Univerzitet u Beogradu

Matematički fakultet

Katedra za računarstvo i informatiku

Istraživanje o mladima

Seminarski rad u okviru kursa Istraživanje podataka

Student: Jovana Pejkić 2016/435

Profesor: Nenad Mitić
Asistent: Mirjana Maljković

Beograd, 2018.

Sažetak

U ovom radu biće predstavljeno nekoliko algoritama za istraživanje podataka. Na samom početku biće opisan skup podataka nad kojim će istraživanje biti sprovedeno, a onda će nad tim skupom biti obavljena analiza i pretprocesiranje (poglavlje 2). Nakon toga će nad tim podacima biti primenjena pravila pridruživanja (poglavlje 3), zatim klaster analizom proveriti da li ljudi čine klaster na osnovu sličnog ponašanja ukoliko posmatramo vrstu muzike koju slušaju (poglavlje 4) i na samom kraju demonstriraćemo nekoliko algoritama klasifikacije (poglavlje 5).

Sadržaj

1. Uvod

U ovom radu biće demonstriran rad algoritama iz oblasti istraživanja podataka kao i način pripremanja podataka pre njihove primene. Algoritmi će biti prikazani odgovaranjem na sledeća pitanja:

- Da li plašljivi ljudi izbegavaju neki odredjeni žanr filmova?
- Imajući u vidu koju muziku preferiraju, da li mogu da se izdvoje različite grupe ljudi?
- Kako starosno doba utice na muziku koja se slusa?
- Koji muzički žanrovi (ne) idu jedni uz druge?
- Da li se može pre predvideti pol osobe na osnovu podataka o visini, težini i u kojoj meri se plaši paukova?
- Na sta mladi troše novac?
- Da li se žene više plaše od muškaraca?

Skup podataka koji koristimo preuzet je sa: https://www.kaggle.com/miroslavsabo/youngpeople-survey/data.

2. Analiza i pretprocesiranje podataka

2.1. Analiza podataka

2.1.1. O podacima

Skup podataka “Young people survey” sastoji se iz dve datoteke: responses.csv i columns.csv. Iako će biti obradjivani podaci isključivo iz datoteke responses.csv, datoteka columns.csv objašnjava značenje atributa prve datoteke, te su obe datoteke značajne za ovo istraživanje.

Fajl responses.csv sadrži 1010 redova i 150 kolona (139 numeričkih atributa i 11 kategoričkih). U ovom fajlu, originalna imena atributa su skraćena (puna imena se nalaze u fajlu columns.csv). Podaci sadrže nedostajuće vrednosti. Svi ispitanici su slovačkog porekla, starosti izmedju 15 i 30 godina.

Korišćeni atributi mogu biti podeljeni u sledeće grupe:

- Izbor muzike (19 stavki)
- Izbor filmova (12 stavki)
- Hobiji i interesovanja (32 stavke)
- Strahovi (10 stavki)
- Zdrave navike (3 stavke)
- Osobine, pogled na svet i mišljenja (57 stavki)
- Potrošačke navike (7 stavki)
- Demografija (10 stavki)

Opisi datoteka su dati u sledećim tabelama:

original	**short**
Skraćen naziv za isti atribut	Pun naziv atributa koji predstavlja pitanje

Tabela columns.csv

Music	Uživam slušajuci muziku
Slow songs or fast songs	Preferiram laganu ili brzu muziku
Dance	Uživam u muzici za ples
Folk	Uživam u folk muzici
Country	Uživam u kantri muzici
Classical music	Uživam u klasičnoj muzici
Musical	Uživam u mjuziklu
Pop	Uživam u pop muzici
Rock	Uživam u rok muzici
Metal or hardrock	Uživam u metal muzici
Punk	Uživam u pank muzici
Hiphop, rap	Uživam u hiphop, rep muzici
Reggae, Ska	Uživam u rege muzici
Swing, Jazz	Uživam u džez muzici

Rock n roll	Uživam u “rock and roll” muzici
Alternative	Uživam u alternativnoj muzici
Latino	Uživam u latinskoj muzici
Techno, Trance	Uživam u tehno muzici
Opera	Uživam u operi
Movies	Uživam gledajuci filmove
Horror	Uživam u horor filmovima
Thriller	Uživam u trilerima
Comedy	Uživam u komedijama
Romantic	Uživam u romanticnim filmovima
Sci-fi	Uživam u naucno-fantasticnim filmovima
War	Uživam u ratnim filmovima
Fantasy/ Fairy tale	Uživam u fantazijama/ bajkama
Animated	Uživam u animiranim filmovima
Documentary	Uživam u dokumentarnim filmovima
Western	Uživam u kaubojskim filmovima
Action	Uživam u akcionim filmovima
History	Interesujem se za istoriju
Psychology	Interesujem se za psihologiju
Politics	Interesujem se za politiku
Mathematics	Interesujem se za matematiku
Physics	Interesujem se za fiziku
Internet	Interesujem se za internet
PC	Interesujem se za PC
Economy management	Interesujem se za ekonomski menadzment
Biology	Interesujem se za biologiju
Chemistry	Interesujem se za hemiju
Reading	Interesujem se za citanje poezije
Geography	Interesujem se za geografiju
Foreign Languages	Interesujem se za strane jezike
Medicine	Interesujem se za medicinu
Law	Interesujem se za pravo
Cars	Interesujem se za automobile
Art exhibitions	Interesujem se za izlozbe
Religion	Interesujem se za religiju
Countryside, outdoors	Interesujem se za aktivnosti na otvorenom
Dancing	Interesujem se za ples
Musical instruments	Interesujem se za sviranje muzickog instrumenta
Writing	Interesujem se za pisanje poezije
Passive sport	Interesujem se za rekreativno bavljenje sportom
Active sport	Interesujem se za takmicenje u sportu
Gardening	Interesujem se za baštovanstvo
Celebrities	Interesujem se za nacin zivota poznatih
Shopping	Interesujem se za kupovinu
Science and technology	Interesujem se za nauku i tehnologiju
Theatre	Interesujem se za pozorista
Fun with friends	Interesujem se za zabavu sa prijateljima (socijal.)
Adrenaline	Interesujem se za adrenalinske sportove
Pets	Interesujem se za kucne ljubimce
Flying	Plašim se letenja
Storm	Plašim se oluje
Darkness	Plašim se mraka
Heights	Plašim se visine
Spiders	Plašim se pauka
Snakes	Plašim se zmija
Rats	Plašim se miseva

Ageing	Plašim se starenja
Dangerous dogs	Plašim se opasnih pasa
Fear of public speaking	Plašim se javnog nastupa
Smoking	Pušim
Healthy eating	Vodim zdrav zivot
Daily events	Primecujem sta se zbiva oko mene
Prioritising workload	Ne ostavljam obaveze za poslednji minut
Writing notes	Uvek pravim "TODO" listu
Workaholism	Radim cak i u slobodno vreme
Thinking ahead	Dobro razmislim pre nego sto nešto uradim
Final judgement	Verujem da će losi ljudi patiti jednog dana, a da će dobri biti nagradheni
Reliability	Pouzdana sam osoba
Keeping promises	Uvek odrzim onbecanje
Loss of interest	Brzo gubim interesovanje
Friends versus money	Pre bih da imam mnogo prijatelja nego mnogo para
Funniness	Trudim se da budem duhovit/a
Fake	Mogu da budem dvolican/dvolicna ponekad
Criminal damage	U proslosti sam ostećivao stvari kada sam bio ljut
Decision making	Koristim svo potrebno vreme da bi doneo/la odluku
Elections	Uvek se trudim da glasam na izborima
Self-criticism	Samokritican/na sam, zalim za odlukama koje sam doneo/la
Judgment calls	Primecujem kada me ljudi slušaju, a kada ne dok pričam sa njima
Hypochondria	Hipohondrican/na sam
Empathy	Empaticna sam osoba
Eating to survive	Jedem samo zato sto moram, ne Uživam u hrani
Giving	Trudim se da poklanjam ljudima sto više
Compassion to animals	Ne volim da vidim zivotinje kako pate
Borrowed stuff	Cuvam stvari koje sam pozajmio/la od nekog
Loneliness	Osećam se usamljeno u zivotu
Cheating in school	"Podvaljivao/ la" sam u skoli
Health	Brinem o svom zdravlju
Changing the past	Voleo/ la bih da promenim proslost
God	Verujem u Boga
Dreams	Uvek sanjam dobre snove
Charity	Uvek dajem u dobrotvorne svrhe
Number of friends	Imam puno prijatelja
Punctuality	Koliko sam tacan/ tacna
Lying	Da li lazete druge
Waiting	Veoma sam strpljiv/ a
New environment	Brzo se prilagodim novoj sredini
Mood swings	Raspoloženje mi se brzo menja
Appearence and gestures	Imam manire i pazim na svoj izgled
Socializing	Uživam da upoznajem nove ljude
Achievements	Govorim drugima o svojim dostignucima
Responding to a serious letter	Dobro razmislim pre nego odgovorim na vazno pismo
Children	Uživam u drustvu dece
Assertiveness	Ne plašim se da iznesem svoje misljenje ako sam siguran u nešto
Getting angry	Lako se iznerviram
Knowing the right people	Uvek se uverim da li sam povezana sa pravim ljudima

Public speaking	Moram biti dobro spremna pre govora pred publikom
Unpopularity	Kod sebe nalazim mane ako se ne dopadnem drugima
Life struggles	Placem kada stvari ne idu svojim tokom
Happiness in life	100% sam srecna svojim zivotom
Energy levels	Uvek sam pun/a zivota i energije
Small - big dogs	Preferiram velike pse ka malim mirnim psima
Personality	Mislim da su sve moje osobine pozitivne
Finding lost valuables	Ako pronadjem nešto sto mi ne pripada, vraticu
Getting up	Tesko se budim ujutru
Interests or hobbies	Imam mnogo različitih hobija i interesovanja
Parents' advice	Uvek slusam savete mojih roditelja
Questionnaires or polls	Volim da ucestvujem u istrazivanju, anketiranju
Internet usage	Koliko vremena provodite online?
Finances	Cuvam sav novac
Shopping centres	Obozavam da idem u trzni centar
Branded clothing	Preferiram markiranu odecu
Entertainment spending	Trosim mnogo novca na provod
Spending on looks	Trosiim mnogo novca na svoj izgled
Spending on gadgets	Trosim mnogo novca na elektroniku (“gedzete”)
Spending on healthy eating	Rado cu potrositi više novca na zdravu hranu
Age	Godine
Height	Visina
Weight	Tezina
Number of siblings	Broj rodjenih braca i sestara
Gender	Pol
Left - right handed	Levoruk/ desnoruk
Education	Nivo obrazovanja
Only child	Ja sam jedino dete
Village - town	Selo - grad
House - block of flats	Kuca - stan u zgradi

Tabela responses.csv

2.1.2. Znacenje odgovora ankete i tipovi atributa

Izbor muzike

• **Uživam slušajuci muziku**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Preferiram**: Slow paced music 1-2-3-4-5 Fast paced music *(integer)*

• **Dens, Disko, Fank**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Folk muzika**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Kantri**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Clasicna muzika**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Mjuzikl**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Pop**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Rok**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Metal, Hard rok**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Pank**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Hip hop, Rep**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Rege, Ska**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Sving, Džez**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **"Rock n Roll"**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Alternaticna muzika**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Latino muzika**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Tehno, Trans**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Opera**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

Izbor filmova

• **Uživam gledajuci filmove**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Horor filmovi**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Trileri**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Komedije**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Romanticni**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Naucno-fantasticni**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Ratni**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Bajke**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Crtani filmovi**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Dokumentarni filmovi**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Kaubojski filmovi**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

• **Akcioni filmovi**: Uopšte ne uživam 1-2-3-4-5 Mnogo uživam *(integer)*

Hobiji i interesovanja

• **Istorija**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Psihologija**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Politika**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Matematika**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Fizika**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Internet**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **PC Softver, Hardver**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Ekonomija, Menadžment**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Biologija:** Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Hemija**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Čitanje poezije**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Geografija**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Strani jezici**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Medicina**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Prava**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Automobili**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Umetnost**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Religija**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Aktivnosti na otvorenom**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Ples**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Sviranje muzičkog instrumenta**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Pisanje poezije**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Sport i slobodne aktivnosti**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Sport na nivou takmičenja:** Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Baštovanstvo**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Život poznatih**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Kupovina**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Nauka i tehnologija**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Pozorište**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Socijalizacija**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Ekstremni sportovi**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

• **Kućni ljubimci**: Nisam zainteresovan/a 1-2-3-4-5 Veoma sam zainteresovan/a *(integer)*

Fobije

• **Letenje**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Grmljavina, sevanje**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Mrak**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Visina**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Pauci**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Zmije**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Pacovi, miševi**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Starenje**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Opasni psi**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

• **Javni govor (javni nastup)**: Ne plašim se uopšte 1-2-3-4-5 Mnogo se plašim *(integer)*

Zdrave navike

• **Pušačke navike**: Nikad nikad pusio/la - Probao/la sam da pušim - "Bivši" pušač - Pušim (trenutno) *(categorical)*

• **Pijem alkohol**: Nikad - Pijem u drustvu - Puno pijem *(categorical)*

• **Živim veoma zdrav Život**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

Osobine, pogled na svet i mišljenja

• **Primećujem šta se oko mene dešava**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Trudim se da izvršim obaveze na vreme, a ne ostavljam ih za poslednji trenutak**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek napravim spisak, da ne bih nešto zaboravio/la**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek učim ili radim, čak i u slobodno vreme**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Pre nego što nastavim dalje, razmotrim stvari iz više različitih uglova**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Verujem u to da će loši ljudi patiti jednog dana, a da će dobri biti nagrađeni**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Pouzdan/a sam na poslu i svoje zadatke obavljam na vreme**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek držim svoja obećanja**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• Zainteresujem se za nekoga veoma brzo, a onda potpuno izgubim interesovanje: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Pre bih da imam mnogo prijatelja nego mnogo novca:** Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek se trudim da budem najsmešniji/a (najzanimljiviji/a)**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Ponekad sam dvoličan/na**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• U proslosti sam unistavao/la sam stvari kada bih se iznervirao/la: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Koristim svoje vreme da donesem odluku**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Trudim se da uvek glasam na izborima**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek razmišljam i žalim za odlukama koje sam doneo/la**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Mogu da uočim da li me ljudi slušaju dok pričam sa njima ili ne**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Hipohondričan/na sam**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Empatična sam osoba**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Jedem samo zato što moram, ne uživam u hrani i jedem što brže mogu**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Trudim se da poklanjam drugim ljudima za Božić što više mogu**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Ne volim da vidim životinje kako pate**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Pazim na stvari koje sam pozajmio/la od nekoga:** Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Osećam se usamljeno**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Podvaljivao/la sam u školi**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Brinem o svom zdravlju**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Voleo/la bih da promenim prošlost zbog stvari koje sam uradio/la**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Verujem u Boga**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek sam imao/la lepe snove**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek dajem u dobrotvorne svrhe**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Imam mnogo prijatelja:** Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Koliko ste tačni?**: Često poranim. - Uvek na vreme. - Uvek kasnim. *(categorical)*

• **Da li lažete druge?**: Nikad. - Samo da ne bih povredila nekoga. - Ponekad. - Svaki put kada mi odgovara. *(categorical)*

• **Veoma sam pažljiv/a**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Mogu brzo da se prilagodim novoj sredini**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Raspoloženje mi se brzo menja**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Imam manire i vodim računa o svom izgledu**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uživam da upoznajem nove ljude**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek ispričam drugima o mojim dostignucima**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Dobro razmislim pre nego sto odgovorim na važno pismo** : Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uživam da sam u društvu dece**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Ne plašim se da iznesem svoje misljenje ako sam sigurna u nešto**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Lako se iznerviram (lako planem)**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Vodim računa da se povežem sa pravim ljudima**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Dobro se pripermim pre javnog nastupa (govora pred publikom)**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Nalazim sebi mane ako se ne dopadnem drugima**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Plačem kad god sam neraspoložen/a ili stvari ne idu po planu**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **100% sam srećna u zivotu:** Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek sam pun/a života i energije**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Preferiram velike, opasne pse ka malim, mirnijim psima**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Mislim da su sve moje osobine pozitivne**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Ako nađem nešto sto mi ne pripada, vratiću**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Teško mi je da se probudim ujutru**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Imam mnogo različitih hobija i interesovanja**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Uvek slušam savete svojih roditelja**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Volim da učestvujem u anketama, istraživanjima**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Koliko vremena provodite "online"?**: Uopšte ne provodim - Manje od sat vremena dnevno - Nekoliko sati dnevno - Većinu dana *(categorical)*

Potrošačke navike

• **Štedim novac**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Obožavam da obilazim velike tržne cenrte**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Preferiram markiranu odeću ka ne markiranu odecu**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Trošim mnogo novca na žurke i druženje**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Ulažem mnogo novca u svoj izgled**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Trošim mnogo novca na elektroniku**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

• **Pre ću platiti više za dobru, kvalitetnu i zdravu hranu**: Uopšte se ne slažem 1-2-3-4-5 U potpunosti se slažem *(integer)*

Demografija

• **Godine**: *(integer)*

• **Visina**: *(integer)*

• **Težina**: *(integer)*

• **Koliko braća i sestara imate?**: *(integer)*

• **Pol**: ženski - muški *(categorical)*

• **Ja sam**: levoruk/a - desnoruk/a *(categorical)*

• **Nivo obrazovanja**: trenutno sam učenik osnovne skole - osnovna skola (završena) - srednja škola - dipomirao/la na fakultetu *(categorical)*

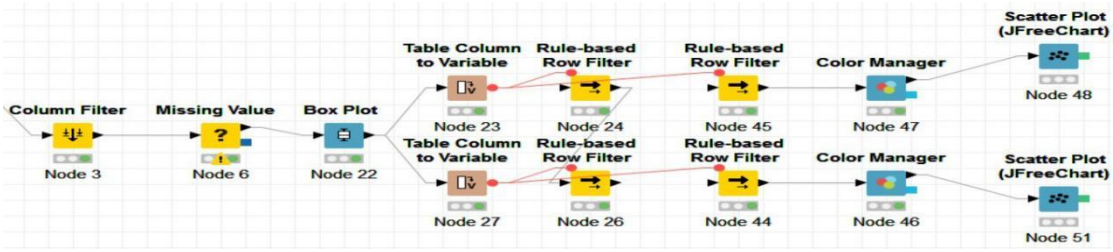
• **Jedino sam dete**: ne - da *(categorical)*

• **Većinu svog detinjstva provela sam u**: gradu - selu *(categorical)*

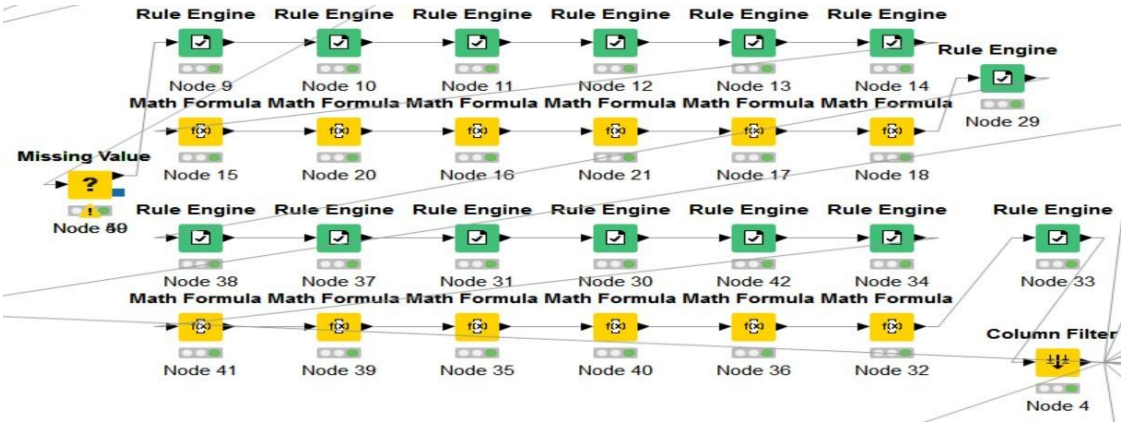
• **Većinu svog detinjstva živela sam u**: kući - stanu *(categorical)*

2.2. Pretprocesiranje podataka

Radi što manjeg gubljenja podataka pretprocesiranje je obavljamo posebno pre primene svakog algoritma. Ono što je uobičajneno za prteprocesiranje koje prethodi svakom primenjenom algoritmu je prvo izbacivanje nepotrebnih kolona, a zatim obrada nedostajućih vrednosti i rad sa nekorektnim podacima (slika 1 i 2).



Slika 1: Izbacivanje autlajera za visinu i težinu



Slika 2: Imputovanje visine i težine

2.2.1. Nedostajuce vrednosti

Polja sa nepoznatom visinom i težinom imputirana su vrednostima na sledeći način. Izračunat je prosek visine i težine odvojeno muškaraca i žena za svaki opseg od 5 godina počev od 15 do 30. Zatim je na mesto nedostajuće vrednosti imputirana odgovarajuća vrednost prema pripadnosti određenoj starosnoj dobi i polu.

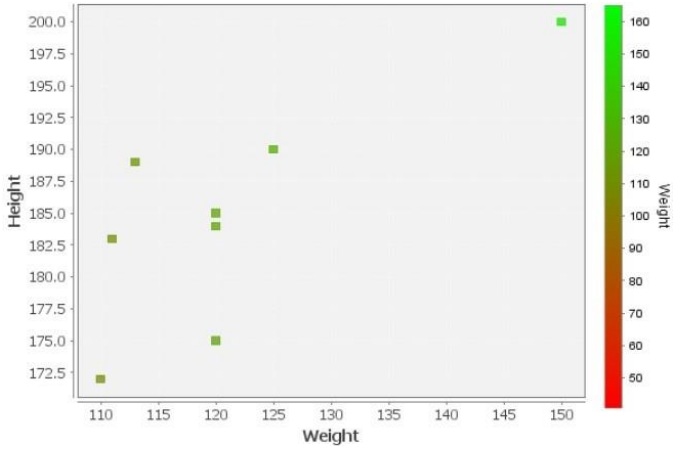
Pre algoritama pravila pridruživanja, gubljenje celokupnih redova zbog nekoliko (ili jedne) nedostajućih vrednosti nije dobro došlo, tako da su nedostajuće vrednosti samo ignorisane pri spajanju odabranih kolona. Detaljniji opis korišćenih čvorova dat je u poglavlju o pravilima pridruživanja. U svim ostalim slučajevima izbačeni su svi redovi koji sadrže nedostajuće vrednosti.

2.2.2. Rad sa nekorektnim podacima

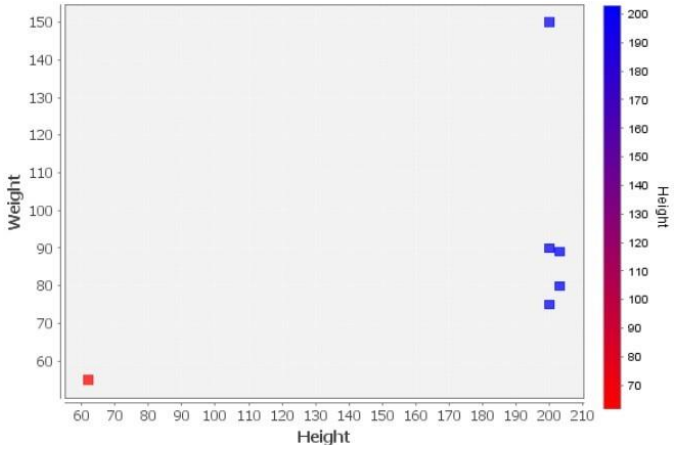
Na većinu pitanja u anketi mogući odgovori su bili brojevi od 1 do 5, tako da je ovo domensko znanje iskorišćeno za proveru neispravnih podataka u okviru kolona koje beleže odgovore na ovaj način. Za proveru je korišćen čvor CSV reader, tabela "spec" i ispostavljeno je da nema podataka van domena.

Kako imamo kolone sa informacijom da li je osoba jedinac i koliko rodjenih braće/sestara ima primećeno je da se podaci ne slažu. Problem je rešen brisanjem celokupnih slogova sa ovom kontradiktornošću.

Podaci kao što su oni o visini i težini, provereni su box-plotom, a celokupni redovi u kojima su se našli autlajeri su izbačeni. Autlajeri su predstavljeni na sledećim slikama. Radi poređenja data je i slika zajedno sa prihvaćenim podacima.



slika height(y-osa), weight (x-osa) - svi podaci



slika weight(y-osa), height (x-osa) - svi podaci

3. Neke zanimljive statistike i vizuelizacija podataka

Nakon prečišćavanja podataka, upotrebljen je čvor Statistics. Pomoću njega se može videti za koja pitanja je bilo najviše pozitivnih, a za koja najviše negativnih odgovora.

Tako, ako se tabela sortira sa statistikama opadajuće po koloni “Mean” i zanemare prva tri reda (jer ti podaci nisu trenutno značajni), može se uočiti da je najviše pozitivnih odgovora bilo na pitanje “Da li uživate u muzici?” sa prosekom 4.763, zatim na pitanje “Da li uživate u filmovima” sa nešto manjim prosekom 4.639, a onda na pitanja “Da li volite da se zabavljate sa prijateljima?” i “Da li uživate u komedijama?” (slika 3).

Row ID	S Column	D Min	D Max	D Mean	D Std. d...	D Varian...
Height	Height	152	203	173.42	9.476	89.789
Weight	Weight	41	150	66.117	13.9	193.218
Age	Age	15	30	20.353	2.733	7.468
Music	Music	1	5	4.763	0.595	0.354
Movies	Movies	1	5	4.639	0.658	0.433
Fun with friends	Fun with f...	2	5	4.552	0.742	0.551
Comedy	Comedy	1	5	4.515	0.746	0.556
Internet	Internet	1	5	4.187	0.909	0.827

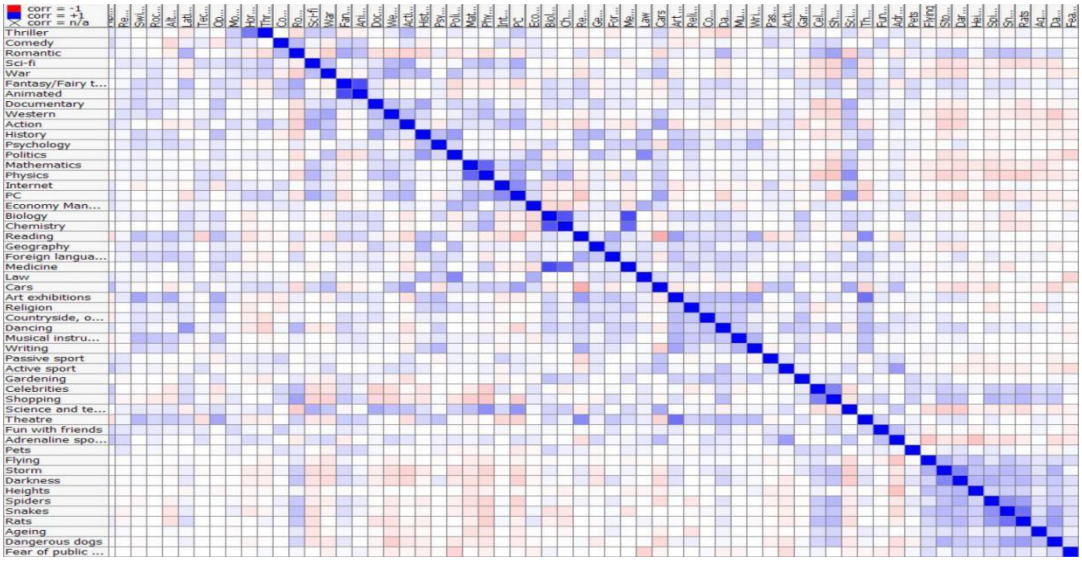
Slika 3: Tabela sa statistikama sortirana opadajuće po atributu “Mean”

Sada, ako se vrednosti atributa kolone “Mean” sortiraju rastuće i pritom zanemari prvi red (jer nije od značaja), dobija se podatak da je najviše negativnih odgovora bilo na pitanje “Da li ste hipohondrični?” sa prosekom 1.853, a slične proseke imaju i pitanja “Da li volite da pišete poeziju?” i “Da li se bavite baštovanstvom?”. Takođe, iz tabele ispod se može zaključiti da se ispitanici uglavnom ne plaše oluje i letenja (slika 4).

Row ID	Column	Min	Max	Mean	Std. d...	Varian...
Number of siblings	Number of...	0	10	1.3	0.993	0.986
Hypochondria	Hypochon...	1	5	1.853	1.129	1.276
Writing	Writing	1	5	1.874	1.289	1.662
Gardening	Gardening	1	5	1.877	1.159	1.344
Storm	Storm	1	5	1.936	1.137	1.293
Flying	Flying	1	5	1.991	1.177	1.385
Charity	Charity	1	5	2.043	1.017	1.034

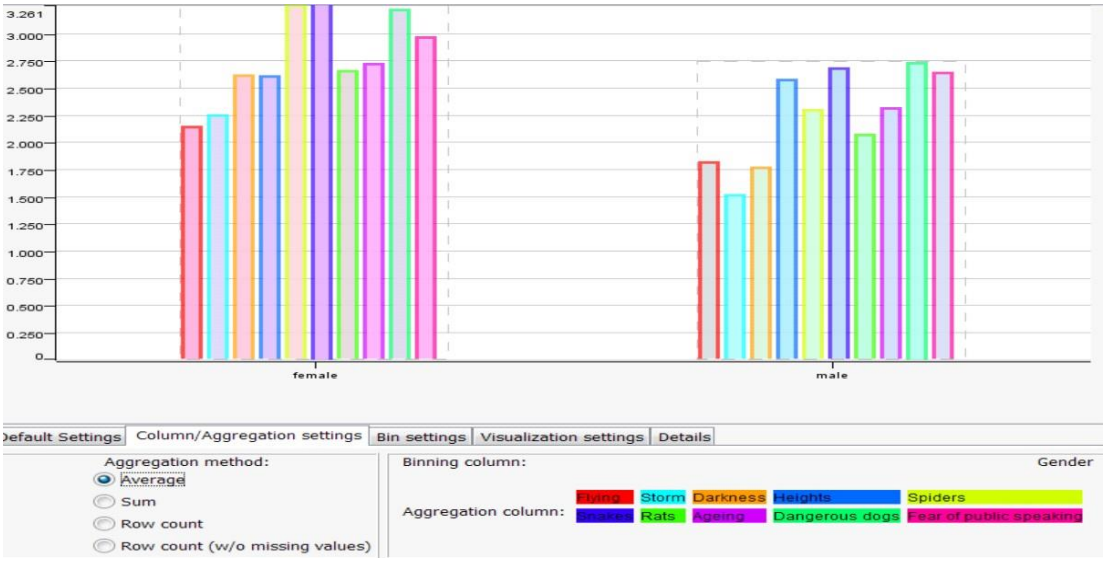
Slika 4: Tabela sa statistikama sortirana rastuće po atributu "Mean"

Dalje, primenom čvora Linear Correlation može se zaključiti da su neki atributi u korelaciji tj. da su povezani (sto je boja tamnije plava to znači da je veza izmedju odgovarajućih atributa jača, dok crvena boja znači da nisu u korelaciji). Na slici ispod prikazan je samo deo korelacione matrice koji je najzanimljiviji jer ima najviše korelisanih atributa. Na primer, može se videti da su vrednosti atributa "Biology" (odnosno "Interesujem se za biologiju") i atributa "Medicine" (odnosno "Interesujem se za medicinu") u jakoj korelaciji, čak 0.7191. Očekivano, u velikoj korelaciji su i vrednosti za atribute "Weight" (tezina) i "Height" (visina), 0.7294 (slika 5).



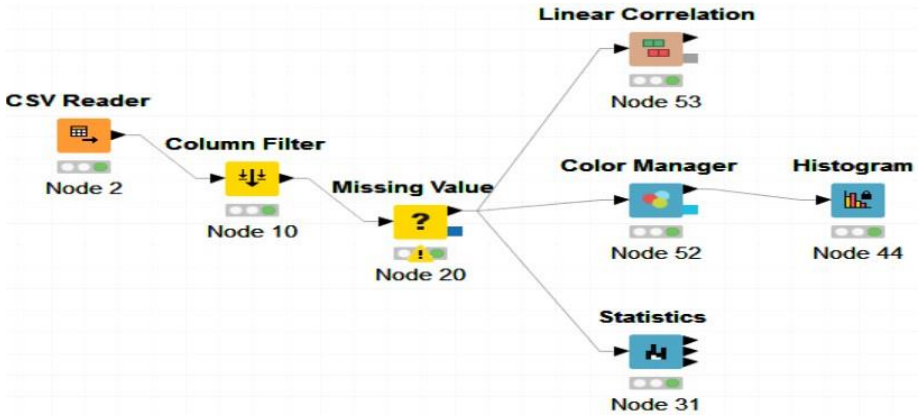
Slika 5: Matrica korelacije

Još jedan zanimljiv podatak do koga se dolazi jeste da se žene svega od ponudjenog više plaše od muškaraca. Za ovo istraživanje upotrebljeni su čvorovi Color Manager i Histogram (slika 6). Na histogramu se jasno vidi da je razlika u strahu od pauka najveća izmedju muškaraca i žena, i to da se žene mnogo više plaše pauka od muškaraca (žuto).



Slika 6: Histogram

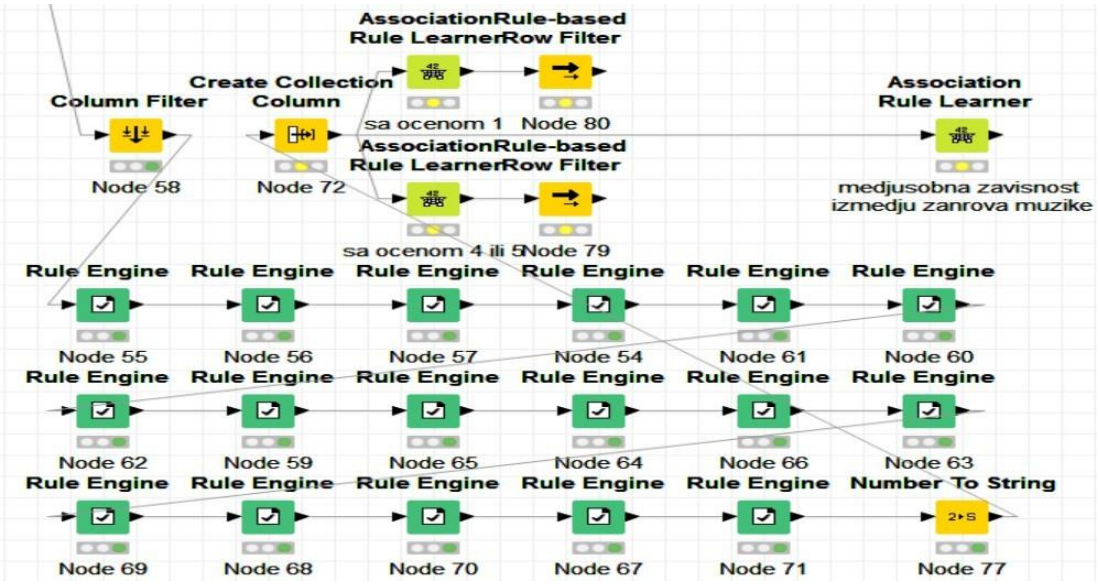
Na slici ispod prikazano je kako su povezani korisćeni čvorovi (slika 7).



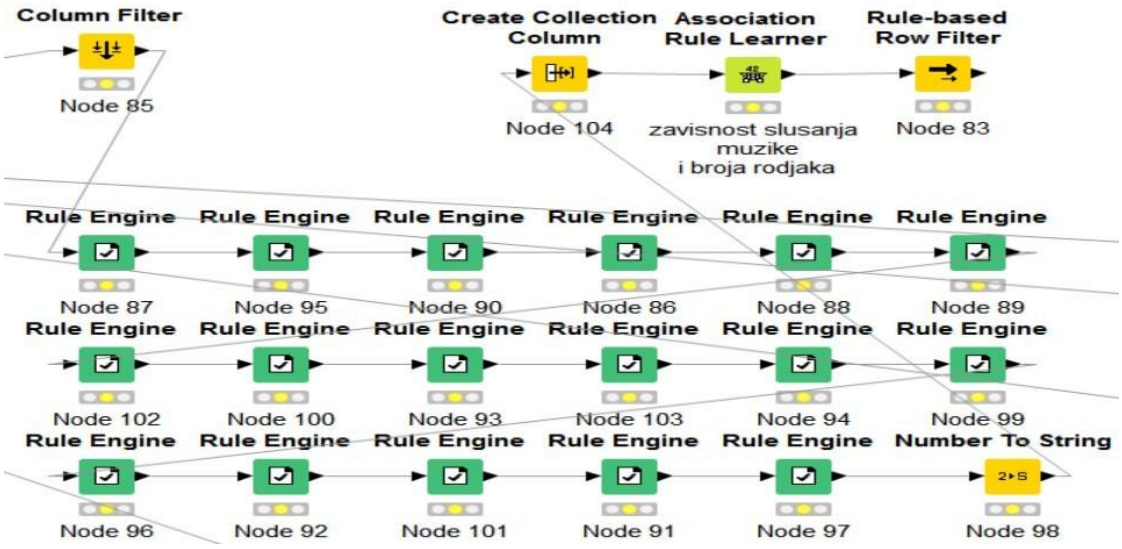
Slika 7: Raspored čvorova

4. Pravila pridruživanja

U ovom delu istraživanja cilj je da se pokaže kako starosna dob utiče na slušanje određene vrste muzike. Primetiće se da među mladima uglavnom postoje omiljeni žanrovi, dok kod odraslih to nije slučaj. Pretpostavka je da je razlog tome to što u današnje vreme mladi mnogo više vremena provode družeći se nego odrasli. Takođe, njihovo druženje često se svodi na izlaske na mesta gde se sluša muzika. Na pitanje da li uživaju u muzici i da li uživaju u druženju sa prijateljima, upravo su mladi ispitanici (18 - 21 god.) ti koji su odgovorili strogo pozitivno (ocenama 4 i 5 u značenju potvrde). Ovo ide u prilog upravo pomenutoj pretpostavci (slike 8 i 9).



Slika 8: Pravila pridruživanja



Slika 9: Pravila pridruživanja gde se gleda uticaj broja rođenih braće i sestara na odabir muzike

Ono što se može zaključiti vezano za starosnu dob i žanr muzike koji je za nju karakterističan je da mladi u 18., 19., 20. i 21. godini izrazito slušaju rok, a u 18. godini i pop. Takođe, dvadesetogodišnjaci izrazito ne uživaju u metal i hardrok muzici, tehno i transu i u operi, dok se za devetnaestogodišnjake može reći da ne uživaju u operi. Ono što se može zaključiti o poveznosti žanrova muzike je sledeće:

- ko ne sluša punk ne sluša ni metal i hard rok
- ko sluša dance sluša i pop muziku
- ko sluša nešto od metal i hardrok, pank, rok en' rol, alternativnu muziku, klasičnu, sving i džez ili rege i ska, sluša i rok

Istraživanje je obavljeno pre svega obradom informacija koje su beležile odgovore na pitanja o uživanju u određenom žanru. Da ne bi došlo do nagomilavanja brojeva od 1 do 5 pri spajanju kolona, prethodno su svi intiger podaci prevedeni u string i to na sledeći način:

- ako je podatak bio “1”, to znači da osoba uopšte ne uživa slušajući odredjenu vrstu muzike, s toga je svako pojavljivanje "1" zamenjeno sa šablonom "ne_imeŽanra"
- kako su najinteresantniji ekstremni slučajevi, tj. kada neko uopšte ne uživa u odredjenoj muzici ili apsolutno uživa u njoj, vrednosti "2" i "3" su zanemarivane (najpre ostavljene kao unknown, a zatim ignorisane)
- vrednosti "4" i "5" označavaju "više nego uobičajeno", s toga se one uzimaju u obzir pri istraživanju i predstavljene su string vrednostima po šablonu "imeŽanra"

Pri spajanju kolona, ignorisane su nedostajuće vrednosti (sva mesta gde su se pojavljivale 2 i 3, kao i gde su inače bile nedostajuće vrednosti), a korišćen je čvor Create Collection Column. Pri upotrebi čvora Association Rule Learner izabrane su sledeće bitne vrednosti:

- minimum suport = 0.06 (budući da ima 15 različitih vrednosti za godinu koja je od interesa, 0.067 je prosek pojavljivanja, i traženo je da se informacije o godini pojavljuje nešto malo manje nego sto bi bio prosečan broj pojavljivanja)
- minimum confidence = 0.5 (kako je za svakog ispitinika verovatnoća da za odredjeni

Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule188	0.073	0.6	1.131	pop	<---	[18]
rule248	0.078	0.642	1.017	rock	<---	[18]
rule112	0.067	0.846	1.341	rock	<---	[19,rock and roll]
rule457	0.136	0.644	1.021	rock	<---	[19]
rule439	0.124	0.642	1.017	rock	<---	[20]
rule306	0.085	0.672	1.065	rock	<---	[21]
rule217	0.075	0.661	1.047	rock	<---	[pop,19]
rule55	0.064	0.63	0.998	rock	<---	[pop,20]
rule141	0.069	0.84	1.33	rock	<---	[rock and roll,20]

žanr ima ekstremno visoku ocenu (4 ili 5) 0.4 budući da ima 5 različitih vrednosti ukupno i pritom verovatnoća da su svi ispitanici istih godina ekstremno visoko ocenili uživanje u nekom žanru je još manja, smatra se da je 0.5 dovoljno visoka vrednost za ovu variablu poverenja)

- minimum confidance = 0.4 (kako je neuživanje u nekom žanru još manja verovatnoća naći, za proveru ovoga korišćen je manji parametar. U ovom slučaju zanemarena su dobijena pravila koja su se odnosila na ocene 4 i 5).
- minimum confidance = 0.7 AND minimum suport = 0.2 (pri uočavanju pravila između žanrova muzike)

rule604	0.08	0.416	1.085	no metal and hardrock	<---	[20]
rule644	0.082	0.426	1.122	no techno, trance	<---	[20]
rule667	0.083	0.432	1.121	no opera	<---	[20]
rule698	0.085	0.404	1.049	no opera	<---	[19]

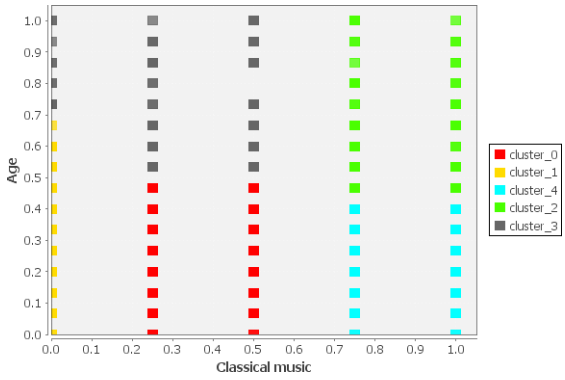
Row ID	Support	Confide...	Lift	Consequent	implies	Items
rule4	0.234	0.737	1.924	no metal and hardr...	<---	[no punk]
rule2	0.215	0.926	1.467	rock	<---	[metal and hardrock]
rule1	0.213	0.917	1.453	rock	<---	[punk]
rule7	0.269	0.728	1.372	pop	<---	[dance]
rule8	0.335	0.84	1.331	rock	<---	[rock and roll]
rule6	0.264	0.825	1.308	rock	<---	[alternative]
rule5	0.249	0.752	1.191	rock	<---	[clasical music]
rule3	0.218	0.736	1.167	rock	<---	[swing, jazz]
rule0	0.204	0.723	1.146	rock	<---	[reggea, ska]

5. Klaster analiza

5.1. Šta u kom uzrastu mladi slušaju

Da bi se došlo do toga šta u kom uzrastu mladi slušaju, neophodno je prvo pripremiti podatke korišćenjem čvorova Column Filter (odabrane su samo dve kolone: prvo godine i klasičnu muziku, a onda godine i tehno, trans), Missing Value (isključeni su redovi koji sadrže nedostajuće vrednosti) i Normalizer (i godine i odgovori su skalirani na isti interval [0.0, 1.0]), a zatim, primenom algoritma K-Means uočeno je da podaci čine odredjene grupe.

Algoritam je primenjen nad 996 ispitanika. U nastavku su prikazani rezultati najpre za klasičnu muziku, a onda za tehno, trans (slike 10 i 11).



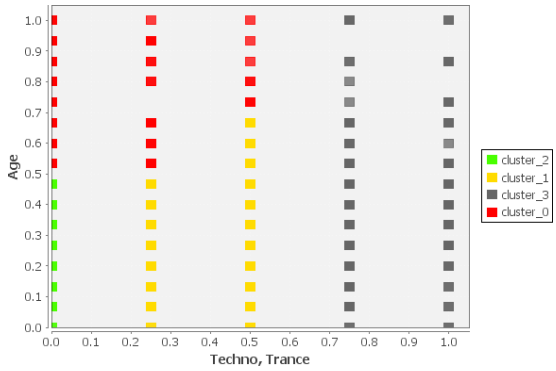
Slika 10

Od ukupnih 996 ispitanika, na osnovu rezultata algoritma K-Means, može se zaključiti da se ispitanici mogu podeliti u pet grupa (pet klastera) prema tome koliko vole da slušaju klasičnu muziku:

1. **Nulti klaster** (*crveno*) - 450 ispitanika, umereno slušaju, godine: od 15 do 22, najbrojniji
2. **Četvrti klaster** (*plavo*) - 236 ispitanika, vole da slušaju, godine: od 15 do 21, brojniji od zelenih
3. **Prvi klaster** (*žuto*) - 127 ispitanika, ne slušaju uopšte, godine: od 15 do 25
4. **Drugi klaster** (*zeleno*) - 96 ispitanika, vole da slušaju, godine: od 22 do 30
5. **Treći klaster** (*sivo*) - 87 ispitanika, umereno slušaju, godine: od 23 do 30

Zaključak 1: najviše ima mladih od 15 do 21 godine koji umereno slušaju klasičnu muziku (45% ispitanika)

Zaključak 2: medju onima koji vole da slušaju klasičnu muziku ima više mladih u uzrastu od 15 do 22 godine (23.7% ispitanika) nego u uzrastu od 22 do 30 godina (9.6% ispitanika)



Slika 11

Od ukupnih 996 ispitanika, na osnovu rezultata algoritma K-Means može se zaključiti da se ispitanici mogu podeliti u četiri grupe (četiri klastera) prema tome koliko vole da slušaju tehno, trans:

1. **Prvi klaster** (*žuto*) - 363 ispitanika, umereno slušaju, godine: od 15 do 23, najbrojniji
2. **Drugi klaster** (*zeleno*) - 303 ispitanika, ne slušaju uopšte, godine: od 15 do 22
3. **Treći klaster** (*sivo*) - 221 ispitanika, vole da slušaju, godine: od 15 do 30 (svih uzrasta)
4. **Nulti klaster** (*crveno*) - 109 ispitanika, ne vole do umereno slušaju, godine: od 23 do 30

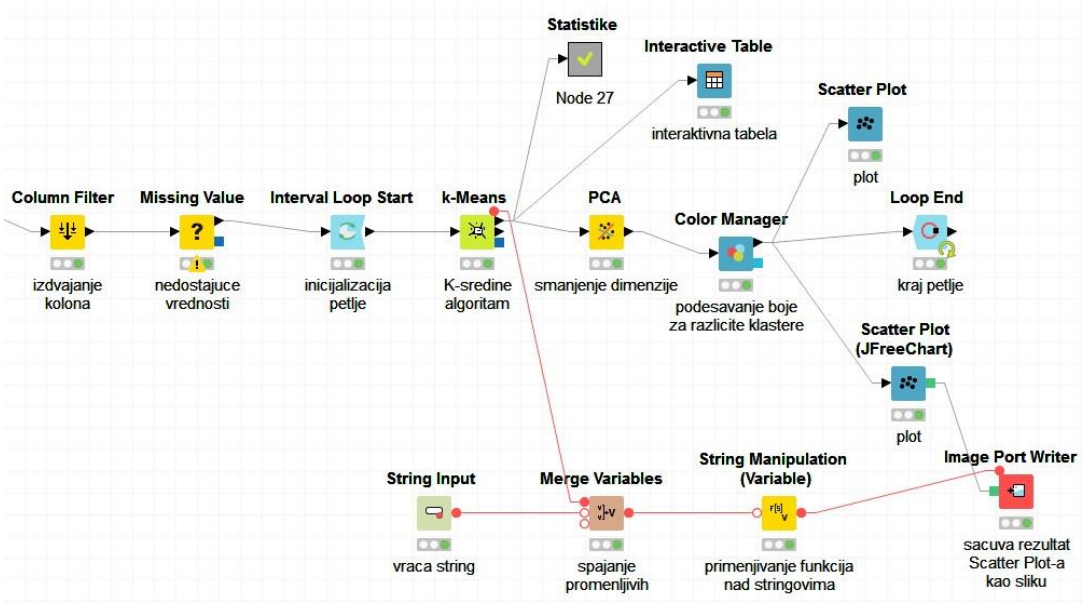
Zaključak: najviše ima onih (36.4%) koji umereno slušaju, uzrasta od 15 do 23 godine i onih (30.4%) koji ne slušaju uopšte uzrasta od 15 do 22 godine.

5.2. Na sta mladi troše novac

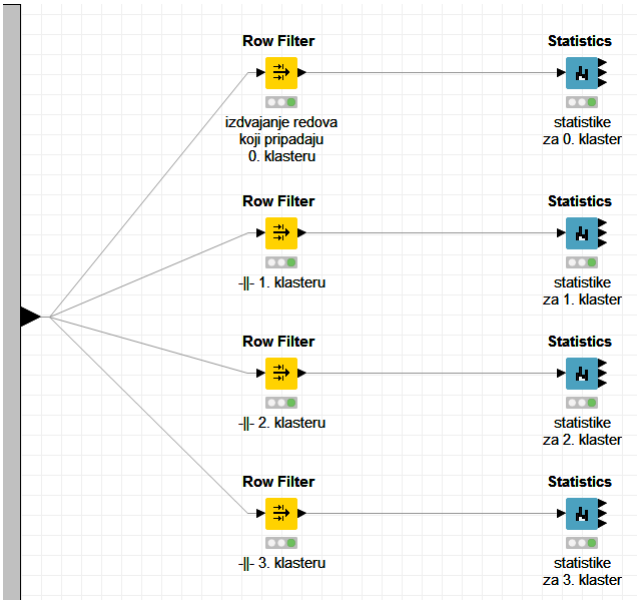
Zelimo da klasterujemo ispitanike po tome na sta najviše troše novac. Za ovo smo koristili samo atribute iz sekcije “potrošačke navike” (opisane u delu analize podataka). korišćenjem čvora Column Filter odabrali smo atribute: Finances, Shopping centres, Branded clothing, Entertainment spending, Spending on looks, Spending on gadgets, Spending on healthy eating, a preko čvora Missing Value isključili smo redove koji sadrže nedostajuce vrednosti. S obzirom da su vrednosti svih korišćenih atributa u istom intervalu, nije bilo potrebe za normalizacijom podataka. Konačno, primenom algoritma K-Means uvideli smo da podaci čine određenje grupe i da se najbolje klasifikuju za k=4.

Algoritam je primenjen nad 996 ispitanika.

U nastavku su prikazani najpre raspored čvorova a onda rezultati dobijeni primenom algoritma K-means za k=4 (slike 12 i 13).



Slika 12: Raspored čvorova



Slika 13: Metanode "Statistike"

Nakon klasifikovanja, izdvojili smo sve redove koji pripadaju nultom klasteru u jednu tabelu, sve redove koji pripadaju prvom klasteru u drugu tabelu, sve redove koji pripadaju drugom klasteru u treću tabelu i sve redove koji pripadaju trećem klasteru u četvrtu tabelu (prikazano na slici 13).

Preko čvora Statistics za svaku od ovih četiri tabela dobijamo sledeće rezultate:

Nultom klasteru pripadaju uglavnom one osobe (njih 340) koje u proseku najviše troše novac na kupovinu u trznim centrima, a najmanje troše na kupovinu uredjaja. Ovo je najbroniji klaster (slika 14).

Row ID	Column	Min	Max	Mean
Shopping c...	Shopping centres	1.25	5	3.625
Spending o...	Spending on healthy eating	0	5	2.989
Finances	Finances	0	5	2.919
Spending o...	Spending on looks	0	5	2.879
Branded clo...	Branded clothing	0	5	2.563
Entertainm...	Entertainment spending	0	5	2.349
Spending o...	Spending on gadgets	0	5	1.654

Slika 14

Prvom klasteru pripadaju uglavnom one osobe (njih 213) koje čuvaju novac (štede), a najmanje troše na markiranu odecu (slika 15).

Row ID	Column	Min	Max	Mean
Finances	Finances	0	5	3.216
Spending o...	Spending on healthy eating	0	5	2.84
Shopping c...	Shopping centres	0	5	1.379
Entertainm...	Entertainment spending	0	5	1.373
Spending o...	Spending on gadgets	0	5	1.25
Spending o...	Spending on looks	0	5	1.068
Branded clo...	Branded clothing	0	3.75	0.869

Slika 15

Drugom klasteru pripadaju uglavnom one osobe (njih 210) koje najviše troše na provod, a najmanje na kupovinu u tržnim centrima (slika 16).

Row ID	Column	Min	Max	Mean
Entertainm...	Entertainment spending	1.25	5	3.482
Spending o...	Spending on healthy eating	0	5	3.065
Spending o...	Spending on gadgets	0	5	2.982
Branded clo...	Branded clothing	0	5	2.673
Spending o...	Spending on looks	0	5	2.054
Finances	Finances	0	5	1.905
Shopping c...	Shopping centres	0	3.75	1.381

Slika 16

Trećem klasteru pripadaju uglavnom one osobe (njih 232) koje mnogo novca ulažu u svoj izgled, i to su uglavnom osobe koje ne čuvaju novac (slika 17).

Row ID	Column	Min	Max	Mean
Spending o...	Spending on looks	1.25	5	4.219
Shopping c...	Shopping centres	0	5	4.176
Branded clo...	Branded clothing	0	5	4.019
Entertainm...	Entertainment spending	0	5	3.933
Spending o...	Spending on healthy eating	0	5	3.917
Spending o...	Spending on gadgets	0	5	3.777
Finances	Finances	0	5	1.891

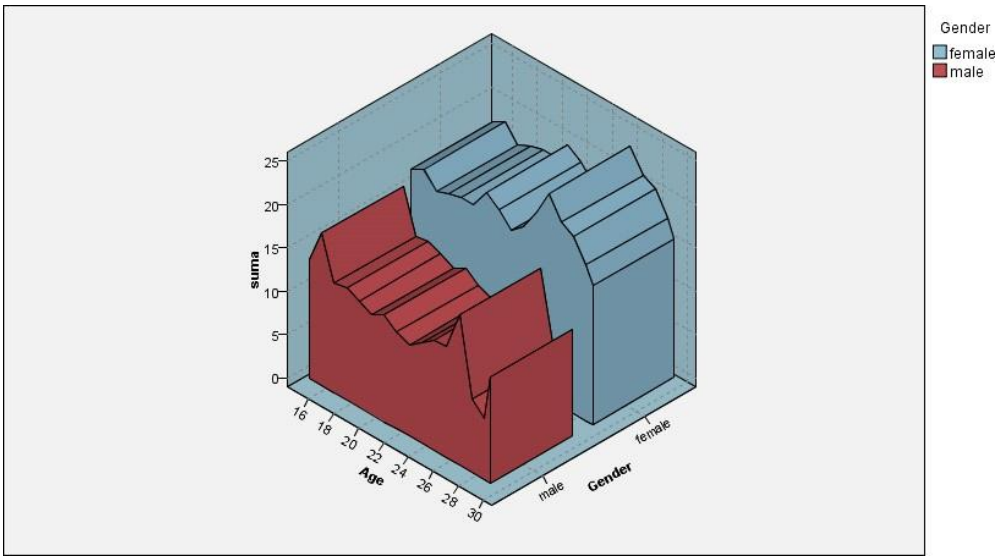
Slika 17

5.3. Podela mladih u zavisnosti koliko se plaše

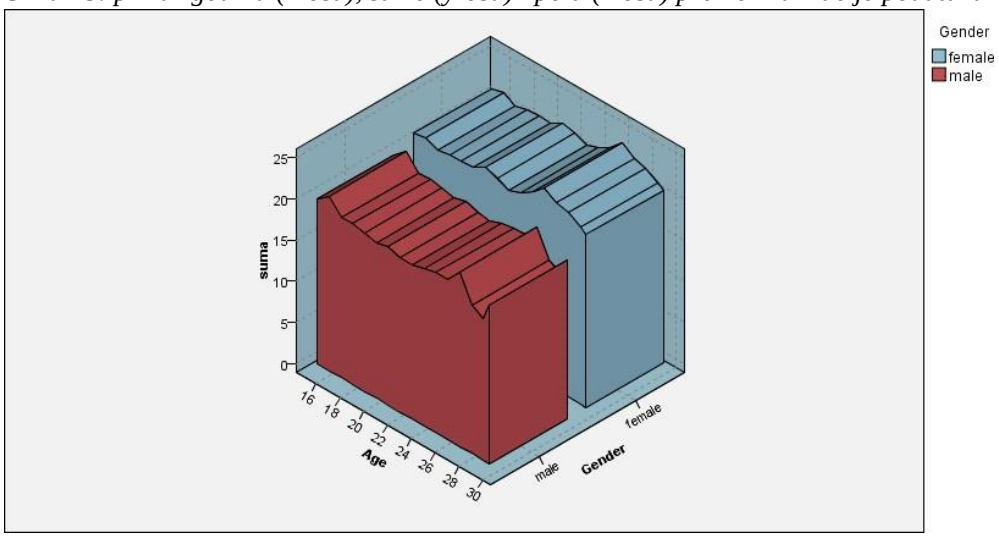
Za ovo klasterovanje korišćen je alat SPSS (ostalo je radjeno u KNIME). Nad podacima su sprovedeni algoritmi *"Kohonen"* i *"K-sredine"*.

Pre klasterovanja iz fajla "*responses.csv"* su izdvojene (korišćenjem čvora "Column Filter") sledeće kolone: *Letenje, Grmljavina, Mrak, Visina, Pauci, Zmije, Miševi, Starenje, Opasni psi, Javni nastup, Pol, Godine.*

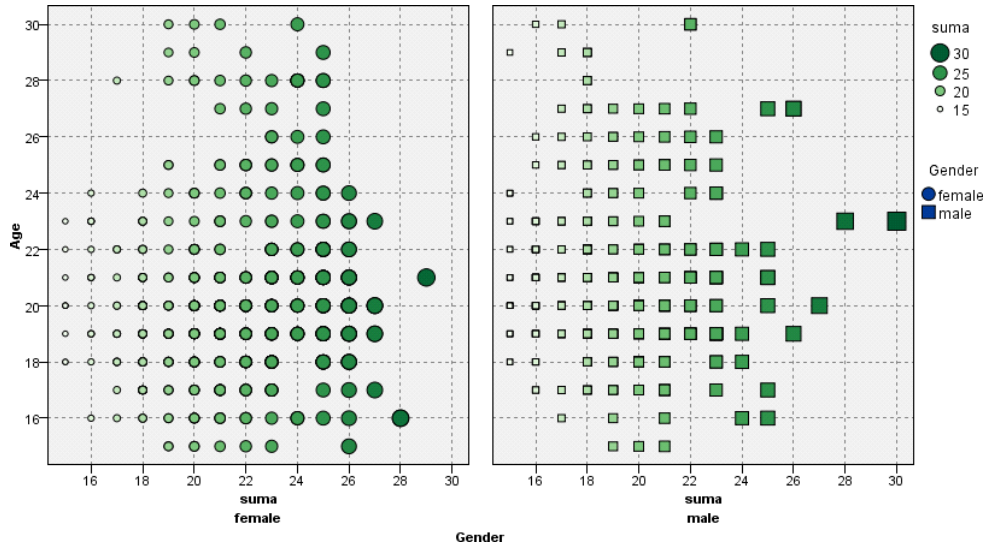
Korišćenjem čvora “Math Formula” izračunat je zbir svih odgovora (ocena) koje je osoba dala na pitanja koliko se plaši odredjenih stvari/ pojava i rezultati su prikazani u dodatnoj koloni “*suma*” koja je nadovezana na postojecu tabelu. Zatim, iz novonastale tabele su izdvojene kolone: *Godine*, *Pol* i *Suma*. Onda su normalizovane vrednosti za atribute *Godine* i *Suma* i primenjeni su gore pomenuti algoritmi za klasifikaciju.



Slika 18: prikaz godina (x-osa), sume (y-osa) i pola (z-osa) pre normalizacije podataka



*Slika 19: prikaz godina (x-osa), sume (y-osa) i pola (z-osa) posle normalizacije podataka_*



Slika 20

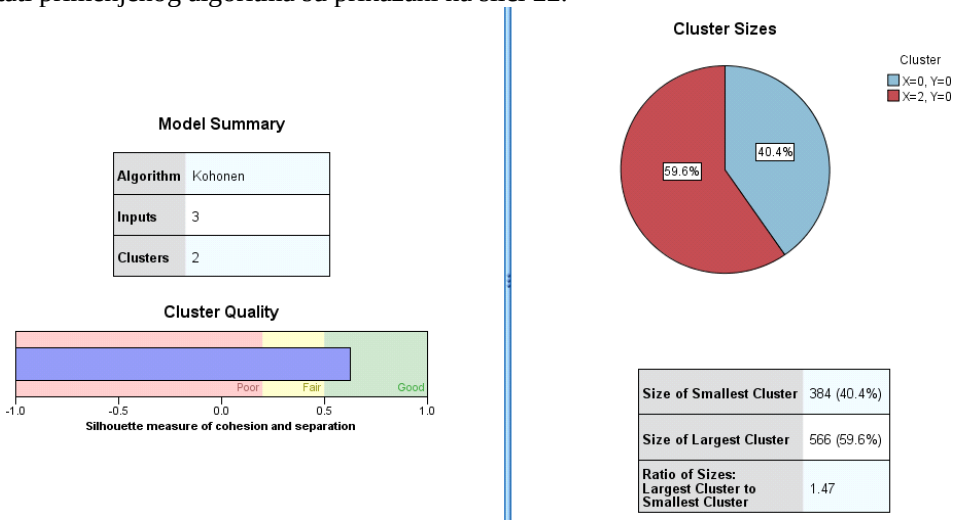
Primenom čvora *Audit Table* uočava se (na slici 21) da ispitanici u proseku imaju 20 i po godina i da se "osrednje" plaše (jer je minimum za atribut *suma* 0, a maksimum 39, a prosek je 15.568).

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Age		Continuous	15	30	20.417	2.819	1.114	--	950
Gender		Flag	--	--	--	--	--	2	950
suma		Continuous	0	39	15.568	7.531	0.184	--	950

Slika 21

5.3.1. Kohonen

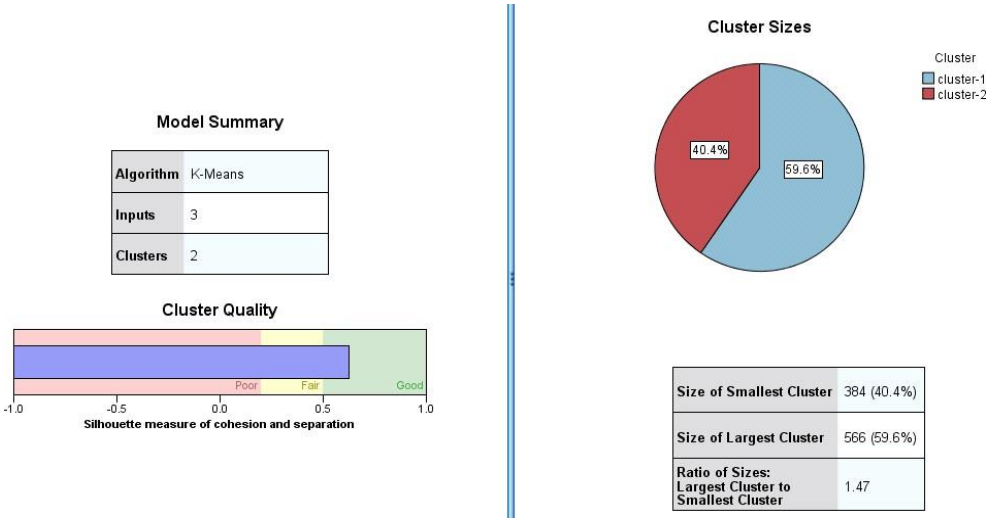
Rezultati primenjenog algoritma su prikazani na slici 22.



Slika 22

5.3.2. K-sredine

Rezultati primenjenog algoritma su prikazani na slici 23.



Slika 23

5.4 Klasterovanje na osnovu strahova

Predmet ispitivanja u ovom odeljku jesu strahovi. Kako u okviru tabele imamo 10 strahova, podelili smo ih na racionalne i iracionalne. To su ujedno postale i dve nove kolone tabele koje su popunjene jednostavnim sabiranjem vrednosti iz kolona odgovarajućih strahova koji im pripadaju. Na kraju je oduzeto 5, čisto da bi se grafik sveo na koordinatni početak, a i zato što ljudi koji su odgovorili sa 1, nisu plašljivi, pa nije u redu da to dodajemo na skali za globalni srah. Naravno podaci su normalizovani radi jednakog uticaja.

Prvo obrađeno pitanje je da li plašljivi ljudi izbegavaju neki određeni žanr filmova. Ispostavlja se da ljudi koji imaju izrazito veći strah od oba, iracionalnog i racionalnog, ne uživaju u gledanju ratnih filmova. To se najbolje vidi pri korišćenju 3 klastera.

[slika onog pogleda čvora kmean] [slika grafika klastera]

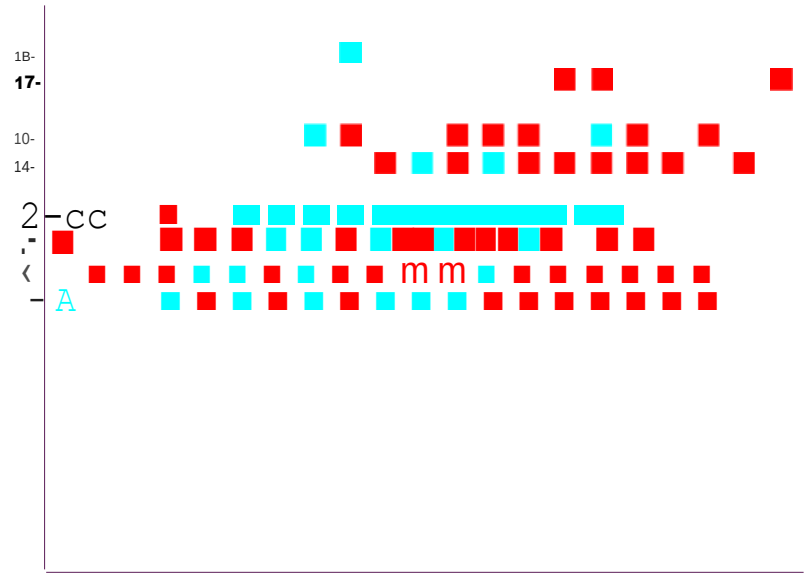
Naredno pitanje koje nas je zanimalo je da li ljudi sa nadprosečnim strahom od oba, iracionalnog i racionalnog, mogu da uživaju u horor filmovima. Očekivano, ljudi velikih strahova imaju prosečnuo najmanju ocenu za uživanje u horor filmovima, međutim, ono što je neočekivano kada se broj klastera poveća je da zapravo grupa ljudi koja ima najveću prosečnu ocenu po pitanju strahova, na drugom su mestu po visini ocene užitka u horor filmovima. Desno je umesto grafika klastera predstavljena raspodela na osnovu uživanja u horor filmovima.

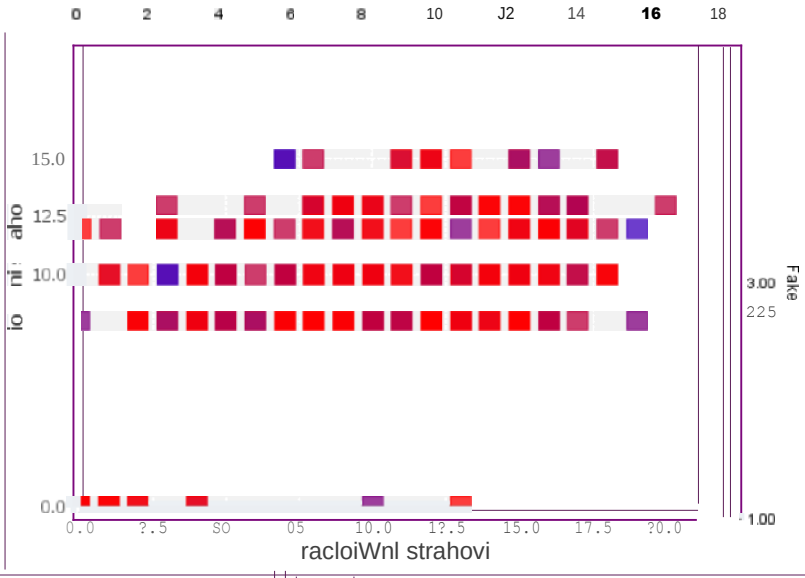
[slika horora prema kmean] [slika horora po strahu]

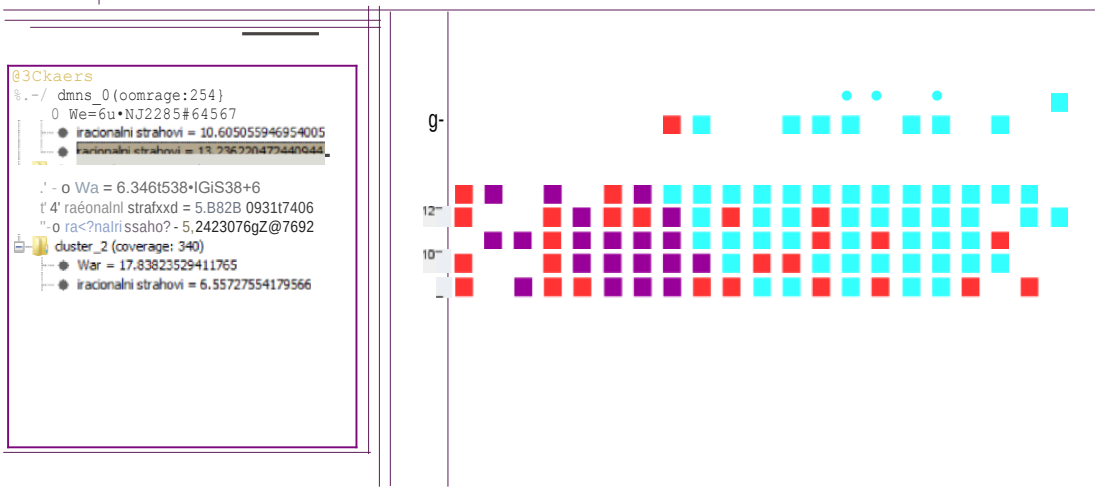
Klasterovanje iracionalnog, racionalnog straha, trošenja novca na vizuelno lepe stvari, osetljivosti (ako tako nazovemo iskaz "plačem kada stvari ne idu kako treba") dovelo nas je do zaključka da su ove osobine skoro proporcionalne. Tj. kada se jedna visoko izražena u grupi, verovatno su i ostale, i obrnuto. Kad su ekstremumi u pitanju, ovo je znatno primetno što se vidi prema klasterima 3 i 1. Između ta dva ekstremna slučaja (koja zapravo cine više od trećine ukupnog uzorka) javlja se mala disproporcionalnost kada je u pitanju atribut koji smo malo pre nazvali "osetljivost".

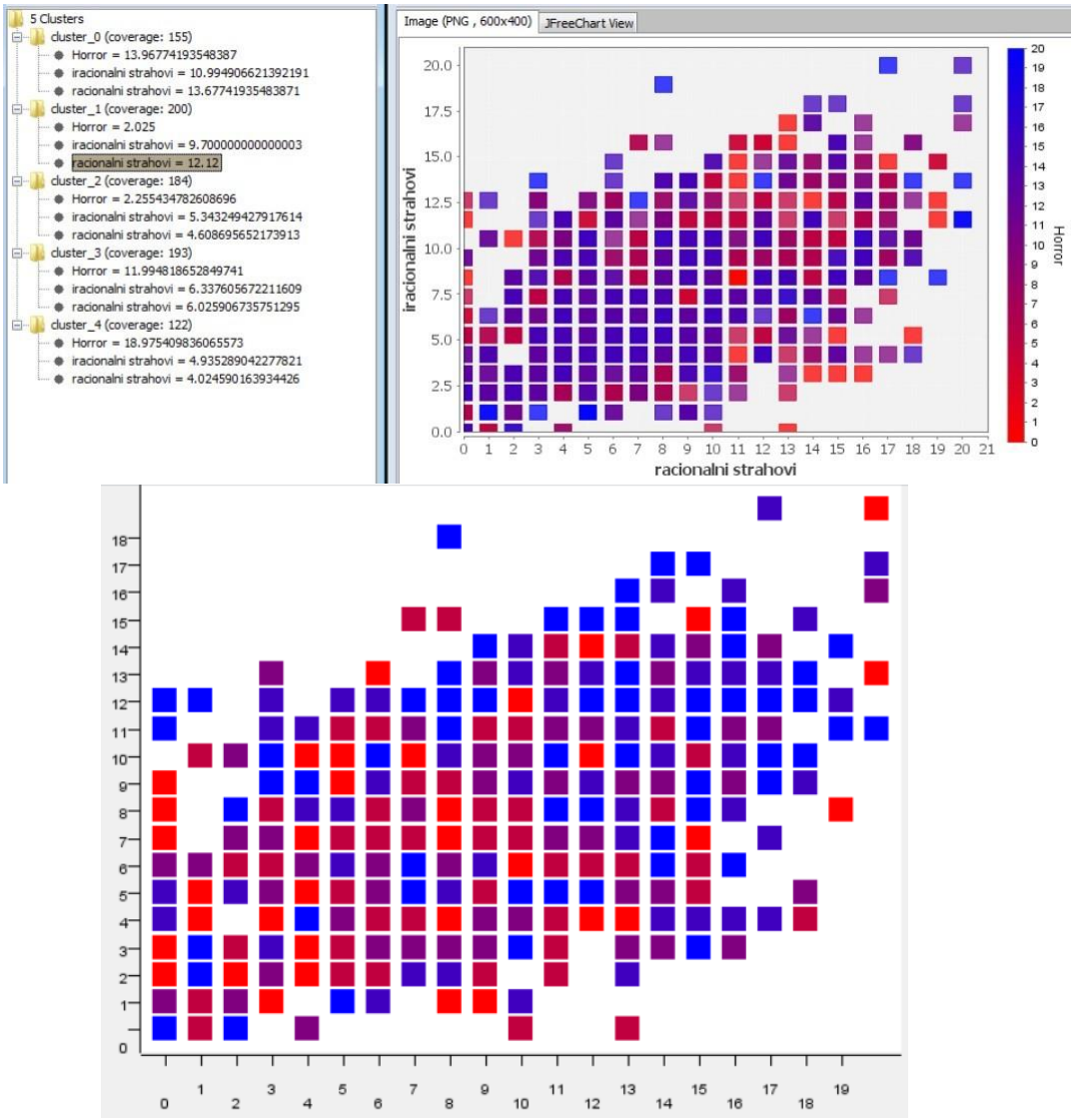
[slika kmean rez] [slika life stragl na xy osi] [slika spending money on look na xy osi]

Slede slike grafika neklasterovanih podataka koje same po sebi govore da grupa postoji, te klasterovanje nije rađeno posebno.


1B-
17-
10-
14-
2
cc
mm
A


0
2
4
6
8
10
J2
14
16
18
15.0
12.5
10.0
0.0
aho
ni
io
3.00
225
1.00
Fake
racloiWnl strahovi


@3Ckaers
dmns_0(oomrage:254}
0 We=6u•NJ2285#64567
iracionalni strahovi = 10.605055946954005
.' - o Wa = 6.346t538•IGiS38+6
t' 4' raéonalnl strafxxd = 5.B82B 0931t7406
''-o ra<?nalri ssaho? - 5,2423076gZ@7692
cluster_2 (coverage: 340)
War = 17.83823529411765
iracionalni strahovi = 6.55727554179566
g-

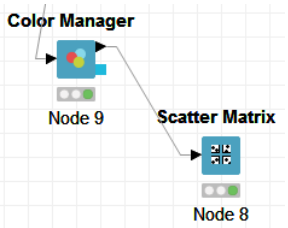


6 Klasifikacija

6.1. Klasifikacija korišćenjem stabla oducivanja

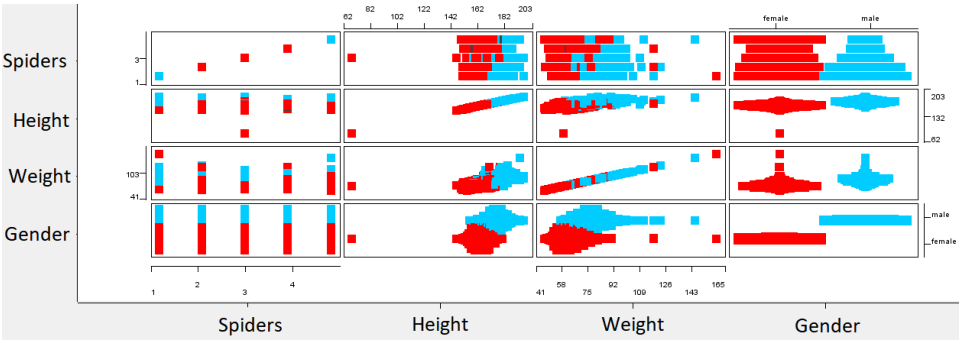
Stablo odlucivanja je hijerarhijska struktura koja se sastoji iz čvorova i direktnih grana. Čvorovi predstavljaju pitanja, a grane odgovore.

Da bi se uočilo koliko su odabrani atributi koji opisuju osobu povezani sa atributom koji predstavlja klasu, najpre je muški pol obojen plavo, a ženski crveno koristeći čvor Color Manager, a onda preko čvora Scatter Matrix vizuelno je predstavljena zavisnost vrednosti ovih atributa (slika 24).



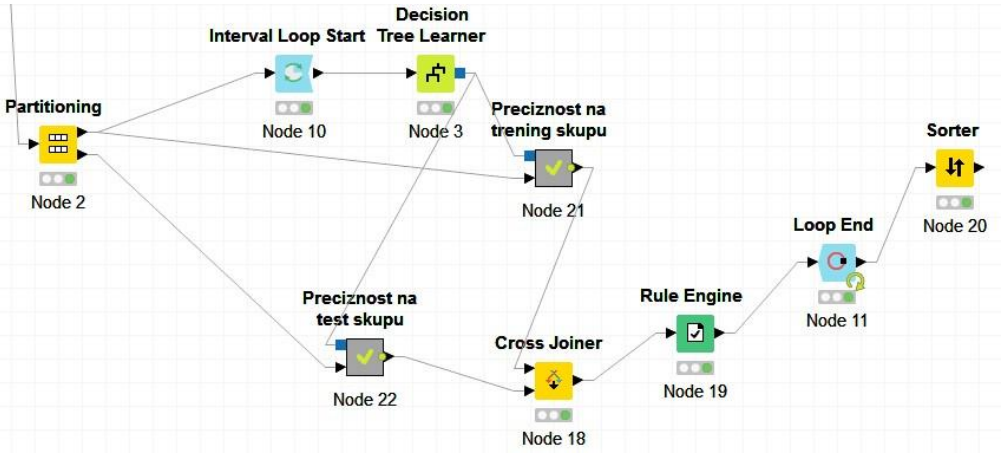
Slika 24: Čvorovi "Color Manager" i "Scatter Matrix"

Uočava se da postoji korelacija izmedju vrednosti odabranih atributa i pola. Primećuje se da je jasnije videti da li je osoba ženskog ili muškog pola na osnovu težine i visine, dok je atribut "Da li se osoba plaši pauka" od manjeg značaja za ovu klasifikaciju (slika 25).



Slika 25: "Scatter" matrica

Dobar model mora korektno da klasifikuje i trening podatke i unapred nepoznate podatke (test podatke). Model koji vrši ukalupljivanje trening podataka previše dobro može imati mnogo veću gresku pri uopstavanju nego model sa većom greskom pri treniranju - preprilagodavanje. Da do ovoga ne bi doslo, ulazna tabela je razdvojena na dve particije: za trening i test. Dobijene dve particije se nalaze u dve izlazne tabele čvora Partitioning (slika 26).



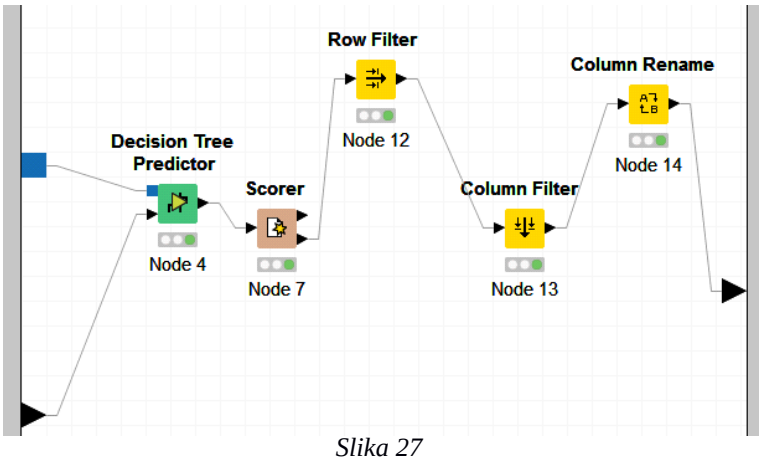
Slika 26

U konfiguraciji čvora Partitioning podešeno je da 70% podataka ulazne tabele pripadnu skupu za trening, a preostalih 30% podataka skupu za test. Izabrana je opcija "Stratified sampling" kako bi raspodela vrednosti kolone Gender (selektovane kolone) bila približno zadržana u izlaznim tabelama. Stratifikacija podrazumeva da se prilikom deljenja podataka obezbedi da delovi imaju istu raspodelu kao i ceo skup podataka.

Dalje, korišćeni su čvorovi Decision Tree Learner i Decision Tree Predictor (slika 27). U konfiguraciji prvog čvora kao target atribut odabran je atribut Gender. S obzirom na to da je ovaj atribut (atribut Gender) nominalni atribut, on može biti izabran. Algoritam obezbedjuje dve mere kvaliteta za računanje podele: "gini index" i "gain ratio". Prvo je odabran "gini index", s toga u nastavku sledi prikaz rezultata dobijenih korišćenjem ove mere nečistoce, a potom rezultata dobijenih korišćenjem mere "gain ratio" (odnos dobiti).

Ginijev indeks predstavlja meru nečistoce koja se koristi za izračunavanje dobiti nekog čvora. Izračunava se formulom:

$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$



Slika 27

Izlaz iz čvora Decision Tree Predictor je prosleđen čvoru Scorer. Ovaj čvor pruža matricu konfuzije (prvi izlaz čvora) i statistiku preciznosti (drugi izlaz čvora).

Sada se može izračunati i preciznost klasifikacije na skupu podataka za trening. Nakon izdvajanja samo poznate vrednosti kolone "Accuracy" i preimenovanja iste dobija se podatak da je preciznost klasifikacije na trening skupu 0.9 (slika 28).

Row ID	D AccuracyTrening
Overall	0.9

Slika 28

Iz Matrice konfuzije (slika 29) može se videti da je većina podataka dobro klasifikovana, ali i da su neki muškarci pogresno klasifikovani kao žene (45), i žene pogresno klasifikovane kao muškarci (23). Kada se izlazna tabela čvora Decision Tree Predictor sortira rastuće po atributu Height, može se uočiti da klasifikacija najviše zavisi od visine, zbog čega su niži muškarci pogrešno klasifikovani kao žene.

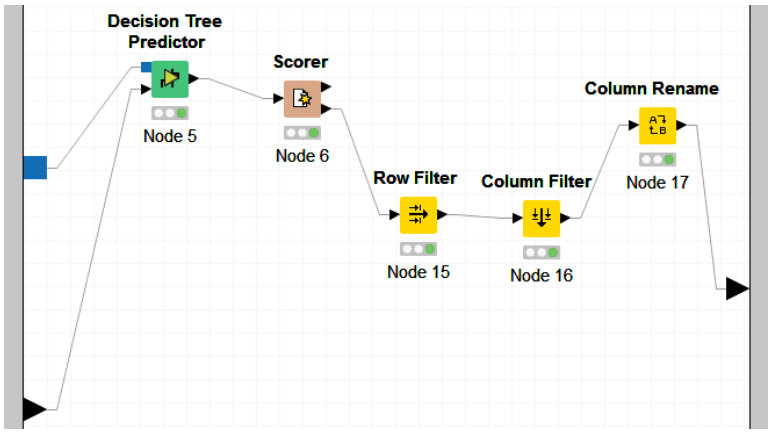
Row ID	female	male
female	376	23
male	45	235

Slika 29

Sve je isto urađeno i za test skup podataka (slika 31). Preciznost na ovom skupu je nešto veća i iznosi 0.904 (slika 30).

Row ID	D AccuracyTest
Overall	0.904

Slika 30



Slika 31

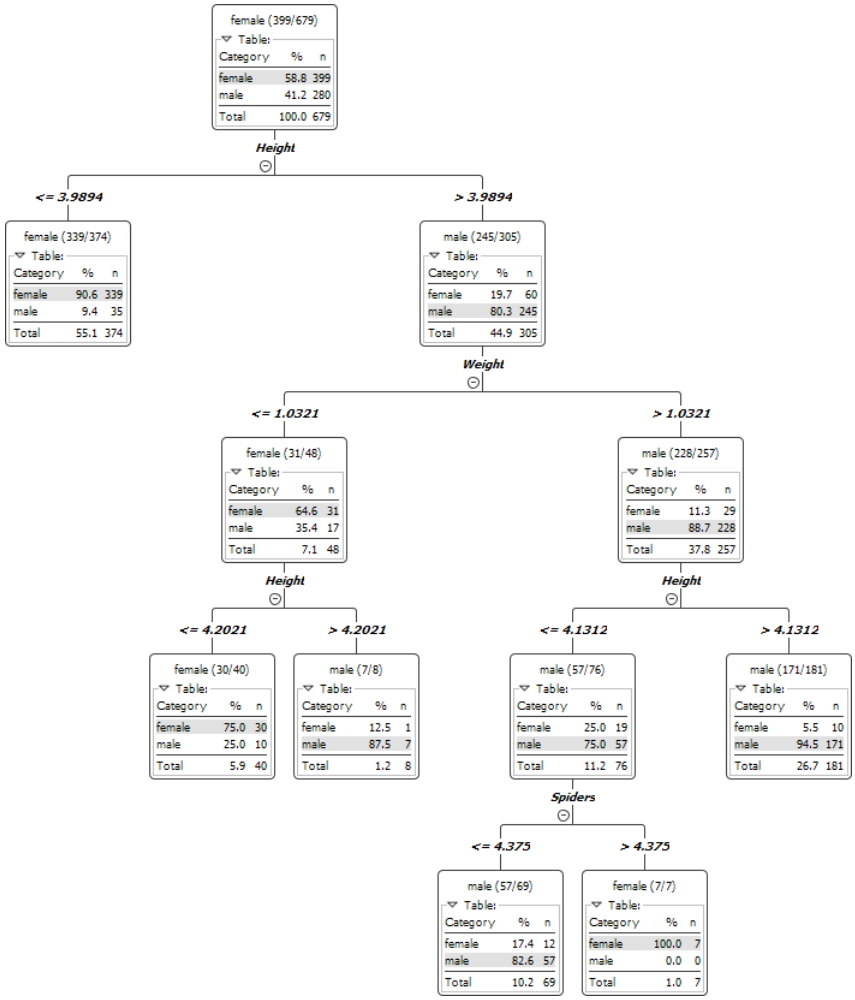
Zatim su ove preciznosti (za trening i test skup) spojene korišćenjem čvora Cross Joiner, kako bi se podaci bolje analizirali. Uz pomoć čvora Rule Engine, prethodnoj tabeli sa preciznostima, dadata je još jedna kolona u kojoj je naveden broj iteracije (slika 32).

Row ID	AccuracyTrening	AccuracyTest	minRec	Iteration
Overall_Ov...	0.9	0.904	1	0
Overall_Ov...	0.9	0.904	3	1
Overall_Ov...	0.9	0.904	5	2
Overall_Ov...	0.9	0.904	7	3
Overall_Ov...	0.881	0.893	19	9
Overall_Ov...	0.887	0.89	9	4
Overall_Ov...	0.887	0.89	11	5
Overall_Ov...	0.887	0.89	13	6
Overall_Ov...	0.887	0.89	15	7
Overall_Ov...	0.887	0.89	17	8

Slika 32

Do sada opisani proces ponavljan je u petlji kako bi se dobila sto veća preciznost. U petlji je menjan minimalan broj osoba po čvoru (stabla odlucivanja), kako se stablo ne bi dalje granalo (menjan je kriterijum zaustavljanja od 1 do 20 sa korakom 2). Optimalna granica za broj čvorova je 1, 3, 5 ili 7, jer je tada preciznost najveća.

Stablo odlucivanja koje je dobijeno koristeći ginijev indeks prikazano je na slici 33. Može se uočiti da se podela najpre vrši po atributu Height, pa po atributu Weight, zatim ponovo po atributu Height i na kraju po atributu Spiders. Ovo je otprilike i očekivano, jer je još na početku, na grafičkom prikazu (Scatter matrix) uočeno da pol najviše zavisi od visine. Tako da je najbolje prvo postaviti pitanje “Koliko je osoba visoka?”, pa ako je visina manja ili jednaka 3.9894 (sto je 174.5 cm), onda je opravdano tvrditi (u 90.6% slučajeva) da je osoba ženskog pola. Inače, postavlja se još pitanja.



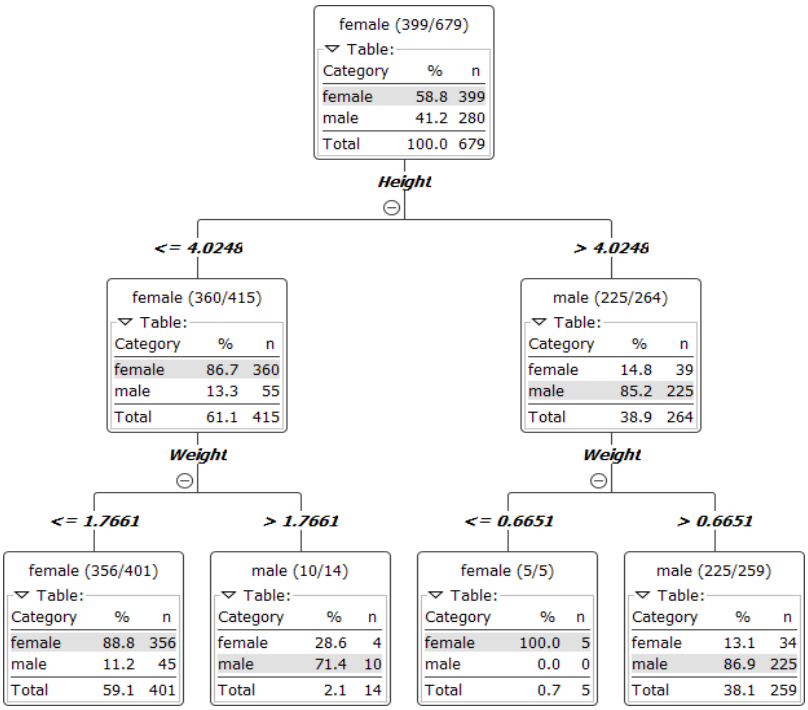
Slika 33: Drvo odlučivanja

Odnos dobiti je mera koja se koristi za određivanje valjanosti podele stabla odlučivanja. Definisana je formulom:

$$GainRatio_{split} = \frac{Gain_{split}}{SplitInfo}$$

$$SplitInfo = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

Stablo odlučivanja je jako slično slablu odlučivanja koje se generisalo kada je za meru odabran Ginijev indeks (slika 34), pa nije detaljno analizirano i ovo stablo.



Slika 34

Row ID	AccuracyTrening	AccuracyTest	minRec	Iteration
Overall_Ov...	0.89	0.893	11	5
Overall_Ov...	0.879	0.89	13	6
Overall_Ov...	0.891	0.887	15	7
Overall_Ov...	0.879	0.887	17	8
Overall_Ov...	0.888	0.883	19	9
Overall_Ov...	0.879	0.883	7	3
Overall_Ov...	0.879	0.883	9	4
Overall_Ov...	0.878	0.883	1	0
Overall_Ov...	0.878	0.883	3	1
Overall_Ov...	0.878	0.883	5	2

Slika 35: Preciznost na trening i test skupu

Ovde je, isto kao kod Ginijevog indeksa, menjan kriterijum zaustavljanja od 1 do 20 sa korakom 2. U ovom slučaju, optimalna granica za broj čvorova je 11, ali se dobro ponaša i sa 15.

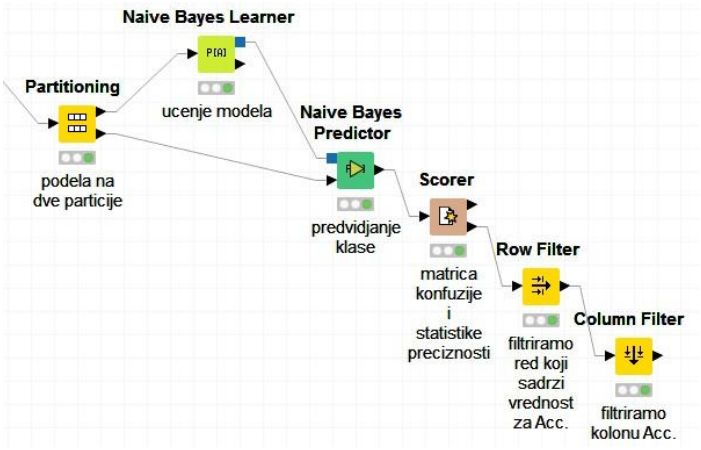
Može se zaključiti da Ginijev indeks ipak daje nešto bolje rezultate. Njegova preciznost je veća za 1.1%.

6.2. Klasifikacija naivnim Bajesovim algoritmom

Ovo je pristup modeliranju probabilistickih veza izmedu skupa atributa i klasa. Naime, primenjuju se verovatnosne i statističke metode i teoreme kao što su Bajesova teorema i teorema o uslovnoj verovatnoći.

Bajesova teorema je definisana ovako:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



Slika 36: Raspored čvorova

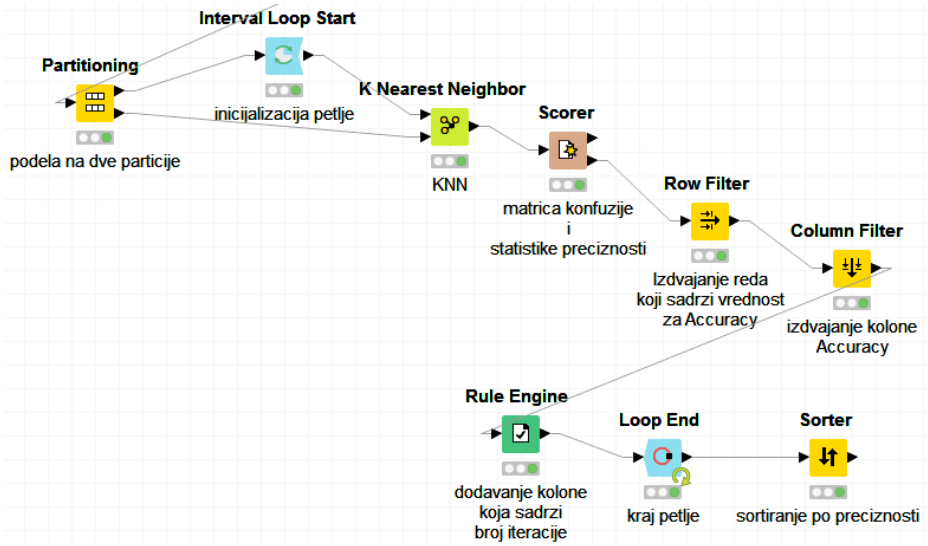
Particionisanjem skupa na test i trening skup, a zatim primenom čvorova Naive Bayes Learner za kreiranje modela odnosno Naive Bayes Predictor za klasifikovanje podataka, dobijeno je da je preciznost predvidjanja klase (iz čvora Scorer) 0.883 (slika 37).

Row ID	D Accuracy
Overall	0.883

Slika 37

6.3. Klasifikacija metodom K najbližih suseda

Osnovna ideja algoritma “K najbližih suseda” je da se na osnovu k najbližih suseda date instance odredi klasa te instance. Najpre se odrede susedi, zatim susedi glasaju, pa se glasovi prebroje. Ona klasa koja je dobila najviše glasova predstavlja konačan odgovor na pitanje kojoj klasi instanca pripada. Zato je važno koja će vrednost biti odabrana za parametar k. Ovaj algoritam ne radi sa kategoričkim vrednostima, ali ovde svakako nema kategoričkih atirbuta.



Slika 38: Raspored čvorova

Pomoću čvora Partitioning particionisani su podaci na test i trening skupu na isti način kao kod konstruisanja stabla odlučivanja. Konfiguracija čvora Interval Loop Start podešena je tako da broj suseda u petlji ide od 2 do 100 sa korakom 1. Za primenu algoritma korišćen je čvor K Nearest Neighbor. Zatim je računata preciznost klasifikacije pomoću čvora Scorer koji, kao što je već pomenuto, daje matricu konfuzije i statistike preciznosti.

Po zavrsetku petlje, pomoću čvora Sorter, sortirane su preciznosti koje su dobijane kroz iteracije i zaključeno je da je optimalan broj suseda za ove podatke k=4 (postiže se u 2. iteraciji). Nešto lošija preciznost se dobija za $k \in \{8,9,10,11\}$.

Row ID	Accuracy	k	Iteration
Overall#2	0.9	4	2
Overall#6	0.897	8	6
Overall#7	0.897	9	7
Overall#8	0.897	10	8
Overall#9	0.897	11	9
Overall#4	0.893	6	4
Overall#11	0.893	13	11
Overall#16	0.893	18	16
Overall#18	0.893	20	18
Overall#21	0.893	23	21
Overall#3	0.89	5	3
Overall#5	0.89	7	5

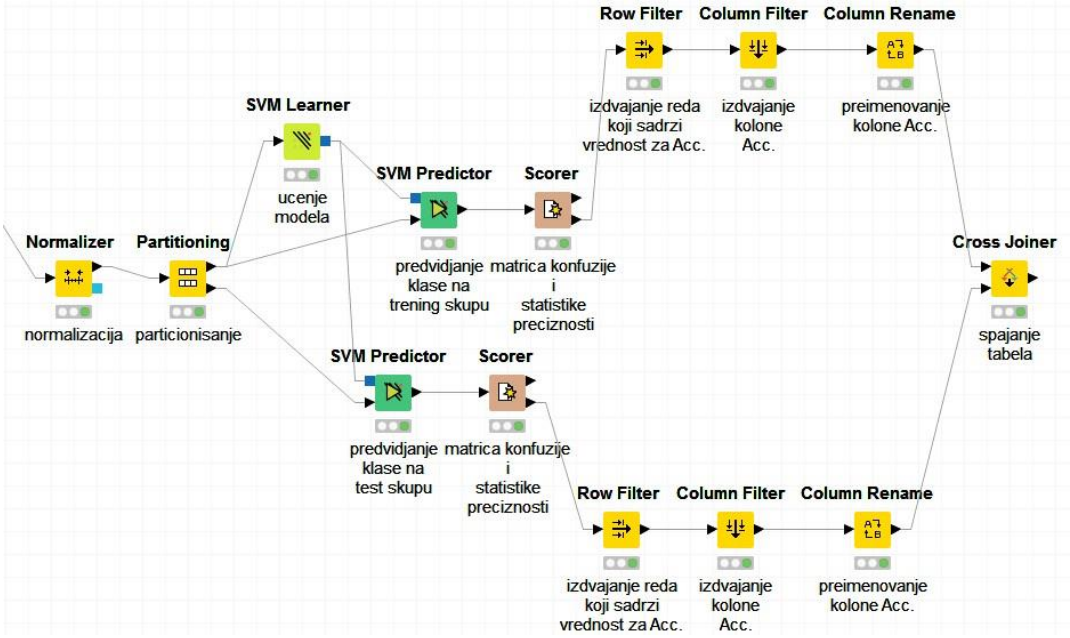
Slika 39

Maksimalna preciznost korišćenjem algoritma KNN je 0.9 (slika 39), što se veoma malo razlikuje od prethodno izračunate preciznosti dobijene korišćenjem stabla odlučivanja koja iznosi 0.904 (za test skup).

6.4. Klasifikacija metodom potpornih vektora (SVM)

Osnovni SVM algoritam je namenjen za binarnu klasifikaciju (dve klase: 1 i -1). Kada postoje više od dve klase, postoje dva pristupa: jedan protiv svih i jedan protiv jednog. Pošto ovaj konkretan problem ima samo dve klase: Male i Female, nema potrebe za nekim od ova dva pristupa.

Ideja ovog metoda je nalaženje razdvajajuće hiperravni tako da su podaci koji pripadaju istoj klasi sa jedne strane ravni, a podaci koji pripadaju drugoj klasi sa suprotne strane ravni. Ovo je lako izvesti kada su podaci linearno razdvojivi, ali kako u ovom skupu podaci nisu linearno razdvojivi, to se postiže uvođenjem Kernel funkcije koja odgovara skalarnom proizvodu u nekom višedimenzionalnom prostoru. Odlučeno je korišćenje bas ovog algoritma jer se u teoriji dobro ponaša za numericke podatke, pa je pretpostavka da će i ovde dati preciznije rezultate.



Slika 40: Raspored čvorova

Pre particionisanja skupa, podaci su normalizovani korišćenjem čvora Normalizer, jer je jedan od uslova za primenu ovog metoda klasifikacije upravo taj da su podaci normalizovani. Zatim su primenjeni čvorovi Partitioning, SVM Learner i SVM Predictor, i dobijeni su klasifikovani podaci za trening i test skup. Da bi se dobili podaci koji su najbolje moguće klasifikovani, korišćeni su različiti kerneli i upoređivani su dobijeni rezultati.

6.4.1. Polinomijalni kernel (Polynomial)

Polinomijalni kernel je definisan formulom:

$$K(x,y) = (x^T y + c)^d$$

Primenom ovog kernela preciznost kojom se trening i test skup klasifikuju je odlična, čak 0.909 (slika 41). Za polinomijalni kernel drugog stepena dobija se preciznost 0.915 (slika 42). Ista preciznost na trening skupu se dobija i za polinomijalni kernel trećeg i šestog stepena.

Row ID	D trening_accuracy	D test_accuracy
Overall_Ov...	0.909	0.88

Slika 41: Prvi stepen

Row ID	D trening_accuracy	D test_accuracy
Overall_Ov...	0.915	0.873

Slika 42: Drugi stepen

6.4.2. Sigmid kernel (Hyper Tangent)

Sigmoid kernel je definisan formulom:

$$K(x, y) = \tanh(\gamma x^T y + r)$$

Primenom ovog kernela preciznost kojom se trening i test skup klasifikuju je takođe jako dobra. Najveću preciznost dobijamo za kappa = 0.9 i delta = 0.1 (slika 43).

Row ID	D trening_accuracy	D test_accuracy
Overall_Ov...	0.829	0.79

Slika 43

6.4.3. Gausov (RBF) kernel

Gausov kernel je definisan formulom:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

Primenom ovog kernela preciznost kojom se trening i test skup klasifikuju je velika. Što je vrednost za sigma manja, dobija se veća preciznost. Na slici 44 su rezultati dobijeni za sigma = 0.1.

Row ID	D trening_accuracy	D test_accuracy
Overall_Ov...	0.922	0.873

Slika 44

Sigma = 0.1

Naravno, bilo je moguće i posmatrati matricu konfuzije (na isti način kao što je to urađeno kod drveta odlučivanja) nakon primene algoritama KNN i SVM, ali zbog velike sličnosti u preciznosti to nije neophodno, jer bi se svuda dobilo slično.

Na osnovu podataka do sada, može se zaključiti da razlike u preciznosti izmedju primenjenih algoritama klasifikacije nisu velike. Ipak, ističe se algoritam KNN koji ima najveću preciznost, čak 0.9.

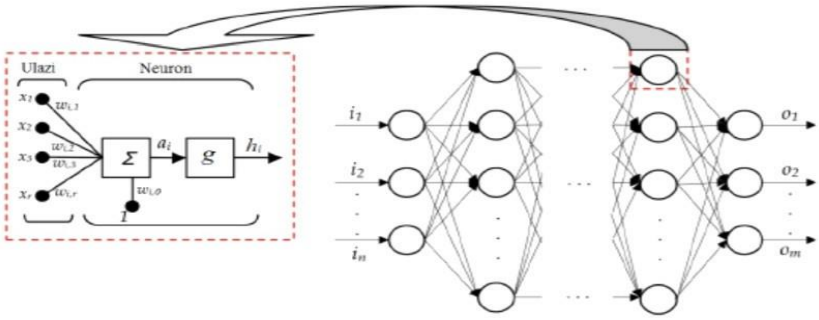
6.5. Klasifikacija algoritmom RProp

Rprop (skraceno od *"resilient backpropagation"*) je algoritam (tačnije, heuristika) nadgledanog učenjenja korišćenjem neuronskih mreža.

Neuronska mreža se sastoji od jedinica (ili neurona), koje predstavljaju jednostavne parametrizovane funkcije. Svaka jedinica računa linearnu kombinaciju svojih argumenata i nad njom računa neku nelinearnu transformaciju (u nastavku *aktivacionu funkciju*). Ove jedinice su organizovane u slojeve, tako da jedinice jednog sloja primaju kao svoje argumente (ulaze) vrednosti svih jedinica prethodnog sloja i sve jedinice prosleđuju svoje izlaze samo jedinicama narednog sloja. Svi slojevi čije jedinice prosleđuju svoje izlaze drugim jedinicama se nazivaju skrivenim slojevima. Ulazi jedinica prvog sloja se nazivaju ulazima mreže. Slično, izlazi jedinica poslednjeg sloja se nazivaju izlazima mreže. Formalno, model se definiše na sledeći način:

$$\begin{aligned} h_0 &= x \\ a_i &= W_i h_{i-1} + \omega_{i0} \qquad i = 1, 2, \ldots, L \\ h_i &= g(a_i) \qquad i = 1, 2, \ldots, L \end{aligned}$$

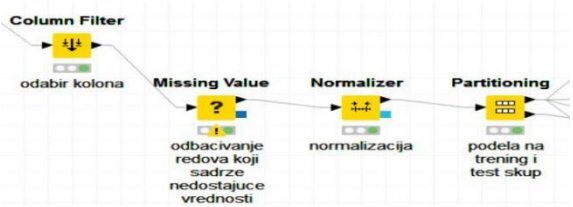
gde je x vektor ulaznih promenljivih, L je broj slojeva, Wi je matrica čija j-ta vrsta predstavlja vektor vrednosti parametara jedinice j u sloju i, wi0 predstavlja vektor slobodnih članova linearnih kombinacija koje jedinice i-tog sloja izračunavaju, a g je nelinearna aktivaciona funkcija (slika 45).



Slika 45: Struktura neuronske mreže sa propagacijom unapred

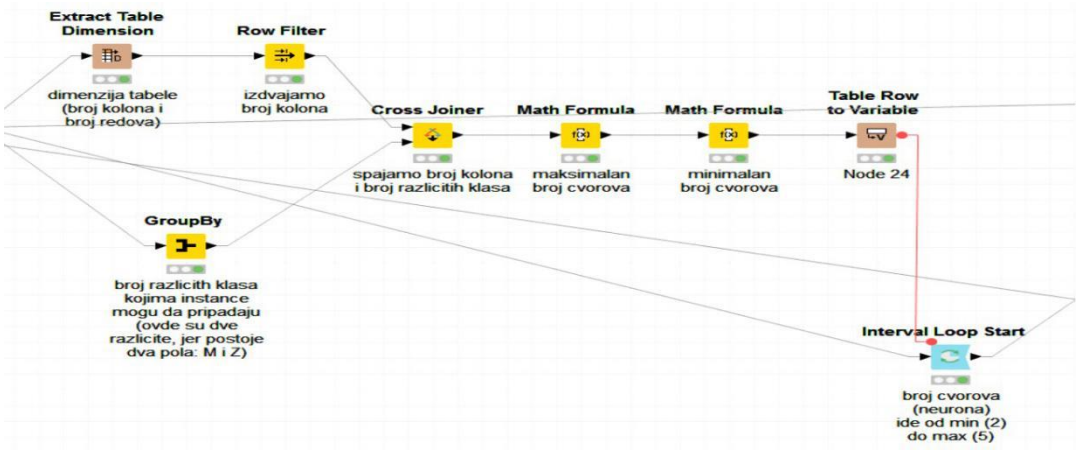
Ukratko, ovde je zadatak da se odredi broj skrivenih slojeva i broj čvorova na svakom sloju. Da bi se dobio što bolji model, vrednosti ovih parametra su inkrementirani u petlji (dve petlje: jedna reguliše broj čvorova tj. neurona, a druga broj skrivenih slojeva).

Ceo proces je izgledao tako što su prvo filtrirane kolone (iste koje su korišćene za ostale algoritme klasifikacije: “Height”, “Weight”, “Spiders”, “Gender”), zatim su odbačeni redovi koji su sadržali nedostajuće vrednosti, a onda su svi atributi (osim atributa “Gender” koji nije numerički pa se ne može normalizovati) normalizovani (skalirali na interval [1.0, 5.0]) jer RProp algoritam to zahteva. Na kraju je podeljen početni skup podataka na skup za trening i skup za test (slika 46 prikazuje raspored ovih čvorova).



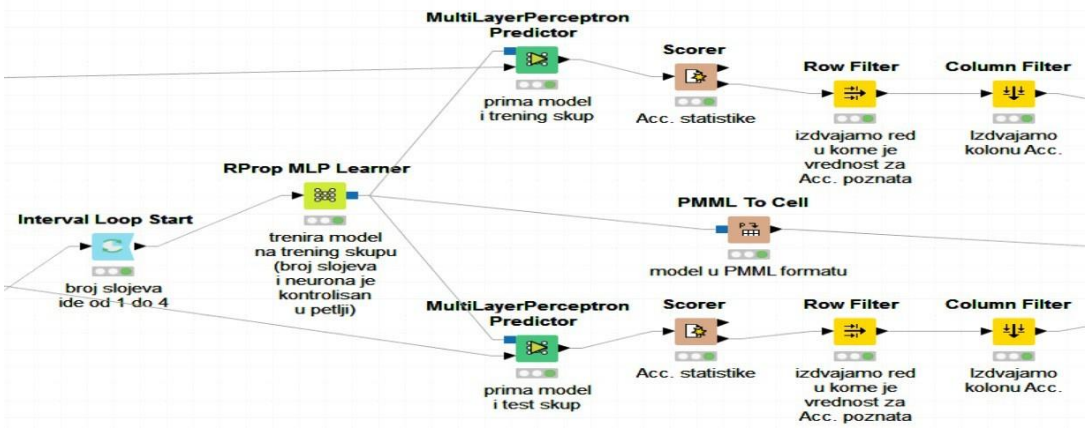
Slika 46

Dalje, izračunate su dimenzije tabele u kojoj se nalazi trening skup, odnosno broj redova i broj kolona. Odatle je izvučen samo broj kolona i spojen sa brojem različitih vrednosti za *target* atribut (atribut koji predstavlja klasu), koji je dobijen preko čvora “GroupBy”. Izračunat je minimalan i maksimalan broj čvorova preko čvorova “Math Formula”, gde minimum predstavlja manju od vrednosti broja kolona umanjenu za jedan (zbog *target* atributa *“Gender”*, koji nije uračunat jer predstavlja klasu) i broja različitih vrednosti *target* atributa, a maksimum predstavlja zbir broja kolona umanjenog za jedan i broja različitih vrednosti *target* atributa. Dobijene vrednosti koje su se smestile u jedan red tabele, prebačene su u promenljivu i ona je povezana sa čvorom “Interval Loop Start” koji će biti kontrolisan vrednostima te promenljive (na slici 47 nalazi se raspored čvorova).



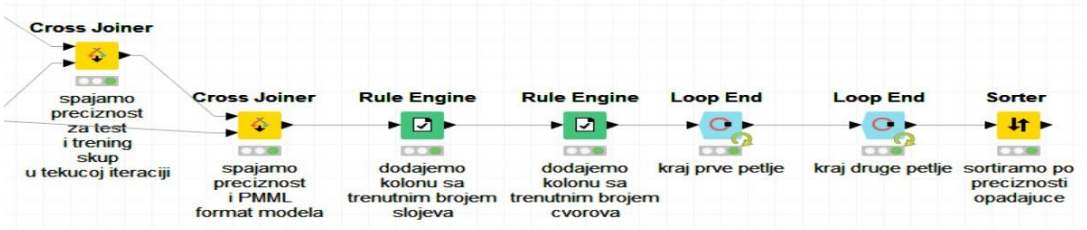
Slika 47

Dalje, dodat je još jedan čvor “Interval Loop Start” koji inicijalizuje unutrašnju petlju i inkrementira broj skrivenih slojeva. Njega smo povezali sa čvorom “RProp MLP Learner” koji trenira model na trening skupu podataka. U svakoj iteraciji, model koji je nastao kao izlaz iz ovog čvora prosleđen je čvoru “PMML To Cell” koji ga čuva u PMML formatu. Ovaj format je inače pogodan za prenošenje modela iz jednog programa za obradu podataka u drugi, jer predstavlja standard. Izvršavanjem čvorova “MultiLayerPerceptron Predictor” na trening i test skupu, računa se preciznost (koliko je dobro predvideo klasu podacima jednog odnosno drugog skupa).



Slika 48: Raspored čvorova

Zatim je spojena preciznost sa modelom u PMML formatu u jednu tabelu kojoj su dodate još dve kolone: jedna sadrži tekući broj slojeva, a druga tekući broj čvorova (neurona) na njemu. Po završetku petlje dobijeni rezultati su sortirani po preciznosti (atributu Accuracy) opadajuće (prvo po test, pa onda po trening skupu).



Slika 49: Raspored čvorova

Dobija se podatak da je preciznost najbolja ako se koriste 3 skrivena sloja i 4 neurona po svakom od slojeva (slika 50).

Row ID	Trening_Accuracy	Test_Accuracy	PMML	broj slojeva	broj neurona (cvorova) po sloju	Iteration	Iteration (#1)
Overall_Ov...	0.879	0.907	<?xml v... <PMML v... <He...	3	4	2	2
Overall_Ov...	0.863	0.89	<?xml v... <PMML v... <He...	1	5	0	3
Overall_Ov...	0.862	0.887	<?xml v... <PMML v... <He...	4	5	3	3

slika 50

7. Zaključak

U ovom radu prikazane su osnovne metode za istraživanje podataka, od pripreme, obrade i analize podataka, do implementacije konkretnih algoritama za klasifikaciju, kao što su stabla odlučivanja, metod K najblizih suseda, Naivni Bajesov algoritam, algoritam potpornih vektora (SVM) i RProp algoritam, i klasterovanje tih podataka. Cilj ovog istraživanja, osim demonstracije pomenutih algoritama jeste da se istraži šta mladi najviše slušaju, na šta najviše troše novac, kao i to koja su njihova interesovanja, razmišljanja i strahovi.