablo odluke je tehnika predviđanja koja u svakoj grani stabla sadržava klasifikacijsko pitanje na temelju kojeg, ovisno o odgovoru na njega, odgovore dijeli u čvorove. Stablo odlučivanja se može promatrati kao segmentacijska podjela podatka. Segmentiranje se koristi u svrhu predviđanja važnog dijela informacije, a predviđeni segmenti iz stabla nose obilježja tih segmenata. Odlika stabala odlučivanja unatoč njihovoj kompleksnosti i pozadini algoritama korištenih za odluku su jednostavnost i transparentnost rezultata.

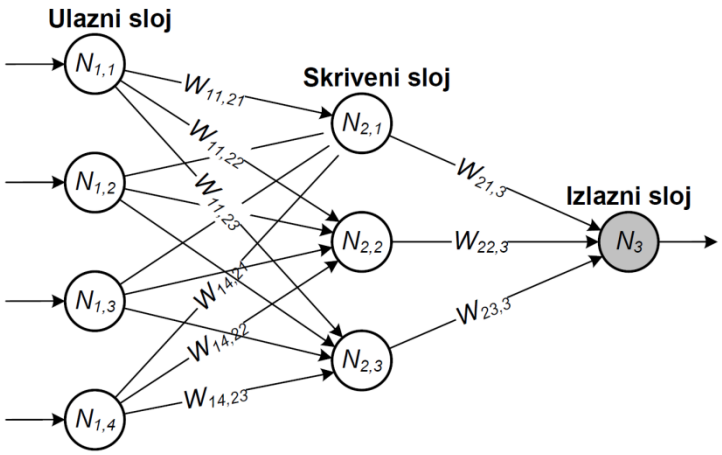
Algoritmi koji koriste ovaj oblik odlučivanja imaju dva stanja, stanje rasta stabla u kojem se podrazumijeva dijeljenje stabla u manje podskupove kroz ponavljanje procesa i stanje čišćenja stabla kako se ne bi granalo u nepotrebne dubine. Pravila čišćenja imaju obilježja maksimalne dubine stabla, minimalni broj segmenata stabla te minimalan broj segmenata u novom čvoru. Osim analitičkog pristupa čišćenju stabla odlučivanja, može se primijeniti i algoritam mjere točnosti odluke meta-stabla na uzorku za ispitivanje.



Slika 2.6. Prikaz stabla odlučivanja

2.4.4. Neuronske mreže

Neuronske mreže su najsloženija tehnika podržana vrlo kompleksnim algoritmima uz vrlo veliku pouzdanost, uštedu u pogledu vremena, ali jednako tako i kompleksnost rezultata izlaznih varijabli. Iako tehnike koje koriste modele zasnovane na neuronskim mrežama obrađuju podatke u realnom vremenu, njihova kompleksnost predstavlja velik problem pri korištenju, pa se i zbog potrebe za prethodnom obradom i konverzijom ulaznih podataka, nerijetko poseže za nekim jednostavnijim tehnikama regresije i klasifikacije. Osnova ideja neuronskih mreža je zasnovana na prikupljanju ulaznih podataka iz vrlo velikog broja ulaza koji bi zatim u što kraćem vremenu s ciljem obrade u realnom vremenu dale izlazni parametar na jednom izlazu. Bitno je napomenuti da su neuronske mreže zasnovane na vrlo apstraktnom pojmu razlučivanja te je težnja njihovo samostalno učenje i nadogradnja, te naposljetku njihovog konačnog stadija razvoja i samostalnog definiranja kao umjetne inteligencije. Iako su u svakom pogledu neuronske mreže napredovale uvelike u svom razvoju, ukoliko je moguće jasno definirati proces odlučivanja od ulaznih do izlaznih parametara i dalje ih se ne može smatrati ničim više od skupa algoritama ili modela za odlučivanje i predviđanje, ili statističke metode.

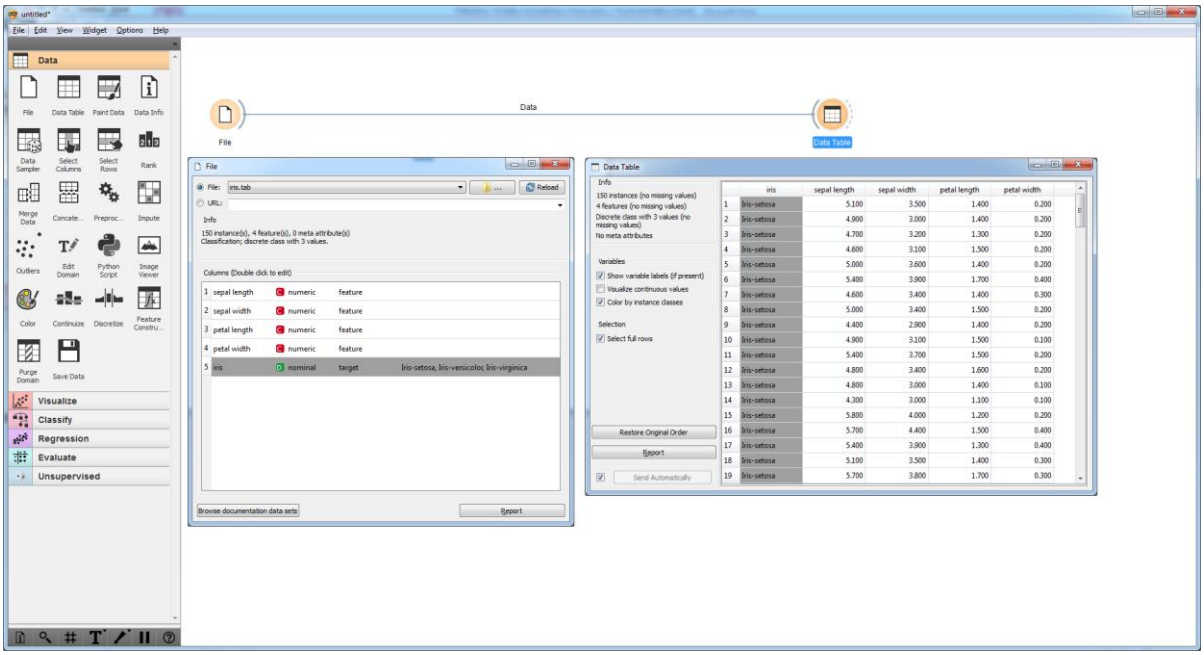


Slika 2.7. Prikaz jednostavne neuronske mreže

3. PREGLED OSNOVNIH MOGUĆNOSTI ORANGE ALATA

3.1. Razvoj tijeka rada za analizu na temelju učitanih podataka i njihovog prikaza

Alatne jedinice (eng. *widgets*) su dio Orange alata za učitavanje, obradu, grupiranje, prikaz, te istraživanje podataka kao i razvoj prediktivnih modela, te raznih drugih modela za rudarenje podataka i njihovo daljnje istraživanje. Alatne jedinice je moguće međusobno povezivati stvaranjem veza između ulaznih i izlaznih kanala, odnosno stvoriti komunikacijski kanal između istih. Ulazni kanal (eng. *Input chanel*) se nalazi s lijeve strane alatne jedinice, odnosno izlaz kanal (eng. *Output chanel*) s desne strane alatne jedinice.

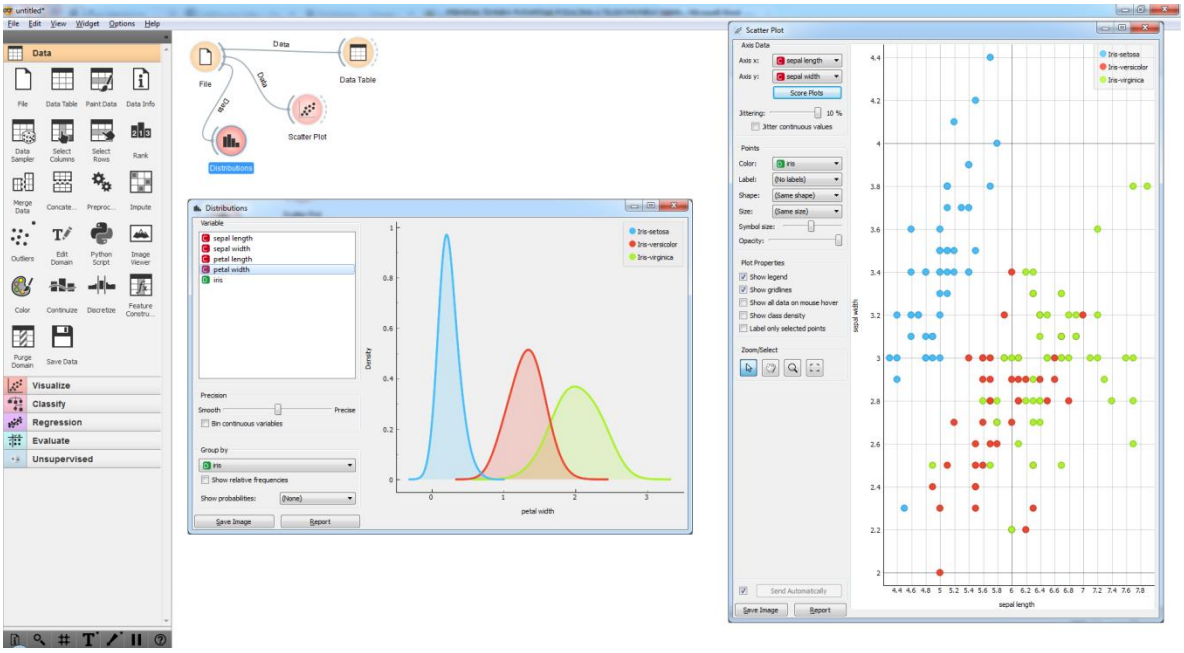


Slika 3.1. Prikaz *„File“* i *„Data Table“* alatnih jedinica u *„Orange“* alatu.

Većina tijekova rada (eng. *workflows*) modela za analizu podataka uobičajeno počinju s *„File“* alatnom jedinicom (eng. *widget*) pomoću koje učitavamo sirove podatkovne skupove. Na samom početku prilikom učitavanja podataka već su dostupni predlošci koji sadrže predefinirane skupove podataka (eng. *Data Sets*) za uvod u mogućnosti alata i rad s njime.

„Data Table“ alatna jedinica predviđena je za tablični prikaz sirovih podataka, te preinake unutar početnog skupa podataka. Vizualizaciju učitanih podataka moguće je ostvariti primjenom *„Scatter Plot“* alatne jedinice predviđene za impulsni prikaz podataka u

koordinatnom sustavu. Prikaz frekvencije ponavljanja podataka ili gustoće nakupljanja podataka je moguć pomoću *„Distributions“* alatne jedinice.



Slika 3.2. Prikaz *„Scatter Plot“* i *„Distributions“* vizualizacijskih alatnih jedinica.

3.2. Priprema, učitavanje i prikaz podataka

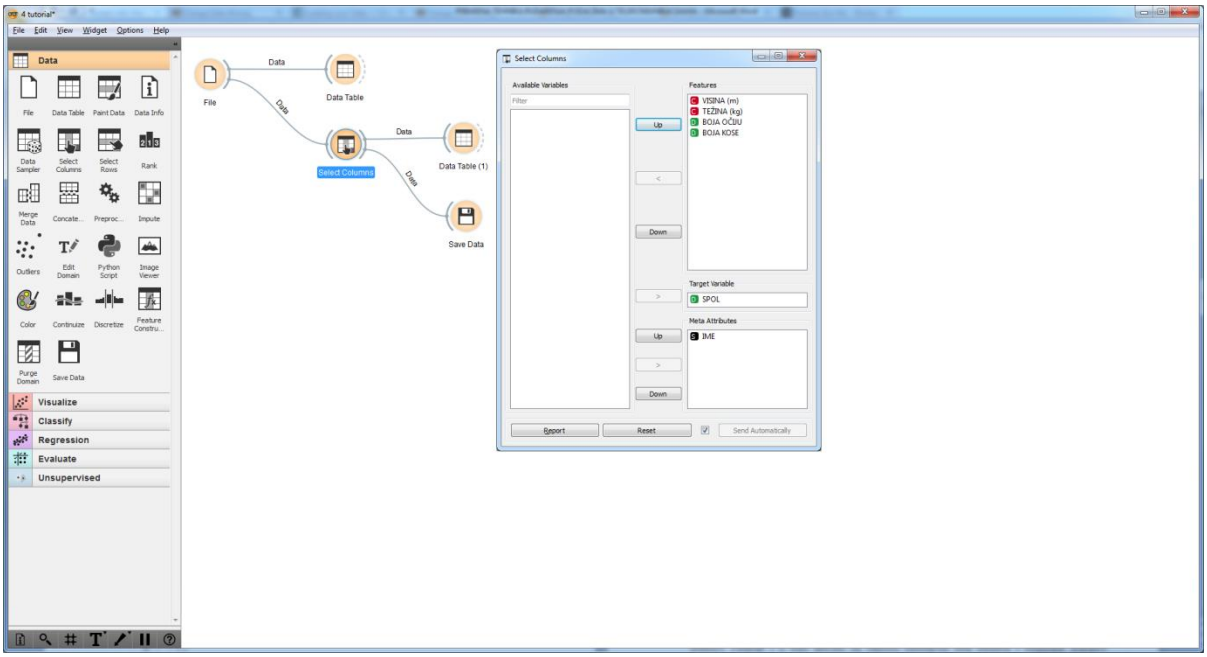
Kod rudarenja podataka najvažniji zadatak je ispravna priprema podataka prije primjene neke od metoda rudarenja podataka, te njihovo spremanje u pravilan format dokumenta kako bi se nesmetano mogli učitati. *Orange* alat ima mogućnost čitanja predodređenih formata dokumenata kao što su .tab koji je također i prirodni format datoteka u *Orangeu,* .xls/.xlsx, te tekstualne .txt datoteke odvojene zarezom. Uobičajeno strukturiranje podataka unutar dokumenta je u obliku tabličnog prikaza gdje se u stupcima nalaze atributi, dok se u redovima nalaze podatkovne instance.

Tablica 3.1. Prikaz primjera tablice za učitavanje s atributima i njihovim vrijednostima.

IME	SPOL	VISINA (m)	TEŽINA (kg)	BOJA OČIJU	BOJA KOSE
Ivana	Ž	1,6	50	zelena	plava
Marko	M	1,8	95	smeđa	crvena
Josip	M	1,75	80	plava	smeđa
Ana	Ž	1,65	65	zelena	crna
Stjepan	M	1,9	90	smeđa	crvena
Tomislav	M	1,85	90	smeđa	smeđa
Katarina	Ž	1,8	65	plava	smeđa
Marija	Ž	1,75	80	zelena	plava
Ivan	M	2	100	plava	crna

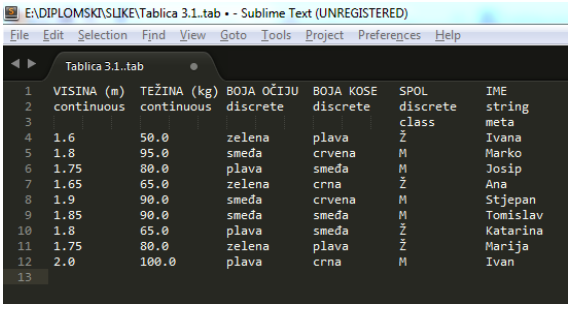
Nakon učitavanja dokumenta kroz *„File“* jedinicu *Orange* alata, dodavanjem *„Data Table“* jedinice i njihovim međusobnim povezivanjem na način opisanim ranije u tekstu. Otvaranjem *„Data Table*“ jedinice *Orange* izvršava automatsko strukturiranje na osnovu učitanih podataka i pretpostavke kako bi mogao definirati meta-podatak i varijablu klase. Ukoliko je alat krivo percipirao neke od podataka, te na osnovu krive pretpostavke jedan od definiranih atributa krivo postavio kao varijablu klase npr. postavljena boja očiju kao varijabla klase umjesto spola. Unatoč tome nije potrebno prepravljanje učitane datoteke koja sadržava inicijalni slog podataka i njeno ponovno učitavanje kako bi se ispravila greška, već je dovoljno povezivanje dodatne *„Select Columns“* alatne jedinice u tijek rada (eng. *workflow*) pomoću koje se može napraviti izmjena postavljanjem ispravnog atributa u *„Target Variable“* polje, kako je i prikazano na slici niže. Nakon toga je potrebno na *„Select Columns“* alatnu jedinicu povezati novu „*Data Table“* jedinicu koja će nadalje čitati ispravno podešene podatke iz tablice. Ispravljeni slog podataka moguće je direktno pohraniti

na računalo pomoću *„Save Data“* alatne jedinice, a najprikladnije je to učiniti za *Orange* prirodnom .tab formatu zbog njegove mogućnosti anotacije atributa unutar zaglavlja.



Slika 3.3. Prikaz modela za ispravljanje učitane datoteke i *„Select Columns“* alatne jedinice u kojoj se može napraviti ispravak.

Moguće je i lokalno definiranje tipa i vrste varijabli sirovih podataka dodavanjem dvaju novih redova ispod prvog reda u kojem su definirani atributi, odnosno nazivi atributa, a iza njega kao drugi red će biti definiran kao vrsta varijable, odnosno vrsta atributa. Prvi red bit će definiran kao tip varijable , u ovom slučaju tip atributa, te ga označiti slovom C kao *continious* ili atribut s numeričkim vrijednostima, D za diskretne atribute ili atribute s kategorijskim vrijednostima i S za string atribute ili atribute s tekstualnim vrijednostima. Drugi red će biti definiran kao vrsta varijable, odnosno vrsta atributa, u kojem će ispod atributa klase, odnosno ispod atributa meta-podataka biti to i naznačeno.

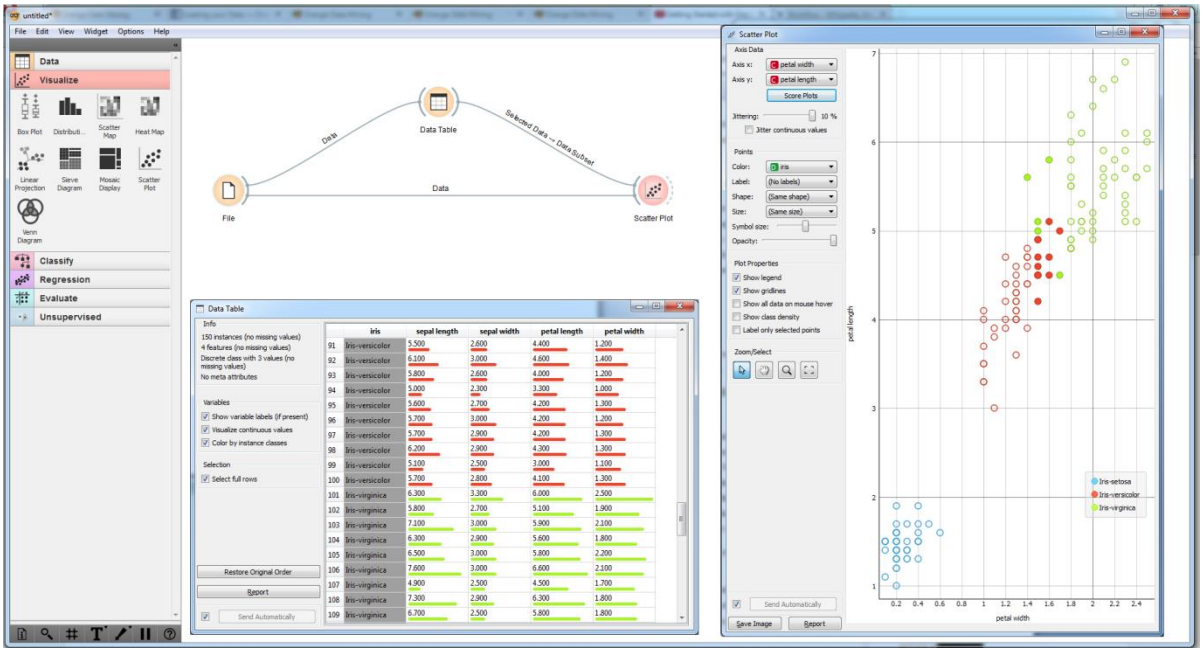


Slika 3.4. Prikaz klase i meta-podataka tablice s atributima i njihovim vrijednostima

3.3. Funkcionalnosti Orange-ovih alata za rudarenje podataka i njihovih komunikacijskih kanala

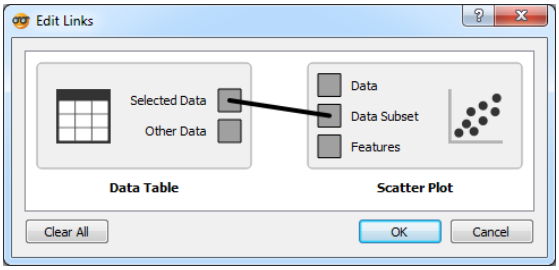
Radni tokovi (eng. *workflows*) se sastoje od skupa samostalnih nezavisnih alatnih jedinica organiziranih u radne jedinice (eng. *jobs*) povezane uzročno-posljedičnom vezom u radni proces koji tvori međuovisni model.

Prilikom konstrukcije željenog modela u *Orange* alatu dostupna je pomoć u obliku automatskog odabira kompatibilnih alatnih jedinica kako bi se olakšala izrada radnih jedinica i dizajn modela radnog toka. Korisniku se također nudi mogućnost određivanja tipa veze između alatnih jedinica, te i uvid u sami organizirani skup podataka koji će biti poslan iz te alatne jedinice dalje prema sljedećoj alatnoj jedinici kroz radnu jedinicu te dalje u druge radne jedinice radnog toka. Objektivan primjer ovakve primjene je prikaz odabranih instanci iz osnovnog skupa podataka unutar *„File“* alatne jedinice, kojeg će alatna jedinica podatkovne tablice (eng. *„Data Table“ widget*) tada slati prema alatnoj jedinici grafičkog prikaza podataka (eng. *„Scatter Polt“ widget*) koja će grafički istaknuti odabrane podatke za lakši prikaz relevantnih instanci, odnosno tako modelirani radni tok (eng. *workflow*) tvori vizualni podatkovni preglednik (eng. *Data Browser*). Također potrebno je napomenuti da se takav radni tok sastoji samo od jedne radne jedinice koja je sastavljena od tri alatne jedinice.



Slika 3.5. Prikaz modela vizualnog podatkovnog preglednika

Odabirom veze između alatnih jedinica podatkovne tablice i jedinici grafičkog prikaza podataka daje mogućnost odabira podataka koji će biti poslani prema sljedećoj alatnoj jedinici. *Orange* i u tom slučaju na osnovu predikcije tipa modela i odabrane instance podataka ili redoslijeda dodavanja računalnih jedinica na radnu površinu, ispravno zaključuje kako odabrani podaci iz inicijalnog skupa podataka su pripadajući podskup prikazanih podataka koje će prikazati.



Slika 3.6. Prikaz postavki veze između *„Data Table“* i *„Scatter Plot“* alatne jedinice vizualnog podatkovnog preglednika

Izmjena tipa veze između dvaju jedinica se direktno odražava na prikaz i prikazani podskup podataka.

3.4. Hijerarhijsko grupiranje

Uz spoznaju kako se podaci mogu dijeliti u pod-skupove, te kako se mogu vizualizirati, pruža se mogućnost grupiranja podatkovnih instanci u logičke skupove. Kako bi podjela u logičke skupove bila točna, bitan je ispravan odabir značajke podskupa podataka koji određuje pripadnost nekom skupu, odnosno odabir podatka koji je zajednički svim pod-skupovima. Otkrivanje skupova podataka, te njihove pripadajuće pod-skupove i ukoliko postoji hijerarhijska podjela pod-skupova u dodatne niže razine pod-skupova. Takva razdioba podataka u logičke skupove se naziva hijerarhijsko grupiranje (eng. *Hierarchical clustering*).

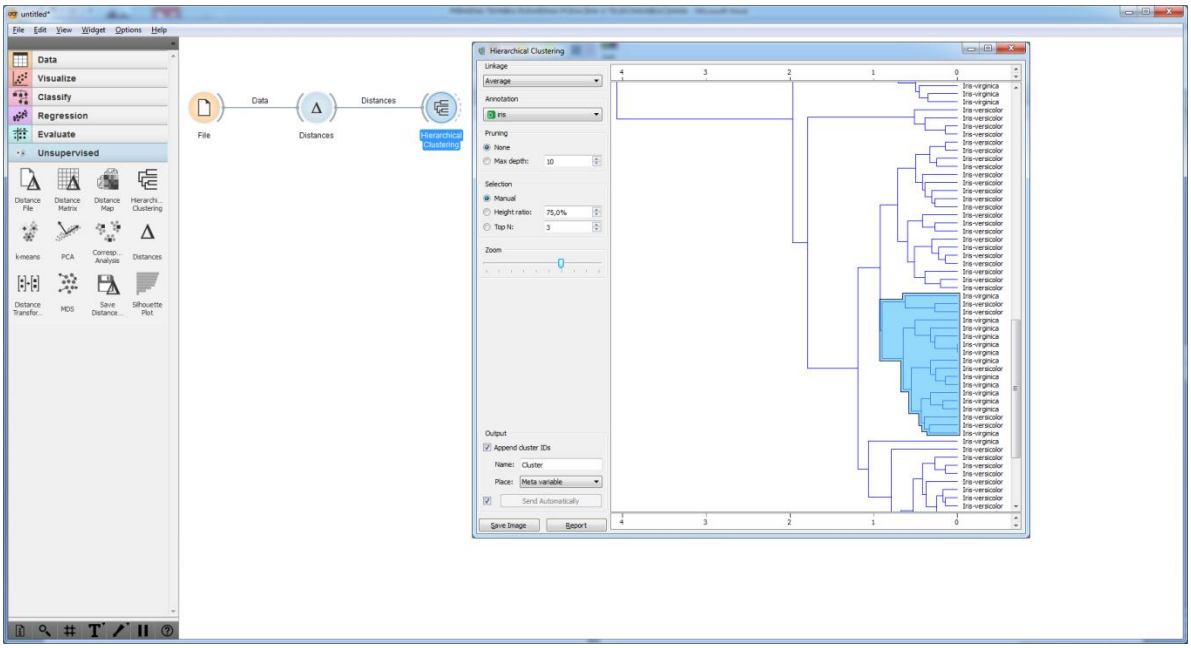
Prirodni skup su grupirani podaci koji dijele zajedničku varijablu u zajedničkoj grupi, npr. ukoliko se radi istraživanje ljudskog izgleda ovisno o geografskom položaju na kojem žive. Osnovni skup podataka sastoji se od njihovih obilježja, a kao zajedničke varijable prirodno se nameću boja kose, boja očiju, visina ili geografska pripadnost.

Prikazom vrijednosti u koordinatnom sustavu, te primjenom Euklidovog algoritma za razdiobu, izraz (3-1) čija je formula prikazana niže u tekstu, na skup varijabli udaljenosti između koordinata točaka učitanih podataka sadržanih u matrici udaljenosti, se vrlo lako prikazuje pripadnost podatka pojedinoj skupini.

$$d_{qp} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{3-1}$$

Hijerarhijsko grupiranje je moguće jednostavno prikazati povezivanjem *„Distances“* alatne jedinice za formiranje matrice i mjerenje udaljenosti između koordinatnih točaka s izvornim skupom podataka u *„File“* alatnoj jedinici, čime se dobiva informacija o grupiranim podacima, njihovim međusobnim udaljenostima u koordinatnom sustavu, te ekvivalentno udaljenostima informaciju o stvorenim prirodnim grupama podataka. Manjim razmakom ili kraćom udaljenosti između koordinata točke ili vrijednosti koja nosi informaciju znači da je veća sličnost ta dva podataka. U *„Distances“* alatnoj jedinici potrebno je podesiti parametar računanja udaljenosti između redova zadanog inicijalnog skupa podataka, te Euklidov algoritam računanja udaljenosti. Nadalje povezivanje *„Hierarchical Clustering“* alatne jedinice s *„Distances“* alatnom jedinicom otkriva dendogram koji i grafički prikazuje raspodjelu u prirodne grupe.

Dendogram je granati prikaz strukture hijerarhijski grupiranih podataka, te udaljenosti između prirodno grupiranih podataka. Odabirom ispravnog diskretnog atributa ili atributa s kategorijskom vrijednosti kao anotaciju vrijednosti prikazanih podataka u dendogramu, vrlo brzo je moguće uvidjeti pripadnost istoj prirodnoj grupi podataka, odnosno sličnost između podataka.

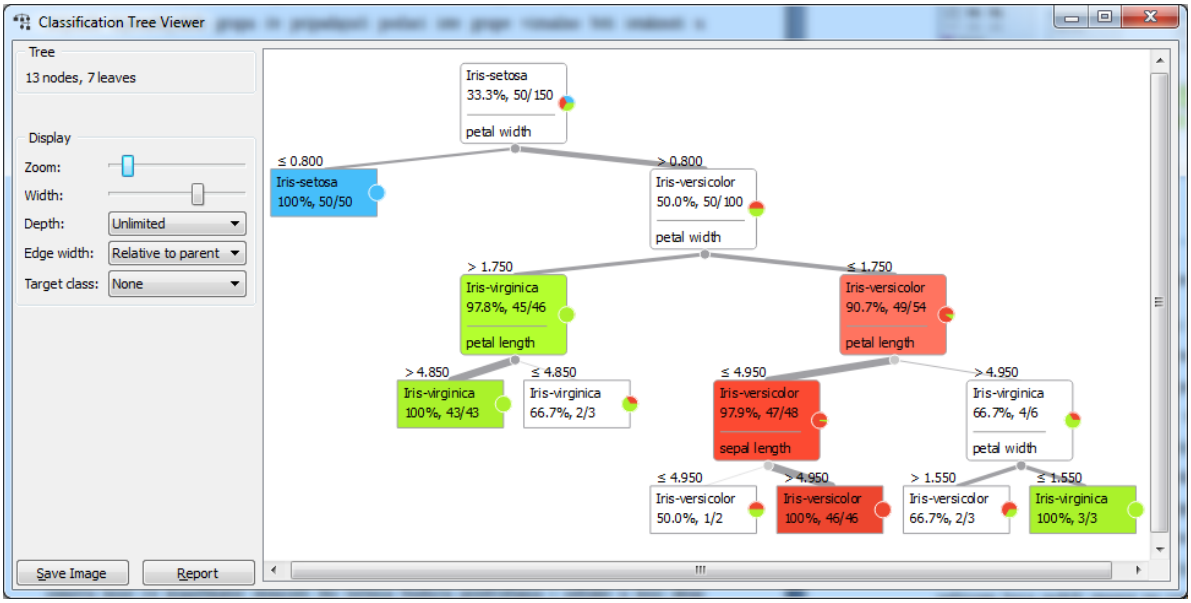


Slika 3.7. Prikaz modela hijerarhijskog grupiranja, te pripadajućeg dendograma *„Hierarchical Clustering“* alatne jedinice.

Ukoliko se u prikazu mogu primijetiti pogrešno grupirani podaci ili preklapanje između dvaju ili više grupa podataka, odabirom te hijerarhijske grupe se podaci šalju na izlazni kanal *„Hierarchical Clustering“* alatne jedinice na koju je potrebno povezati *„Data Table“* alatnu jedinicu kako bi se mogla točno uočiti razlika, odnosno približna sličnost odabranih podataka koji jednoznačno ne pripadaju hijerarhijskoj grupi u koju su smješteni. Nažalost ukoliko se radi o hijerarhijskoj grupi s većim brojem instanci, ponekad nije vrlo lako uočiti razlike, te je potrebno detaljnije razmotriti podatke. Vrlo jednostavna i učinkovita metoda je njihovim grafičkim prikazom, što je moguće izvesti povezivanjem *„Scatterplot“* alatne jedinice s izvornim skupom podataka u *„File“* alatnoj jedinici i s *„Hierarchical Clustering“* alatnom jedinicom. Otvaranjem *„Scatterplot“* i *„Hierarchical Clustering“* alatne jedinice i odabirom željenih hijerarhijskih grupa će pripadajući podaci iste grupe vizualno biti istaknuti u *„Scatterplot“* alatnoj jedinici. Tako organizirani model s automatiziranim hijerarhijskim grupiranjem, subjektivnom korekcijom korisnika i njegovim reprezentativnim prikazom podataka na izlaznom kanalu postaje model vizualnog podatkovnog preglednika za istraživanje hijerarhijskog grupiranja (eng. *Visual Data browser for Hierarchical clustering*).

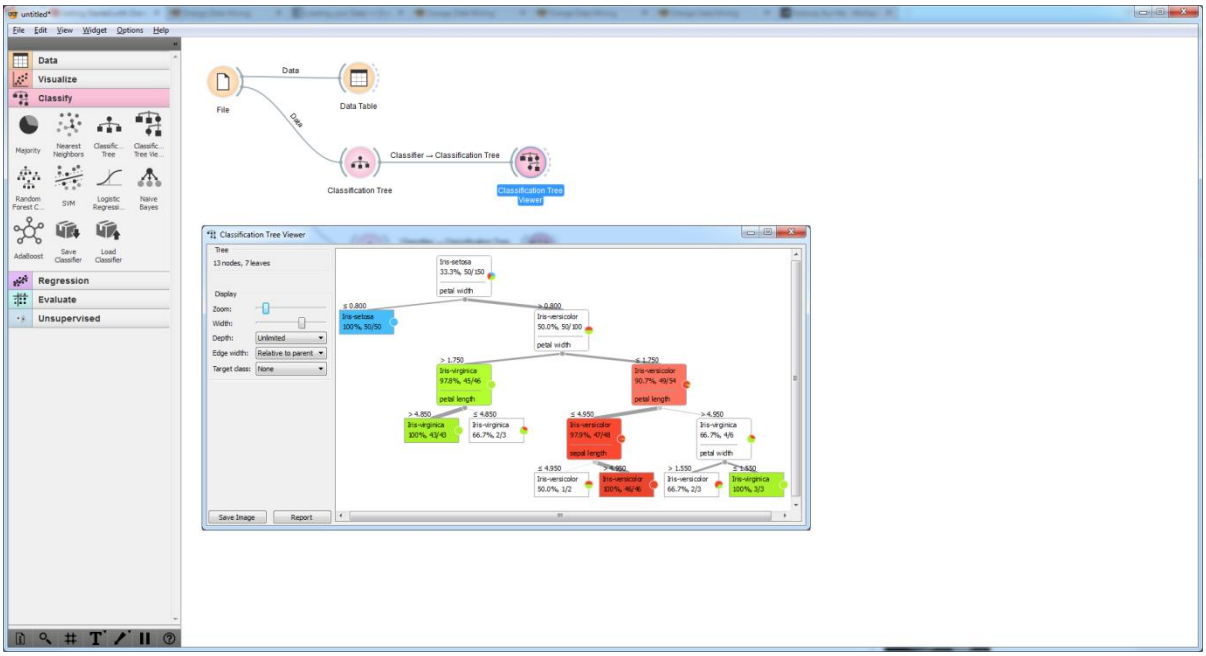
3.5. Klasifikacija i predviđanje na temelju podataka

Na temelju danih podataka moguće je napraviti njihovu klasifikaciju stablom odluke pomoću *„Classification Tree“* alatne jedinice predviđenom da bude klasifikator, odnosno izvršiti predviđanje oznaka klasa i instanci s određenom dozom sigurnosti unutar podatkovnog skupa. Prethodno klasifikaciji željenog skupa podataka, potrebno je omogućiti klasifikatoru period učenja ili ukoliko se radi o manjem podskupu od osnovnog skupa podataka, vrijeme treninga klasifikatora na osnovnom skupu podataka. Svaki podatkovni skup odlikuju njegove vrijednosti oznaka atributa na osnovu kojih će u novom setu podataka trenirani klasifikator donositi odluku stablom odluke kojem skupu pripadaju, te kojoj skupini zadana instanca pripada. Kako bi se napravio ispravan odabir oznaka atributa kojim se dobiva pouzdana informacija o razlici na osnovu koje će klasifikator donositi što točnija buduća predviđanja i odluke u koji skup pojedina instanca pripada se vidi pomoću *„Classification Tree Viewer“* alatne jedinice za prikaz stabla odluke.



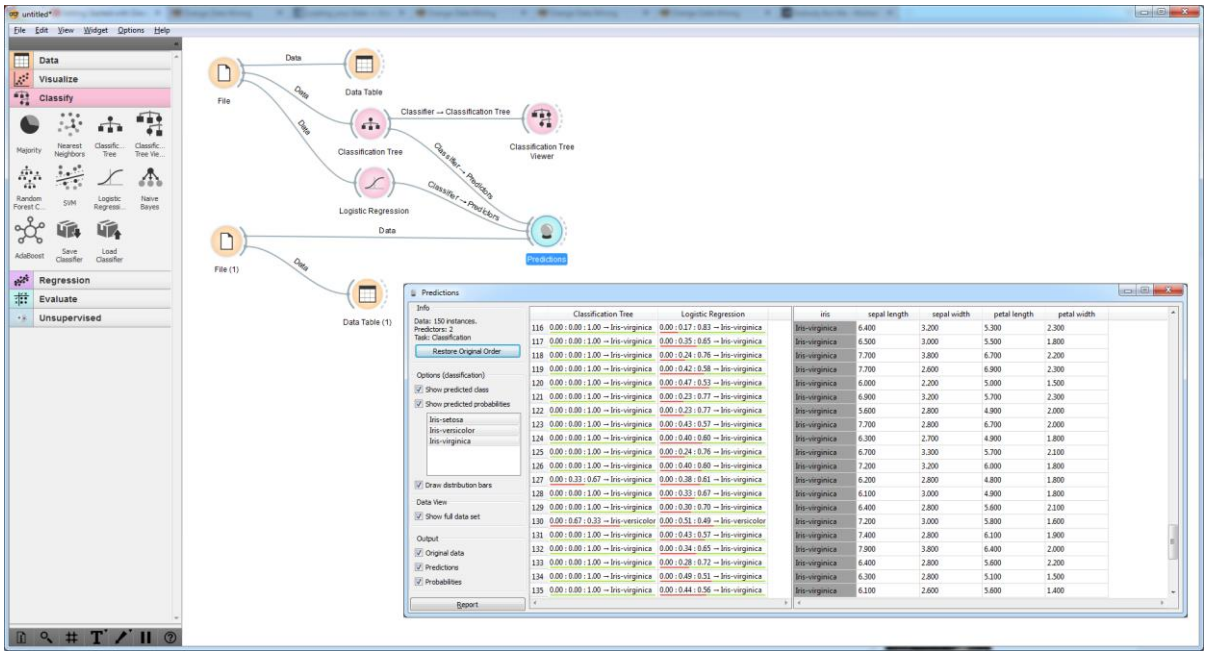
Slika 3.8. Prikaz strukture klasifikacijskog modela.

„Classification Tree Viewer“ alatna jedinica jasno prikazuje koji atributi i koje vrijednosti oznaka na osnovu njihovog grupiranja dijele inicijalni skup podataka u njegove jasne pod-skupove. Tako organizirani skup alatnih jedinica se zasebno može gledati ili se koristiti kao klasifikacijski model.



Slika 3.9. Prikaz strukture klasifikacijskog modela (6 video 1:04)

U dobiveni naučeni ili istrenirani klasifikacijski model je potrebno učitati prošireni skup podataka na kojem je klasifikator treniran s novim pod-skupom ili novo dodanim instancama, odnosno učitati novi sličan skup podataka, pri čemu je potrebno pripaziti da nazivi atributa oznaka budu jednaki kao i kod skupa podataka za treniranje klasifikatora kako bi ih klasifikator mogao ispravno primijeniti. Tako novi učitani skup podataka u *„File“* alatnoj jedinici potrebno je prvo pregledati s *„Data Table“* alatnom jedinicom da li su polja zadovoljavajuće popunjena i u ispravnom formatu. Nakon toga je potrebno „*Data Table“* alatnu jedinicu spojiti s *„Predictions“* alatnom jedinicom za predviđanje klase novih podatka, te joj dodati razvijeni klasifikacijski model s treniranom „*Classification Tree“* alatnom jedinicom koja sadrži znanje na osnovu kojeg će to i učiniti. Tako organizirani i povezani model naziva se modelom za predviđanje, a njegova funkcija predikcijska metoda ili metoda predviđanja.



Slika 3.10. Prikaz klasifikacijskog modela, te rezultata njegovog predviđanja

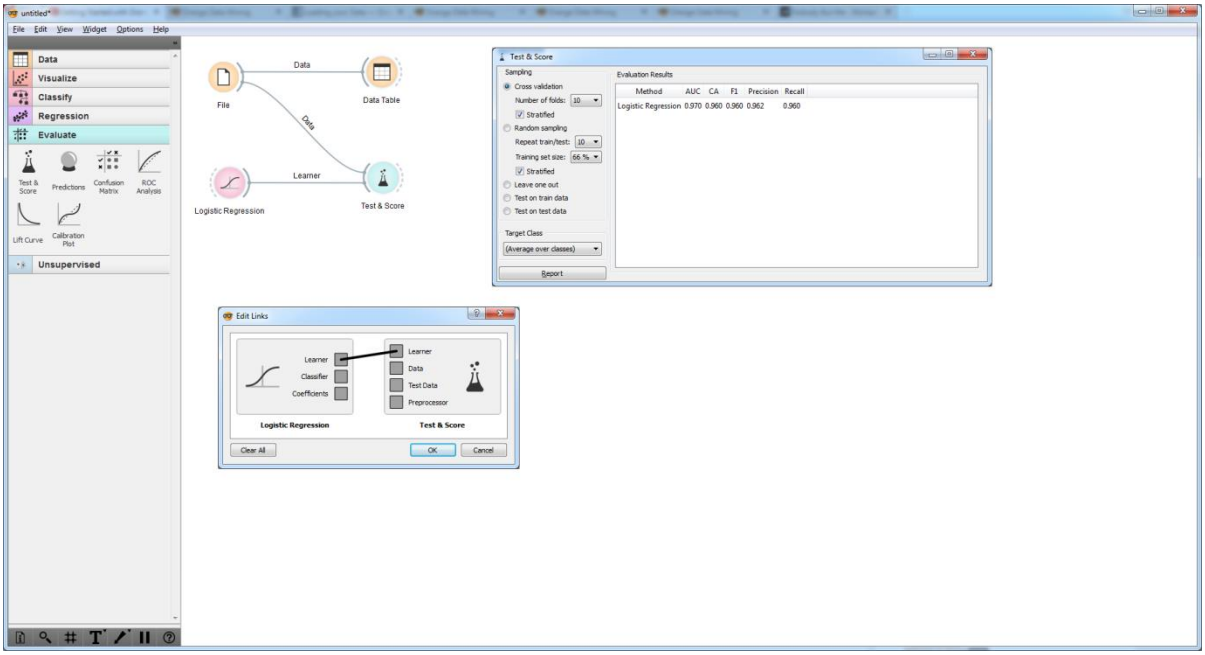
Dakako, moguće je i korištenje i drugih klasifikatora, jedan brz i jednostavan primjer klasifikatora bi bila logistička regresija, tj. *„Logistic Regression“* alatna jedinica koju je moguće također primijeniti dodatno i uz drugi postojeći klasifikacijski model kako bi se izbjegla eventualno postojeća kriva predviđanja u prijelomnim dijelovima ili se primijenila logistička regresija na instance koje se nalaze na graničnim dijelovima.

Nakon dovoljnog broja realnih instanci u inicijalnom podatkovnom skupu klasifikacijskog modela, te eventualno dodatno potpomognutim s relevantnim matematičkim modelom za predviđanje kao pomoć pri odlučivanju kako bi se izbjegle manje pogreške, moguće je donositi realna predviđanja novih instanci u stvarnom svijetu, a tako i njihova realna primjena.

3.6. Evaluacija i rangiranje predikcijskih metoda klasifikacijskih modela

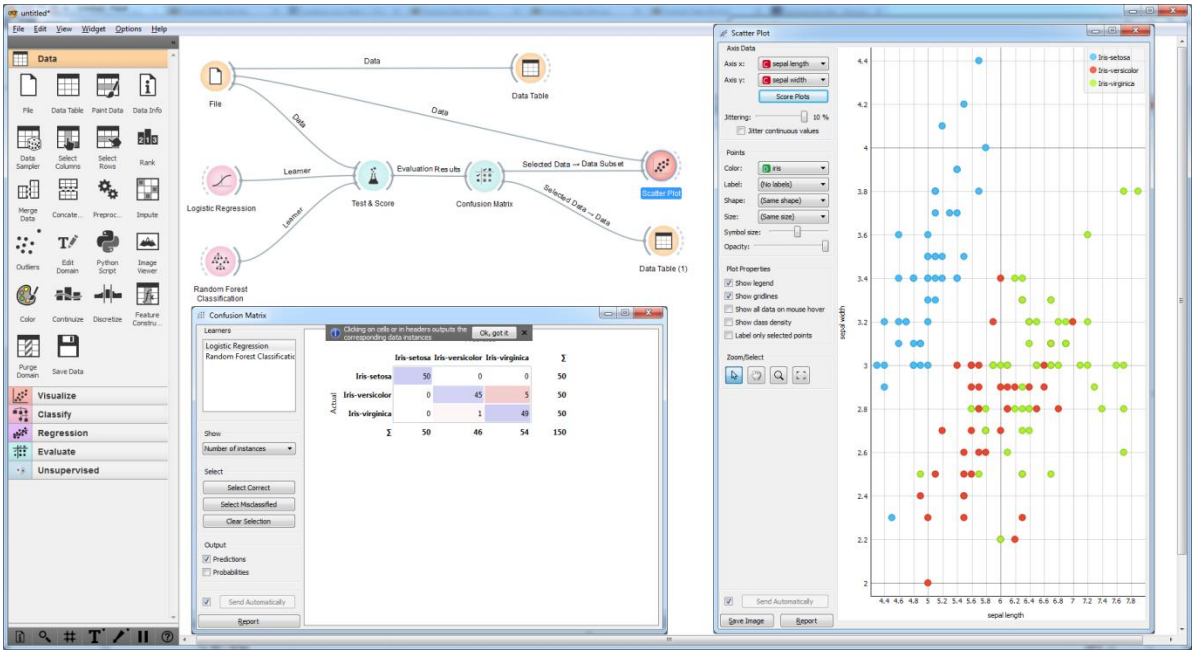
Dok se prošlo poglavlje bavilo razvojem pravilnog klasifikacijskog modela i njegovo povezivanje sa predikciskim dijelom modela, ovo poglavlje će se bazirati na procjeni kvalitete razvijenog klasifikacijskog modela kako bi cjelokupni predikcijski model mogao donijeti što točnije predviđanje.

Na prvom mjestu je potrebno kvalitetno treniranje klasifikatora, što je moguće vrlo jednostavno učiniti uz logističku regresiju, odnosno povezivanjem *„Logistic Regression“* alatne jedinice u tijek rada. Klasifikacijski model se gradi na podskupu podataka ili instanci koji služe za trening ili učenje klasifikatora, dok testiranje klasifikacijskog modela se vrši na drugom, zasebnom podatkovnom podskupu ili instanci kako bi zagarantirali različitost vrijednosti podataka podatkovnih pod-skupova i time ne bi prouzrokovali preklapanje vrijednosti podataka koje bi u konačnici dovele do krivih predviđanja predikcijskog modela. Takav proces treninga na osnovu konačnog broja instanci u podatkovnom skupu dijeli ih u dva dijela, od čega je podatkovni podskup za trening se sastoji od 90% inicijalnog skupa podataka, a test se vrši na preostalih 10% dijela inicijalnog skupa, te će takav ciklus ponavljati unutar podskupa dok je moguće testirati na dotad nepoznatim podacima. Takav proces unakrsne validacije je potrebno napraviti u što više ciklusa, te nakon svakog ciklusa zabilježiti točnost klasificiranja. Podatak točnosti kod unakrsne validacije klasifikatora na osnovu danih podataka je moguće pronaći u „Test & Score“ alatnoj jedinici pod oznakom „CA“ koja predstavlja točnost klasifikacije (*eng. Clasification Acurasy*) i predstavlja broj točno klasificiranih podatkovnih instanci iz testnog podskupa danih podatka.



Slika 3.11. Prikaz modela unakrsne klasifikacije, te relevantnih postavki i prikaz rezultata

Ukoliko točnost klasifikacije nije 1, odnosno 100%, moguće je provjeriti eventualnu krivu klasifikaciju ili područje krive klasifikacije unutar podatkovnog skupa s *„Confusion Matrix“* alatnom jedinicom. Pregledom alatne jedinice *„Confusion Matrix*“ moguće je vidjet kojem inicijalno učitanom skupu podataka pripadaju krivo klasificirane instance, te ih odabrati klikom na *„Misclassified“* s lijeve strane izbornika alatne jedinice koji će ih tada poslati na izlazni kanal iste. Nakon čega je moguće povezati s *„Data Table“* alatnom jedinicom koju tada dalje je potrebno povezati s *„Scatter Plot“* alatnom jedinicom povezanu također na ulaznom kanalu s cjelovitim podatkovnim skupom na kojem je izvršen trening klasifikatora kako bi bila moguća jasna vizualizacija pogrešno klasificiranih instanci.

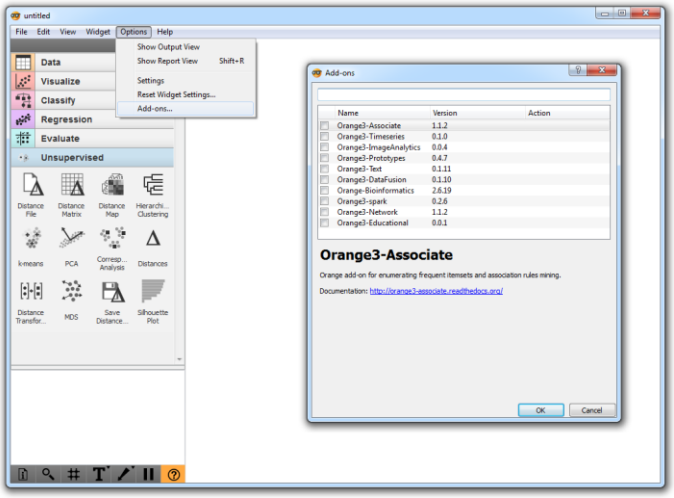


Slika 3.12. Prikaz klasifikacijskog modela

Daljnje moguće poboljšanje klasifikacijskog modela je dodavanjem drugih klasifikacijskih alatnih jedinica iz *„Classify“* skupine alatnih jedinica u klasifikacijski model kao i njihova kombinacija ovisno o predmetu klasifikacije ili istraživanja.

3.7. Specijalizirani alati rudarenja podataka

Ukoliko postoji potreba usmjerenog istraživanja određenog znanstvenog područja ili specijalizirane znanosti, kao npr. bio-informatike, proučavanja teksta ili biometriju, odnosno analizu slike i slične potrebe, tada mogućnost korištenja specijaliziranih dodataka je ključna. Orange alat nudi mogućnost korištenja specijaliziranih alata, a njih je moguće naći u glavnom izborniku pod *„Options“,* te *„Add-ons…“* kako je prikazano i na slici niže.



Slika 3.13. Prikaz glavnog izbornika u koji sadrži alate za specijalizirana istraživanja

Potrebno je također napomenuti kako je nakon odabira željenog dodatka i njegove instalacije potrebno je ponovno pokrenuti *Orange* alat kako bi se nove alatne jedinice uspješno primijenile unutar alata za korištenje.

Vrlo korisna alatna jedinica je *„GEO Data Sets“* koja omogućava pristup velikom repozitoriju prikupljenih podataka putem anketa ne ciljanog tipa u svrhu rudarenju podataka i statističkog proučavanja.

4. PRIMJERI PRIMJENE TEHNIKA I METODA RUDARENJA U TELEKOMUNIKACIJAMA

Zadatak rudarenja podataka u području upravljanja odnosa s korisnicima temelji se na pretpostavci da podaci iz prošlih događaja sadrže informacije o budućim događajima. Taj zaključak proizlazi iz tog da korisničko ponašanje nije niz slučajnih događaja već kompleksni uzorak održavanja navika, potreba i sklonosti, te međusobni odnos između korisnika i pružatelja telekomunikacijskih usluga. Cilj je pronalaženje tih uzoraka koji jasno prikazuju korisnikove navike i potrebe, te sklonosti zbog kojih korisnik koristi pružene usluge. Upravo u tom dijelu dolazi se do prvih prepreka postavljenih pred rudarenje podataka. Radi se o labilnim podacima upućenim od strane korisnika koji su nedovoljno činjenično konkretni, a navike i potrebe korisnika nisu posve jasne, dok su čak u nekim slučajevima zbunjujuće jer se šalju kontradiktorne informacije. Prilikom prikupljanja inicijalnog skupa sirovih podataka proces rudarenja podataka ima glavnu zadaću pročišćavanja bitnih od nebitnih ili manje bitnih podataka, te jasno odvajanje od nekonzistentnosti ili jednostavnije mogućnost prepoznavanja osnovnih uzoraka i njihovih varijacija.

Rudarenje podataka značajni porast primjene dobiva porastom prostora za pohranu podataka s povećanja kapaciteta tvrdih diskova, te povećanjem rasprostranjenosti podatkovnih centara (eng. *data center*), kao i eksponencijalnim rastom procesorske moći za obradu podataka. Osim razvoja u pogledu fizičke opreme, značajnu primjenu rudarenje podataka dobiva i širim rasprostiranjem pristupačnih alata za izvještavanje poslovnih subjekata kojima prirodni korak je u smjeru dubinskih analiza svojih poslovnih podataka ili analiza na razini korporacije kako bi kompanija napredovala u konkurentnosti sa svojim uslugama i proizvodima te si osigurala veći udio na tržištu.

Telekomunikacijska industrija postaje svakim danom sve utjecajnija grana industrije koja duguje svoj uspjeh ubrzanom razvoju novih komunikacijskih tehnologija kao i povećanju broja uređaja na tržištu predviđenih za korištenje istih. Povećanjem utjecajnosti telekomunikacijske grane industrije ekvivalentno donosi kao posljedicu i povećanje konkurentnosti u borbi za zaradom i prostorom za daljnjim širenjem poslovanja, te tako daljnjim povećanjem zarade i nastavkom daljnjeg svog širenja. Stoga nastaje vrlo velika potražnja za rudarenjem podataka koja bi identificirala uzorke u korištenju usluga, korisničkim preferencijama, pronalasku potražnje za novim ponudama

usluga i prilagodbi cijena postojećih usluga. Analiziranjem podataka se također dolazi do boljeg shvaćanja samih poslovnih procesa, detekcijom propusta, sprječavanjem prijevara i cjelokupnim poboljšanjem usluge.

Kao osnovna prepreka svih telekomunikacijskih pružatelja usluga je zadržavanje korisnika u svojoj mreži i stvaranje konstantne zainteresiranosti korisnika uslugama. Zadržavanje korisnika je znatno jeftinije nego li pridobivanje novih korisnika koji okvirno nepisanim pravilima iznosi 1:4 [9], odnosno za svakog novog pridobivenog korisnika moguće je zadržati 4 postojeća korisnika u svojoj mreži bez dodatnih troškova. Rudarenjem podataka se dolazi do spoznaje rizičnosti odlaska korisnika drugom pružatelju usluge te je moguće uložiti dodatne napore i resurse pri njegovom zadržavanju. Moguće je razlikovati, korisnike koji će biti zadržani bez ulaganja dodatnih resursa te korisnike koji su teret korporativno gledano i treba im dopustiti odlazak drugom pružatelju usluge ili isključenje usluge koju koristi. U nabrojanim scenarijima rudarenje podataka dolazi do izražaja u modelima za predviđanje korisničkog ponašanja, kao što je distinktno predviđanje koji je korisnik potencijalni kandidat za odlazak, koja je vjerojatnost odlaska, vremenski period u kojem će korisnik napraviti takav korak, razlog odlaska te procjenu dobiti ukoliko se takav korisnik zadrži ili izgubi. Analizom i predikcijom korisnicima se nude pogodnosti za ostanak u mreži, razmatranjem razloga potencijalnog odlaska, a korigiranjem usluga smanjiti postotak odlazaka iz mreže i povećanja zadovoljstva korisnika. Procjenom dobiti se kalibriraju pogodnosti kojima je moguće zadržati i ostaviti pozitivan dojam na korisnike. Svakako je potrebno odstraniti i negativnu stavku nezadovoljstva korisnika koje nije moguće zadržati, imaju malu vrijednost ili generiraju veću količinu troškova nego dobit. To su korisnici koji troše resurse koje bi bilo moguće usmjeriti na zahtjevnije zadatke ili opremu koja može podržavati veći broj korisnika, a time i povećati dobit, odnosno oni korisnici koji ostvarenom pogodnosti ne ispunjavaju očekivanja operatora kroz naplatu. Osim dobiti istim načinom detektiranja ili predviđanjem se mogu odrediti korisnici koji su manje zahtjevni, odnosno strpljiviji korisnici kojima je moguće produljiti vrijeme čekanja na korisničku podršku ili tehničku intervenciju. Posljednji tip korisnika u telekomunikacijskoj industriji su oni korisnici kojima je primarni cilj izbjegavanje plaćanja usluge ili neka druga vrsta ostvarivanja benefita bez poštivanja sklopljenih ugovora ili obveza s njihove strane. Takvi korisnici se najčešće predstavljaju lažnim podacima, krivom identifikacijom, a tradicionalno su bili otkriveni pomoću neplaćenih

nekoliko računa za vrijeme korištenja usluga. Rudarenjem podataka takvu vrstu korisnika koji su već ranije imali prijestup tog ili sličnog oblika i proučavanjem njihovih podataka je moguće unaprijed uz određenu vjerojatnost spriječiti u ponavljanju istog ili bar umanjiti nastalu štetu.

Osim analize korisnika rudarenje podataka se koristi prilikom provođenja raznih kampanja, omjerima uloženih resursa i dobivenih rezultata ili smanjenju resursa bez značajnog utjecaja na kampanju. To dovodi do uvođenja novih usluga na tržište i njihovu prezentaciju, koja nerijetko može biti napadna,a ne nužno korisna jer se određeni skup korisnika opire konstantnom marketingu nebrojenim reklamnim metodama. Tu rudarenje nudi segmentaciju korisnika kako bi se segmentirali u skupove korisnika zajedničkih značajki te usmjereno djelovalo na njih. Usmjerenim djelovanjem se prezentiraju određene usluge određenim skupovima korisnika. Na osnovu prethodnih kampanja moguće je odrediti također i vjerojatnost odziva korisnika na ponuđenu uslugu te time smanjiti troškove same kampanje.

4.1. Metode rudarenja podataka u telekomunikacijama

Alate ili metode je moguće podijeliti u osnovne skupine primjene kao što su alati za učitavanje i sadržavanje skupa podataka (npr. razne vrste tablica s predefiniranim primjenama, osnovnim i dodatnim ključevima).

Klasifikacija je raspodjela nestrukturiranih podataka u čvrsto definirane klase koje su definirane na osnovu tih nestrukturiranih podataka. Nameće se kao vrlo čest i koristan postupak u rudarenju podataka, klasifikacijski alati za grupiranje podataka, regresijski alati za procjenu i oblik povezanosti podataka, te alati za vizualni prikaz u obliku grafova i dijagrama.

Regresija i razvrstavanje je najzahtjevnija tehnika prilikom razvoja modela pomoću skupova alata koji bi mogao regresijom previđati vrijednosti, odnosno razvrstavanjem predviđati kojoj klasi bi pojedina vrijednost pripadala. Primjena alata tih vrsta je uvelike prisutna kod predviđanja budućih troškova ili predviđanja rizika tržišta, te samog trajanja isplativosti korisnika. U neke tehnike klasifikacije i regresije uključena su i stabla odluke, neuronske mreže.

Redoslijed i uparivanje alata određuje opisne modele pomoću kojih se tvore pravila po kojim se buduće varijable uparuju. Primjenom spomenutih alata u metodama regresije i razvrstavanja te primjena istih u određenom redoslijedu i kod uparivanja dobiva se osnova za razvoj predikcijskih modela. Iz skupova alata unutar takvih modela proizlaze omjerne varijable korelacija između dvaju ili više faktora. Primjer bi bili korisnici usluga ograničenih paketa Internet prometa kao što su 1 GB, 15 GB i sl. Oni vrlo često nisu zainteresirani za proširenjem korisničke lepeze usluga, stoga je tim korisnicima potrebno ponuditi mogućnost prelaska na neograničeni paket internet prometa te ukoliko prihvate ponudu, tek u tom slučaju ih uključiti u kampanju s ponudom dodatnih usluga.

Grupiranje je tehnika kojoj je zadaća grupiranje sličnih entiteta zajedničkih obilježja. Kod ove vrste tehnike metoda pronalaska korelacija se isključivo zasniva na udaljenosti dvije točke jedne od druge, odnosno udaljenosti od određene referentne točke pa je vrlo svojstvena primijenjenom algoritmu. Stoga ukoliko se radi o neutvrđenom skupu ili skupu s većim brojem podskupova, moglo bi se reći da je i nužno koristiti više od jednog algoritma. Između dva algoritma kod kojih postoji mogućnost da će doći do različitih točaka grupiranja i parametara povezivanja entiteta također je potreban stručni ljudski faktor koji ima mogućnost razlučivanja i prosudbe da li su novo nastali grupirani pod-skupovi korisni ili ne. Ukoliko su rezultati proizašli iz dva različita modela ili algoritma isti, takva provjera potvrđuje točnost grupiranja i zaključaka proizašlog iz njih. Uvelike korisna odjelima marketinga i prodaje koji iz izvještaja izvučenih pomoću metode ove vrste i metoda njoj sličnih određuju korisnike zajedničkih obilježja i sličnih interesa, te pogodne korisnike za ciljanu kampanju.

4.2. Tehnike rudarenja podataka u telekomunikacijama

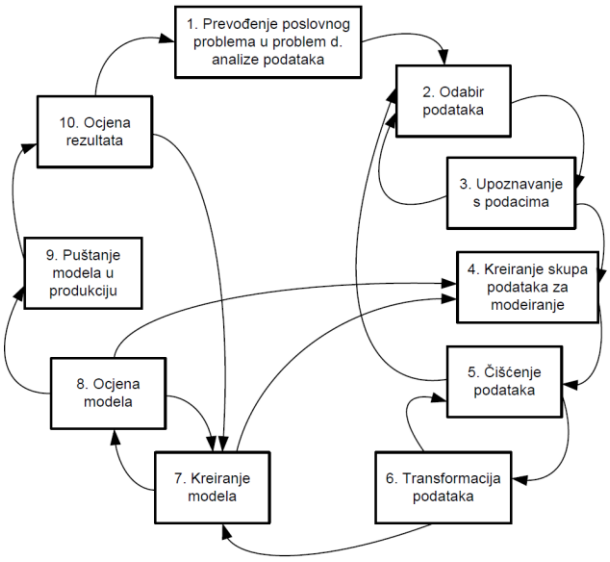
Postupci spomenuti u prethodnom poglavlju se najčešće koriste kod modeliranja pomoću tehnika rudarenja podataka kao što su stabla odluke, tehnika grupiranja, tj. najmanje udaljenosti ili tehnike najbližih susjeda, tehnici analize mreža, a naravno primjenjivi su i u tehnici neuronskih mreža, te u tehnikama analize veza (eng. *Link analysis*). U telekomunikacijskoj industriji prilikom potpisivanja novih ili produljenja postojećih ugovora korisnik dobiva pogodnost u obliku umanjene cijene mobilnog uređaja. U toj prilici pružatelj telekomunikacijskih usluga svjesno se stavlja u nepovoljan položaj, ali kroz period trajanja ugovorne obveze korisnik koristi usluge pružatelja usluge te redovno plaća iznos tarife na koju je korisnik potpisao ugovor, odnosno jednako tako otplaćuje vrijednost uređaja koji je dobio po sniženoj cijeni. Klasifikacija telekomunikacijskih korisnika u visoki, srednji ili niski segment rizičnosti korisnika omogućuje prilagodbu ponude usluga prilikom potpisivanja ugovora na prikladnu vrijednost čime pružatelj usluge smanjuje rizik.

Grupiranje prema zajedničkim obilježjima ili pravilima pridruživanja (eng. *Association rules*), kako i sama riječ kaže, povezuje zajedničke atribute. Ukoliko dva ili više produkata imaju zajednički ili sličan atribut bit će grupirani.

Utvrđivanje uspjeha poslovanja se zasniva na količini korištenja ponuđenih usluga, odnosno na grupiranju količina različitih usluga koje korisnici koriste. Jedan način je otkrivanje koje usluge pojedini korisnik koristi u kombinaciji s drugim uslugama, odnosno stvaranje šire slike korištenja usluga. Prilikom određivanja faktora korištenja usluga neizbježno je korištenje rudarenja podataka, te primjena tehnike redoslijeda i uparivanja za analizu podataka kako bi se utvrdili obrasci. Izlazne varijable ove vrste analize bi bili omjerne varijable, a nosit će informaciju korelacije između usluga koje korisnici koriste. Alati uparivanja će definirati pravila po kojim će buduće varijable biti grupirane. Takav tip analize pomoću tehnike redoslijeda i uparivanja se popularno zove analiza potrošačke košarice (eng. *Market basket analysis*). Analiza potrošačkih košarica se primjenjuje u telekomunikacijskoj industriji kod dodjeljivanja popusta, kreiranja promocija i kampanja za ponudu. Nakon završene analize i kreirane nove ponude, tehnikama regresije i razvrstavanja se može predvidjeti njena uspješnost, a samim time i opravdanost njenog pokretanja. Razvrstavanje daje odgovore na segmentiranje korisnika kao što su visoka,srednja ili niska rizičnost korisnika, je li korisnik kombinirao dvije

usluge ili nije. Dok se regresijom dobivaju kontinuirane varijable i koristi se za slučajeve gdje se izlazna varijabla može predvidjeti za neograničeni broj vrijednosti kao što su "isplativost po pojedinog korisnika", odnosno predviđanje zarade od njega. U poslovanju pravila pridruživanja uz analizu potrošačke košarice definiraju uzročno posljedičnu vezu između dva ili više atributa. Kod telekomunikacijskih usluga ona se koristi prilikom širenja ponude usluga korisniku s novim uslugama ili detekciju potencijalnog uzroka greške.

Kako je na slici (Sl. 4.1.) je prikazan postupak rudarenja podataka je transparentnije i najbolje prikazati kao skup petlji koje su međusobno povezane te njihovi koraci kroz postupak imaju svoj redoslijed. Iako postoji redoslijed, sami koraci započinju prije nego je prethodni korak u potpunosti završen. Takvom odradom se dobiva revidirajuće stanje koraka postupka koje činjenice naučene u kasnijim koracima primjenjuje na prethodne korake i tako ih revidira.



Slika 4.1. Postupak rudarenja podataka

5. PRAKTIČAN PRIMJER

Kako je već objašnjeno u prethodnim poglavljima, rudarenje podataka (eng. *Data mining*) je postupak dobivanja određenih relevantnih i korisnih informacija nakon izvršenog sortiranja, organiziranja te grupiranja velikog broja podataka i njihovih korelacija, te prepoznavanja uzoraka koje oni tvore. Postupak se sastoji od više koraka koje je kroz proces potrebno više puta ponavljati i ne preskakati zbog mogućeg kompromitiranja zaključaka. Uz razumijevanje procesa, također je potrebno dobro razumjeti zadatak poslovnog zahtjeva koji se nameće kao početni uvjet, te odabir postupka rudarenja podataka. Oni mogu biti usmjereni odnosno nadgledani (eng. *supervised*) postupak rudarenja podataka ili neusmjereni odnosno nenadgledani (eng. *unsupervised*) postupak rudarenja podataka.

Tako se usmjerenim postupkom rudarenja podataka prikazuju odnosi između izlazne varijable i ulaznog skupa podataka prilikom klasifikacije ulaznog skupa podataka, procjene kvalitete klasifikacijskih modela na osnovu danih izlaznih varijabli, te provedbe predviđanja iz njih.

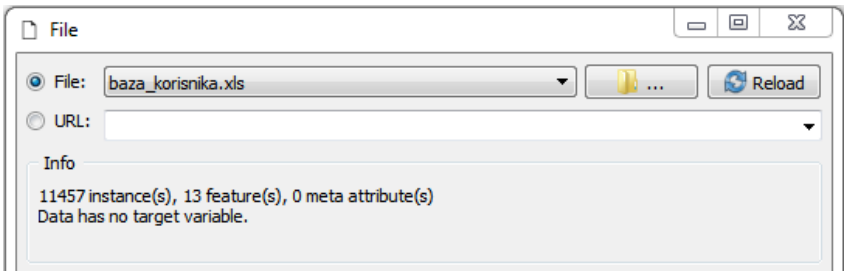
Kod neusmjerenog postupka rudarenja podataka rezultat ne postoji, on se bavi unutarnjom korelacijom i strukturalnim vezama unutar ulaznog skupa podataka kako bi ih prikazao u obliku grupiranja, odnosno definirao pod-skupove kako bi bili jasnije vidljivi pomoću hijerarhijskog grupiranja u logičke skupove (eng. *hierarhical clustering*), grupiranja prema zajedničkim obilježjima (eng. *affinity grouping*), te profiliranja.

U slijedećim poglavljima ovog rada cilj je prikazati praktičnu primjenu procesa, te tehnika i metoda rudarenja podataka u svrhu profiliranja tipova korisnika telekomunikacijskih usluga ovisno o uslugama koje koriste primjenom Orange alata.

5.1. Priprema podataka za obradu

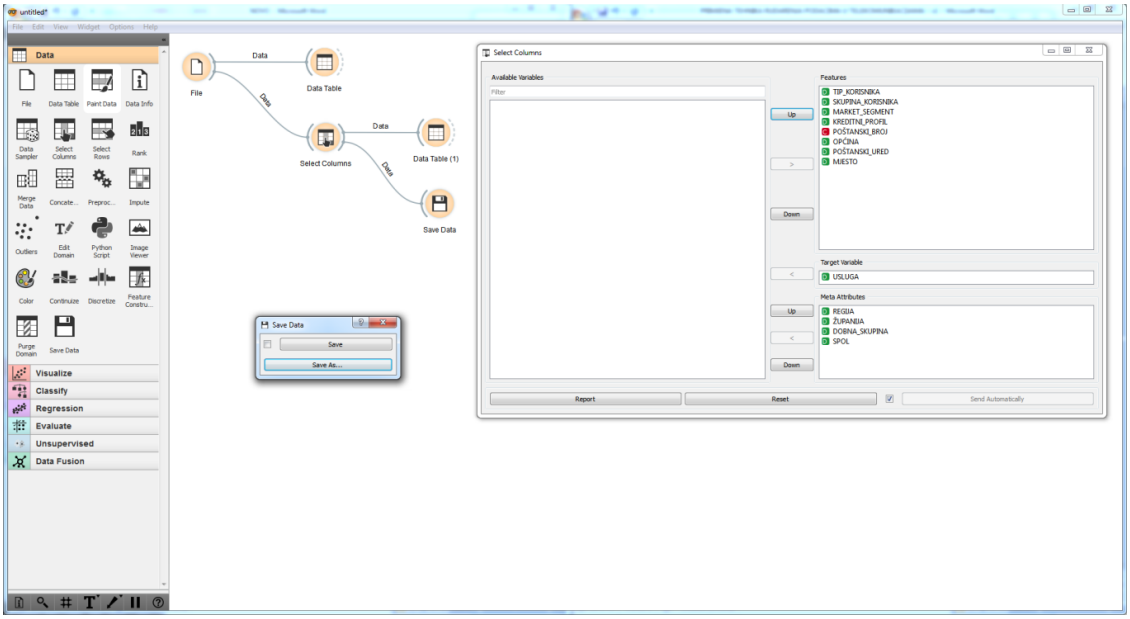
Priprema podataka kako bi bili pogodni za korištenje, te da bi ih alati mogli učitati možda je jedan od najvažnijih zadatka u rudarenju podataka. *Orange* alat ima mogućnost čitanja predodređenih formata dokumenata kao što su .tab koji je također i nativni format datoteka u *Orangeu*, .xls/.xlsx, te tekstualne .txt datoteke odvojene zarezom. Prije obrade bilo kojeg nestrukturiranog skupa podataka potrebno ga je pravilno pripremiti i prilagoditi za obradu. Uobičajeno strukturiranje podataka unutar dokumenta je u obliku tabličnog prikaza gdje se u stupcima nalaze atributi, dok se u redovima nalaze njihove vrijednosti. *Orange* po učitanju podataka izvršava automatsko strukturiranje na osnovu vrijednosti učitanih podataka i postavlja pretpostavku kako bi mogao definirati meta-podatak i varijablu klase. Pretpostavka alata prilikom definiranja meta-atributa i varijable klase nije uvijek točna, već odabire najpogodnije za tu zadaću, što ne mora uvijek biti i cilju poslovnog zadatka.

Baza podataka s informacijama o korisnicima preuzete su iz javno dostupnih izvješća i anketa Hrvatske regulatorne agencije za mrežne djelatnosti [10], te iz izvješća globalnih geografskih ICT podataka za Republiku Hrvatsku (eng. *Global and Regional ICT data*) ITU-Agencije UN za informacije i komunikacije [11]. Prvenstveno nestrukturirane podatke na kojima se želi izvršiti analiza potrebno je spremiti u pogodan format kao što su .xls/.xlsx, te kratkim pregledom potvrditi da su tekstualne, numeričke i postotne vrijednosti u ispravnim veličinama i s ispravno postavljenim decimalnim mjestima, te s dozvoljenim dijakritičkim oznakama. Nakon provjere podataka tako spremljeni dokument se učitava u *Orange* alat kroz *„File“* alatnu jedinicu odabirom iz lokalno spremljene datoteke ili URL lokacije ukoliko se radi o javno dostupnom dokumentu na mreži ili internetu. Ukoliko *Orange* nije pretpostavio ispravno meta-atribute i varijablu klase (Sl. 5.1.), podatke je moguće ispraviti lokalno ili pomoću *„Selected Columns“* alatne jedinice kako je opisano u poglavlju 3.2. ovog rada.



Slika 5.1. Prikaz *„File“* i neispravno određenim meta-atributima i varijablom klase tablice

Kako u praktičnom primjeru nisu eksplicitno određeni meta-atributi i varijabla klase, potrebno ih je odrediti individualno prema poslovnom zadatku kroz *Orange* alat. Potrebno je razviti model za ispravak učitane datoteke (Sl. 5.2.).

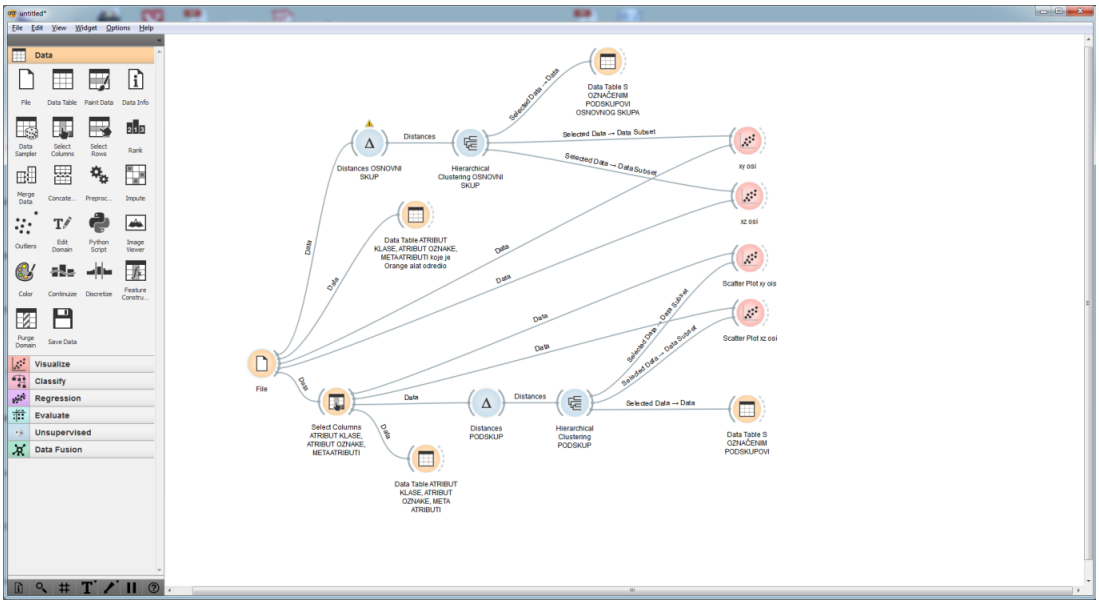


Slika 5.2. Prikaz modela za ispravljanje učitane datoteke praktičnog primjera

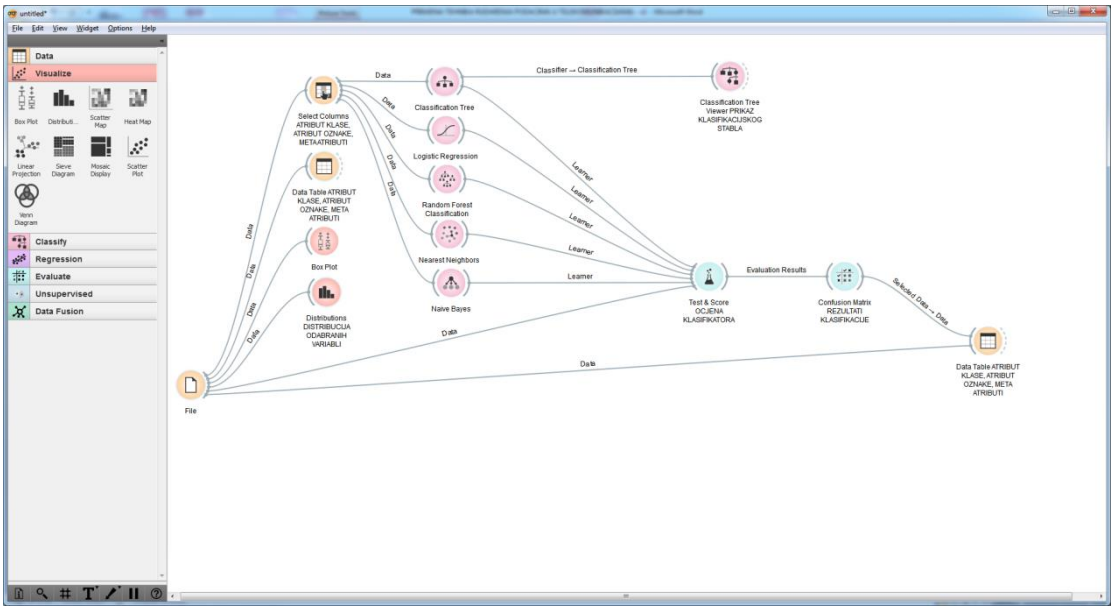
Poslovni zadatak nalaže profiliranje tipova korisnika telekomunikacijskih usluga ovisno o uslugama koje koriste, stoga je kao ciljanu varijablu klase (eng. *Target Variable*) na osnovu koje će biti izvršena klasifikacija korisnika postavljen atribut „USLUGA“ tekstualnog tipa koji sadrži tipove usluga kao vrijednosti koje korisnici koriste. Kao meta-atribute (eng. *Meta Attributes*) prikladno je odabrati diskretne atribute obilježja korisnika po kojim će biti definirana klasifikacija kao što su „SPOL“, „DOBNA_SKUPINA“, „REGIJA“ i „ŽUPANIJA“ koji sadrže kategorijske vrijednosti. Nakon završenog strukturiranja podataka određivanjem relevantnih atributa tako pripremljen skup je potrebno spremiti lokalno kako bi se mogao ponovno učitati u klasifikacijskom modelu. Spremanje dokumenta s definiranim atributima najpogodnije je pohraniti u predefiniranom (eng. *native*) .tab formatu *Orange* alata.

Za grupiranje korisnika u ovisnosti o uslugama koje koriste korišteni skup podataka dobiven je iz većeg broja godišnjih izvješća Hrvatske regulatorne agencije za mrežne djelatnosti [10] i globalnih ITU izvješća za Republiku Hrvatsku [11]. Također u navedenom skupu podataka potrebno je individualna odredba varijable klase i meta-atributa. Obrađen je

isti skup podataka u dva slučaja, od čega je u prvom slučaju ciljana varijabla klase atribut „POPULACIJA“ s meta-atributima koji su direktno vezani uz korištenje usluga po pojedinačnom broju korisnika. U drugom slučaju je korišten kao ciljana varijabla klase atribut „KUĆANSTVA“ te meta-atributi vezani uz raspodjelu usluga po kućanstvu te koeficijentu korištenja pojedinih usluga na 100 stanovnika.



Slika 5.3. Prikaz modela hijerarhijskog grupiranja osnovnog skupa i pod-skupova s individualno definiranim atributima klase

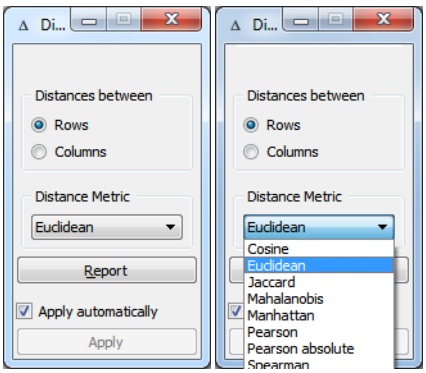


Slika 5.4. Prikaz modela klasifikacije skupa u ovisnosti o odabranim atributima klase

5.2. Grupiranje korisnika u ovisnosti o uslugama koje koriste

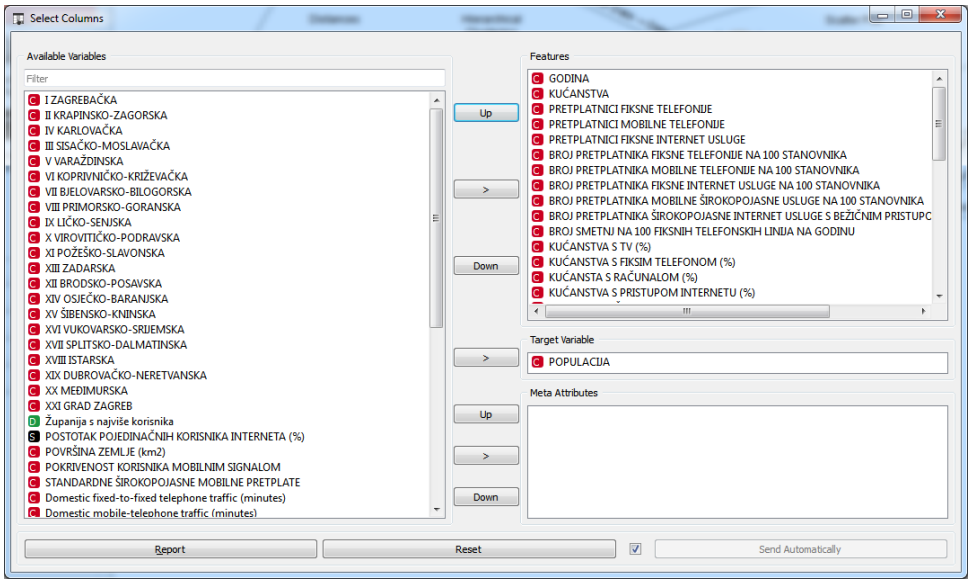
Kao poslovni zahtjev ovog rada je postavljen zadatak grupiranja i klasifikacije korisnika u ovisnosti o uslugama koje koriste. Kroz naslovljeno poglavlje će biti obrađen dio zadatka koji se bavi grupiranjem korisnika u grupe, podgrupe, te podgrupe definirane i opisane višim razinama grupiranja i razdiobe. Proces grupiranja korisnika je potpunjen grafičkim prikazima grupiranja, opisima alatnih jedinica, postupaka i obrada, kao i razvojni tijek rada na temelju učitanih podataka. Grupiranje se određuje mjerenjem udaljenosti između vrijednosti, te njihovo grupiranje provodi na osnovu mjera u pripadajuće skupine. Ukoliko su rezultati proizašli iz dva različita modela ili algoritma isti, takva provjera potvrđuje točnost grupiranja i zaključaka proizašlih iz njih. Grupiranje je moguće podijeliti u dvije skupine hijerarhijsko grupiranje i grupiranje s podjelom prema obilježjima.

Uzevši u obzir poslovni zahtjev, iako vrlo jasno opisan, potrebno ga je prevesti u zadatak obrade i definirati tijek rada (eng. *Workflow*). Logičan odabir je hijerarhijsko grupiranje jer nudi izlaznu varijablu u obliku dendograma koji jasno prikazuje udaljenosti pridruživanja ili razdvajanja pomoću kojih je moguće jasno definirati broj skupova bez preklapanja. Svaka skupina predstavlja jasnu grupu, te se logično se da zaključiti kako će broj inicijalnih skupina biti jednak broju grupa. Prilikom izračuna udaljenosti Euklidovom razdiobom stvara se simetrična dijagonalna matrica udaljenosti čija simetričnost označava da se radi o istoj udaljenosti između dva podatka. Oznaku razdiobe po kojoj će biti napravljeno grupiranje nalazi se u *„Distances“* alatnoj jedinici s dovedenim ulaznim podacima iz *„File“* alatne jedinice s učitanom ITU bazom podataka za Republiku Hrvatsku.

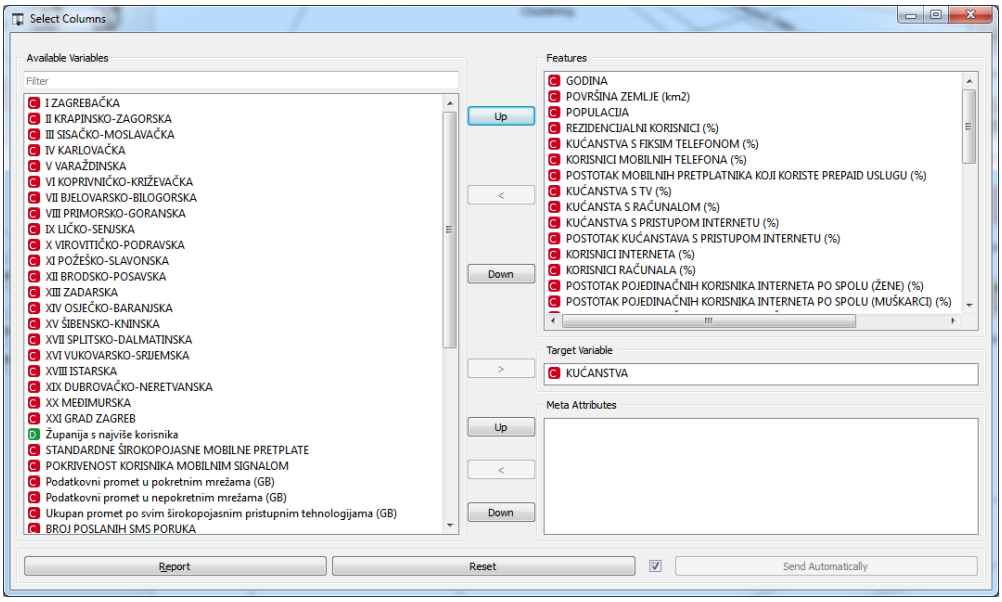


Slika 5.5. Prikaz *„Distances“* alatne jedinice s oznakom korištene razdiobe dendograma

Iako je moguće napraviti hijerarhijsko grupiranje na osnovnom skupu podataka bez individualnog određivanja klase, velika je vjerojatnost za donošenjem krive pretpostavke obilježja grupiranja, meta-atributa pa čak i određivanjem krive klase. Do takvog scenarija je moguće vrlo lako doći iako imamo jasno definiran tijek rada, a iz njega bi proizašli potpuno krivi, pa možda i kontradiktorni zaključci. Za izbjegavanje poteškoća kao poput navedene i sličnih vrsta, osnovni skup podataka je podijeljen varijabli klase i vrijednostima po kojima se žele grupirati podaci kako bi se dobila jasna slika. U osnovnom skupu podataka sadržani su podaci o populaciji, broju korisnika pojedinih usluga, postotku korištenja usluga, broju kućanstava i postotku usluga korištenih po kućanstvima. Bit će obrađen isti skup podataka u dva tijeka rada, od čega je u prvom tijeku rada u *„Select Columns“* alatnoj jedinici ciljana varijabla klase atribut „POPULACIJA“ uz atribute oznaka čije vrijednosti se direktno vežu uz korištenje usluga po broju korisnika (Sl. 5.6.). Drugi tijek rada ima u *„Select Columns“* alatnoj jedinici ciljanu varijabla klase atribut „KUĆANSTVA“ uz atribute oznaka čije vrijednosti sadrže raspodjelu usluga po kućanstvu (Sl. 5.7.). Cilj stvaranja dva jednaka podskupa podataka iz jednog osnovnog skupa podataka je u svrhu provjere tijeka rada i njegove ispravnosti prilikom razlikovanja grupiranja i stvaranja grupa pod-skupova podataka.



Slika 5.6. Prikaz *„Select Columns“* alatne jedinice s atributom „POPULACIJA“ kao varijablom klase

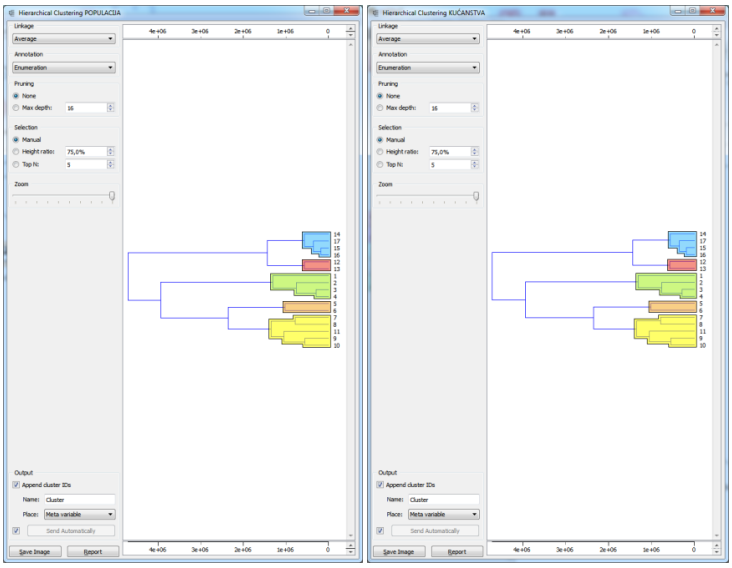


Slika 5.7. Prikaz *„Select Columns“* alatne jedinice s atributom „KUĆANSTVA“ kao varijablom klase

Tako pripremljeni podaci iz *„Select Columns“* alatne jedinice prosljeđuju se u *„Distances“* jedinicu na osnovu pruženih podataka definira udaljenosti Euklidovom razdiobom objašnjenom u ranijem poglavlju 2.4.2. Pomoću *„Distances Matrix“* alatne jedinice moguć je prikaz matrice udaljenosti dobivenih između podatkovnih instanci (Sl.5.8.). Tako dobivene udaljenosti su zatim proslijeđene prema alatnoj jedinici prikaza dendograma hijerarhijskog grupiranja *“Hierarchical Clustering“ (*Sl. 5.9.).

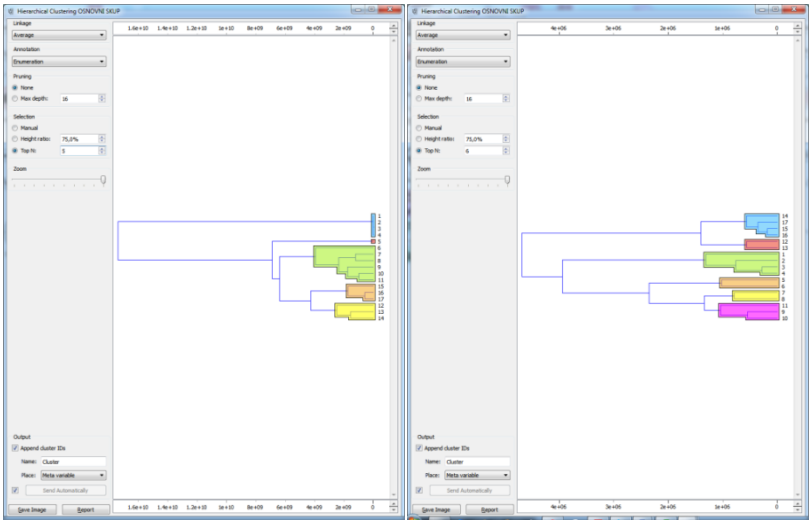
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1		809437.051	1434401.376	1692963.871	2716605.339	3689538.437	4596360.431	5355966.835	5479884.159	5646321.046	6083136.541	6064527.878	6013996.836	6267463.274	6160296.103	6109713.020	6239595.825
2	809437.051		624966.559	887736.409	2022387.416	2955211.998	3847906.572	4600025.788	4772592.635	4948503.653	5396944.749	5444313.345	5425885.600	5749228.622	5679637.704	5638577.198	5784481.676
3	1434401.376	624966.559		283172.263	1564074.442	2421145.703	3289703.747	4030469.030	4251392.537	4435630.589	4892868.231	5002479.199	5014048.502	5398170.477	5362722.882	5330485.151	5488988.927
4	1692963.871	887736.409	283172.263		1344957.770	2169176.322	3031622.766	3769617.037	4009676.377	4195254.024	4659811.326	4806637.684	4839072.174	5250574.176	5229824.663	5201787.422	5364669.441
5	2716605.339	2022387.416	1564074.442	1344957.770		1009108.408	1941739.552	2720204.591	2827885.230	3034883.153	3554212.928	4167285.453	4326698.139	4833532.228	4868749.040	4866083.201	5047436.021
6	3689538.437	2955211.998	2421145.703	2169176.322	1009108.408		936667.178	1717790.609	1893613.963	2136150.625	2709745.770	3663100.456	3918109.328	4537032.973	4644347.014	4662689.416	4858549.095
7	4596360.431	3847906.572	3289703.747	3031622.766	1941739.552	936667.178		784027.610	1118364.919	1396021.598	2025849.269	3319813.885	3664007.285	4367253.558	4538782.377	4575845.052	4777868.274
8	5355966.835	4600025.788	4030469.030	3769617.037	2720204.591	1717790.609	784027.610		821588.850	1056057.694	1655915.059	3181201.444	3591080.898	4341451.068	4564347.939	4616713.059	4816052.631
9	5479884.159	4772592.635	4251392.537	4009676.377	2827885.230	1893613.963	1118364.919	821588.850		552529.271	1303911.813	3242393.445	3670059.194	4381191.879	4597315.095	4659218.672	4857723.727
10	5646321.046	4948503.653	4435630.589	4195254.024	3034883.153	2136150.625	1396021.598	1056057.694	552529.271		789969.643	2952995.447	3411252.470	4138278.463	4366104.894	4433842.585	4632458.991
11	6083136.541	5396944.749	4892868.231	4659811.326	3554212.928	2709745.770	2025849.269	1655915.059	1303911.813	789969.643		2812897.078	3301441.468	4043479.824	4298303.445	4379047.190	4577598.312
12	6064527.878	5444313.345	5002479.199	4806637.684	4167285.453	3663100.456	3319813.885	3181201.444	3242393.445	2952995.447	2812897.078		571448.478	1351497.773	1663897.989	1766707.231	1948400.595
13	6013996.836	5425885.600	5014048.502	4839072.174	4326698.139	3918109.328	3664007.285	3591080.898	3670059.194	3411252.470	3301441.468	571448.478		844881.522	1160327.298	1262975.546	1451342.566
14	6267463.274	5749228.622	5398170.477	5250574.176	4833532.228	4537032.973	4367253.558	4341451.068	4381191.879	4138278.463	4043479.824	1351497.773	844881.522		418383.056	565405.355	712616.229
15	6160296.103	5679637.704	5362722.882	5229824.663	4868749.040	4644347.014	4538782.377	4564347.939	4597315.095	4366104.894	4298303.445	1663897.989	1160327.298	418383.056		183821.440	405322.704
16	6109713.020	5638577.198	5330485.151	5201787.422	4866083.201	4662689.416	4575845.052	4616713.059	4659218.672	4433842.585	4379047.190	1766707.231	1262975.546	565405.355	183821.440		302871.888
17	6239595.825	5784481.676	5488988.927	5364669.441	5047436.021	4858549.095	4777868.274	4816052.631	4857723.727	4632458.991	4577598.312	1948400.595	1451342.566	712616.229	405322.704	302871.888	

Slika 5.8. Prikaz udaljenosti između podatkovnih instanci pomoću *„Distances Matrix“* alatne jedinice



Slika 5.9. Prikaz *„Hierarhical Clustering“* alatne jedinice s dobivenim grupiranjem pod-skupova

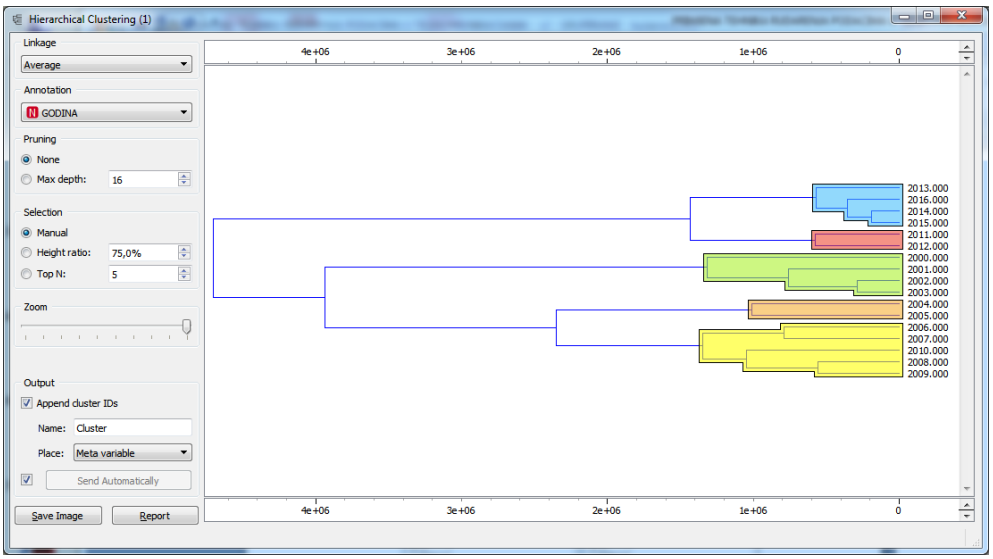
Kako bi dokazali opravdanost razrade modela te semantičke koristi ljudskog utjecaja na zadani model hijerarhijskog grupiranja, osnovni skup podataka je propušten ne ažuriran na obradu u *„Distances“* alatnu jedinicu pod istim uvjetima kao i prethodna dva podskupa. Za izlaznu varijablu su dobivene različite vrijednosti udaljenosti u dendogramu iz kojih su proizašla različita grupiranja (Sl. 5.10.).



Slika 5.10. Prikaz *„Hierarhical Clustering“* alatne jedinice s dobivenim grupiranjem osnovnog skupa

Nažalost, iz prikaza dendograma moguće je vidjeti ograničen broj informacija, ali *Orange* programski alat nudi veći broj alatnih jedinica za vizualizaciju rezultata i njihovu analizu. U ovom slučaju je vrlo korisna *„Scatter Plot“* alatna jedinica za prikaz podataka u dvodimenzionalnom koordinatnom sustavu u odnosu na odabrane atribute. *„Scatter Plot“* alatna jedinica u produžetku će biti spojena na „Hierarhical Clustering“ alatnu jedinicu vezom prikaza podskupa odabranih podataka u dendogramu u odnosu na cjeloviti skup podataka sadržano u dendogramu kojeg će pružati druga veza spojena na *„Data Table“* ili *"File"* alatnu jedinicu poput vizualnog podatkovnog preglednika opisanog ranije u radu.

Dendogram jasno prikazuje 5 grupiranja podskupa osnovnog skupa podataka sastavljenog od 41 atributa oznaka te jednog atributa klase definiranih s 17 instanci podataka (Sl. 5.8.). Za grupirane instance u samom je početku neovisno koji je atribut klase u pitanju. Jedan atribut oznake je u konstantnom linearnom porastu u odnosu na ostale, a to je atribut „GODINA“. Za daljnju obradu grupiranja, detaljno promatranje grupa, shvaćanje grupiranja i donošenje zaključaka iz njega je najpogodniji taj atribut oznake. Svakako bitan korak u donošenju zaključaka je provjera točnosti dobivenih informacija, tako će se na referentno ili anotacijsko (eng. *Annotation*) polje *“Hierarhical Clustering*“ alatne jedinice za prikaz dendograma biti postavljen atribut „GODINA“ (Sl. 5.11.).



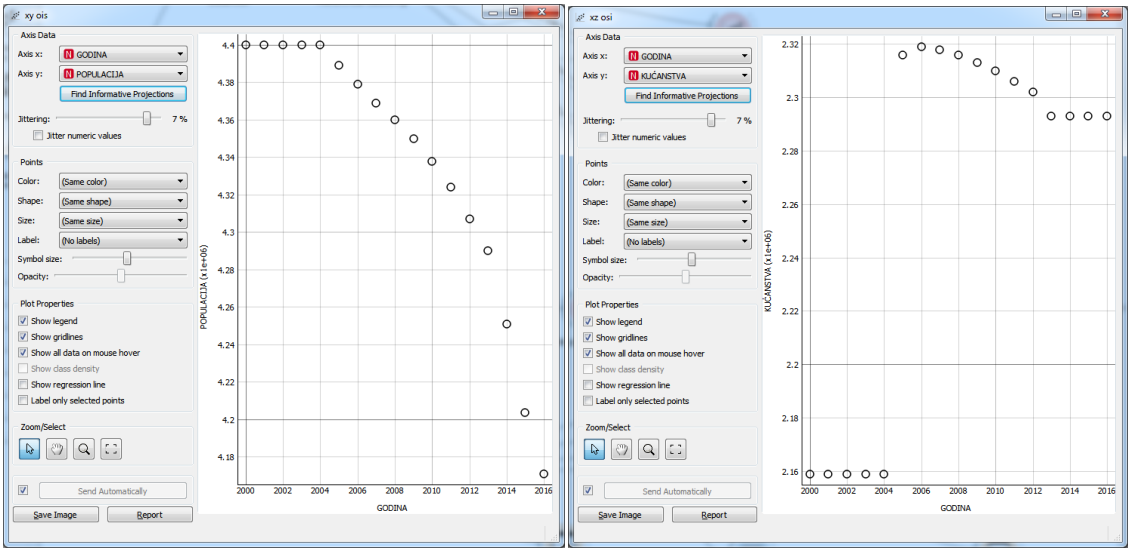
Slika 5.11. Prikaz *„Hierarhical Clustering“* alatne jedinice s referencom na atribut „GODINA“

Primjetno je kako su grupirane zajedno uključujući godine 2000. do 2003., 2004. i 2005., 2006. do 2010., 2011. i 2012., te 2013. do 2014.. Za prvi podskup godina 2000. do 2003. jasan je zaključak kako se radi o ranim godinama gleda li se s aspekta telekomunikacijskih usluga, njihove ponude, cijene i dostupnosti široj populaciji, kao i percepcija populacije o potrebi za korištenjem više telekomunikacijskih usluga osim fiksnih telefonskih linija, malog broja mobilne telefonije i jednako malog broja pristupa internetu putem *„Dial-Up“* i ISDN kablovskih modema.

Ispravno grupirane 2004. i 2005. godine donose veliku prekretnicu na tržište telekomunikacijskih usluga, kako uslugom tako i cijenom, a to su širokopojasna ADSL usluga pristupa internetu i IPTV usluge kablovske televizije. Naravno dostupnošću usluga uz prihvatljive cijene mijenja se i psihološki faktor okoline trenutnih i budućih korisnika. Tako se dolazi do trećeg podskupa označenog između 2006. i 2009. godine koji donosi drastični procvat ponude i potražnje cijelog spektra telekomunikacijskih usluga kao i proporcionalno proširenje kapaciteta za nove usluge te poboljšanje postojećih usluga. Tako u svega manje od 5 godina korisnicima od pristupa internetu putem *„Dial-Up“* i ISDN kablovskih modema postaje dostupna usluga pristupa internetu putem FTTH (eng. *Fiber to the home*) ili FTTB (eng. *Fiber to the building*) s velikom razlikom u pristupnoj brzini. U četvrtom grupiranju, tokom 2011. i 2012. godine ne dolazi do promjene u ponudi telekomunikacijskoj industriji, ali do pada potražnje obilježene padom broja pretplata svih usluga prouzročene ekonomskom krizom nestabilnošću financijskog tržišta, te dovođenjem većeg broja korisnika u nepovoljniji položaj s financijskog gledišta smanjenjem mjesečnih primanja. Također taj faktor je Orange alat u ovom hijerarhijskom grupiranju primijetio zbog pada broja svih vrijednosti instanci atributa usluga u odnosu na relativno ne promijenjene vrijednosti podatkovnih instanci atributa populacije i broja kućanstava. Navedene promjene je lako vidjeti u narednim dijagramima odnosa broja kućanstava s pristupom internetu i broja korisnika interneta kao individualaca u odnosu na godinu. U zadnjem petom grupiranom podskupu dolazi do oporavka broja pretplatnika svih telekomunikacijskih usluga zbog vraćanja stabilnosti, još veće dostupnosti usluga, ali i pretvaranja kroz niz godina korištenje usluga u svakodnevicu.

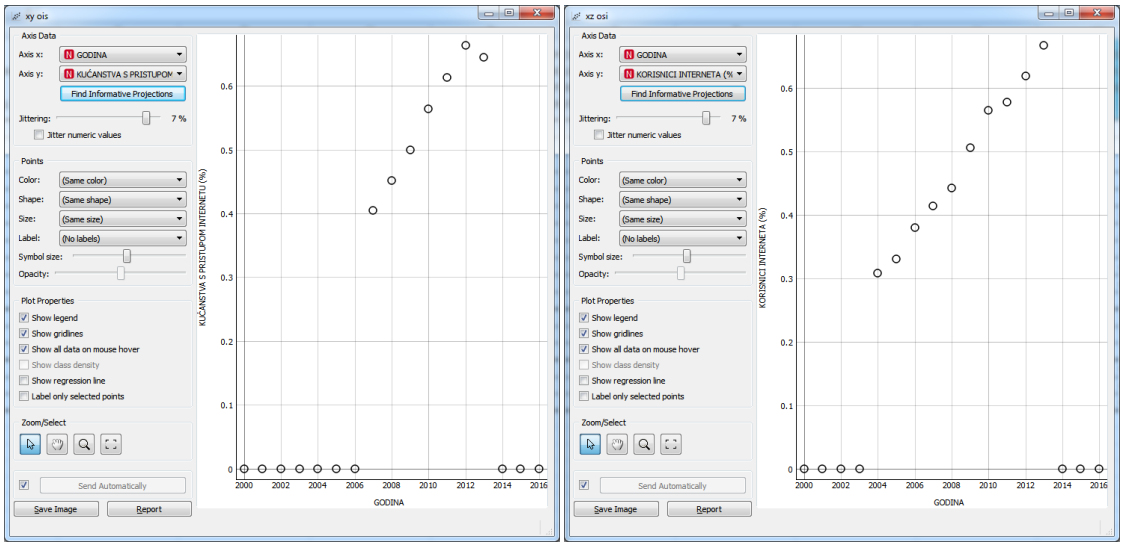
Opisane značajke grupiranja pod-skupova su vidljive u grafičkim prikazima i osnovna raspodjela individualnih korisnika i kućanstava u odnosu na usluge koje koriste. Vrlo zanimljiv prikaz je postotak korištenja osnovnih telekomunikacijskih usluga fiksne i mobilne telefonije, pretplatnika Internet pristupa i slično. Kako bi prikaz davao što više informacija u

odnosu na zajednički atribut koji nije podložan promjenama kao što je atribut oznake godina, na izlaz *„Hierarhical Clustering“* alatne jedinice su spojene dvije *„Scatter Plot“* alatne jedinice kako bi dobili trodimenzionalni prikaz kako će na x osi oba prikaza biti atribut oznake godina, a druge dvije osi služiti za prikaz korelirajućih podataka (Sl 5.12.).



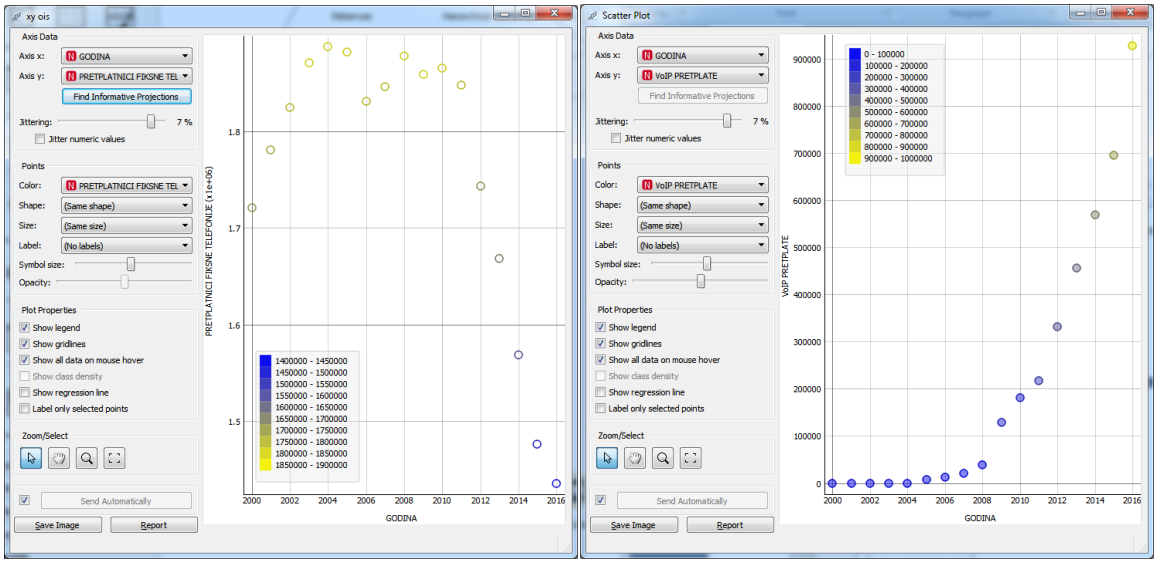
Slika 5.12. Prikaz odnosa broja populacije i broja kućanstava u RH

Iz koreliranja grupiranih podataka o postotku kućanstava s pristupom internetu na 2 300 000 kućanstava i postotka pojedinačnih korisnika interneta može se primijetiti kako iako tokom smanjenja broja kućanstava s pristupom internetu, postotak pojedinačnih korisnika interneta ne doživljava nikakvo smanjenje, čak naprotiv dolazi do porasta broja korisnika interneta istom tendencijom rasta kao i prije (Sl 5.13.). Iz prikazanog grafa se može donijeti zaključak kako korisnicima je korištenje interneta postao bitan dio svakodnevice.



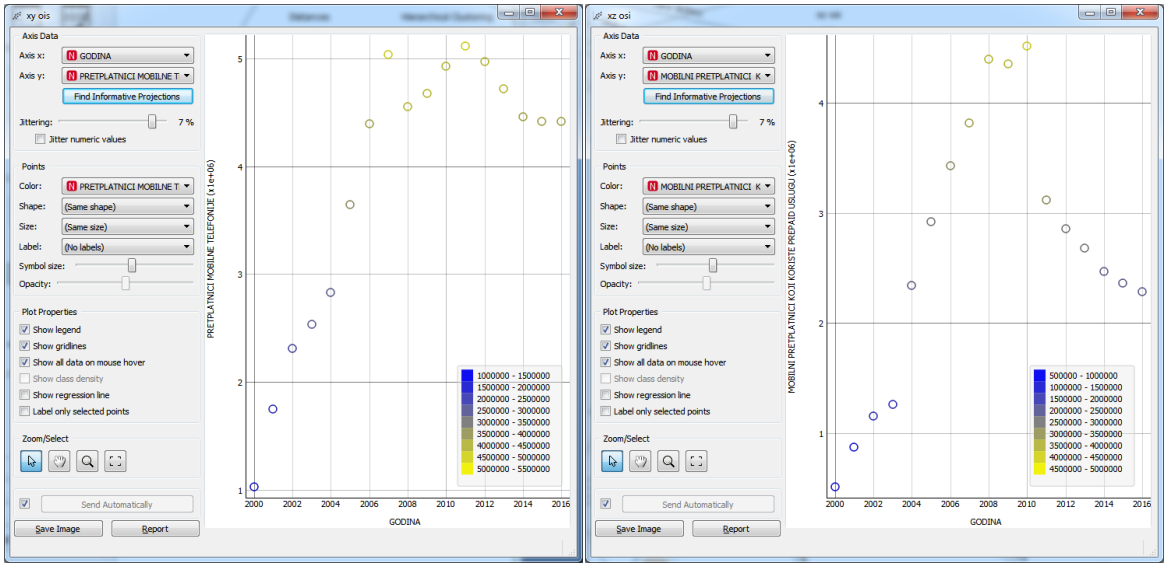
Slika 5.13. Prikaz odnosa postotka kućanstava s pristupom internetu i postotka pojedinačnih korisnika interneta

Iako fiksna telefonija se može smatrati zastarjelom telekomunikacijskom uslugom i u konstantnom opadanju je, VoIP tehnologija pruža novu perspektivu toj usluzi i s gotovo 1 000 000 pretplata u 2016. godini pokriva 2/3 ukupne fiksne telefonije (Sl 5.14.).



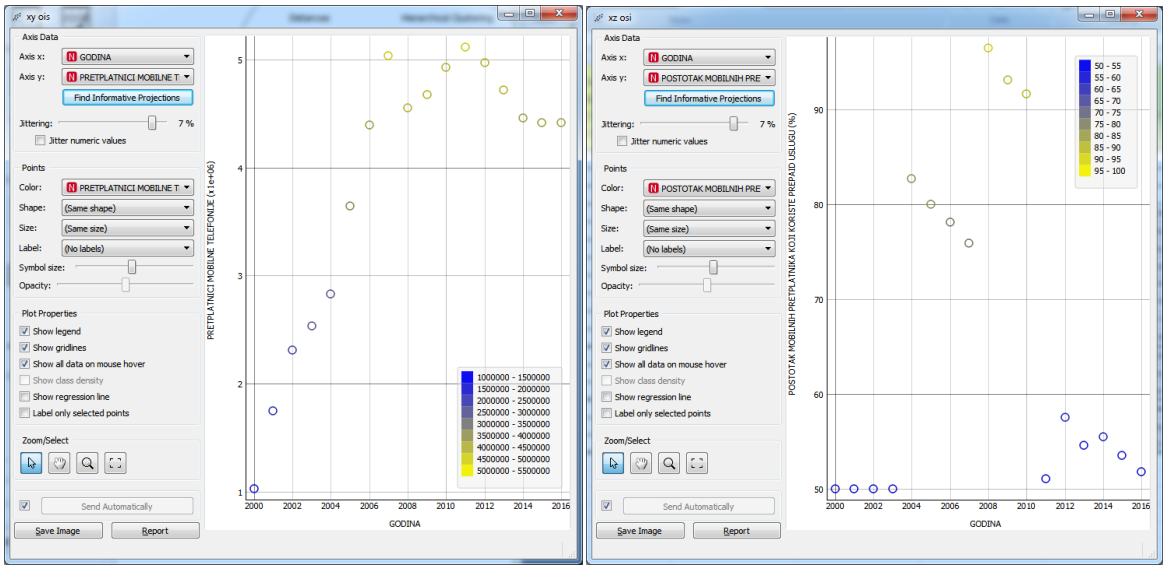
Slika 5.14. Prikaz odnosa broja pretplatnika fiksne telefonije i broja pretplata fiksne telefonije na VoIP-u

Kako postoji zapis broja korisnika mobilne telefonije, broja korisnika koji koriste pretplatu te postotka dobivenog tom korelacijom. Jasno je vidljivo povećanje broja pretplatnika mobilne telefonije u periodima od dvije godine.



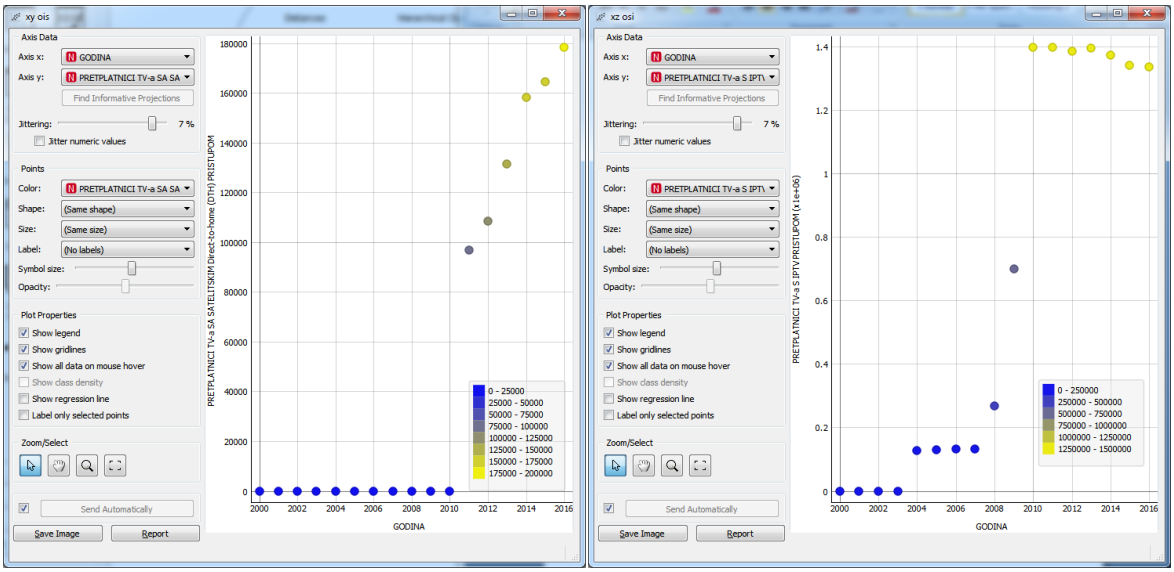
Slika 5.15. Prikaz odnosa broja pretplatnika mobilne telefonije i broja pretplata mobilne telefonije koji koriste prepaid uslugu

Tokom četvrtog grupiranog podskupa je vidljivo drastično opadanje broja pretplatnika s vjerojatnom korelacijom opadanja potražnje za telekomunikacijskim uslugama uslijed ekonomske krize (Sl 5.16.).



Slika 5.16. Prikaz odnosa broja pretplatnika mobilne telefonije i postotka pretplatnika mobilne telefonije koji koriste prepaid uslugu

Korisnici u TV usluge su podijeljeni po tehnologiji pristupa sadržaju, stoga je moguće reći kako zbroj korisnika ovih tehnologija predstavlja ukupan broj korisnika TV usluge (Sl 5.17.).



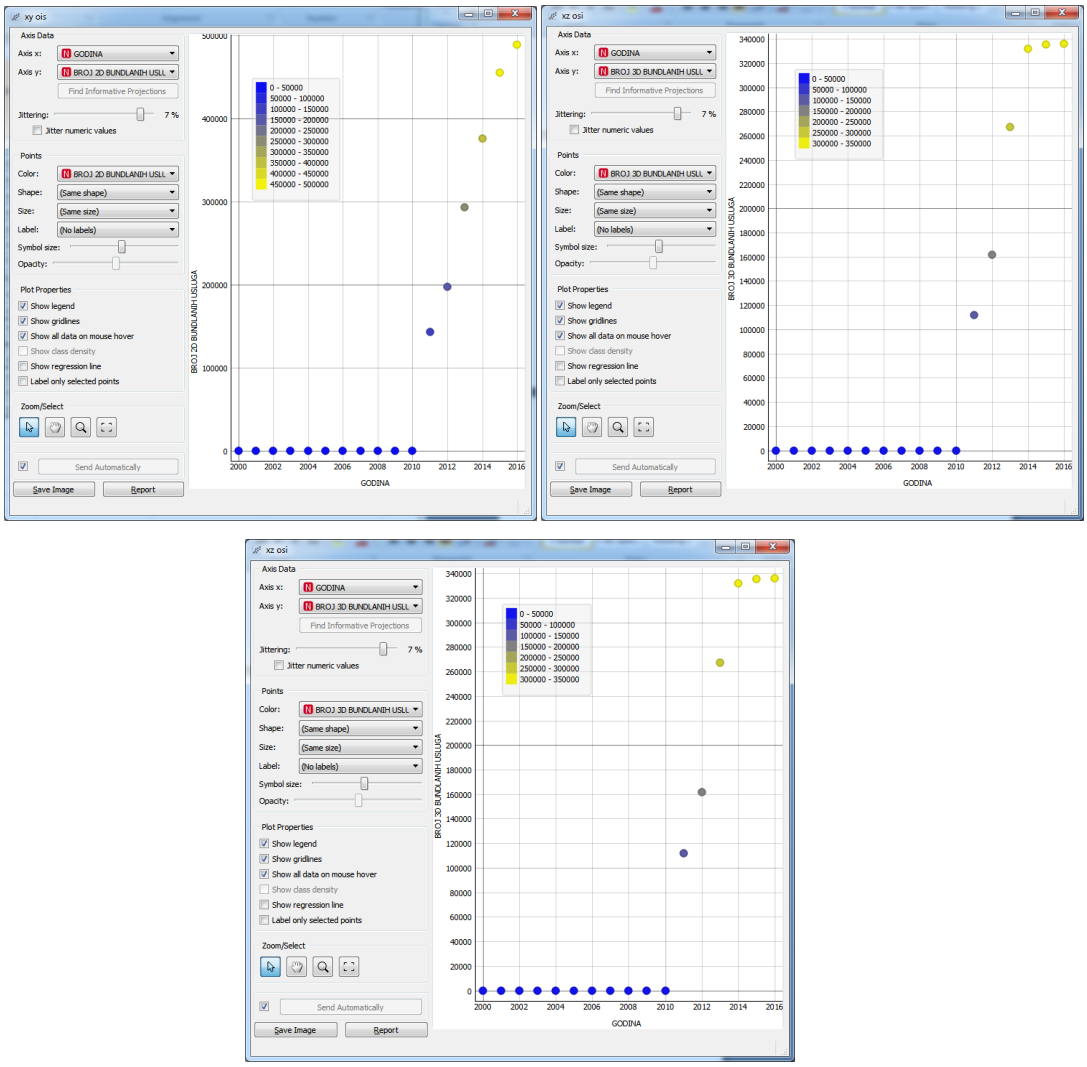
Slika 5.17. Prikaz broja pretplatnika TV usluge ovisno o vrsti tehnologije koju koriste (IPTV i DTH)

Korisnike je moguće grupirati u tri osnovna podskupa (Sl 5.13.) :

- 2D skup korisnika čiji pripadnici su korisnici dvije telekomunikacijske usluge jednog operatera povezane korisničkom oznakom (npr. Fiksni tel.+Internet ili Internet+TV+Mobilni tel. ili Internet+TV)

- 3D skup korisnika čiji pripadnici su korisnici tri telekomunikacijske usluge jednog operatera povezane korisničkom oznakom (npr. Fiksni tel.+Internet+TV ili Internet+TV+Mobilni tel.)

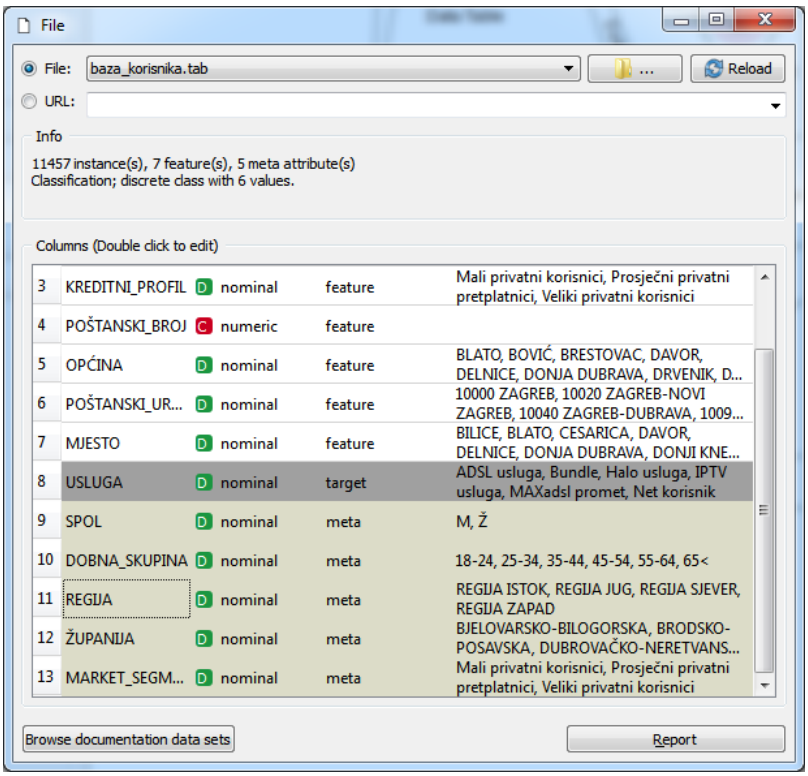
- 4D skup korisnika čiji pripadnici su korisnici četiri telekomunikacijske usluge jednog operatera povezane korisničkom oznakom (npr. Fiksni tel.+Internet+TV+Mobilni tel.)



Slika 5.18. Prikaz grupiranih korisnika u 2D, 3D i 4D grupe korisnika na osnovu broja usluga koje koriste.

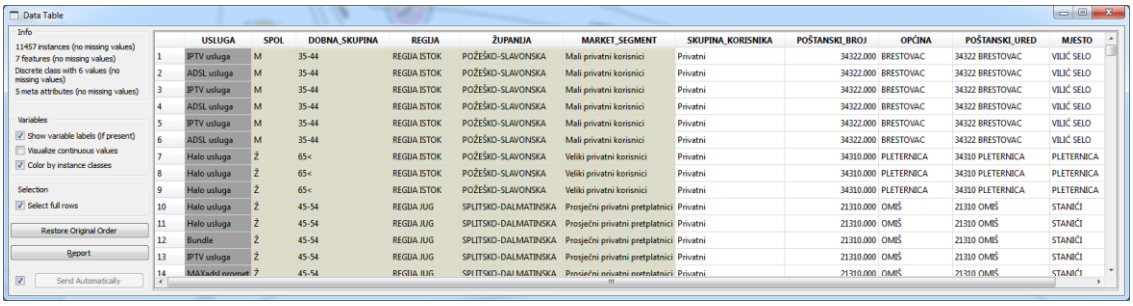
5.3. Klasifikacija korisnika u ovisnosti o uslugama koje koriste

Kada se govori o rudarenju podataka (eng. *Data mining*) u većini se ono odnosi na proces klasifikacije i stvaranje predikcija iz prikupljenih podataka nekom anketom, skupa podataka nastalog nekom vrstom proučavanja određenog područja ili nekog sličnog skupa konkretnih podataka s određenom razinom pouzdanosti. Kako je već ranije napomenuto u tekstu, za praktični primjer upotrjebljen je skup podataka dobiven iz većeg broja godišnjih izvješća Hrvatske regulatorne agencije za mrežne djelatnosti [10] i globalnih ITU izvještaja za Republiku Hrvatsku [11]. Iz prikupljenih podataka kreirana je baza korisnika ekvivalentna prikupljenim izvješćima (Sl. 5.19.).



Slika 5.19. Prikaz učitane baze korisnika

Baza korisnika za klasifikaciju u svojem konačnom obliku sastoji se od 5 opisnih atributa s 11457 podatkovnih instanci, te diskretnog atributa klase „USLUGA“ s 6 pripadajućih vrijednosti. Detaljniji opis informacija koje određeni atribut sadrži, način klasifikacije objašnjen je u narednom poglavlju (Sl. 5.20.).



	USLUGA	SPOL	DOBNA_SKUPINA	REGIJA	ŽUPANIJA	MARKET_SEGMENT	SKUPINA_KORISNIKA	POŠTANSKI_BROJ	OPĆINA	POŠTANSKI_URED	MJESTO
1	IPTV usluga	M	35-44	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Mali privatni korisnici	Privatni	34322.000	BRESTOVAC	34322 BRESTOVAC	VILIĆ SELO
2	ADSL usluga	M	35-44	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Mali privatni korisnici	Privatni	34322.000	BRESTOVAC	34322 BRESTOVAC	VILIĆ SELO
3	IPTV usluga	M	35-44	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Mali privatni korisnici	Privatni	34322.000	BRESTOVAC	34322 BRESTOVAC	VILIĆ SELO
4	ADSL usluga	M	35-44	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Mali privatni korisnici	Privatni	34322.000	BRESTOVAC	34322 BRESTOVAC	VILIĆ SELO
5	IPTV usluga	M	35-44	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Mali privatni korisnici	Privatni	34322.000	BRESTOVAC	34322 BRESTOVAC	VILIĆ SELO
6	ADSL usluga	M	35-44	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Mali privatni korisnici	Privatni	34322.000	BRESTOVAC	34322 BRESTOVAC	VILIĆ SELO
7	Halo usluga	Ž	65<	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Veliki privatni korisnici	Privatni	34310.000	PLETERNICA	34310 PLETERNICA	PLETERNICA
8	Halo usluga	Ž	65<	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Veliki privatni korisnici	Privatni	34310.000	PLETERNICA	34310 PLETERNICA	PLETERNICA
9	Halo usluga	Ž	65<	REGIJA ISTOK	POŽEŠKO-SLAVONSKA	Veliki privatni korisnici	Privatni	34310.000	PLETERNICA	34310 PLETERNICA	PLETERNICA
10	Halo usluga	Ž	45-54	REGIJA JUG	SPLITSKO-DALMATINSKA	Prosječni privatni pretplatnici	Privatni	21310.000	OMIŠ	21310 OMIŠ	STANIĆI
11	Halo usluga	Ž	45-54	REGIJA JUG	SPLITSKO-DALMATINSKA	Prosječni privatni pretplatnici	Privatni	21310.000	OMIŠ	21310 OMIŠ	STANIĆI
12	Bundle	Ž	45-54	REGIJA JUG	SPLITSKO-DALMATINSKA	Prosječni privatni pretplatnici	Privatni	21310.000	OMIŠ	21310 OMIŠ	STANIĆI
13	IPTV usluga	Ž	45-54	REGIJA JUG	SPLITSKO-DALMATINSKA	Prosječni privatni pretplatnici	Privatni	21310.000	OMIŠ	21310 OMIŠ	STANIĆI
14	MAXadsl paket	Ž	45-54	REGIJA JUG	SPLITSKO-DALMATINSKA	Prosječni privatni pretplatnici	Privatni	21310.000	OMIŠ	21310 OMIŠ	STANIĆI

Slika 5.20. Prikaz informacija koje ona sadržava učitane baze korisnika

Glavna zadaća klasifikacijskog modela je svrstavanje pojedine podatkovne instance koja predstavlja podatak koji opisuje korisnika s vrijednostima koje su zabilježene u atributima u određenu grupu korisnika opisanu skupom atributa. Takvu raspodjelu najbolje je moguće vidjeti korištenjem više klasifikacijskih alatnih jedinica kako bi dobili prvenstveno željenu informaciju o točnosti klasifikacije dobivenu iz pojedine klasifikacijske alane jedinice u Orange alatu, a na kraju i točnosti cjelokupnog klasifikacijskog modela. Nakon izvršene klasifikacije dobiveni rezultati su prikazani kroz vizualizacijske alatne jedinice u kojima se vide jasne klase usluga koje korisnici koriste.

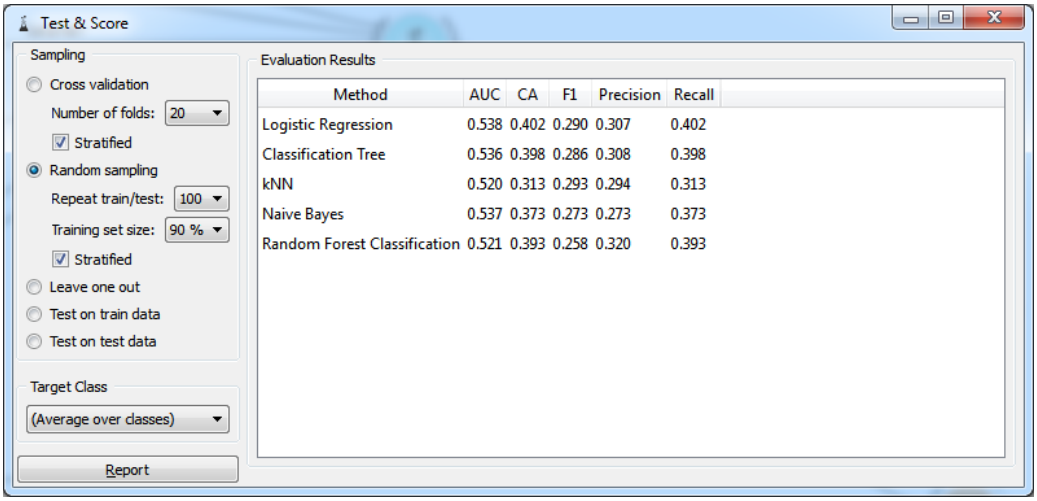
Za izradu klasifikacijskog modela korišteno je više različitih alatnih jedinica klasifikatora kako bi odgovorili na pitanje koji od atributa su ključni pri klasifikaciji instanci skupa podataka, odnosno koji je klasifikator najpogodniji za klasificiranje korisnika po uslugama koje koriste. Kako bi provjerili koji klasifikatora klasifikacijskog modela radi najbolje potrebnoj je koristiti „Test & Score“ alatnu jedinicu za obradu usporedbe i procjene klasifikacijskih metoda.

Unutar klasifikacijsog modela korištene alatne jedinice su:

- *Classification Tree*
- *Logistical Regression*
- *Random Forest Classification*
- *Naive Bayes*
- *Nearest Neighbors*

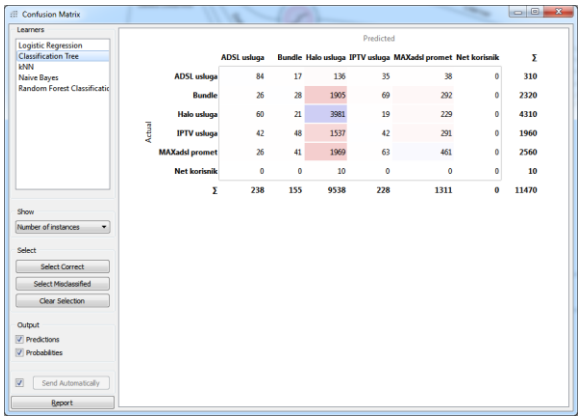
Klasifikacijski model se gradi na podskupu podataka ili instanci koji služe za trening ili učenje klasifikatora, dok se testiranje klasifikacijskog modela se vrši na drugom, zasebnom podatkovnom podskupu ili instanci kako bi zagarantirali različitost vrijednosti atributa

podatkovnih pod-skupova i time ne bi prouzrokovali njihovo preklapanje koje u konačnici može dovesti do krivih zaključaka klasifikacijskog modela. Proces testiranja na podatkovnom skupu dijeli ih u dva dijela, od čega se podatkovni podskup za trening sastoji od 90% inicijalnog skupa podataka, test se vrši na preostalih 10% dijela inicijalnog skupa, a ciklus se ponavlja unutar podskupa dok je moguće testirati na dotad nepoznatim podacima. Takav proces unakrsne validacije je napravljen u 100 ciklusa 10 puta, te nakon svakog ciklusa zabilježena je točnost klasificiranja. Podatak točnosti unakrsne validacije klasifikatora se nalazi u *„Test & Score“* alatnoj jedinici pod oznakom „CA“ koja predstavlja točnost klasifikacije (*eng. Clasification Acurasy*) i predstavlja broj točno klasificiranih podatkovnih instanci iz testnog podskupa danih podatka (Sl. 5.20.).



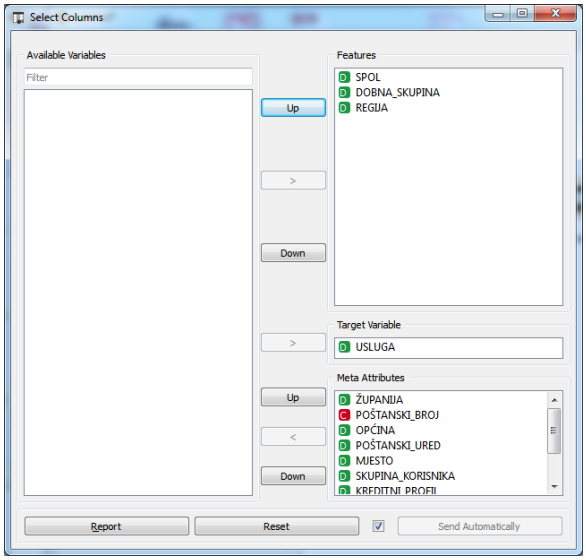
Slika 5.20. Prikaz *„Test & Score“* alatne jedinice s rezultatima procjene klasifikacijskih metoda

Iako najveću točnost klasifikacije daje metoda logističke regresije s 40,2%, najveću preciznost i najmanje odstupanje rezultata, a time i najbolji prikaz klasifikacije s 39,8% pruža *„Classification Tree“* alatna jedinica koja daje klasifikaciju na osnovu metode stabla odluke, te raspodjele prema ključnim atributima koji najbolje definiraju pod-skupove. Iako postoji određeni dio krive klasifikacije na broju telefonskih pretplata, uzrok pogreške je lako pripisati velikom skupu korisnika koji imaju telefonsku pretplatu, iako nije došlo do preklapanja pod-skupova (Sl. 5.21.). Klasifikator tako doživljava velik broj korisnika usluge u odnosu na druge. Ukoliko se uzme prosjek broja fiksnih telefonskih pretplata od 1 800 000 na prosječan broj kućanstava od 2 250 000 kućanstava, dolazi se do zaključka kako samo 1/5 kućanstava nema pretplatu na fiksnu telefonsku liniju.



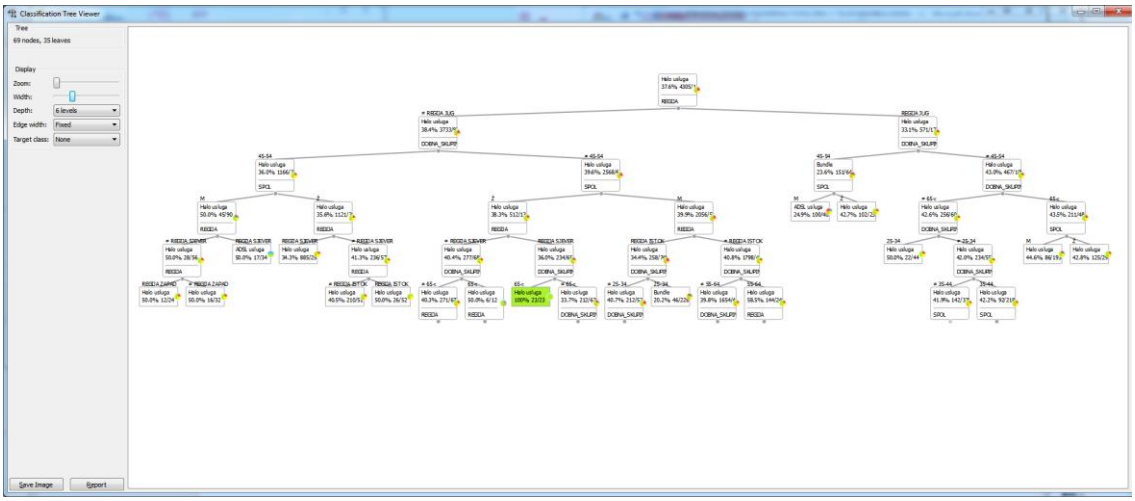
Slika 5.21. Prikaz *„Confusion Matrix“* alatne jedinice s prikazom matrice krivo klasificiranih instanci

Grafički prikaz stabla odluke je izveden pomoću *„Classification Tree Viewer“* alatne jedinice. Klasifikacijsko stablo koje koristi sve atribute podskupa, iako daje vrlo detaljnu klasifikaciju cijelog skupa podataka, nažalost nije lako shvatljivo i pogodno za jasan prikaz. Stoga u *„Classification Tree“* alatnu jedinicu će biti dostavljeni samo ciljani atributi za prikaz kako bi bilo lakše napraviti analizu i prikaz, a napravljeno je pomoću *„Select Columns“* alatne jedinice s kojom je moguće manipulirati brojem atributa i njihovom kardinalnosti u skupu (Sl. 5.22.).



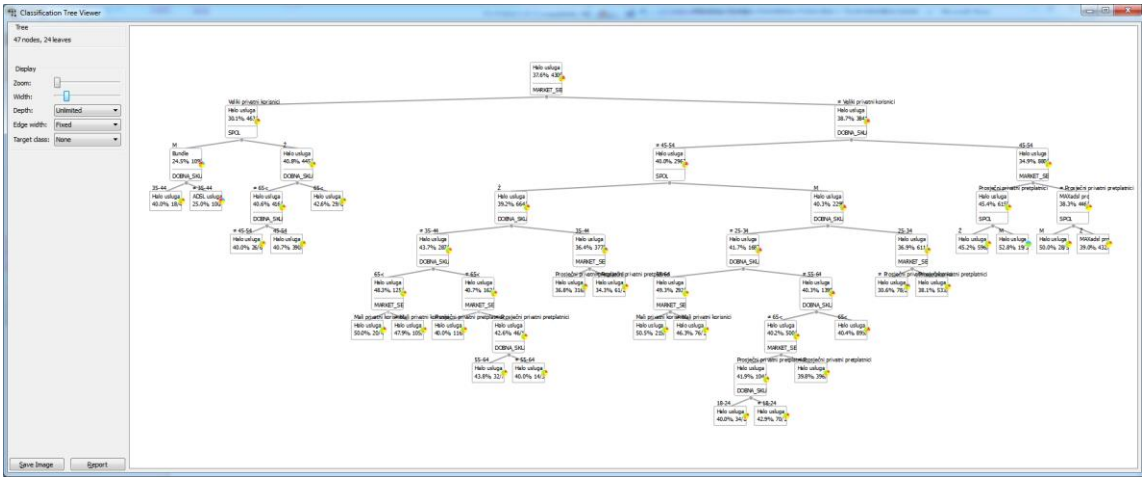
Slika 5.22. Prikaz *„Select Columns“* alatne jedinice s odabranim atributima proslijeđenima prema stablu odlučivanja

Vrijedan prikaz raspodjele korisnika svih usluga je prikaz u odnosu na spol, dobnu skupinu i regiju u kojoj žive. Tako je dobiveno ispod prikazano stablo (Sl. 5.23.).



Slika 5.23. Prikaz stabla odlučivanja u *„Classification Tree Viewer*“ alatnoj jedinici korištenja usluga u odnosu na spol,dobnu skupinu i regiju korisnika

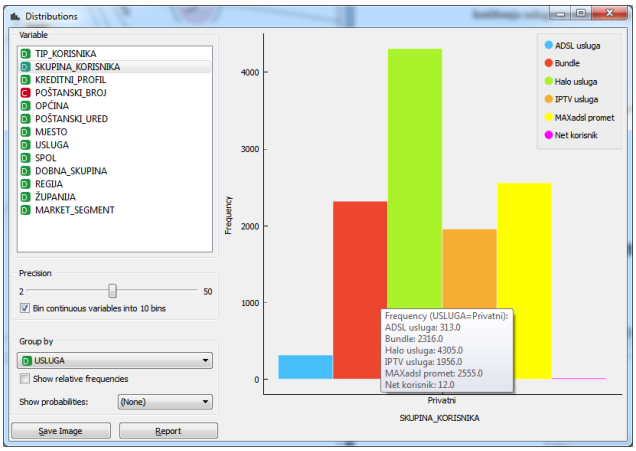
Klasifikaciju korisnika po marketinškom segmentu nam govori o potencijalnoj vrijednosti korisnika, kao i broju korisnika koji su pogodni za proširenje broja usluga koje koriste (Sl. 5.24.).



Slika 5.24. Prikaz stabla odlučivanja u *„Classification Tree Viewer“* alatnoj jedinici korištenja usluga u odnosu na spol,dobnu skupinu i marketinški segment korisnika

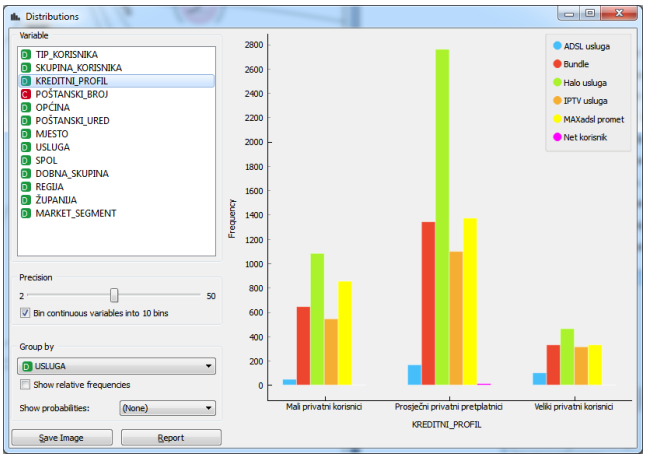
„Distributions“ alatna jedinica nudi prikaza grupiranja korisnika po usluzi koju koriste. Tako je vidljivo da od 4300 korisnika fiksne telefonske linije navedenih u bazi, 2300 korisnika grupirane *„Bundle“* usluge u obliku 2D ili 3D paketa, bez 4D paketa koji bi uključivao uslugu mobilne telefonije koju nije moguće jasno prikazati u skupu jer postoji veći broj pretplata mobilne telefonije nego broja populacije, lako je doći do zaključka kako tek nešto manje od 50% fiksne telefonije nije grupirano u jednu od grupnih usluga. Kako je

vrlo lako zaključiti da većina korisnika koji imaju pristupnu komponentu telekomunikacijske usluge za pristup internetu će posjedovati i prometnu komponentu paketa Internet prometa zaključuje se da svi korisnici ADSL usluge pristupa internetu su grupirani unutar 2D ili 3D paketa što odnosi 300 Bundle paketa. Preostaje još 2000 telefonskih paketa koji su grupirani unutar 2D ili 3D paketa što odgovara proporciji pretplata IPTV usluge. Net korisnik koji predstavlja pristupni paket za spajanje putem kablovskih modema te se koristi isključivo zbog e-mail adrese zauzima tek neznatan broj korisnika (Sl. 5.25.).



Slika 5.25. Prikaz ukupne skupine korisnika u odnosu na broj usluga koje koriste

Ukoliko uzmemo u obzir marketinški segment korisnika u odnosu na usluge koje koriste vidljiva je pravilna distribucija „Bundle“ paketa između segmenata korisnika dolazi se do potvrde ranije donesenog zaključka o većini grupiranih usluga u zajedničke 2D i 3D pakete (Sl. 5.26.).



Slika 5.26. Prikaz marketinškog segmenta korisnika u ovisnosti o broju korištenih usluga.

6. ZAKLJUČAK

Cilj rada je prikazati kako se primjenom rudarenja podataka može unaprijediti poslovno odlučivanje i kvaliteta donesenih odluka što direktno poboljšava poslovne rezultate. Prikazana je primjena tehnika i proces rudarenja podataka u telekomunikacijama koji se sastoji od više koraka. Početni korak u postupku rudarenja podataka je dobro razumijevanje zadatka poslovnog zahtjeva, nakon čega je tek moguće njegovo definiranje i prilagodba u tehnički proces rudarenja podataka. Nakon precizno definiranog procesa ključan korak je modeliranje postupka rudarenja podataka, ocjena modela i primjena najbolje ocijenjenog modela. Primjenom računalnih alata za istraživanje kao Orange alat za rudarenje podataka korišten u ovom radu identificirane su najbitnije primjene rudarenja podataka u telekomunikacijama prikazanih kroz preference korisnika i korelacija između korištenih usluga. Anketno istraživanje i godišnja izvješća regulatora čine vrlo detaljnu bazu podataka za primjenu postupka rudarenja na njoj kroz klasifikaciju i grupiranje korisnika na osnovu usluga koje koriste.

Iz prikazanog dendograma hijerarhijskog grupiranja korisnika jasno su vidljivi uzorci korelacije korištenja usluga kroz godine u ovisnosti o nizu sociooloških i faktora kao što je stav okoline o nužnosti korištenja određene telekomunikacijske usluge, te cjenovne pristupačnosti usluge širim populacijama. U radu je obrađen vremenski period od 16 godina kroz koji su se događale velike promjene u telekomunikacijskoj industriji. Tako obrađeni podaci osnovnog skupa odnose se na populaciju od 4 200 000 stanovnika u 2 300 000 kućanstava, je vidljivo kako korisnici u Republici Hrvatskoj prate svjetske trendove u korištenju telekomunikacijskih usluga. S vrlo visokom geografskom pokrivenošću telekomunikacijskom infrastrukturom. 2011. godine broj korisnika dostiže 118,3 pretplatnika mobilnih usluga na 100 stanovnika, te kroz naredne 3 godine uz prisutnost ekonomske krize taj broj pada na 104,94 pretplatnika na 100 stanovnika u 2014. godini što je ekvivalent broju korisnika u 2008 godini. Zanimljiv podatak je kako broj pretplatnika usluga pristupa internetu se smanjuje u istom periodu, postotak pojedinačnih korisnika interneta ne doživljava nikakvo smanjenje, čak naprotiv dolazi do porasta broja korisnika interneta istom tendencijom rasta kao do tada te se može donijeti zaključak kako korisnicima je korištenje interneta postao bitan dio svakodnevice. Grupiranjem također prikazuje kako 72% korisnika interneta u odnosu na ukupnu populaciju s 76% kućanstava s pristupom Internetu daje da su u 75% kućanstava svi ukućani korisnici usluge pristupa internetu od čega su 76%

pojedinačnih korisnika žene, 68% njih muškarci. Također u uvidom u broj pretplata grupiranih usluga u 2D,3D i 4D pakete jasno se zaključuje kako je raspodjela usluga proporcionalna broju pretplata svih usluga što je jasna poruka da korisnici u velikom broju usluge s definiranim pretplatama na taj način zbog veće pristupačnosti cijena.

Iako *Orange* alat za rudarenje podataka je edukativni alat za predodžbu potencijala i svrhe upuštanja u takvu vrstu analize, te nije dovoljno moćan za analizu nad velikim osnovnim skupom podataka i dolazilo je do poteškoća kod učitavanja 11 000 podatkovnih instanci u „File“ alatne jedinice za učitavanje podataka te je bilo prijeko potrebno spremanje u .tab format koji je prirodan *Orange* alatu te nosi kontekste klase, meta atributa i atributa oznaka u svom zaglavlju za razliku od .xls/.xlsx kod kojeg je to ručno potrebno napraviti dodavanjem dva nova reda na vrh seta podataka. Uključivanjem akademske zajednice, pružanjem podrške i davanjem interesa za rad u njemu, te pružanjem povratnih informacija razvojnom timu alat nudi veliki potencijal za razvoj.

LITERATURA

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

[2] U. M. Fayyad, G. P. Shapiro, P. Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol. 3, No. 17, 1996, pp. 37-54.

[3] Introduction to Data Miningand Knowledge Discovery, http://www.twocrows.com/intro-dm.pdf?bcsi_scan_998937D63ACB8A9E=0&bcsi_scan_filename=intro-dm.pdf

[4] The DBMS Guide to Data Mining solutions, http://www.dbmsmag.com/9807m00.html,

[5] Daniel T. Larose: Data Mining Methods and Models, John Wiley & Sons, Inc. USA, 2006.

[6] Andrew H. Karp, "Getting Started with PROC LOGISTIC", Statistics, Data Analysis, and Data Mining, članak 248-26,: www.emerald-library.com

[7] D. Lowd, P. Domingos, Naive Bayes Models for Probability Estimation, http://www.cs.washington.edu/ai/nbe/nbe_icml.pdf, 21. Lipnja 2009.

[8]K.Tekonomo, Hierarchical Clustering Tutorial, http://people.revoledu.com/kardi/tutorial/Clustering/index.html

[9] S. Sumathi, S.N. Sivanandam, Introduction to Data Mining and its Applications, str. 620,.Springer, Berlin, Njemačka, 2006.]

[10] Hrvatska regulatorna agencija za mrežne djelatnosti,

https://www.hakom.hr/default.aspx?id=323

[11] ITU-Agencije Ujedinjenih Naroda za informacije i komunikacije

http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

SAŽETAK

Cilj rada je prikazati kako se primjenom rudarenja podataka može unaprijediti poslovno odlučivanje i kvaliteta donesenih odluka što direktno poboljšava poslovne rezultate. Prikazana je primjena tehnika i proces rudarenja podataka u telekomunikacijama kroz klasifikaciju i grupiranje korisnika na osnovu usluga koje koriste napravljen je pomoću Orange alata za rudarenje podataka

Ključne riječi: Rudarenje podataka, grupiranje, klasifikacija, modeliranje, regresija, neuronske mreže, stablo odluke

ABSTRACT

Goal of the paper is to demonstrate that data mining techniques could use existing data in telecomunications for increasing quality of decision making, which in turn improves business performance. Analysis of possible applications of data mining in telecomunications through classification and grouping of users depending on the telecomunication service that they are using is done by the Orange data mining tool.

Key words: Data mining, grouping, classification, modeling, regression, neuron network, classification tree

PRILOZI

baza_korisnika.tab

ITU podaci_populacija.tab

baza_korisnika.xls

ITU PODACI.xls

PRIMJENA TEHNIKA RUDARENJA PODACIMA U TELEKOMUNIKACIJAMA - GRUPIRANJE.ows

PRIMJENA TEHNIKA RUDARENJA PODACIMA U TELEKOMUNIKACIJAMA - KLASIFIKACIJA.ows

ŽIVOTOPIS

Zoran Kundek rođen je 22. ožujka 1988. u Požegi. Živi u Zagrebu, gdje trenutno radi u Hrvatskom telekomu na radnom mjestu Specijalista za razvoj informacijskih sustava. Osnovnu školu završio je u Požegi. Upisao je Srednju školu Elektrotehničkog smjera također u Požegi te je tijekom srednjoškolskog obrazovanja sudjelovao na natjecanju Loginet Europe. Sve razrede završava vrlo dobrim uspjehom. 2007. godine upisuje Elektrotehnički fakultet u Osijeku, gdje je odabrao smjer Komunikacije i informatika. 2011. godine završava preddiplomski studij s temom završnog rada Tehnologija mikro-elektromehaničkih sustava (MEMS) u izradi poluvičkih akcelerometara. Iste godine upisuje diplomski studij na Elektrotehničkom fakultetu u Osijeku.