Klasifikacija anketa o programerskim veštinama

Seminarski rad u okviru kursa
Istraživanje podataka 1
Matematički fakultet

Nikola Živković, Dušica Krstić
nmzivkovic@gmail.com, dusicamkrstic@gmail.com

19. juni 2019

Sažetak

U ovom radu je prikazana klasifikacija odgovora na anketu o programerskim veštinama. Ilustrovana je primena određenih algoritama, a razvrstavanje je vršeno prema kriterijumima: pol, starost, zaposlenje i zemlja ispitanika. Dobijeni rezultati su upoređeni i zaključeno koji su algoritmi dali najbolje rezultate u kojoj podeli i šta bi bilo neophodno kako bi se preciznost klasifikacija povećala.

Sadržaj

1 Uvod

Problem klasifikacije je problem učenja, gde se nepoznate instance razvrstavaju u jednu od unapred ponuđenih grupa[6]. Ove grupe se nazivaju klase ili kategorije. Učenje ovih kategorija se ostvaruje pomoću modela, koji se kasnije koristi za procenu odgovarajuće klase do sada neviđenih podataka[1]. Klasifikacija pripada problemu učenja koji se naziva nadgledano učenje. Nadgledano mašinsko učenje se karakteriše time da su za sve podatke poznate vrednosti ciljne promenljive.

U ovom radu smo se bavili klasifikacijom anketa o programerskim veštinama. Klasifikacija je rađena u programu ***IBMM SPSS Modeler***, a korišćeni skup podataka može biti preuzet sa sledećeg linka: https://www.kaggle.com/hackerrank/developer-survey-2018.

2 Korišćeni skup podataka

Skup podataka *"HackerRank Developer Survey 2018"* sadrži 25090 odgovora od strane studenata i profesionalaca, vezano za njihove veštine, obrazovanje, trenutne pozicije itd...

Sastoji se od pet datoteka:

- *HackerRank-Developer-Survey-2018-Codebook.csv* sadrži sva pitanja koja su se pojavila u anketi. U daljem tekstu će biti referisana kao Codebook.csv.
- *Country-Code-Mapping.csv* sadrži imena i kodove svih zemalja koje su se pojavile u anketi.
- *HackerRank-Developer-Survey-2018-Numeric.csv* sadrži odgovore ispitanika na pitanja. Većina odgovora su numerički. U daljem tekstu će biti referisana kao Numeric.csv.
- *HackerRank-Developer-Survey-2018-Numeric-Mapping.csv* sadrži objašnjenje za svaki numerički odgovor koji se pojavljuje u anketi. U daljem tekstu će biti referisana kao Numeric-Mapping.csv.
- *HackerRank-Developer-Survey-2018-Values.csv* sadrži odgovore na pitanja ankete u tekstualnom formatu. U daljem tekstu će biti referisana kao Values.csv.

3 Analiza i preprocesiranje podataka

U ovom odeljku ćemo se baviti analizom, pripremom i obradom podataka kako bi bili pogodniji za dalju analizu i primenu algoritama za klasifikaciju.

3.1 Analiza datoteka

Opisi određenih datoteka su dati u tabelama 1, 2, 3.

Data Field	Imena kolona u tabelama Numeric.csv i Values.csv
Survey question	Tekst pitanja
Notes	beleške

Tabela 1: Opis datoteke *Codebook.csv*

Value	Numerička vrednost
Label	Ime države koju numerička vrednost predstavlja

Tabela 2: Opis datoteke *Country-Code-Mapping.csv*

Data Field	Imena kolona u tabelama koje sadrže odgovore na anketu
Value	Numerička vrednost
Label	Opis šta numerička vrednost predstavlja

Tabela 3: Opis datoteke *Numeric-Mapping.csv*

U tabelama *Numeric.csv* i *Values.csv* kolone odgovaraju redovima tabele Codebook.csv.

Primarna datoteka koju smo koristili za klasifikaciju podataka je Numeric.csv, dok su ostale korišćene za tumačenje pročitanih podataka. Datoteka koju uopšte nismo koristili je Values.csv.

3.2 Preprocesiranje

Analizirajući datoteku Numeric.csv uočili smo kolone koje se mogu zanemariti zbog irelevantnih podakata (identifikator, vreme početka i završetka popunjavanja ankete), prevelikih količina praznih slogova ili tekstualno neadekvatnih odgovora (mahom sva *other* polja). Ovo je urađeno korišćenjem *Filter* čvora.

Korišćenjem *Filler* čvora smo sve tekstualne "#NULL!" vrednosti prebacili u sistemske nedostajuće vrednosti "$null$", kako bismo očuvali konzistentnost među podacima (ne želimo da pola polja bude nedostajuća vrednost, a druga polovina tekstualna vrednost "#NULL!").

Nakon filtriranja smo sve nedostajuće vrednosti zamenili sa nulom, pri čemu nam ta vrednost predstavlja indikator da je ispitanik na određeno pitanje umesto konkretnog ponuđenog odgovora odabrao opciju "other".

Nakon ovoga su svi atributi kategorizovani (pretvoreni u nominalne, redne ili binarne), zbog potreba algoritama korišćenih za klasifikaciju.

Na slici 1 se nalazi prikaz korišćenih čvorova.

Slika 1: Obrada nedostajućih vrednosti i kategorizacija

4 Klasifikacija

U ovom odeljku je ilustrovana klasifikacija po polu, starosti ispitanika, državama iz kojih dolaze ispitanici i poslovnim pozicijama na kojima se nalaze. Klasifikacija je rađena primenom različitih algoritama nad podacima koji su podeljeni u skup za treniranje (70%) i za testiranje (30%). Podela je izvršena korišćenjem "Partition"čvora, a ulazni atributi i ciljna klasa po kojoj je vršena klasifikacija su određeni pomoću "Type"čvora.

4.1 Algoritmi za klasifikaciju

U ovom odeljku je napravljen pregled korišćenih algoritama.

4.1.1 C5.0

Algoritam C5.0[3] može da se koristi za pravljenje stabla odlučivanja ili određivanja pravila pridruživanja. Radi tako što pravi podelu po atributu koji ima najveću informacionu dobit. Naknadno se vrši potkresivanje drveta kako bi se procenila i smanjila greška i kako bi se odredili intervali poverenja.

Neophodno je da ciljni atribut bude kategorički, dok druga polja mogu biti bilo kog drugog tipa. Svi atributi čija je uloga obeležena sa *"both"* ili *"none"* se ignorišu.

Ovaj algoritam češće grupiše kategorije na osnovu sličnosti, kada radi sa imenskim i rednim atributima.

Prednosti ovog algoritma su stabilnost kada radi sa nedostajućim vrednostima ili velikim brojem ulaznih polja. Brzo se izvršava.

Najbolje rezultate smo dobili koristeći *boosting*, *gobal pruning* i *winnow attributes* opcije. *Boosting* je do najboljeg modela došao posle 10 koraka, *global pruning* je po završetku potkresao najslabija podstabla, a *winnow attributes* je sprečio da dođe do preprilagođavanja, iako su razlike na test skupu oba modela bile minimalne. Nismo koristili *group symbolics* da bismo izbegli spajanje atributa u toku podele stabla, a *cross-validate* nam nije bio neophodan jer je naš skup podataka dovoljno veliki da može da se lako podeli na trening i test skup.

4.1.2 C&R Tree

Algoritam C&R Tree[2] kao rezultat daje binarno stablo odlučivanja. Koristi metod rekurzivnog particionisanja kako bi podatke za treniranje podelio u kategorije sa istim ciljnim vrednostima. Za grananje koristi meru nečistoće. Kao najbolja mera pokazao se Ginijev indeks.

Za izvršavanje mu je potrebno najmanje jedan ulazni i tačno jedan ciljni atribut. Ti atributi mogu biti bilo kontinualni bilo kategorički.

Prednosti su iste kao kod C5.0 algoritma.

4.1.3 CHAID

CHAID algoritam[4] je metod izrade stabala odlučivanja koji koristi hi-kvadrat statistiku za određivanje optimalnih podela. Biraju se statistički značajna polja za unos. Ako unos ima više od dve kategorije, one se analiziraju i kategorije koje ne pokazuju razlike u krajnjem ishodu se pripajaju drugim poljima. Za nominalne atribute, sve vrednosti mogu biti spojene, dok se kod rednih atributa mogu samo susedne vrednosti spojiti. Ciljna i polja unosa mogu biti kontinualna ili kategorička. Za razliku od C&R stabala odlučivanja, ovo stablo ne mora biti binarno.

4.1.4 Neural Net

Neuronska mreža[5] se koristi za temeljnu istragu svih devijacija od norme i klasifikaciju datih podataka. Sastoji se od slojeva neurona. Postoje ulazni, izlazni i skriveni slojevi. Za aktivaciju mreže koristi se aktivaciona funkcija (sigmoid, ReLU, softmax...). Za proveru greške pri treniranju modela se koristi funkcija greške (npr. entropija). Model se trenira određeni broj epoha, sve dok preciznost ne iskonvergira.

4.1.5 Ostali algoritmi

Osim nabrojanih algoritama za klasifikaciju se mogu koristiti i drugi algoritmi (npr. SVM i KNN), međutim zbog veličine korišćenog skupa podataka, za njihovo izvršavanje je potrebno mnogo više vremena, a dobijeni rezultati nisu bolji od onih koji su prikazani u ovom radu. Iz tog razloga njih nismo koristili u završnoj verziji.

4.2 Klasifikacija po polu

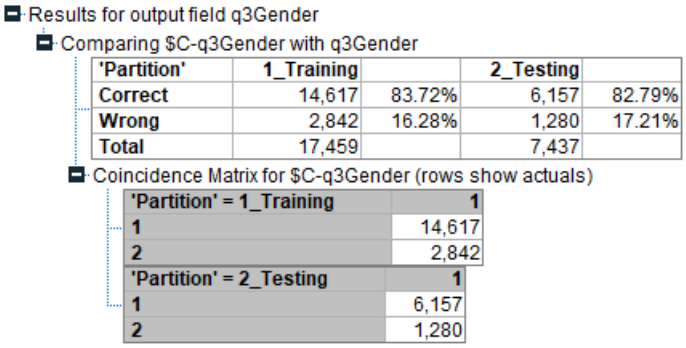
Na anketi je na pitanje o polu bilo moguće odabrati jednu od tri opcije:

- 1 - muško
- 2 - žensko
- 3 - treći (non-binary)

Pošto je ljudi sa obeleženom opcijom 3, kao i ispitanika koji nisu popunili ovo polje, bilo zanemarljivo malo, oni nisu uzeti u obzir za klasifikaciju.

4.2.1 C5.0

Rezultati klasifikacije korišćenjem C5.0 algoritma se nalaze na slici 2.


Results for output field q3Gender

Comparing $C-q3Gender with q3Gender

'Partition'	1_Training		2_Testing	
Correct	14,617	83.72%	6,157	82.79%
Wrong	2,842	16.28%	1,280	17.21%
Total	17,459		7,437	

Coincidence Matrix for $C-q3Gender (rows show actuals)

'Partition' = 1_Training	1
1	14,617
2	2,842

'Partition' = 2_Testing	1
1	6,157
2	1,280

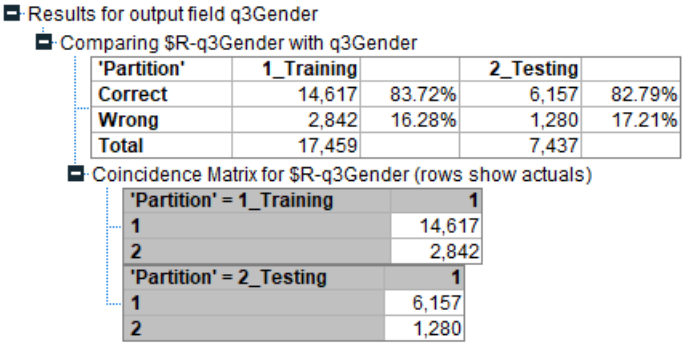
Slika 2: Klasifikacija po polu koristeći C5.0 algoritam

Dostignuta preciznost na trening skupu iznosi 83.72%, a na test skupu 82.79%.

Klasifikacija po polu nije dobro izvršena. Kao što se može videti na slici 2 sve žene su klasifikovane u muškarce, što dalje implicira da dostignuta preciznost na oba skupa predstavlja sve dobro klasifikovane muškarce. Ovo se dešava zbog toga što nema dovoljno anketiranih žena.

4.2.2 C&R Tree

Rezultati C&R Tree algoritma se nalaze na slici 3.


Results for output field q3Gender

Comparing $R-q3Gender with q3Gender

'Partition'	1_Training		2_Testing	
Correct	14,617	83.72%	6,157	82.79%
Wrong	2,842	16.28%	1,280	17.21%
Total	17,459		7,437	

Coincidence Matrix for $R-q3Gender (rows show actuals)

'Partition' = 1_Training	1
1	14,617
2	2,842

'Partition' = 2_Testing	1
1	6,157
2	1,280

Slika 3: Klasifikacija po polu koristeći C&R Tree algoritam

Dostignuta je preciznost od 83.72% na trening, i 82.79% na test skupu.

Nažalost i ovde algoritam prepoznaje samo jednu klasu (muškarce) zbog manjka ženskih ispitanika.

4.2.3 CHAID

Rezultati CHAID algoritma se nalaze na slici 4.

Results for output field q3Gender

Comparing $R-q3Gender with q3Gender

'Partition'	1_Training		2_Testing	
Correct	14,625	83.77%	6,155	82.76%
Wrong	2,834	16.23%	1,282	17.24%
Total	17,459		7,437	

Coincidence Matrix for $R-q3Gender (rows show actuals)

'Partition' = 1_Training	1	2
1	14,497	120
2	2,714	128

'Partition' = 2_Testing	1	2
1	6,114	43
2	1,239	41

Slika 4: Klasifikacija po polu koristeći CHAID algoritam

Dostignuta je preciznost od 83.72% na trening, i 82.76% na test skupu. Ovaj algoritam prepoznaje dve ciljne kategorije iako drugu loše klasifikuje kao i prethodni algoritmi.

4.3 Klasifikacija po starosti

Na anketi je na pitanje o godinama bilo moguće odabrati devet opcija:

- 1 - ispod 12 godina
- 2 - između 13 i 18 godina
- 3 - između 19 i 24 godina
- 4 - između 15 i 34 godina
- 5 - između 35 i 44 godina
- 6 - između 45 i 54 godina
- 7 - između 55 i 64 godina
- 8 - između 65 i 74 godina
- 9 - preko 75 godina

I ovde smo želeli da koristimo što srazmernije podatke, stoga smo koristili samo kategorije 3, 4 i 5, jer one čine preko 93% svih odgovora.

4.3.1 C5.0

Rezultati klasifikacije korišćenjem C5.0 algoritma se nalaze na slici 5.

Results for output field q2Age

Comparing $C-q2Age with q2Age

'Partition'	1_Training		2_Testing	
Correct	13,007	79.25%	5,369	76.44%
Wrong	3,405	20.75%	1,655	23.56%
Total	16,412		7,024	

Coincidence Matrix for $C-q2Age (rows show actuals)

'Partition' = 1_Training	3	4	5
3	7,877	854	26
4	1,180	4,729	132
5	75	1,138	401

'Partition' = 2_Testing	3	4	5
3	3,396	423	13
4	524	1,851	98
5	31	566	122

Slika 5: Klasifikacija po starosti koristeći C5.0 algoritam

Dostignuta preciznost na trening skupu iznosi 79.25%, a na test skupu 76.44%. Razlika između ovog i preprilagođenog modela je na trening podacima bila oko 20%, dok na test podacima razlike skoro uopšte nije bilo.

Klase 3 i 4 se dobro klasifikuju, dok se klasa 5 mahom klasifikuje kao klasa 4. To znači da su njihovi odgovori bili najsličniji anketiranim osobama koje pripadaju klasi 4. Da bi se mogla napraviti razlika između ove dve kategorije bilo bi nam potrebno da anketiramo još ljudi iz kategorije 5 i na taj način prikupimo neophodne podatke, ili da spojimo kategorije 4 i 5 i zapravo vršimo klasifikaciju u dve ciljne klase (između 19 i 24, i između 25 i 44 godine).

4.3.2 C&R Tree

Rezultati C&R Tree algoritma se nalaze na slici 6.

Results for output field q2Age

Comparing $R-q2Age with q2Age

'Partition'	1_Training		2_Testing	
Correct	12,366	75.35%	5,300	75.46%
Wrong	4,046	24.65%	1,724	24.54%
Total	16,412		7,024	

Coincidence Matrix for $R-q2Age (rows show actuals)

'Partition' = 1_Training	3	4
3	7,877	880
4	1,552	4,489
5	136	1,478

'Partition' = 2_Testing	3	4
3	3,447	385
4	620	1,853
5	50	669

Slika 6: Klasifikacija po starosti koristeći C&R Tree algoritam

Dostignuta je preciznost od 75.35% na trening, i 75.45% na test skupu. Ovaj algoritam pravi razliku između dve klase, a to su kategorije 3 i 4. Sve osobe koje pripadaju kategoriji 5 su većinom svrstane u kategoriju 4. Došli smo do istog zaključka kao i kod C5.0 algoritma.

4.3.3 CHAID

Rezultati CHAID algoritma se nalaze na slici 4.

Results for output field q2Age

Comparing $R-q2Age with q2Age

'Partition'	1_Training		2_Testing	
Correct	12,530	76.35%	5,289	75.3%
Wrong	3,882	23.65%	1,735	24.7%
Total	16,412		7,024	

Coincidence Matrix for $R-q2Age (rows show actuals)

'Partition' = 1_Training	3	4	5
3	7,620	1,124	13
4	1,191	4,670	180
5	66	1,308	240

'Partition' = 2_Testing	3	4	5
3	3,303	518	11
4	492	1,903	78
5	24	612	83

Slika 7: Klasifikacija po starosti koristeći CHAID algoritam

Dostignuta je preciznost od 76.35% na trening, i 75.3% na test skupu. Ovaj algoritam se ponaša identično prema kategoriji broj 5 kao i prethodni algoritmi.

4.4 Klasifikacija po zaposlenju

Na pitanje o zaposlenju je bilo devetnaest mogućih opcija, međutim zbog lakše klasifikacije mi smo pozicije reorganizovali na sledeći način:

- 1 - Web Developer
- 5 - Data Analyst
- 7 - Mobile Developer
- 9 - Software Engingeer
- 12 - Engineer
- 15 - Administrator
- 18 - Student

Zbog malog broja anketiranih osoba koje pripadaju kategorijama 5, 7, 12 i 15, one su zanemarene.

4.4.1 C5.0

Rezultati klasifikacije korišćenjem C5.0 algoritma se nalaze na slici 8.

Results for output field q9CurrentRole

Comparing $C-q9CurrentRole with q9CurrentRole

'Partition'	1_Training		2_Testing	
Correct	12,477	82.65%	4,626	71.44%
Wrong	2,620	17.35%	1,849	28.56%
Total	15,097		6,475	

Coincidence Matrix for $C-q9CurrentRole (rows show actuals)

'Partition' = 1_Training	1	18	9
1	4,117	536	570
18	76	5,694	83
9	996	359	2,666

'Partition' = 2_Testing	1	18	9
1	1,400	285	550
18	74	2,423	37
9	749	154	803

Slika 8: Klasifikacija po zaposlenju koristeći C5.0 algoritam

Dostignuta preciznost na trening skupu iznosi 82.65%, a na test skupu 71.44%. Iako se kategorija 9 dobro klasifikuje, uočavamo da najveći deo pogrešno klasifikovanih developera ove kategorije potpadne pod kategoriju 1, što znači da uočava izvesnu sličnost u odgovorima ispitanika ove dve klase. Da bi algoritam mogao da napravi bolju podelu između ove dve kategorije trebalo bi povećati broj ispitanika druge kategorije.

4.4.2 C&R Tree

Rezultati klasifikacije korišćenjem C5.0 algoritma se nalaze na slici 9.

Results for output field q9CurrentRole

Comparing $R-q9CurrentRole with q9CurrentRole

'Partition'	1_Training		2_Testing	
Correct	10,876	72.04%	4,619	71.34%
Wrong	4,221	27.96%	1,856	28.66%
Total	15,097		6,475	

Coincidence Matrix for $R-q9CurrentRole (rows show actuals)

'Partition' = 1_Training	1	18	9
1	3,737	730	756
18	95	5,700	58
9	2,201	381	1,439

'Partition' = 2_Testing	1	18	9
1	1,566	321	348
18	49	2,467	18
9	958	162	586

Slika 9: Klasifikacija po zaposlenju koristeći C&R Tree algoritam

Dostignuta preciznost na trening skupu iznosi 72.04%, a na test skupu 71.34%. Algoritam se ponaša isto kao algoritam C5.0.

4.4.3 CHAID

Rezultati CHAID algoritma se nalaze na slici 10.

Results for output field q9CurrentRole

Comparing $R-q9CurrentRole with q9CurrentRole

'Partition'	1_Training		2_Testing	
Correct	11,017	72.97%	4,656	71.91%
Wrong	4,080	27.03%	1,819	28.09%
Total	15,097		6,475	

Coincidence Matrix for $R-q9CurrentRole (rows show actuals)

'Partition' = 1_Training	1	18	9
1	3,301	730	1,192
18	47	5,700	106
9	1,624	381	2,016

'Partition' = 2_Testing	1	18	9
1	1,366	321	548
18	18	2,467	49
9	721	162	823

Slika 10: Klasifikacija po zaposlenju koristeći CHAID algoritam

Dostignuta je preciznost od 72.79% na trening, i 71.91% na test skupu. Ovaj algoritam se ponaša identično prema kategoriji 9 kao i prethodni algoritmi.

4.5 Klasifikacija po zemljama

Ukupan broj različitih zemalja je 152. Nemamo dovoljno podataka kako bismo mogli da razlikujemo 152 ciljne klase. Najveći broj ispitanika je iz Indije i SAD. Raspodela po preostalim zemljama je drastično mala, stoga smo ih sve svrstali u jednu kategoriju koju smo nazvali "Other".

4.5.1 CNN

Rezultati rada neuronske mreže se nalaze na slici 11.

Results for output field Label

Comparing $N-Label with Label

'Partition'	1_Training		2_Testing		3_Validation	
Correct	9,974	67.55%	2,092	65.87%	2,081	65.9%
Wrong	4,791	32.45%	1,084	34.13%	1,077	34.1%
Total	14,765		3,176		3,158	

Coincidence Matrix for $N-Label (rows show actuals)

'Partition' = 1_Training	India	Other	United States
India	4,595	696	405
Other	1,039	3,808	781
United States	424	1,446	1,571

'Partition' = 2_Testing	India	Other	United States
India	964	165	95
Other	222	813	180
United States	109	313	315

'Partition' = 3_Validation	India	Other	United States
India	934	158	76
Other	226	829	176
United States	114	327	318

Slika 11: Klasifikacija po zemljama koristeći neuronsku mrežu

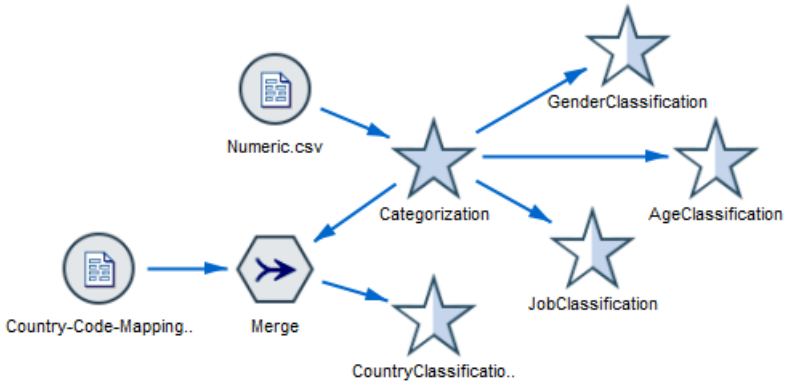
Za rad neuronske mreže, skup podataka je podeljen na tri dela: podatke za trening (70%), podatke za validaciju (15%) i podatke za testiranje (15%). Validacioni skup se koristi kako bismo se u toku samog treninga imao uvid u to kako se model ponašao nad podacima nad kojima još nije trenirao. Dostignuta je preciznost od 67.55% na trening, 65.87% na test skupu i 65.9% na validacionom skupu.

Najbolje se klasifikuje Indija, dok se najlošije klasifikuje SAD. Broj dobro klasifikovanih ispitanika iz SAD je skoro jednak broju ispitanika koji su pogrešno klasifikovani u "Other".

5 Zaključak

U klasifikaciji po polu se najbolje pokazao CHAID algoritam iako po ovom atributu klasifikacija nije dobra, jer ne postoji dovoljan broj ženskih ispitanika ove ankete. U klasifikacijama po godinama i poziciji se najbolje pokazao algoritam C5.0. Najveći broj ljudi koji se bave programiranjem su osobe starosti između 19 i 44 godine, gde najmanje ima osoba od 35 do 44 godine. One su mahom klasifikovane u grupu od 25 do 34 godine. Što se tiče zaposlenja, možemo zaključiti da je klasifikacija po ovom atributu dala najbolje rezultate. Kod klasifikacije zemalja je postojao problem klasifikovanja osoba iz SAD. Uzrok ovakvom lošem klasifikovanju je to što nakon spajanja ispitanika koji nisu iz SAD i Indije, osoba iz SAD ima upola manje nego u preostale dve kategorije.

Prikaz celog toka je dat na slici 12.



Slika 12: Ceo SPSS tok

Literatura

[1] Charu C. Aggarwal. *Data Mining The Textbook*. Springer, New York, 2015.

[2] IBM. C& r tree node, 2012. on-line at: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/cartnode_general.htm.

[3] IBM. C5.0 node, 2012. on-line at: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm.

[4] IBM. Chaid node, 2012. on-line at: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/chaidnode_general.htm.

[5] IBM. Neural network node, 2012. on-line at: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/trainnetnode_general.htm.

[6] Mladen Nikolić Predrag Janičić. Veštačka inteligencija, 2018. on-line at: http://poincare.matf.bg.ac.rs/~janicic//courses/vi.pdf.