Klasifikacija mobilnih aplikacija

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Vukajlović Nikoleta

14. avgust 2019

Sažetak

U ovom radu su prikazani rezultati istraživanja skupa podataka *Mobile App Statistics*. Primarni cilj je bio klasifikovati ocenu aplikacije na osnovu ostalih atributa. Radi upoznavanja čitaoca sa datim skupom podataka, na početku je dat opis strukture podataka, a zatim i vizuelizacija nekih značajnih atributa, kao i preprocesiranje.

Sadržaj

1 Opis skupa podataka

Skup podataka *Mobile App Statictics*, koji se može naći na linku https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps je skup koji se sastoji od oko 7200 slogova koji sadže detalje o Apple iOS mobilnim aplikacijama. Svaki slog je opisan sa 16 atributa, od kojih je 5 kategoričkih. Detaljan opis atributa je dat u tabeli 1.

Naziv atributa	Opis atributa
id	ID aplikacije
track_name	ime aplikacije
size_bytes	veličina aplikacije (u bajtovima)
currency	valuta
price	cena aplikacije
rating_count_tot	broj ocena korisnika za sve verzije
rating_count_ver	broj ocena korisnika za trenutnu verziju
rating_rating	prosečna ocena korisnika za sve verzije
rating_rating_ver	prosečna ocena korisnika za trenutnu verziju
ver	kod najnovije verzije
cont_rating	ocena sadržaja
prime_genre	primarni žanr
sup_devices.num	broj podržanih uređaja
ipadSc_urls.num	broj screenshot-ova prikazanih na ekranu
lang.num	broj podržanih jezika
vpp_lic	da li je omogućeno licenciranje na Vpp uređaju

Tabela 1: Opis atributa skupa Mobile App Statistics

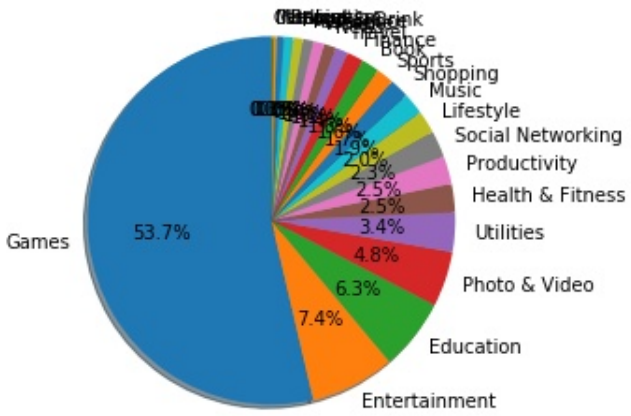
čđšžć

2 Preprocesiranje i vizuelizacija

U ovom odeljku biće diskutovano na koji način je obrađen početni skup podataka kako bi na njega mogli da se primene algoritmi klasifikacije prikazani u daljem tekstu. Vizuelizacija, kao i statistički podaci su takođe pomogli da se dodje do određenih zaključaka koji su uticali na dalje istraživanje. Čitavo preprocesiranje vršeno je u *pajton*-u.

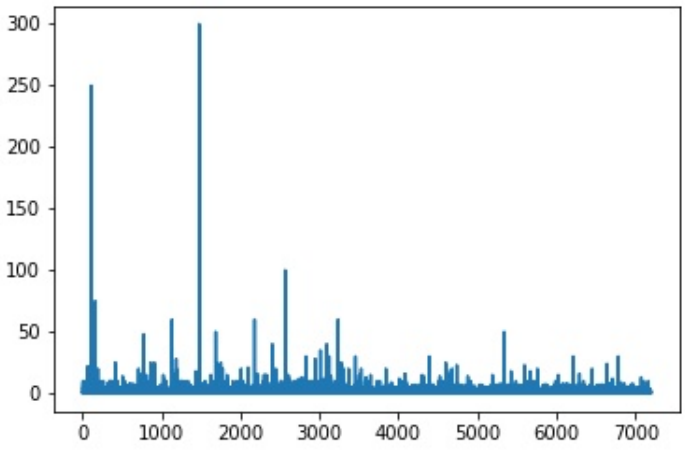
Za početak, skup nije imao nedostajućih vrednosti. Prostojalo je 5 atributa sa kategoričkim vrednostima i to su: id, cur, track_name, prim_genre i ver. *Id* je jedinstvena vrednost aplikacije i on kao takav nije od interesa za proces klasifikacije, pa je izbačen iz skupa. *Cur* je valuta koja je kod svih slogova ista (USD), zbog toga je i ovaj atribut izbačen iz skupa. *Track_ name* je ime aplikacije koje je isto za samo 2 sloga, s toga ni ovaj atribut nije neophodan. *Prime_ genre* je atribut koji opisuje primarni žanr aplikacije i ima 23 različite kategorije. Vizuelni prikaz ovoga atributa može se videti na slici 1. Zbog toliko velikog broja kategorija za žanr, ostavljene su samo one koje obuhvataju vise od 4% svih slogova, tako da je broj žanrova smanjen na 5. Atribut *ver* predstavlja kod najnovije verzije i ima preko 1000 kategorija. Bilo bi besmisleno da se ovaj atribut ostavi u skupu s obzirom na toliko veliki broj kategorija. Da bi to bilo potvrđeno isprobana je klasifikacija i sa ovim atributom koja je dala

čak lošije rezultate, tako da ni ovaj atribut nije korišćen. Takođe atribut *Unnamed: 0* je izbačen iz supa jer predstavljao indeks sloga.



Slika 1: Procenat zastupljenosti žanrova

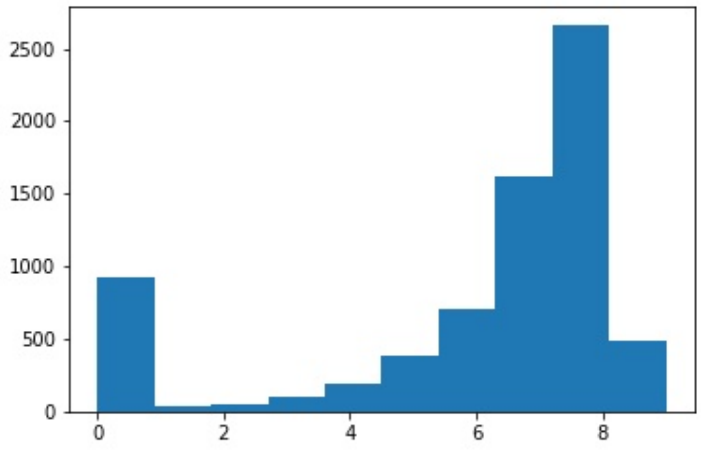
Vrednosti atributa *price* se mogu videti na slici 2. Ustanovljeno je da veoma mali broj aplikacija poseduje vrednosti ovog atributa vece od 50 i mogu se smatrati outlier-ima. Međutim kako ovaj atribut nema toliko značajan uticaj na ciljnu promenljivu i ove vrednosti su ostavljene u skupu.



Slika 2: Vrednosti atributa *price*

Ciljni atribut je bio iz intervala [0, 5] sa korakom 0.5. Da bi mogli da

se primene algoritmi, ovaj atribut je mapiran u vrednosti od 0 do 9 sa korakom 1 i tako smo dobili 10 klasa u koje treba da razvrstamo podatke. Vizuelizacija ovog atributa nam je ukazala da više od pola instanci pripada klasama 7 i 8, sto će kasnije imati velikog uticaja na proces klasifikacije. Ovo se može videti na slici 3.



Slika 3: Vizuelizacija atributa *user_rating*

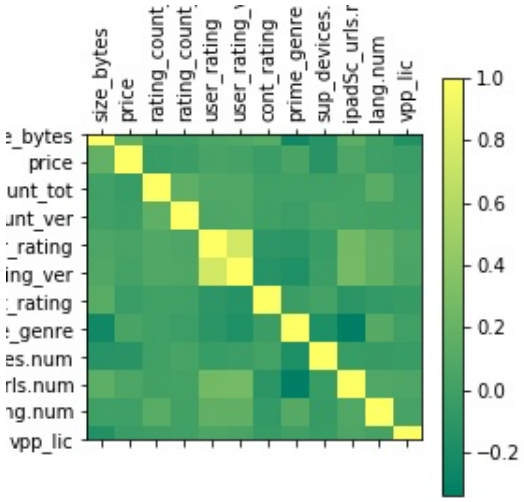
Vizuelizacija ostalih atributa nije dovela do nekih zanimljivijih zaključaka. S toga je izvršena i analiza statističkih podataka dobijenih od atributa koji su nama od važnosti. Uvideli smo da postoji velika korelacija između atributa *user_rating_ver* i ciljnog atributa *user_rating*, što je i očekivano jer ne bi trebalo da se ocena znatno promeni pri prelasku na drugu verziju. Korelacije između ostalih atributa prikazane su na slici 4.

3 Klasifikacija

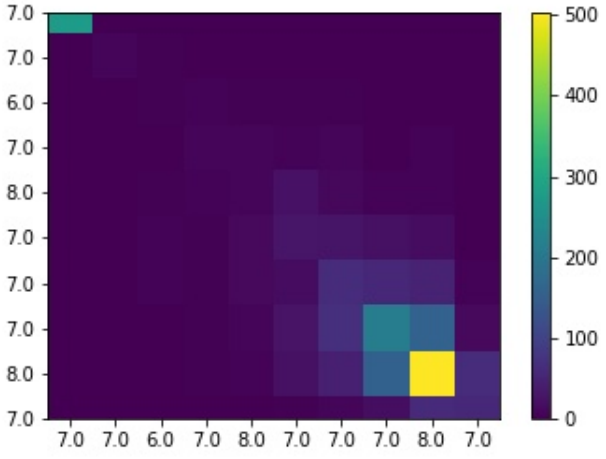
Klasifikacija je rađena metodom stabla odlučivanja, k najbližih suseda, neuronskom mrežom i metodom slučajne šume. Isprobane su i neke druge metode, ali ove su izdvojene kao one koje su imale najviše uspeha. Neuronske mreže su rađene u *SPSS Modeler*-u, dok su ostale metode odrađene u *python*-u. Prilikom klasifikacije korišćeni su svi atributi dobijeni preprocesiranjem, jer izdvajanje najznačajnijih atributa nije dovelo do boljih rezultata.

3.1 Klasifikacija stablom odlučivanja

Model dobijen stablom odlučivanja je bio preprilagođen, pa je preciznost na test podacima bila veoma loša, 54%. Nije ni očekivan bolji rezultat s obizorom da postoji 10 klasa, među kojim više od polovine instanci pripada dvema klasama i da se stabla odlučivanja bolje ponašaju pri binarnoj klasifikaciji, te je ova metoda urađena radi poređenja sa ostalim modelima. Matrica konfuzije je prikazana na slici 5 .



Slika 4: Korelacija između atributa

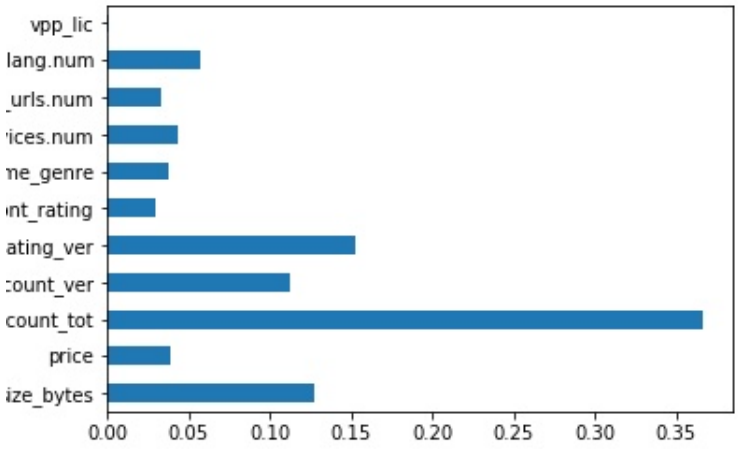


Slika 5: Matrica konfuzije dobijena stablom odlučivanja

Atribut sa najvećom važnošću kod ovog modela je bio *rating_ count_ tot*, što i nije bilo očekivano s obzirom da je korelacija između ovog atributa i ciljne promenljive statistički gledano bila 0.08. Važnosti ostalih atributa mogu se videti na slici 6 .

3.2 Klasifikacija metodom K najbližih suseda

Metoda K najbližih suseda je dala neznatno bolje rezultate u odnosu na stablo odlučivanja. U petlji je izvršavano klasifikovanje sa različitim



Slika 6: Vaznost atributa kod stabla odlučivanja

parametrima kako bi se došlo do najboljeg rezultata. Broj suseda je išao od 3 do 10, za taj parametar vršena je klasifikacija kad svi susedi imaju podjednak uticaj, a zatim i kada bliži imaju veći uticaj, a takođe korišćeno je i Euklidsko i Menhetn rastojanje. Najbolju preciznost smo dobili u dva slučaja. Prvi kada je korišćeno Euklidsko rastojanje sa podjednakim uticajem 7 suseda. Drugi najbolji slučaj je dobijen kada je korišćeno Manhatn rastojanje, gde su bliži susedi imali veći uticaj i bilo ih je 8. Preciznost je oba puta bila 56%.

Na slici 8 je prikazano kako se menjala greška sa različitim vrednostima za parametar k. Za vizuelizaciju ovih podataka k je išlo od 1 do 50 sa korakom 2, kako bi se dobila što šira slika.

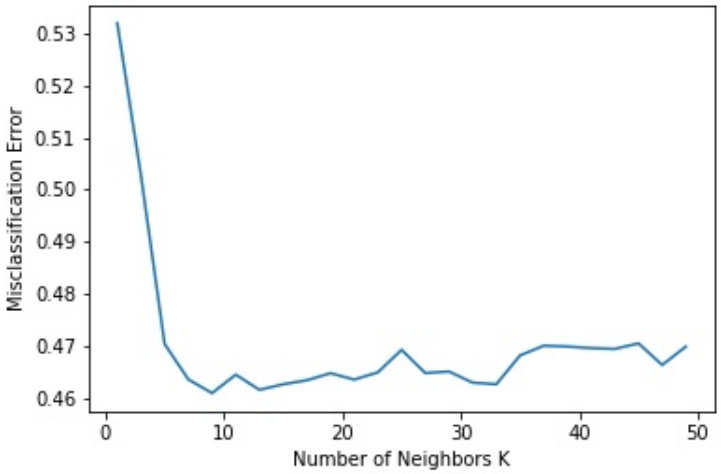
3.3 Klasifikacija metodom slučajne šume

U izostajanju dobrih rezultata, isproban je i algoritam slučajne šume. Iako je zasnovan na stablima odlučivanja, ovaj algoritam bi trebalo da daje bolje rezultate jer je zasnovan na ideji da više stabala može doneti bolju odluku nego jedno. Kontrolisan je parametar *n_estimators* koji odgovara broju stabala koji učestvuju u glasanju. Takođe su poređeni i rezultati koji su dobijeni pri korišćenju *Gini* kriterijuma i *entropije*.

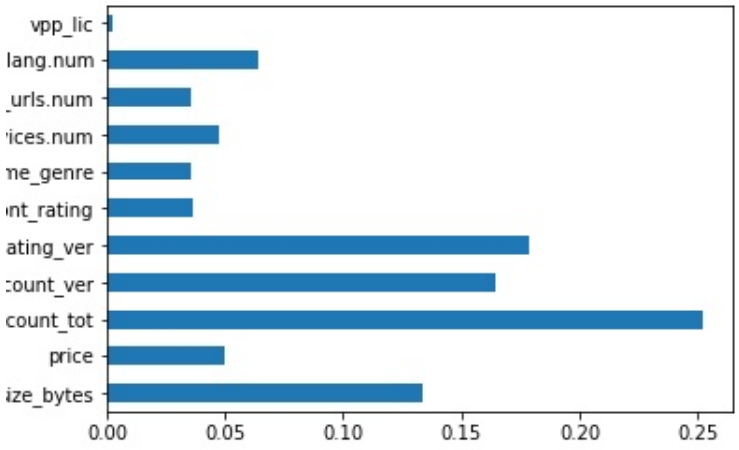
Kao što je je i očekivano, najbolja preciznost kod test skupa je dobijena primenom ove metode i ona je iznosila i do 62% za na primer n_estimators = 28 i criterion='entropy'. Važnost atributa je prikazana na slici 10 i ona je veoma slična kao i kod stabala odlučivanja.

4 Klasifikacija neuronskim mrežama

Klasifikacija neurnskim mrežama je rađena u *SPSS Modeler*-u. Ova metoda je neočekivano dala slične rezultate kao i stablo odlučivanja. Preciznost na test skupu je bila 54.44%, što se može videti na slici 9.



Slika 7: Greske za različite parametre k



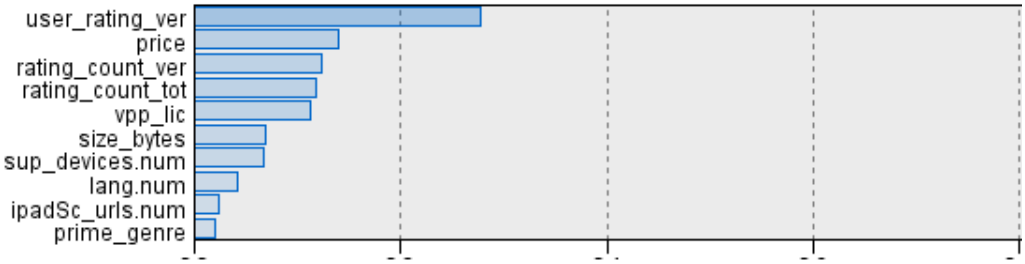
Slika 8: Vaznost atributa kod slučajnih šuma

Comparing $N-user_rating with user_rating

'Partition'	1_Training		2_Testing	
Correct	3,234	56.18%	784	54.44%
Wrong	2,523	43.82%	656	45.56%
Total	5,757		1,440	

Slika 9: Preciznost na trening i test skupu kod neuronskih mreza

U preprocesiranju smo došli do zaključka da najveći uticaj na ciljnu promenljivu ima atribut *user_ rating_ ver*, što je ova metoda i potvrdila. Kod ostalih metoda to nije bio slučaj. Važnost atributa kod neuronskih mreža se može videti na slici 10 .



Slika 10: Važnost atributa kod neuronskih mreža

5 Zaključak

Iako su podaci na prvi pogled bili dobri, nije bilo nedostajućih vrednosti i nekonzistentnosti, rezultati nisu zadovoljavajući, iako su isprobane mnoge metode sa različitim vrstama ulaznih podataka. Po mom mišljenju, to je zbog neravnomerne rasprostranjenosti instanci po klasama, odnosno 2 od 10 klasa obuhvataju skoro 3/4 instanci, s toga nije moglo puno toga da se zaključi i uspešnost procesa klasifikacije nije očekivana.