Istraživanje o mladim ljudima

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Gorana Vučić
mi13181@alas.matf.bg.ac.rs

24. jun 2018.

Sažetak

U svetu postoji sve veće interesovanje za istraživanje, obradu, rukovanje podacima u različite svrhe, stoga je ovo primer rada u kome je vršeno istraživanje podataka mladih osoba starosti između 15 i 30 godina. Preuzeti su prikupljeni podaci o njihovim interesovanjima, pogledu na život, hobijima, stavovima itd. Adekvatnim pretprocesiranjem, vizuelizacijom, primenom najpoznatijih algoritama iz oblasti istraživanja podataka prikazani su različiti zanimljivi rezultati, kao rezultat njihove primene.

Sadržaj

1 Uvod **2**

2 Podaci **2**
2.1 Opšti podaci o osobama 2
2.2 Nedostajuće vrednosti 4
2.3 Korelacija između filmskih žanrova 4
2.4 Korelacija između muzičkih žanrova 5

3 Pravila pridruživanja **6**

4 Klasterovanje podataka **9**
4.1 Muzika 9
4.2 Filmovi 10
4.3 Hobiji 10
4.4 Fobije 11
4.5 Osobine ličnosti i pogled na život 11
4.6 Potrošačke navike 11

5 Klasifikacija podataka **12**

Literatura **16**

1 Uvod

Grupa mladih ljudi iz Slovačke je 2013. godine učestvovala u istraživanju. Naime, prikupljeni su podaci o njihovim interesovanjima kao što su muzika i filmovi, fobijama, zdravim navikama, pogledu na život, potrošačkim navikama i demografski podaci. Uz pomoć KNIME *"Konstanz Information Miner"*, SPSS *"Statistical Package for the Social Sciences"* alata, kao i python jezika koji služe za istraživanje podataka, u radu će biti predstavljeni različiti rezultati o podacima, koji su dobijeni primenom odgovarajućih algoritama pravila pridruživanja, klasterovanja i klasifikacije.

2 Podaci

Podaci se mogu preuzeti sa linka i smešteni su u dve datoteke. U datoteci *"responses.csv"* nalaze se odgovori na različita pitanja. Datoteka ima 1010 redova i 150 kolona, od toga 139 kolona su integer vrednosti, a 11 kolona su kategoričke vrednosti. Svi atributi integer tipa koji predstavljaju hobije i interesovanja, pogled na život, potrošačke navike itd. imaju vrednosti od **1** do **5**. Vrednost **1** predstavlja nezainteresovanost osobe za nešto ili neslaganje sa nekim stavom, dok vrednost **5** označava suprotno. Deo podataka se može videti na slici 1. U drugoj datoteci *"columns.csv"* nalaze se originalna imena atributa i njihova skraćena imena koja su korišćena u datoteci *"responses.csv"*.

Music	Slow songs or fast songs	Dance	Folk	Country	Classical music	Musical	Pop	Rock	Metal or Hardrock	...	Age	Height	Weight	Number of siblings	Gender	Left - right handed
5.0	3.0	2.0	1.0	2.0	2.0	1.0	5.0	5.0	1.0	...	20.0	163.0	48.0	1.0	female	right handed
4.0	4.0	2.0	1.0	1.0	1.0	2.0	3.0	5.0	4.0	...	19.0	163.0	58.0	2.0	female	right handed
5.0	5.0	2.0	2.0	3.0	4.0	5.0	3.0	5.0	3.0	...	20.0	176.0	67.0	2.0	female	right handed
5.0	3.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	...	22.0	172.0	59.0	1.0	female	right handed
5.0	3.0	4.0	3.0	2.0	4.0	3.0	5.0	3.0	1.0	...	20.0	170.0	59.0	1.0	female	right handed
5.0	3.0	2.0	3.0	2.0	3.0	3.0	2.0	5.0	5.0	...	20.0	186.0	77.0	1.0	male	right handed

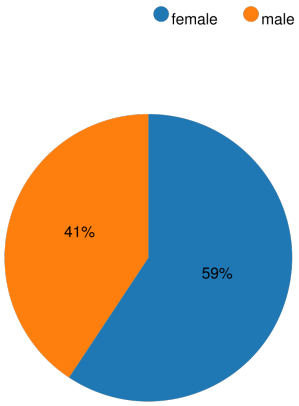
Slika 1: Podaci

2.1 Opšti podaci o osobama

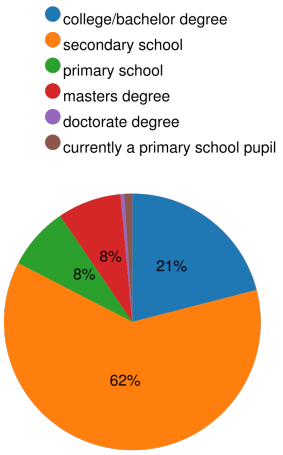
U daljem tekstu uz pomoć dijagrama će biti predstavljeni neki statistički podaci o ispitanicima.

- U istraživanju je učestvovalo 41% osoba muškog pola i 59% osoba ženskog pola (Slika 2).
- 62% ispitanika je završilo srednju školu, 21% je završilo osnovne studije, 8% je završilo osnovnu školu, 8% master studije i ostatak su bili ispitanici koji su bili u osnovnoj školi ili su imali doktorsku disertaciju (Slika 3).
- 70% ispitanika je većinu svog života provela u gradu, dok je 30% njih živelo na selu (Slika 4).
- Po pitanju korišćenja interneta ispitanici su se izjasnili tako da 74% njih je reklo da koristi internet nekoliko sati dnevno, 14% da koristi manje od jednog sata dnevno, dok se 12% izjasnilo da većinu dana provode na internetu. Jedan vrlo mali procenat njih ne koristi internet (Slika 5).
- Kada je konzumacija alkohola u pitanju, 66% ispitanika konzumira alkohol samo kada se nalazi u društvu, 22% se izjasnilo da konzumiraju u velikim količinama, dok 12% ispitanika nikada nije konzumiralo alkohol (Slika 6).

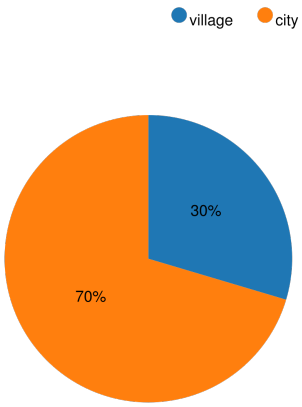
- Što se konzumacije cigareta tiče, 43% se izjasnilo da su probali cigarete, 21% njih ih nikada nije konzumiralo, 19% konzumira i ostatak ispitanika su nekada konzumirali cigarete (Slika 7).



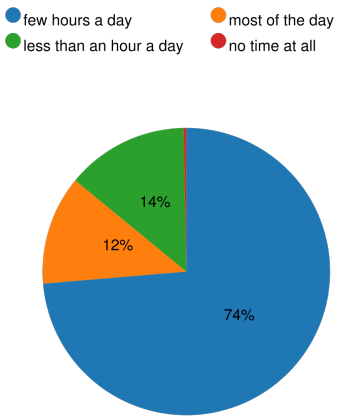
Slika 2: Pol



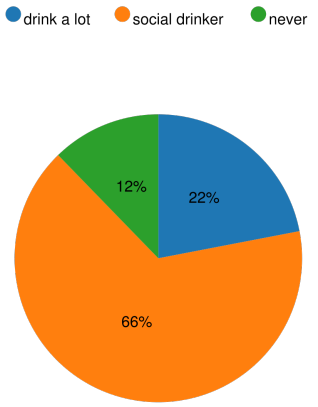
Slika 3: Edukacija



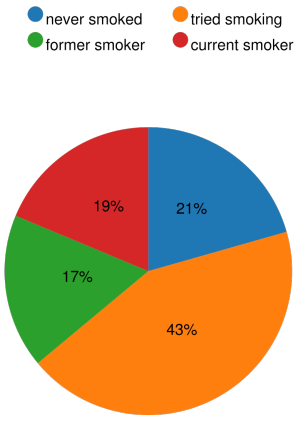
Slika 4: Selo/Grad



Slika 5: Korišćenje interneta



Slika 6: Konzumacija alkohola



Slika 7: Komzumacija cigareta

2.2 Nedostajuće vrednosti

U samim podacima bilo je nedostajućih vrednosti. Iz prikaza koliko nedostajućih vrednosti ima, može se primetiti da je najviše takvih vrednosti bilo u vezi sa atributima gde je trebalo popuniti težinu, visinu, kao i podatke u vezi sa fizičkom aktivnošću. Na slici 8 su prikazani atributi za koje je najviše nedostajalo podataka.

S obzirom da je za većinu atributa bilo jako malo nedostajućih vrednosti, one su obrađene tako da su za atribute tipa integer postavljene zaokružene srednje vrednosti tih atributa, dok su atributi tipa string zamenjeni vrednošću koja se najviše pojavljuje za taj atribut.

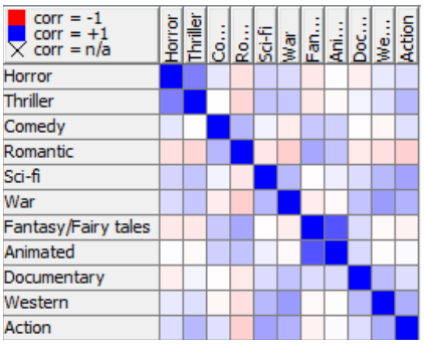
Weight	20
Height	20
Passive sport	15
Chemistry	10
Geography	9
Theatre	8
Smoking	8
Documentary	8
Punk	8
Latino	8
Techno, Trance	7
Criminal damage	7

Slika 8: Nedostajuće vrednosti

2.3 Korelacija između filmskih žanrova

Na slici 9 prikazana je matrica linearne korelacije filmiskih žanrova, na osnovu koje može da se utvrdi da li postoje žanrovi među kojima postoji visoka korelacija. Najveća korelacija se može primetiti između trilera i horor filmova, što pokazuje da osobe koje vole da gledaju trilere, vrlo verovatno vole

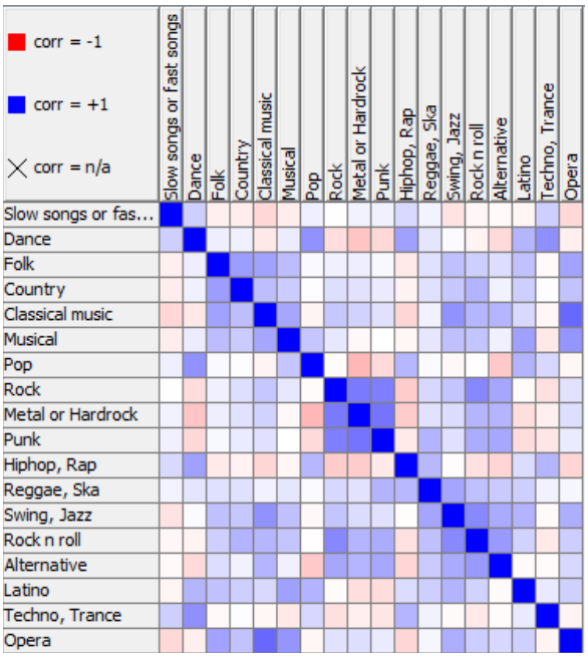
da gledaju i horor filmove (važi i obrnuto). Takođe visoka korelacija se može primetiti i između animiranih i fantazionih filmova. Nešto manja korelacija, ali opet značajna se može primetiti između vestern i ratnih filmova, kao i između akcionih i naučno-fantastičnih filmova. Važi i da osobe koje vole da gledaju fantazione, vole da gledaju i romantične filmove.



Slika 9: Korelacija filmskih žanrova

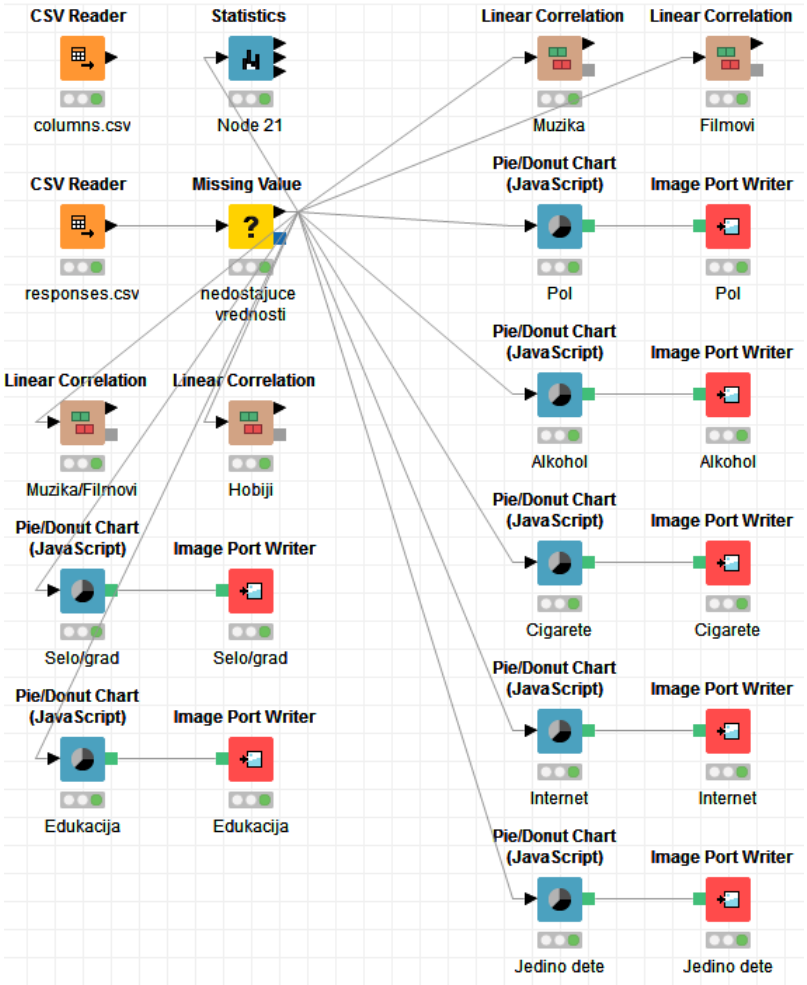
2.4 Korelacija između muzičkih žanrova

Na matrici linearne korelacije muzičkih žanrova, koja se može videti na slici 10, takođe se može primetiti da postoji visoka korelacija između određenih žanrova. Osobe koje slušaju rok muziku takođe slušaju i hardrok kao i pank muziku. Osobe koje slušaju operu naklonjeniji su klasičnoj muzici. Oni koji slušaju pop muziku, slušaju i dens muziku.



Slika 10: Korelacija muzičkih žanrova

Čvorovi koji su korišćeni za prikazivanje opštih podataka o osobama i linearnoj korelaciji u KNIME alatu su prikazani na slici 11.



Slika 11: KNIME

3 Pravila pridruživanja

Kako bi se primenila pravila pridruživanja i otkrila neka zanimljiva pravila, neophodno je pretprocesirati podatke. Naime, kako su podaci delom pogodni za tabelarni format, atributi koji su uzeti u obzir i imali su vrednosti od 1 do 5, zamenjeni su tako da je umesto 4 i 5 postavljena vrednost 1, dok je za ostale vrednosti postavljena vrednost 0. Na taj način izdvojena su samo ona polja atributa za koja su ispitanici dali visoke ocene, što znači da su veoma privrženi značenjima tih atributa. Na samom početku je dodata kolona za transakcioni identifikacioni broj. Za pretprocesiranje podataka u tabelarni format korišćen je python jezik, čiji se kod može videti na slici 12, a deo obrađenih podataka na slici 13.

Kada se na dati skup podataka, gde su u obzir uzeta polja koja su imala samo vrednosti 5 za određeni skup atributa kojima pripadaju svi, sem kategoričkih i atributa vezanih za demografske podatke, primeni *Apriori* i *Carma* algoritam sa podrškom od 10% i pouzdanošću od 80%, *Apriori* algoritam će davati bolje rezultate po lift meri, tako da će njegove vrednosti biti prikazane. Dobije se 5849 pravila, od toga oko 50 pravila koja imaju najveću lift vrednost, u glavi sadrže fantazione filmove i bajke. U telu pravila nalaze se vrednosti kao što su empatija, animirani filmovi, komedije, zabava sa prijateljima, muzika, romantični filmovi itd. i kombinacije ovih vrednosti (slika 14). Kako sva

```
import pandas as pd
df = pd.read_csv("responses.csv")
df.isnull().sum()
df_out1 = df = df.select_dtypes(
    include=['float64','int64']).apply(
    lambda x: x.fillna(x.mean().round(0)))
df.isnull().sum()
df[df<4] = 0
df = df.replace(5.0, 1.0)
df = df.replace(4.0, 1.0)
df.insert(0,'id',range(0, len(df)))
df.to_csv('out.csv')
df
```

Slika 12: Obrada podataka u python jeziku

id	Music	Slow songs or fast songs	Dance	Folk	Country	Classical music	Musical	Pop	Rock	...	Shopping centres
0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	...	1.0
1	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	1.0
2	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	...	1.0
3	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0
4	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	...	0.0
5	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0
6	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0

Slika 13: Deo podataka u tabelarnom obliku

pravila koja imaju veliku lift meru (veću od 2.0) uglavnom prikazuju zavisnost između filmskih žanrova i muzike, ova pravila i nisu preterano zanimljiva.

	Consequent	Antecedent	Lift
1	Fantasy/Fairy tales	Empathy and Animated and Comedy and Fun with friends and Music	2.56
2	Fantasy/Fairy tales	Empathy and Animated and Comedy and Music	2.523
3	Fantasy/Fairy tales	Romantic and Animated and Comedy and Fun with friends	2.521
4	Fantasy/Fairy tales	Romantic and Animated and Comedy	2.52
5	Fantasy/Fairy tales	Romantic and Animated and Comedy and Music	2.513
6	Fantasy/Fairy tales	Empathy and Animated and Comedy and Fun with friends	2.506
7	Fantasy/Fairy tales	Empathy and Animated and Comedy and Movies and Music	2.505
8	Fantasy/Fairy tales	Romantic and Animated and Comedy and Movies and Music	2.499
9	Fantasy/Fairy tales	Romantic and Animated and Comedy and Movies	2.498
10	Fantasy/Fairy tales	Empathy and Animated and Comedy and Fun with friends and Movies	2.496
11	Fantasy/Fairy tales	Romantic and Animated and Fun with friends	2.473

Slika 14: Pravila pridruživanja 1

Ako se iz prethodnog skupa podataka izbace atributi koji predstavljaju filmske i muzičke žanrove, što je moguće uraditi jer postoje nazavisni atributi gde su ispitainici ocenili svoje interesovanje za muziku i film, dobiju se malo drugačija pravila. *Apriori* će dati veći broj pravila od *Carma* algoritma i uz to pravila sa najvećom lift merom su vrlo slična između ova dva algoritma, tako da će ponovo biti prikazani rezultati koji su dobijeni primenom *Apriorija* (slika 15). Kao najzanimljivija pravila izdvajaju se ona koja u glavi sadrže internet, dok u telu pravila sadrže kombinacije atributa PC, zabava sa prijateljima, filmovi, muzika, nauka i tehnologija, što implicira da osobe koje su zainteresovane za ove hobije provode vreme na internetu. Izdvajaju se još neka pravila koja govore da osobe koje vole životinje, empatične su, vole zabavu sa prijateljima, filmove, muziku itd. ne vole da gledaju kako životinje pate.

	Consequent	Antecedent	Lift
1	Internet	PC and Fun with friends and Movies	1.952
2	Internet	PC and Fun with friends and Movies and Music	1.951
3	Internet	PC and Movies and Music	1.937
4	Internet	PC and Fun with friends and Music	1.923
5	Internet	PC and Movies	1.919
6	Internet	PC and Science and technology	1.917
7	Internet	PC and Music	1.915
8	Internet	PC and Fun with friends	1.891
9	Internet	PC	1.87
10	Compassion to animals	Pets and Empathy and Fun with friends and Movies	1.776
11	Compassion to animals	Pets and Empathy and Fun with friends	1.757
12	Compassion to animals	Pets and Judgment calls and Movies and Music	1.756
13	Compassion to animals	Pets and Empathy and Fun with friends and Music	1.749
14	Compassion to animals	Pets and Empathy and Movies and Music	1.749
15	Compassion to animals	Pets and Empathy and Movies	1.743
16	Compassion to animals	Pets and Judgment calls and Fun with friends and Music	1.737
17	Compassion to animals	Pets and Empathy and Music	1.734
18	Compassion to animals	Pets and Empathy	1.729
19	Compassion to animals	Pets and Judgment calls and Music	1.712
20	Fun with friends	Number of friends and Movies	1.375

Slika 15: Pravila pridruživanja 2

Najzanimljivija pravila se dobijaju primenom *Apriori* algoritma na skup podataka u kome su izdvojena polja sa vrednostima 4 i 5, za skup atributa u kome se nalaze svi atributi, sem kategoričkih vrednosti, atributa vezanih za demografske podatke, atributa muzičkih i filmskih žanrova. Podrška je postavljena na 20%, a pouzdanost na 70%. Treba napomenuti da ovde nije bilo moguće poređenje sa rezultatima *Carma* algoritma, jer za data podešavanja nije htela da se izvrši. Na slici 16 se može primetiti da se izdvajaju pravila u kojima:

- Osobe koje vole medicinu, zabavu sa prijateljima, filmove, muziku takođe vole i biologiju;
- Osobe koje vole da troše novac na zabavu, da provode vreme u šoping centrima, zabavu sa prijateljima, vole muziku i filmove, brinu o zdravoj ishrani itd. vole da troše novac na izgled;
- Osobe koje vole umetničke izložbe, zabavu sa prijateljima, muziku i filmove takođe vole da idu u pozorište;

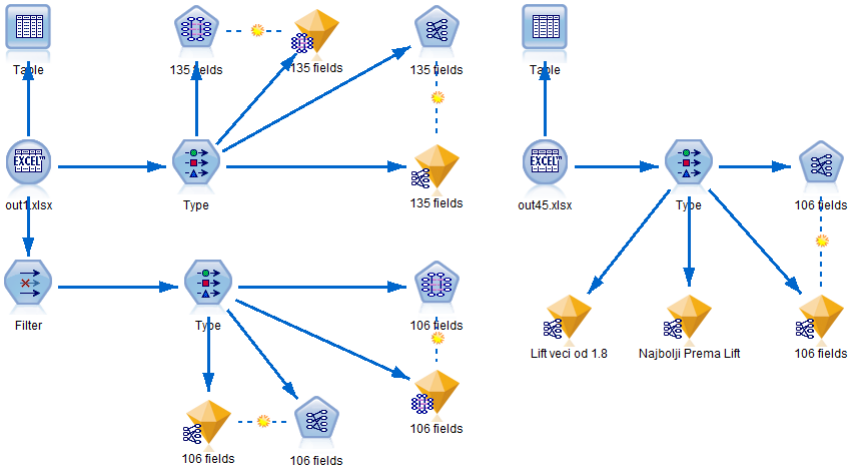
	Consequent	Antecedent	Lift
1	Biology	Medicine and Fun with friends	2.813
2	Biology	Medicine and Movies and Music	2.813
3	Biology	Medicine and Movies	2.812
4	Biology	Medicine and Music	2.797
5	Biology	Medicine	2.789
6	Spending on looks	Entertainment spending and Shopping centres and Fun with friends	2.029
7	Spending on looks	Entertainment spending and Shopping centres and Music	1.991
8	Spending on looks	Entertainment spending and Shopping centres	1.986
9	Spending on looks	Shopping and Shopping centres and Spending on healthy eating and Music	1.982
10	Spending on looks	Entertainment spending and Shopping centres and Movies	1.981
11	Spending on looks	Shopping and Shopping centres and Appearence and gestures and Internet	1.978
12	Spending on looks	Branded clothing and Shopping and Fun with friends	1.975
13	Spending on looks	Shopping and Shopping centres and Spending on healthy eating	1.975
14	Theatre	Art exhibitions and Fun with friends and Music	1.974
15	Theatre	Art exhibitions and Fun with friends	1.973
16	Spending on looks	Branded clothing and Shopping centres and Internet and Fun with friends	1.969
17	Shopping	Spending on looks and Shopping centres and Borrowed stuff	1.965
18	Theatre	Art exhibitions and Fun with friends and Movies	1.96
19	Theatre	Art exhibitions	1.951
20	Theatre	Art exhibitions and Fun with friends and Movies and Music	1.951

Slika 16: Pravila pridruživanja 3

Još neka zanimljiva pravila koja se izdvajaju a nisu prikazana na slici su:

- Osobe koje se lako prilagođavaju novom okruženju, koje imaju različita interesovanja i hobije, koje su pune energije i srećne u životu imaju puno prijatelja;
- Osobe koje vole nauku i tehnologiju, internet i lako se prilagođavaju novom okruženju vole kompijutere;
- Osobe koje vole da idu u pozorište, uvek glasaju na izborima, vole zabavu sa prijateljima i vole muziku takođe vole i da čitaju poeziju;
- Osobe koje veruju da će loši ljudi jednog dana patiti, veruju u Boga;

Svi čvorovi koji su korišćeni za dobijanje pravila pridruživanja primenom *Apriori* i *Carma* algoritma u SPSS modeleru mogu se videti na slici 17.



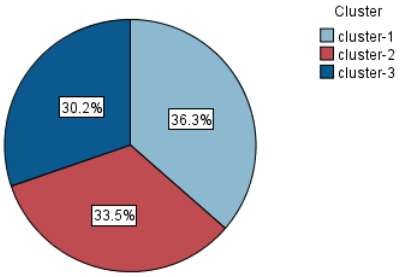
Slika 17: Pravila pridruživanja u SPSS modeleru

4 Klasterovanje podataka

Ideja klasterovanja jeste da se pronađu grupe objekata, takvih da su objekti u grupi međusobno slični (ili povezani) [3, 4]. Kako u datim podacima ima ogroman broj atributa, kada se *K-sredina* ili *Kohonen* algoritam primene na takav skup, dobije se jako loš model čija silueta iznosi 0.1 za broj klastera 3, 4, 5 i 6. S obzirom na to da su podaci podeljeni u određene segmente kao što su muzička interesovanja, interesovanja u vezi sa filmovima, hobiji, fobije, osobine ličnosti i potrošačke navike, u skladu sa tim će biti vršeno klasterovanje i biće prikazani različiti rezultati. Prilikom klasterovanja podataka primenom *Kohonen* algoritma, dobijeni su loši rezultati, tako da u daljem tekstu neće biti razmatrani.

4.1 Muzika

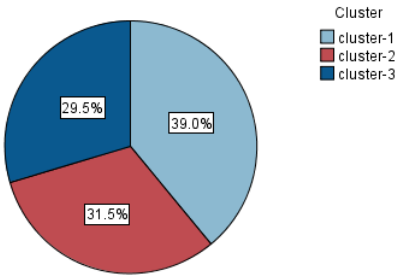
Postoji 18 muzičkih žanrova, primenom *K-sredina* algoritma na ovaj skup atributa najbolji model izdvaja 3 klastera, gde silueta iznosi 0.2 (slika 18). Najznačajni atributi za klasterovanje su metal, rok i rokenrol muzika. Prvi klaster čine ljudi koji najviše vole da slušaju pop, hiphop i rep muziku. Drugi klaster čine osobe koje najviše vole da slušaju rokenrol, sving, džez, dens, operu i pop muziku. Treći klaster čine osobe koje slušaju metal, hardrok, rokenrol, pank i alternativni rok.



Slika 18: Klasterovanje prema interesovanjima za muziku

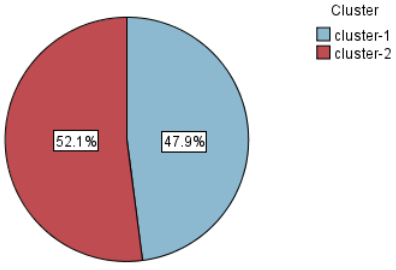
4.2 Filmovi

Primenom *K-sredina* algoritma na 11 filmskih žanrova, najbolji model izdvaja 3 klastera, gde vrednost za siluetu iznosi 0.2. Atributi koji su najznačajniji za klasterovanje su fantazioni i animirani filmovi. Prema rezultatima, prvi klaster čine osobe koje vole da gledaju animirane, fantazione filmove kao i komedije. Drugi klaster čine osobe koje vole da gledaju animirane, naučno-fantastične, ratne, akcione, vestern, trilere i komedije. Treći klaster čine osobe koje najviše vole da gledaju trilere (Slika 19).



Slika 19: Klasterovanje prema interesovanjima za filmove

4.3 Hobiji



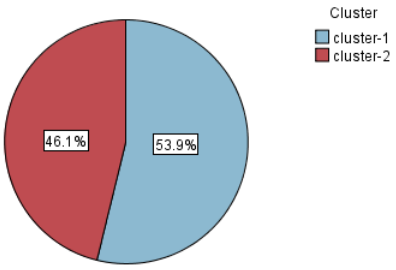
Slika 20: Klasterovanje prema hobijima i interesovanjima

Kada su u pitanju hobiji i različita interesovanja tu postoji 32 atributa, primenom *K-sredina* na ovakav skup najbolji model izdvaja 2 klastera i vrednost za siluetu iznosi 0.2 (Slika 20). Atributi koji su najvažniji za klasterovanje u ovom slučaju su izložbe, pozorište, čitanje, medicina i biologija. Prvi klaster čine osobe koje vole da provode vreme na internetu, zainteresovani su

za menadžment u ekonomiji, vole da se bave sportom i vole zabavu sa prijateljima. Drugi veći klaster čine osobe koje vole izložbe, pozorište, čitanje, medicinu, biologiju, vole da sviraju muzičke instrumente, psihologiju, istoriju, takmičarske sportove i vole zabavu sa prijateljima.

4.4 Fobije

Broj atributa koji opisuju fobije je 10. Neke od fobija su strah od letenja, strah od grmljavine, strah od mraka, visine, paukova, zmija, pacova itd. Najbolji model primenom algoritma *K-sredina* izdvaja 2 klastera. Vrednost za siluetu iznosi 0.3. Atributi koji su najvažniji za klasterovanje su strah od zmija i pacova. Kada se ova dva klastera uporede, manji klaster čine osobe koje se plaše zmija, pacova, paukova, opasnih pasa, mraka i visine, dok drugi veći klaster čine osobe koje nemaju tako veliki strah u vezi sa fobijama ali se može reći da imaju mali strah od javnog govora, starenja i visine (Slika 21).



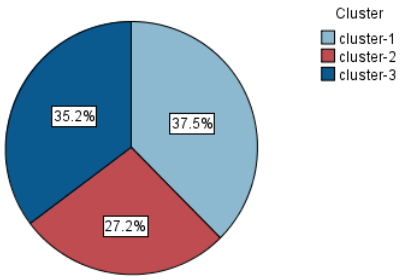
Slika 21: Klasterovanje prema fobijama

4.5 Osobine ličnosti i pogled na život

Za skup atributa kojih ima 57 i predstavljaju osobine ličnosti i pogled na život ispitanika, dobija se i najlošiji model čija silueta iznosi 0.1 i broj klastera je jednak 3 (Slika 22). Atributi koji su najvažniji za klasterovanje su atribut koji opisuje da li osoba piše podsetnike, atribut koji opisuje koliko osobe rade i atribut koji opisuje nivo energije koju osoba poseduje. Prvi klaster čine osobe koje vole da pišu podsetnike, pune su energije, lako se prilagođavaju novom okruženju, imaju veliki broj prijatelja, obavljaju sve svoje zadatke na vreme, izlaze na izbore, žive srećno, slušaju savete svojih roditelja, razmišljaju unapred, ne vole da gledaju patnju životinja, brinu o svojim manirima i izgledu, poznaju prave ljude. Drugi klaster čine osobe koje vole da pišu podsetnike, izdvajaju vreme prilikom donošenja odluka, izražavaju emocije u teškim životnim trenucima, često menjaju raspoloženje, dobro se pripremaju pred javni govor, veruju u Boga, razmišljaju unapred, voleli bi da mogu da promene svoju prošlost, brinu o zdravlju, vole da učestvuju u istraživanjima, lako se naljute, ne vole da gledaju patnju životinja, brinu o manirima i izgledu. Treći klaster čine osobe koje su energične, lako se prilagođavaju novom okruženju, srećne su u životu, u stanju su da polome stvari kada se naljute, varale su u školi, strpljive su.

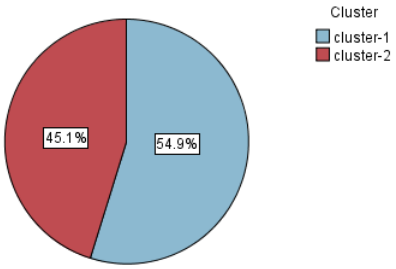
4.6 Potrošačke navike

Kod potrošačkih navika, primenom algoritma *K-sredina* nad 7 atributa, najbolji model izdvaja 2 klastera, gde silueta iznosi 0.3 (Slika 23). Najznačajniji atributi za ovo klasterovanje su atribut koji opisuje da osoba radije nosi brendiranu odeću i atribut koji opisuje koliko osoba troši para na svoj izgled. Kada uporedimo ova dva klastera, prvi klster čine osobe koje ne troše preterano novca na izgled i u šoping centrima dok drugi manji klaster čine osobe



Slika 22: Klasterovanje prema osobinama

koje nose brendiranu odeću, troše puno novca na izgled, provode vreme u šoping centrima, troše novac na uređaje, troše puno novca na žurke i druženje, i vole da troše novac na zdravu ishranu.



Slika 23: Klasterovanje prema potrošnji

5 Klasifikacija podataka

Klasifikacija predstavlja pronalaženje modela koji preslikava svaki skup atributa x u jednu od predefinisanih klasa y [1, 4]. U skupu podataka javlja se 11 kategoričkih atributa, međutim nisu svi zanimljivi za klasifikaciju. Potrebno je napomenuti da je isprobana klasifikacija za atribute kao što su konzumacija alkohola, konzumacija cigareta, upotreba interneta, nivo edukacije, da li osoba laže, da li osoba živi na selu ili u gradu, da li živi u stanu ili u kući primenom različitih klasifikacionih algoritama kao što su *C 5.0*, *CART*, *CHAID*, *K najbližih suseda* [2, 4] i dobijeni modeli su bili preprilagođeni, stoga se neće prikazivati njihovi rezultati.

Za kategorički atribut koji određuje da li je osoba ženskog ili muškog pola dobijaju se dobri modeli. Kada se uporede modeli algoritama *C 5.0*, *CART*, *CHAID* i *K najbližih suseda*, najblji model daje algoritam *C 5.0*, u nastavku će biti prikazano drvo koje je dobijeno njegovom primenom. Podaci su pre primene algoritma particionisani, tako da je 60% podataka uzeto za trening, a 40% podataka za test podatke. Preciznost modela za trening podatke iznosi 98.49%, dok za test podatke iznosi 82.52%. Matrica konfuzije se može videti na slici 24.

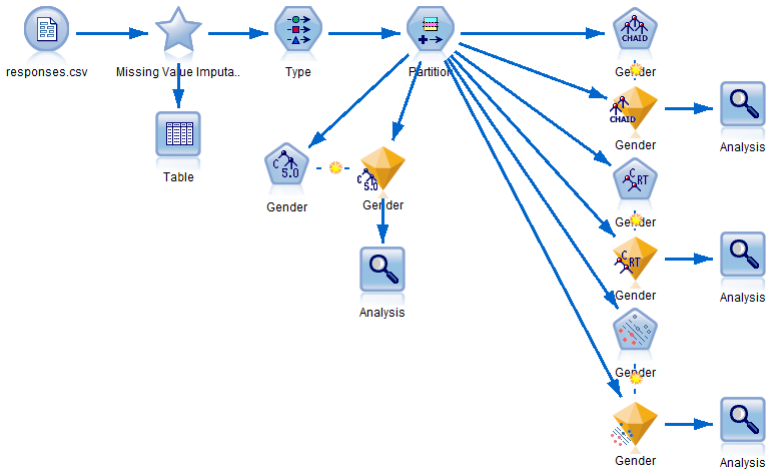
Postmatranjem drveta odlučivanja (Slike: 26, 27, 28) može se primetiti da je prvi atribut po kome se vrši podela visina za vrednost 176 cm. U levom podstablu drveta nalaze se čvorovi u kojima osobe imaju visinu manju od 176 cm, kao naredni atribut za levi deo podstabla izabran je atribut koji predstavlja interesovanje za ratne filmove, gde se u levom podstablu nalaze osobe koje manje vole da gledaju taj žanr. Ako se nastavi sa praćenjem levog podstabla, naredni atribut po kome se vrši podela je težina za vrednost 69

'Partition' = Testing	female	male
female	210	34
male	38	130
'Partition' = Training	female	male
female	353	2
male	7	236

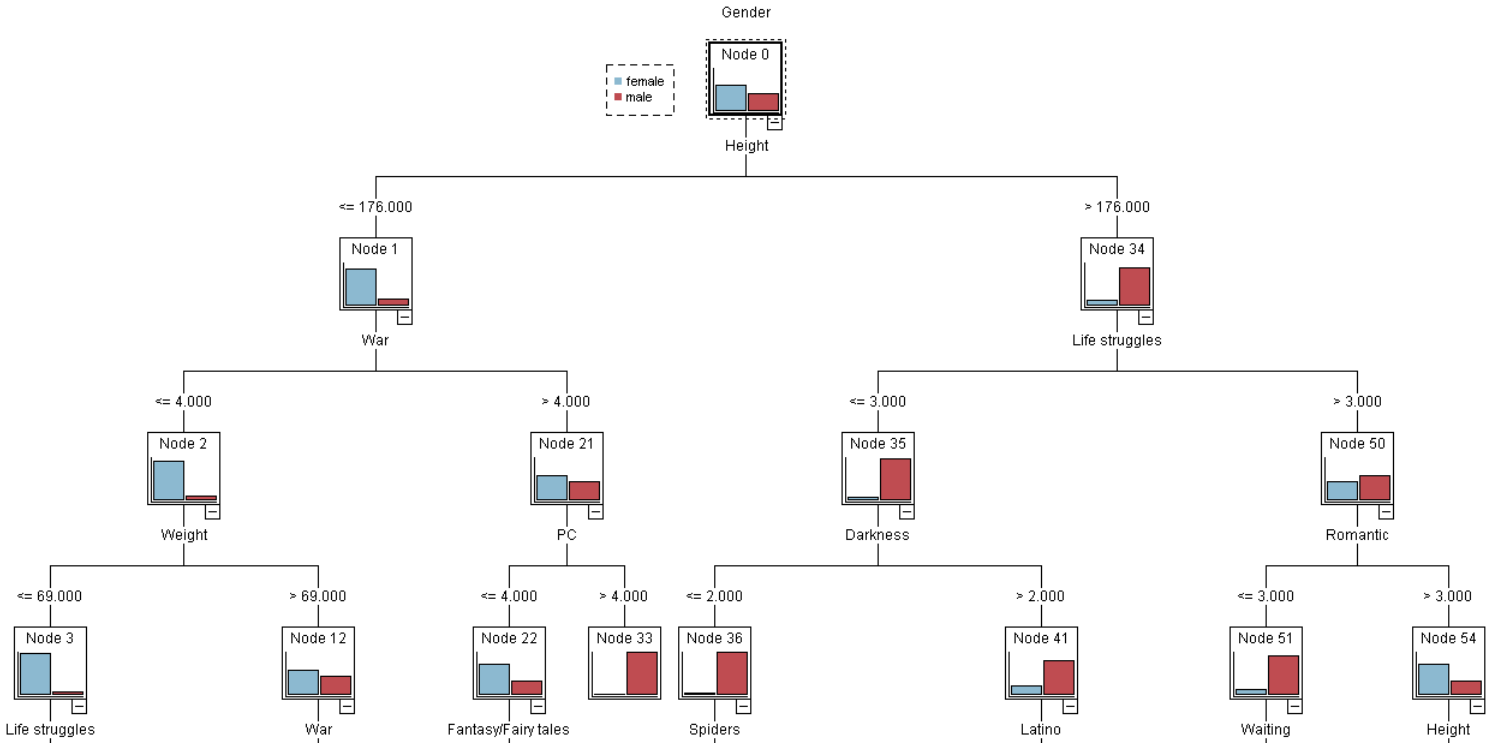
Slika 24: Matrica konfuzije za pol

kg. Potom se podela vrši za atribut koji ocenjuje da li osobe plaču kada imaju životnih problema, koji dovodi i do prvog lista, pa tako osobe koje su niže od 176 cm, manje vole da gledaju ratne filmove, imaju manje od 69 kg i često plaču u teškim situacijama klasifikovane su kao osobe ženskog pola. Dok se u drugom delu istog podstabla nastavlja sa klasifikacijom za atribut koji opisuje kako osobe ocenjuju nivo energije koji poseduju. Osobe koje su niže od 176 cm, manje vole da gledaju ratne filove, lakše su od 69 kg, ne plaču u teškim životnim situacijama, ali takođe opisuju sebe kao osobe sa nedovoljno energije klasifikovane su kao osobe muškog pola. Sa druge strane osobe koje su energičnije, ali nisu preterano zainteresovane za automobile svrstane su u kategoriju ženskog pola, osobe koje vole automobile, ali se osećaju usamljeno svrstane su u kategoriju muškog pola, a osobe koje takođe vole automobile ali se ne osećaju toliko usamljeno kategorisane su kao osobe ženskog pola. Slično i za ostale delove podstabala ovog drveta može se odrediti kojoj klasi pripada neki skup atributa.

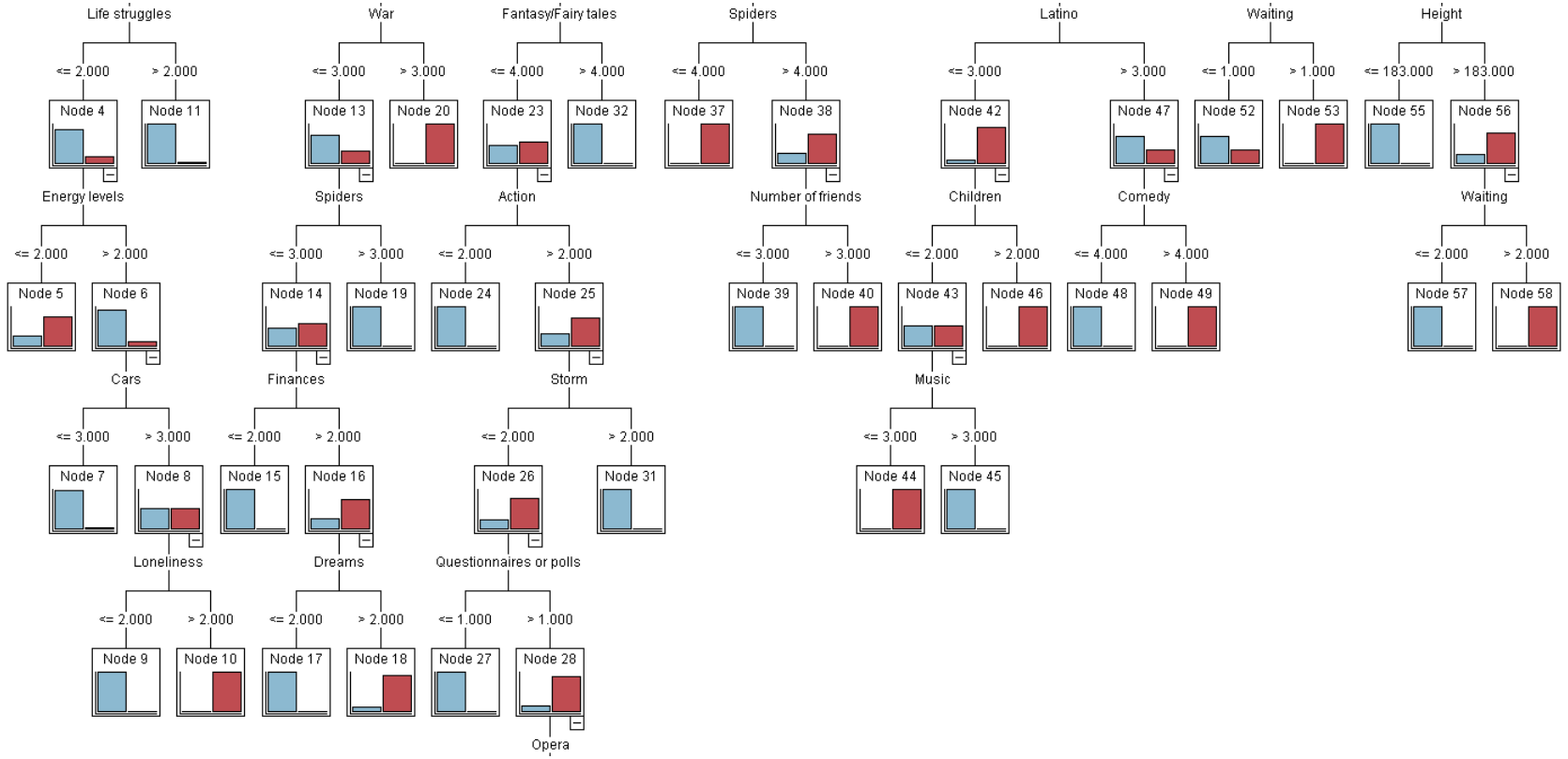
Deo čvorova koji su korišćeni za klasifikaciju u SPSS modeleru mogu se videti na slici 25.



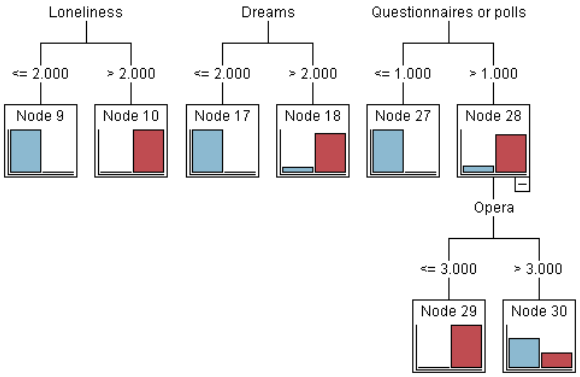
Slika 25: SPSS - klasifikacija



Slika 26: Drvo odlučivanja za pol



Slika 27: Drvo odlučivanja za pol - nastavak



Slika 28: Drvo odlučivanja za pol - nastavak

Literatura

[1] Klasifikacija. http://poincare.matf.bg.ac.rs/~nenad/ip1/7.klasifikacija.pdf.

[2] Klasifikacioni algoritmi. http://poincare.matf.bg.ac.rs/~nenad/ip1/8.klasifikacioni_algoritmi.pdf.

[3] Klasterovanje. http://poincare.matf.bg.ac.rs/~nenad/ip1/6.klasterovanje.pdf.

[4] Charu C. Aggarwal. *Data Mining: The Textbook*. Springer Publishing Company, Incorporated, 2015.