Primena algoritama Istraživanja podataka nad skupom podataka 'baseball dataset'

Matematički fakultet Univerziteta u Beogradu

Relja Pešić 73/2019

August 2023

Sadržaj

Uvod

Ovaj rad je predvidjen za demonstriranje rada brojnih tehnika istraživanja podataka. Baza podataka koja se koristi sadrži podatke o bejzbol igračima i njihovim statistikama koje su postigli.

1.1 Analiza skupa podataka

Ukupan broj instanci je 1340 od kojih njih 20 ima nedostsajuće vrednosti. Broj predvidjenih klasa je 3.
Iz skupa podataka su uklonjeni karakteri navodnika, apostrofa i obnutih kosih crta, a razmaci su zamenjeni karakterom donja crta. Napomenuto je da je kolona 'Player' identifikator koji je potrebno ignorisati prilikom kreiranja modela.

Podaci su opisani narednim atributima:

- ciljni atribut 'Hall_of_Fame': Prilikom klasifikacije instancama se dodeljuje jedna od tri moguće vrednosti: '0', '1', '2'
- 'Player': Naziv igrača, identifikator sa 1339 različitih vrednosti, bez nedostajućih vrendnosti
- 'Number_seasons': Broj sezona tokom kojih je igrač bio aktivan, numerički atribut 17 različitih vrednosti, bez nedostajućih vrednosti
- 'Games_played': Podatak o broju utakcima koje je igrač odigrao, numerički atribut 981 različitih vrednosti, bez nedostajućih vrendnosti
- 'At_bats': Poen koji se ostvaruje ukoliko udarač osvoji bazu pod nekim specifičnim uslovima, numerički atribut 1239 različitih vrednosti, bez nedostajućih
- 'Runs': Poen koji se dobija ukoliko igrač uspešno obidje sve baze i vrati se na početnu poziciju, numerički atribut 812 različitih vrednosti, bez nedostajućih vrendnosti
- 'Hits': Pripisuje se udaraču ukoliko nakon udara loptice stigne do ili prodje prvu bazu, numerički atribut 999 različitih vrednosti, bez nedostajućih vrendnosti
- 'Doubles': Pripisuje se udaraču ukoliko nakon udara loptice stigne do ili prodje drugu bazu, numerički atribut 418 različitih vrednosti, bez nedostajućih vrendnosti
- 'Triples': Pripisuje se udaraču ukoliko nakon udara loptice stigne do ili prodje treću bazu, numerički atribut 180 različitih vrednosti, bez nedostajućih vrendnosti
- 'Home_runs': Poeni koje se ostvaruju ukoliko udarač izbije lopticu van granica terena, numerički atribut 291 različitih vrednosti, bez nedostajućih vrendnosti
- 'RBIs': Poeni koji se dodeljuju udaraču ukoliko nakon što udari lopticu, njegov saigrač ostvari trčanje ('Run'), numerički atribut 795 različitih vrednosti, bez nedostajućih vrendnosti

- 'Walks': Ukoliko bacač četiri puta nepravilo baci lopticu, udarač se može pomeriti na sledeću bazu, numerički atribut 712 različitih vrednosti, bez nedostajućih vrendnosti
- 'Strikeouts': Ukoliko bacač uspe da postigne tri pogotka, numerički atribut 722 različitih vrednosti, 20 nedostajućih vrendnosti
- 'Batting_average': Prosek uspešnih udaraca igrača, numerički atribut 143 različitih vrednosti, bez nedostajućih vrendnosti
- 'On_base_pct': Procenat koji govori koliko često udarač osvaja baze, numerički atribut 176 različitih vrednosti, bez nedostajućih vrendnosti
- 'Slugging_pct': Broj baza koje je igrač osvojio po 'At_bats', numerički atribut 274 različitih vrednosti, bez nedostajućih vrendnosti
- 'Fielding_ave': Procenat koji opisuje koliko puta je odbrambeni igrač uspešno reagovao na udarenu lopticu, numerički atribut 125 različitih vrednosti, bez nedostajućih vrendnosti
- 'Position': Pozicija igrača, kategorički atribut 7 različitih vrednosti, bez nedostajućih vrendnosti

Kategorički atributi 'Player', 'Position' i 'Hall_of_Fame' koji su prvobitno bili sačuvani kao bString su dekodirani radi lakšeg rada.

1.2 Pretprocesiranje

1.2.1 Rad sa nedostajućim vrednostima

Kao što je već napomenuto, dvadeset instanci ima nedostajuće vrednosti za atribut 'Strikeouts'. S'obzirom na manjkost ovakvih podataka u odnosu na ceo skup, razne tehnike dovodile su do sličnih rezultata. Problem rešavamo tako što nepoznate vrednosti menjamo prosečnom za tu kolonu.

1.2.2 Identifikacija elemenata van granica

Anomalije u podacima nisu očigledne, iako su elementi van granica prisutni u velikoj većini. Skup sadrži igrače koji su postigli zavidna dostignuća što upravo mogu biti igrači koji su se upisali u 'Hall of Fame' tj. instance koje pripadaju klasama '1' i '2'.

1.2.3 Enkodiranje kategoričkih atributa

Nominalni atribut 'Position' ima sedam mogućih vrenosti: 'Outfield', 'Second_base', 'Third_base', 'First_base', 'Shortstop', 'Catcher', 'Designated_hitter'. Kolonu enkodiramo dodavanjem novih kolona za svaku od navedenih vrednosti. Novododate kolone mogu imati vrednosti '0' i '1' čime indukuju na kojoj poziciji igra odgovarajući igrač. ('1' označava da instanca ima odgovarajuću vrednost za atribut 'Position', a '0' ne) Navedeni proces ze naziva binarizacija.

1.2.4 Priprema podataka za klasifikaciju

Postoji mnogo veći broj instanci koje pripadaju klasi '0' nego klasama '1' i '2', što dodatno otežava proces klasifikacije. Ukoliko se ovaj problem ne reguliše, naši modeli će favorizovati brojniju klasu i neće utvrditi pravilnosti koje važe za manjinske klase.

- klasa '0': 1215 instanci
- klasa '1': 57 instanci
- klasa '2': 68 instanci

U situacijama kada distribucija instanci po klasama nije jednaka, tačnost odnosno *accuracy* nije dovoljna mera kvaliteta modela. Npr. ako 99% instanci pripradaju većinskoj klasi, možemo napraviti model koji sve instance klasifikuje u većinsku. Tačnost ovakvog modela je 0.99 što predstavlja samo prividnu sigurnost, jer u većinskom slučaju upravo instance koje pripadaju manjinskim klasama želimo ispravno da klasifikujemo.

Postoje brojne tehnike za obradu nebalansiranih klasa uključujući metode za over-sampling (metode koje generišu nove instance manjinskih klasa), metode za uder-sampling (metode koje uzorkovanjem iz brojnije klase izjednačavaju brojnost klasa) kao i metode dodeljivanja odredjenih težinskih vrednosti manjinskim klasama.

Tehnike uzorkovanja (Under-sampling metode) nisu primenjve u ovom slučaju zato što nakon podele podataka na trening i test skup nemamo dovoljno podataka za treniranje modela.
Tehnike generisanja veštačkih instanci poput SMOTE i SMOTENN takodje nisu primenjive, zato što skup podataka poseduje kategoričke atribute.
Ovaj problem rešava varijacija prethodno pomenutog algoritma - SMOTENC kome se može naglasiti koji atributi su kategorički. Bitno je napomenuti da su algoritmi koji generišu vešstačke instance osetljivi na autlajere (*engl - outliers*), jer se nove instance generišu izmedju postojećih, te iz tog razloga pre primene algoritma moramo standardizovati podatke.

Na osnovu grafičkog prikaza primene SMOTENC, možemo pretpostaviti da će najteži zadatak za naše modele upravo biti klasifikacija instanci koje pripadaju klasi 2 (na slici prikazane zelenom bojom). Zato što se najviše mešaju sa ostalim instancama.

1.2.5 Priprema podataka za klasterovanje

Skaliramo numeričke atribute skupa X i nakon toga primenjujemo PCA da bismo mogli da vizualizujemo podatke.

Na grafičkom prikazu pripremnjenih podataka za klasterovanje uočavamo da podaci nisu jasno odvojeni jedni od drugih, medjutim primećujemo da postoji razlika u gustini raspodele podataka koji se nalaze na levom i na desnom kraju slike, što može sugerisati odgovarajuću metriku za računanje klastera.

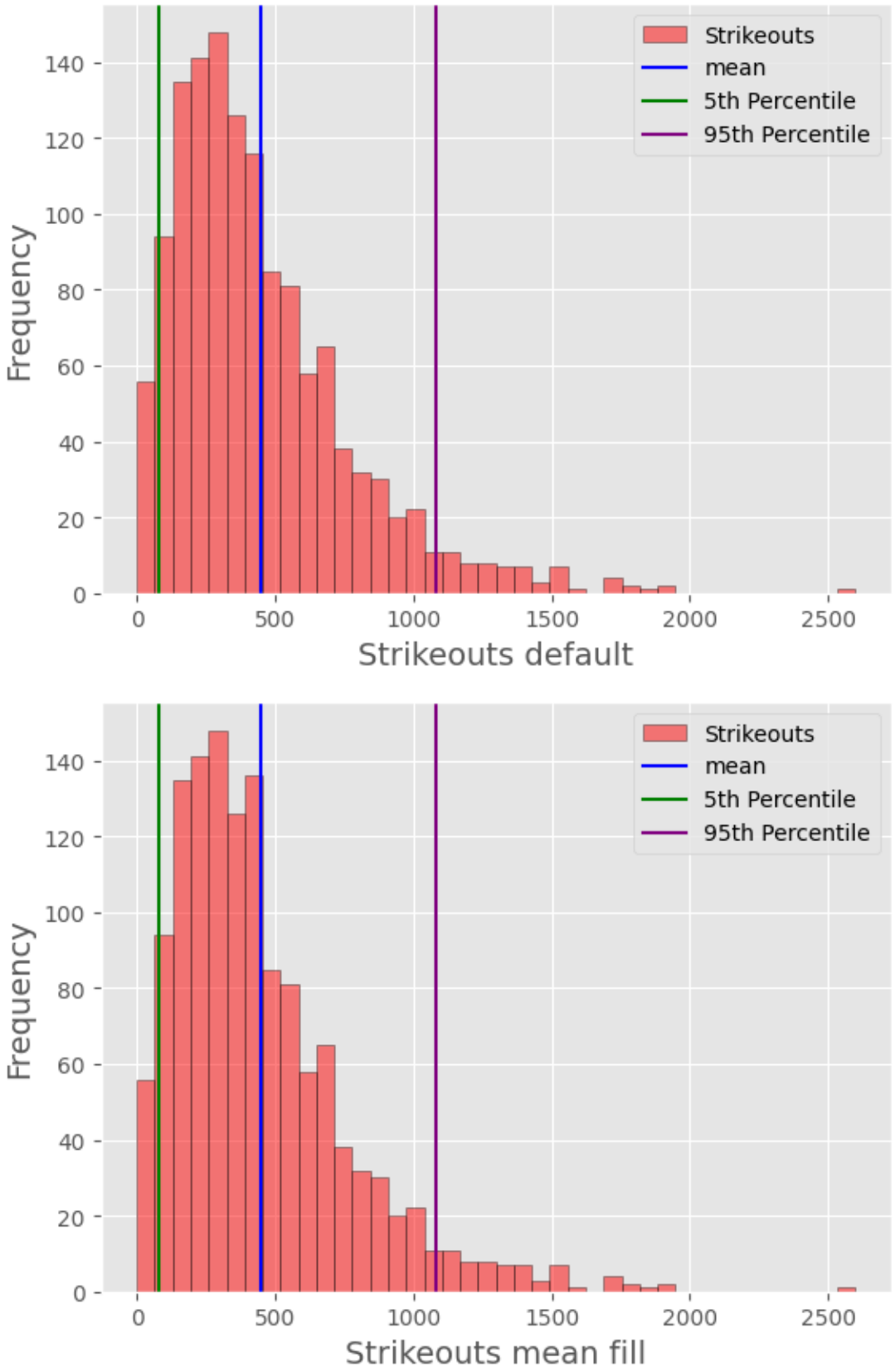


Figure 1.1: Obrada nedostajućih vrednosti

	lower	min	num_lower	upper	max	num_upper	percantage
Number_seasons	5.000000	10.000	0	21.000000	26.000	29	2
Games_played	-80.375000	140.000	0	2690.625000	3562.000	17	1
At_bats	-1274.125000	252.000	0	10068.875000	14053.000	13	1
Runs	-377.375000	20.000	0	1575.625000	2246.000	34	3
Hits	-503.250000	48.000	0	2882.750000	4256.000	24	2
Doubles	-106.000000	6.000	0	486.000000	792.000	33	2
Triples	-48.500000	0.000	0	139.500000	309.000	61	5
Home_runs	-107.000000	0.000	0	237.000000	755.000	106	8
RBIs	-333.500000	21.000	0	1376.500000	2297.000	46	3
Walks	-284.000000	17.000	0	1092.000000	2056.000	43	3
Strikeouts	-344.537500	0.000	0	1149.122500	2597.000	51	4
Batting_average	0.202500	0.161	6	0.334500	0.366	19	2
On_base_pct	0.251625	0.194	8	0.420625	0.483	14	2
Slugging_pct	0.221500	0.201	1	0.545500	0.690	13	1
Fielding_ave	0.918000	0.820	76	1.022000	1.000	0	6

Figure 1.2: Identifikacija elemenata van granica preko IQR

	Position
0	Outfield
1	Second_base
2	Second_base
3	Third_base
4	First_base
...	...
1335	Outfield
1336	Catcher
1337	Third_base
1338	Third_base
1339	Outfield

1340 rows × 1 columns

	Catcher	Designated_hitter	First_base	Outfield	Second_base	Shortstop	Third_base
0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	1.0	0.0	0.0
2	0.0	0.0	0.0	0.0	1.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	1.0
4	0.0	0.0	1.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...
1335	0.0	0.0	0.0	1.0	0.0	0.0	0.0
1336	1.0	0.0	0.0	0.0	0.0	0.0	0.0
1337	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1338	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1339	0.0	0.0	0.0	1.0	0.0	0.0	0.0

1340 rows × 7 columns

Figure 1.3: Binarizacija kolone 'Position'

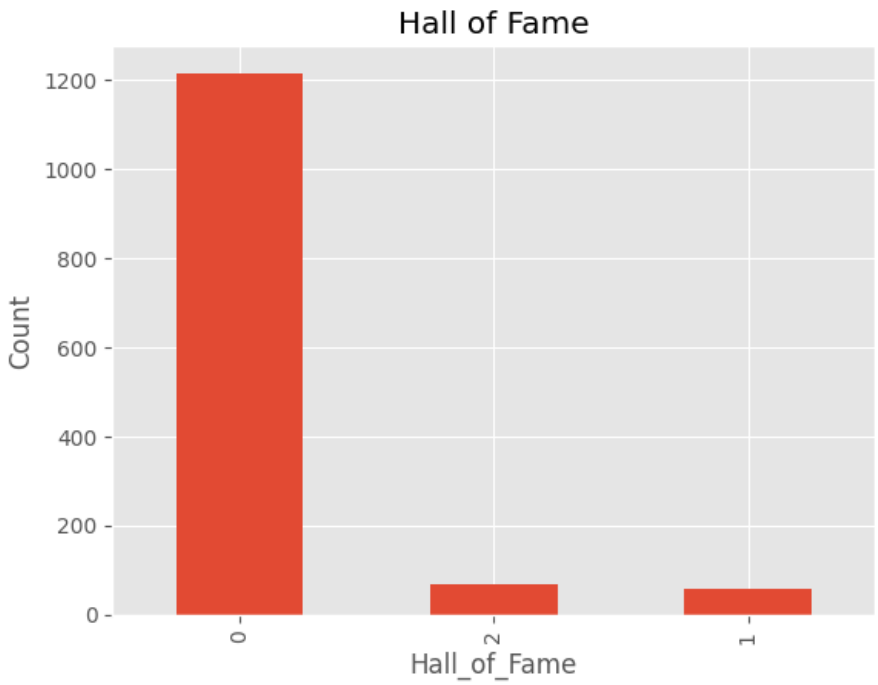


Figure 1.4: Prikaz distribucije podataka po klasama

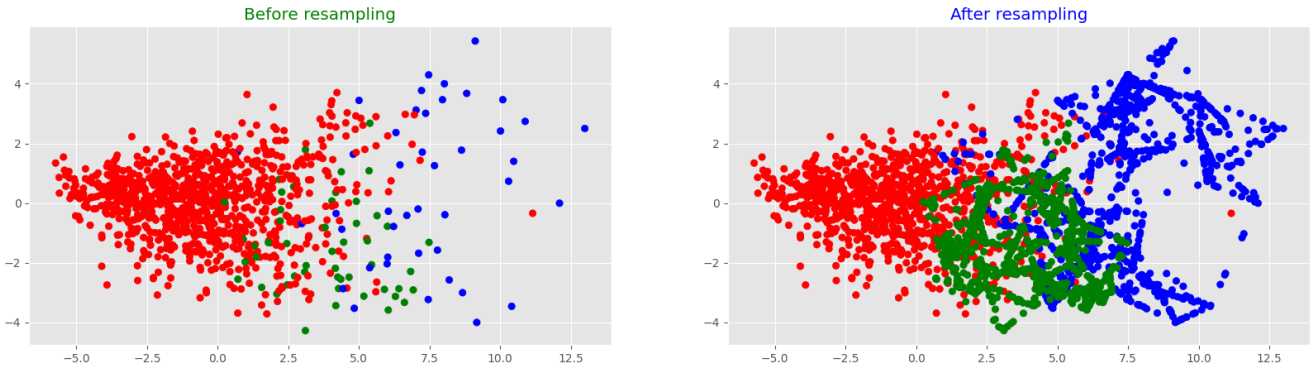


Figure 1.5: Grafički prikaz primene SMOTENC

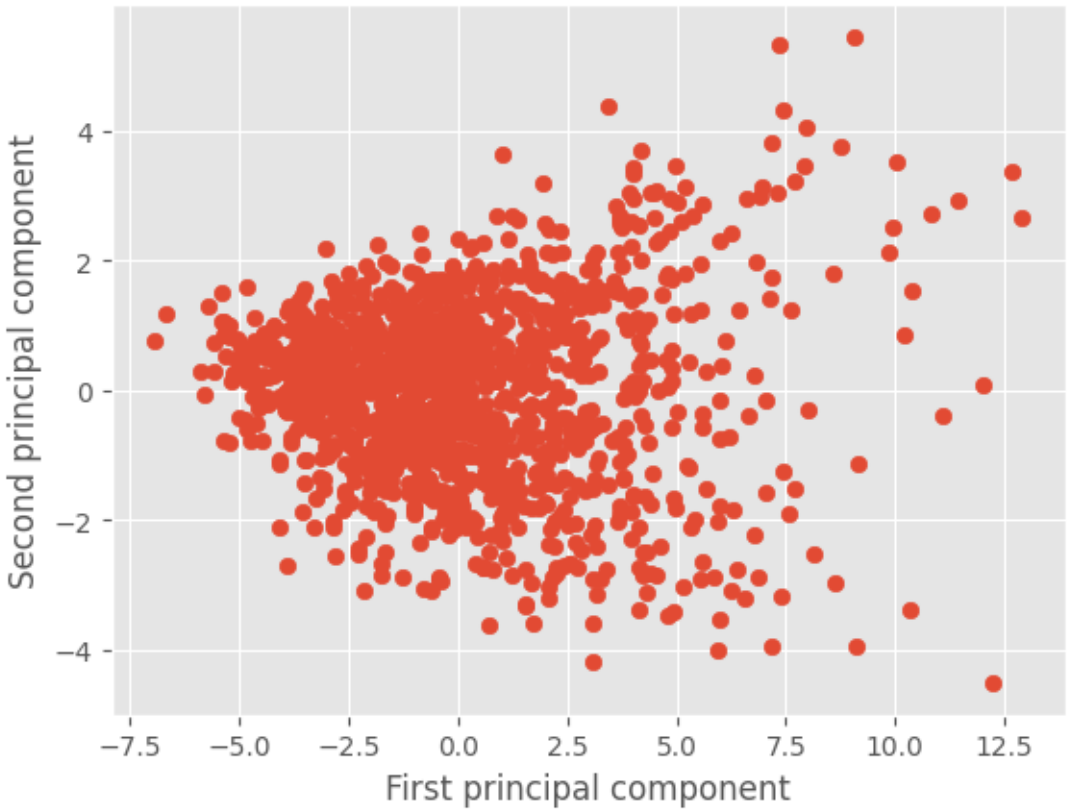


Figure 1.6: Grafički prikaz skupa X nakon skaliranja i primene PCA

Klasifikacija

Zadatak klasifikacije jeste da odredi funkciju (klasifikacioni model) koja kvalitetno preslikava svaki skup atributa X u jedno od predefinisanih vrednosti atributa y (ciljne klase). Klasifikacija pripada metodama nadgledanog učenja. U našem slučaju, predvidjamo koji od igrača pripada odgovarajućoj klasi atributa 'Hall_of_Fame' koja ima moguće vrednosti '0', '1', '2'.

2.1 Stabla odlučivanja

Jedan od osnovnih algoritama za rešavanje problema klasifikacije jesu 'Stabla odlučivanja'. Model se formira u obliku drveta, a u čvorovima koji nisu listovi se nalaze pitanja na osnovu kojih se vrši granjanje. Ulazni podaci se rasporedjuju na osnovu vrednosti njihovih atributa i pitanja u čvorovima. Listovi stabla odredjuju klase kojim instance pripadaju.

Unapredjenu varijaciju ovog algoritma predstavlja algoritam slučajnih šuma, koji kreira više stabala odlučivanja na nasumičnim uzorcima skupa podataka. Finalni zaključak donosi kombinovanjem prethodnih predvidjanja. U našem slučaju korišćen je RFC sa 100 predvidjača.

Za evaluaciju modela posmatramo preciznost, odziv, f1 skor, kao i odgovarajuću matricu konfuzije. Prilikom odabira najboljeg modela treba napomenuti da je biran tako da dobro predvidja sve tri klase.

Algoritam slučajnih stabala ima bolje performanse od algoritma stabala odlučivanja, što je i očekivano. Medjutim, bitno je napomenuti da nismo dobili primetno bolji model dodatnim podešavanjem hiper parametara, jer je model i sa podrazumevanim parametrima davao odlične rezultate.

2.2 K najbližih suseda

Još jedan od osnovnih algoritama klasifikacije predstavlja K najbližih suseda. Ukoliko podatke sa sličnim osobinama preslikamo na ravan, možemo očekivati da će se oni naći jedni pored drugih, što nam daje ideju za ovaj algoritam. Parametri koje možemo podešavati su metrika kojom merimo udaljenost izmedju instanci i broj suseda koji grupišemo.

Možemo da primetimo da je algoritam KNN sa podrazumevanim parametrima postigao bolje rezultate od algoritma stabla odlučivanja sa podrazumevanim parametrima, što smo mogli i da pretpostavimo. Na grafičkom prikazu primene algoritma SMOTENC može se utvrditi pravilost u raspodeli podataka, što objašnjava ovakve rezlutate.

Najbolje rezultate postižu algoritmi Slučajnih šuma i KNN sa optimizovanim parametrima ('n _neighbors': 10, 'p': 1, 'weights': 'uniform').

```
Classification report for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       1.00      1.00      1.00      1.00      1.00      1.00       911
          1       1.00      1.00      1.00      1.00      1.00      1.00       911
          2       1.00      1.00      1.00      1.00      1.00      1.00       911

avg / total       1.00      1.00      1.00      1.00      1.00      1.00      2733

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------------
     0    1    2
0  911    0    0
1    0  911    0
2    0    0  911
--------------------------------------------------------------------------------
Classification report for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.97      0.95      0.74      0.96      0.84      0.72       304
          1       0.57      0.57      0.98      0.57      0.75      0.54        14
          2       0.36      0.53      0.95      0.43      0.71      0.48        17

avg / total       0.93      0.91      0.76      0.92      0.83      0.70       335

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------------
     0  1   2
0  288  3  13
1    3  8   3
2    5  3   9
--------------------------------------------------------------------------------
```

Figure 2.1: DecisionTreeClassifier

```
Classification report for model GridSearchCV on training data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.96      0.96      0.98      0.96      0.97      0.94       911
          1       1.00      1.00      1.00      1.00      1.00      1.00       911
          2       0.95      0.96      0.98      0.96      0.97      0.94       911

avg / total       0.97      0.97      0.99      0.97      0.98      0.96      2733

--------------------------------------------------------------------------------
Confusion matrix for model GridSearchCV on training data
--------------------------------------------------------------------------------
     0    1    2
0  871    0   40
1    1  907    3
2   33    0  878
--------------------------------------------------------------------------------
Classification report for model GridSearchCV on test data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.98      0.91      0.77      0.94      0.84      0.72       304
          1       0.56      0.64      0.98      0.60      0.79      0.61        14
          2       0.32      0.65      0.93      0.43      0.77      0.58        17

avg / total       0.93      0.89      0.79      0.90      0.84      0.71       335

--------------------------------------------------------------------------------
Confusion matrix for model GridSearchCV on test data
--------------------------------------------------------------------------------
     0  1   2
0  278  4  22
1    4  9   1
2    3  3  11
--------------------------------------------------------------------------------
```

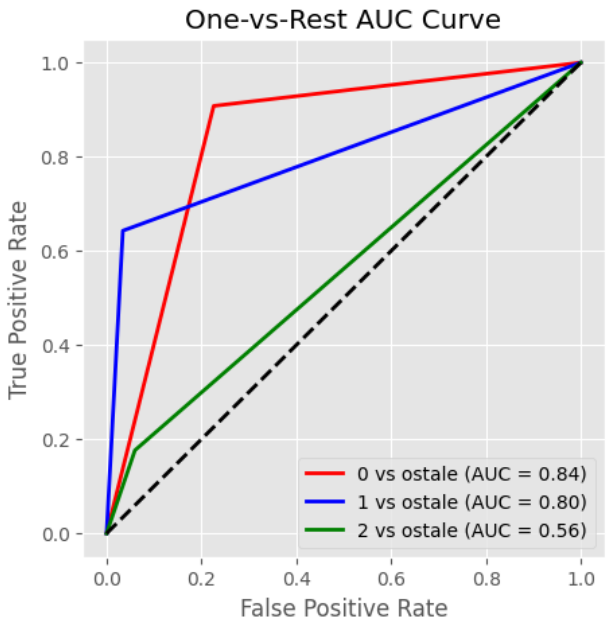
Figure 2.2: DecisionTreeClassifier sa optimizovanim hiper parametrima

```
Classification report for model RandomForestClassifier on training data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       1.00      1.00      1.00      1.00      1.00      1.00       911
          1       1.00      1.00      1.00      1.00      1.00      1.00       911
          2       1.00      1.00      1.00      1.00      1.00      1.00       911

avg / total       1.00      1.00      1.00      1.00      1.00      1.00      2733

--------------------------------------------------------------------------------
Confusion matrix for model RandomForestClassifier on training data
--------------------------------------------------------------------------------
     0    1    2
0  911    0    0
1    0  911    0
2    0    0  911
--------------------------------------------------------------------------------
Classification report for model RandomForestClassifier on test data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.98      0.96      0.77      0.97      0.86      0.75       304
          1       0.60      0.64      0.98      0.62      0.79      0.61        14
          2       0.32      0.41      0.95      0.36      0.63      0.37        17

avg / total       0.93      0.92      0.79      0.92      0.85      0.73       335

--------------------------------------------------------------------------------
Confusion matrix for model RandomForestClassifier on test data
--------------------------------------------------------------------------------
     0  1   2
0  291  2  11
1    1  9   4
2    6  4   7
--------------------------------------------------------------------------------
```

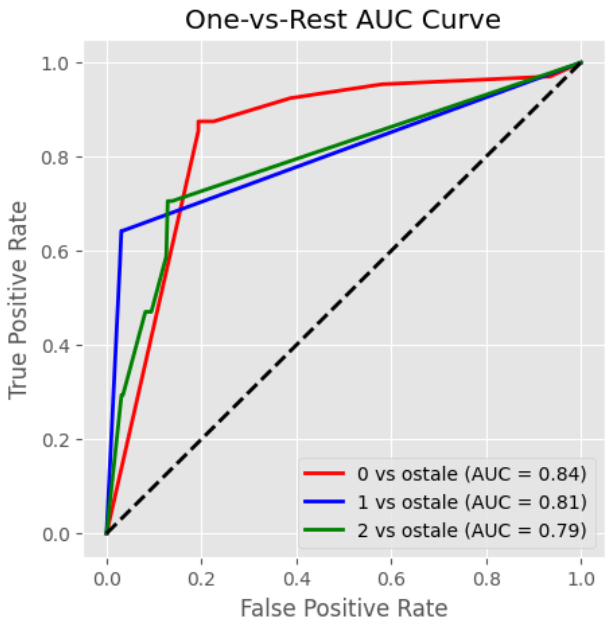
Figure 2.3: RandomForestClassifier

```
Classification report for model GridSearchCV on training data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       1.00      1.00      1.00      1.00      1.00      1.00       911
          1       1.00      1.00      1.00      1.00      1.00      1.00       911
          2       1.00      1.00      1.00      1.00      1.00      1.00       911

avg / total       1.00      1.00      1.00      1.00      1.00      1.00      2733

--------------------------------------------------------------------------------
Confusion matrix for model GridSearchCV on training data
--------------------------------------------------------------------------------
     0    1    2
0  909    0    2
1    0  911    0
2    0    0  911
--------------------------------------------------------------------------------
Classification report for model GridSearchCV on test data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.98      0.95      0.81      0.97      0.88      0.78       304
          1       0.64      0.64      0.98      0.64      0.80      0.61        14
          2       0.36      0.53      0.95      0.43      0.71      0.48        17

avg / total       0.93      0.92      0.82      0.93      0.87      0.76       335

--------------------------------------------------------------------------------
Confusion matrix for model GridSearchCV on test data
--------------------------------------------------------------------------------
     0  1   2
0  290  2  12
1    1  9   4
2    5  3   9
--------------------------------------------------------------------------------
```

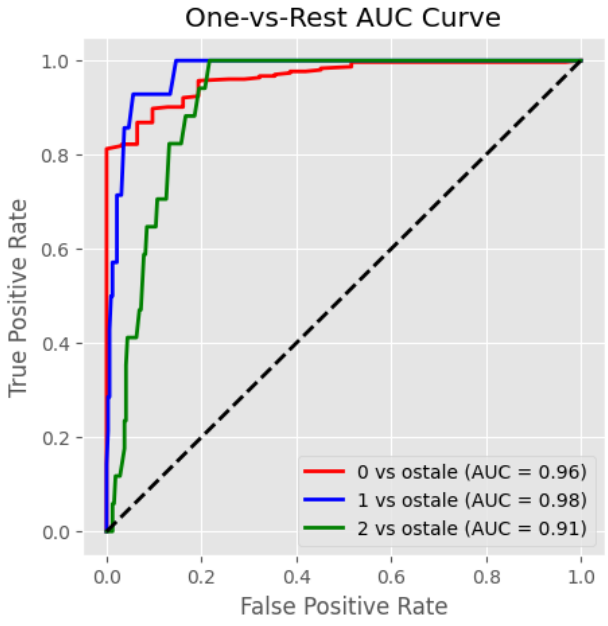
Figure 2.4: RandomForestClassifier sa optimizovanim hiper parametrima



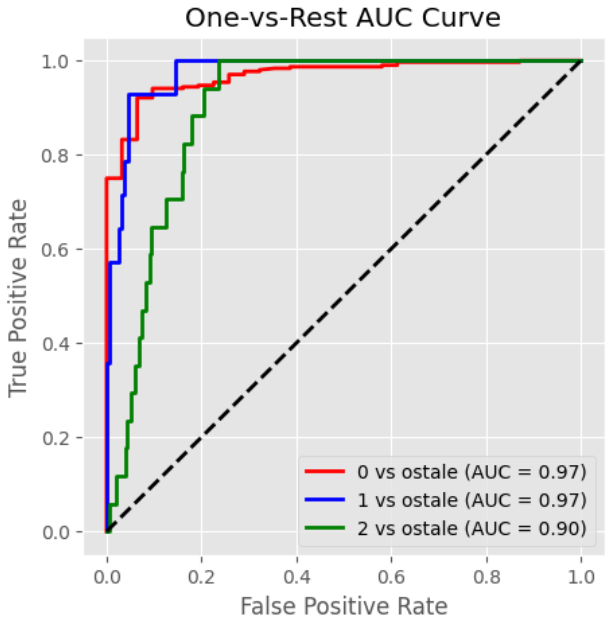
(a) DTC



(b) GridSearchCV nad DTC

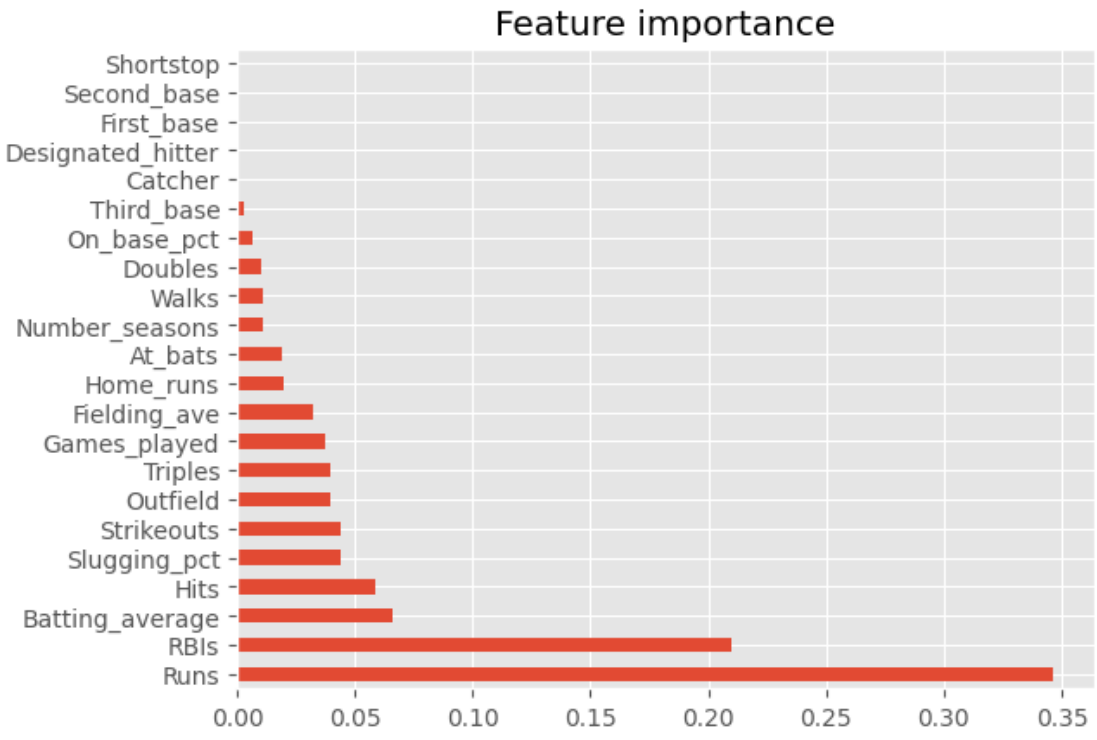


(c) RFC

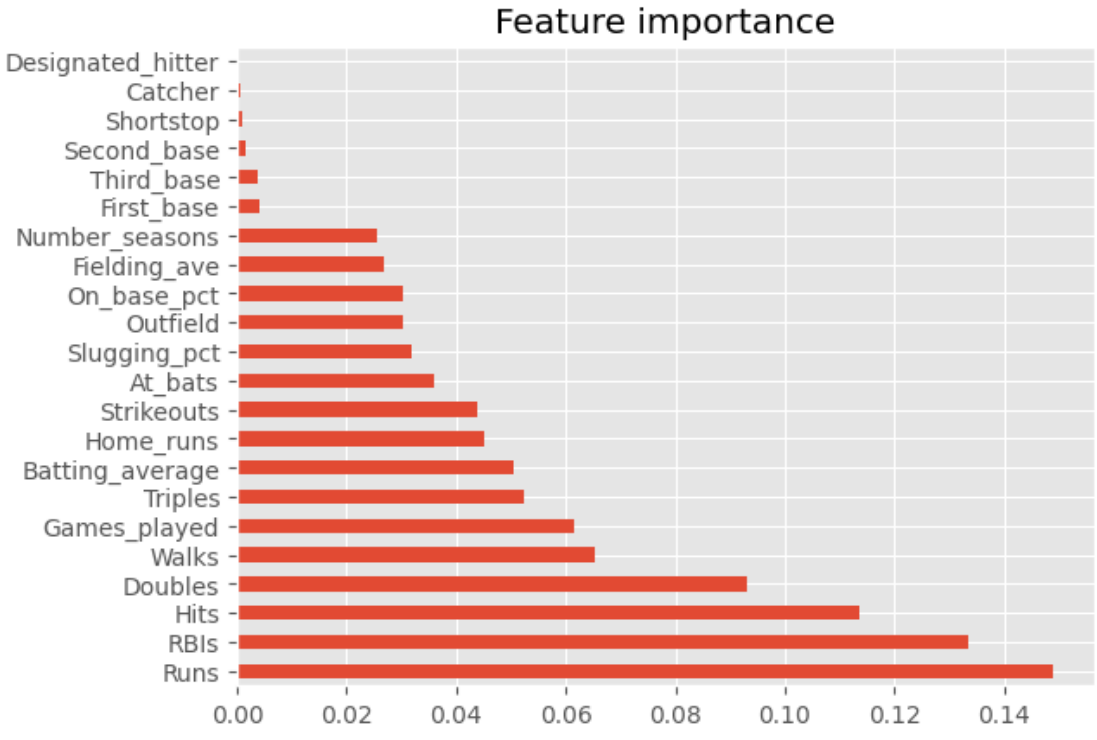


(d) GridSearchCv nad RFC

Figure 2.5: Prikaz ROC krive za modele stabla odlučivanja



(a) Stabla odlučivanja



(b) Slučajne šume

Figure 2.6: Značajnost atributa prilikom klasifikacije

```
Classification report for model KNeighborsClassifier on training data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       1.00      0.89      1.00      0.94      0.95      0.88       911
          1       0.99      1.00      0.99      0.99      1.00      0.99       911
          2       0.91      1.00      0.95      0.95      0.98      0.96       911

avg / total       0.97      0.96      0.98      0.96      0.97      0.94      2733

--------------------------------------------------------------------------------
Confusion matrix for model KNeighborsClassifier on training data
--------------------------------------------------------------------------------
     0    1    2
0  815   11   85
1    1  908    2
2    0    0  911
--------------------------------------------------------------------------------
Classification report for model KNeighborsClassifier on test data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.99      0.88      0.90      0.93      0.89      0.79       304
          1       0.39      0.64      0.96      0.49      0.78      0.60        14
          2       0.23      0.59      0.90      0.33      0.73      0.51        17

avg / total       0.93      0.85      0.91      0.88      0.88      0.77       335

--------------------------------------------------------------------------------
Confusion matrix for model KNeighborsClassifier on test data
--------------------------------------------------------------------------------
     0   1   2
0  266  10  28
1    0   9   5
2    3   4  10
--------------------------------------------------------------------------------
```

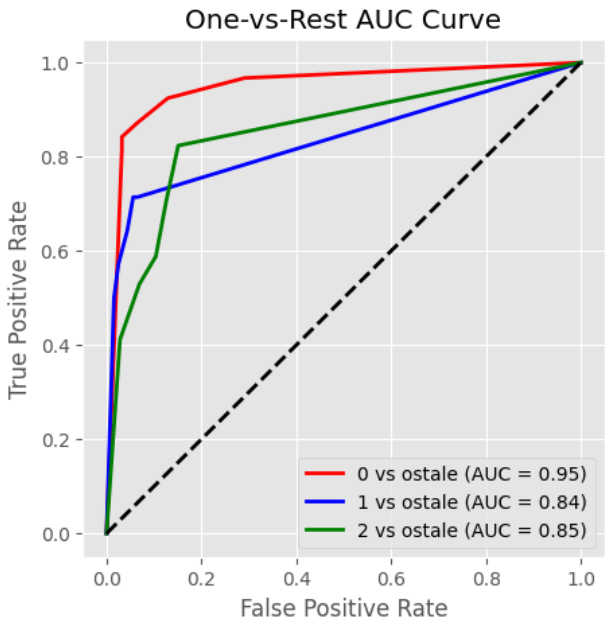
Figure 2.7: K nearest neighbours

```
Classification report for model GridSearchCV on training data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       1.00      0.87      1.00      0.93      0.93      0.86       911
          1       0.98      1.00      0.99      0.99      0.99      0.99       911
          2       0.89      1.00      0.94      0.94      0.97      0.95       911

avg / total       0.96      0.95      0.98      0.95      0.96      0.93      2733

--------------------------------------------------------------------------------
Confusion matrix for model GridSearchCV on training data
--------------------------------------------------------------------------------
     0    1    2
0  790   16  105
1    1  907    3
2    0    0  911
--------------------------------------------------------------------------------
Classification report for model GridSearchCV on test data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.99      0.88      0.94      0.94      0.91      0.82       304
          1       0.41      0.64      0.96      0.50      0.79      0.60        14
          2       0.26      0.65      0.90      0.37      0.76      0.57        17

avg / total       0.93      0.86      0.93      0.89      0.90      0.80       335

--------------------------------------------------------------------------------
Confusion matrix for model GridSearchCV on test data
--------------------------------------------------------------------------------
     0  1   2
0  269  9  26
1    0  9   5
2    2  4  11
--------------------------------------------------------------------------------
```

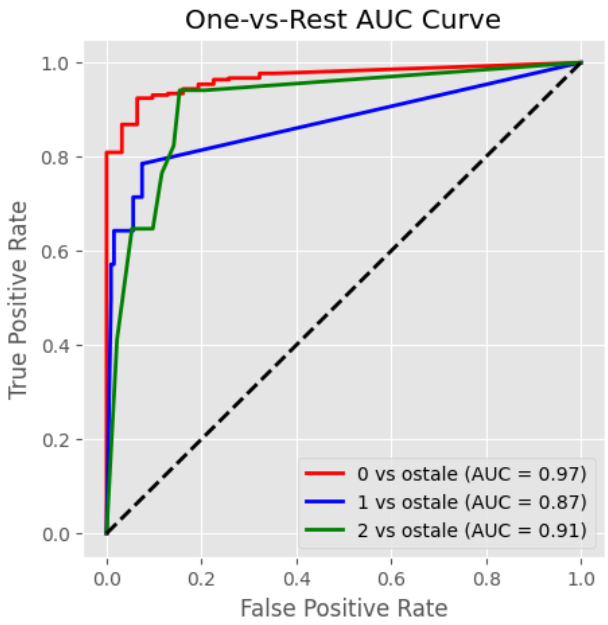
Figure 2.8: K nearest neighbours sa optimizovanim parametrima

```
Classification report for model BaggingClassifier on training data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       1.00      0.90      1.00      0.94      0.95      0.89       911
          1       0.99      1.00      0.99      0.99      1.00      0.99       911
          2       0.92      1.00      0.95      0.96      0.98      0.96       911

avg / total       0.97      0.96      0.98      0.96      0.97      0.95      2733

--------------------------------------------------------------------------------
Confusion matrix for model BaggingClassifier on training data
--------------------------------------------------------------------------------
     0    1    2
0  816   13   82
1    0  909    2
2    0    0  911
--------------------------------------------------------------------------------
Classification report for model BaggingClassifier on test data
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

          0       0.99      0.87      0.90      0.93      0.89      0.78       304
          1       0.36      0.64      0.95      0.46      0.78      0.59        14
          2       0.24      0.59      0.90      0.34      0.73      0.51        17

avg / total       0.92      0.85      0.90      0.88      0.87      0.76       335

--------------------------------------------------------------------------------
Confusion matrix for model BaggingClassifier on test data
--------------------------------------------------------------------------------
     0   1   2
0  265  12  27
1    0   9   5
2    3   4  10
--------------------------------------------------------------------------------
```

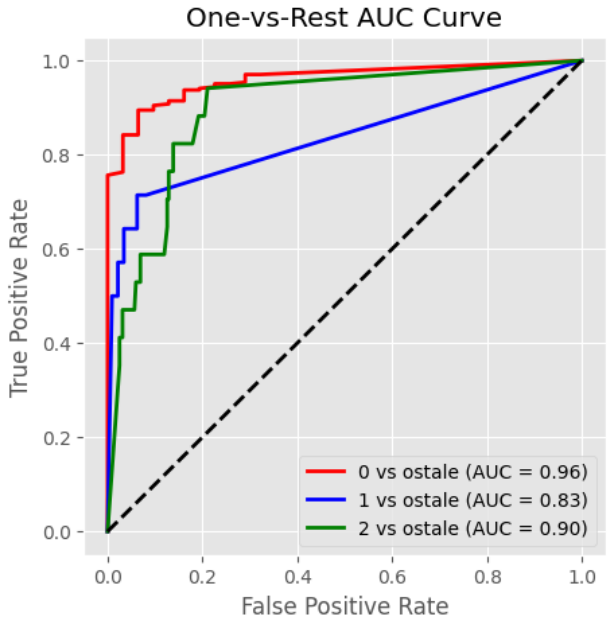
Figure 2.9: BaggingClassifier



(a) KNN



(b) GridSearchCV nad KNN



(c) BaggingClassifier

Figure 2.10: Prikaz ROC krive za modele k najbližih suseda

Klasterovanje

Problem identifikacije grupa u podacima na način da su instance jedne grupe (klastera) medjusobno u velikoj meri slični, a instance iz različitih klastera u velikoj meri jako različite nazivamo 'Klasterovanje'.

Klasterovanje pripada metodama nenadgledanog učenja, iz razloga što ne znamo koje instance treba da pripadaju odgovarajućem klasteru.

S'obzirom da se podaci mogu grupisati na više različitih načina, pojam klasterovanja nije jednoznačno odredjen.

3.1 Algoritam K-sredina

K-sredina (K-means) je jedan od osnovnih algoritama klasterovanja zasnovan na reprezentativnim predstavnicima. Reč je o iterativnom algoritmu koji deli podatke u K klastera, gde je K unapred definisan broj. Svaka tačka pripada tačno jednom klasteru. Broj K se bira nasumično, a potom se ponavljaju sledeći koraci:

- instance se grupišu u odgovarajuće klastere tako da svaka instanca pripada grupi kojoj je predstavnik najbliži centroid
- izračunavaju se novi centroidi kao prosek instanci koje su im pridružene

Koraci se ponavljaju dok centoride ne ostanu iste u dve uzastopne iteracije. Cilj algoritma je minimizacija zbira srednje-kvadratnih udaljenosti izmedju instacni i odgovarajućeg centroida.

Na osnovu 'Metode lakta' i vrednosti koeficijenta siluete potvrdjujemo prethodno donetu pretpostavku da je optimalan broj klastera $k = 2$.

3.2 Hijerarhijsko klasterovanje i DBSCAN

'Agglomerative clustering' je algoritam hijerarhijskog klasterovanja koji prati pristup odozdo naviše. Svaka instanca se na početku inicijalizuje u sopstveni klaster, a zatim se iterativnim postupkom klasteri spajaju na osnovu njihove sličnosti dok se ne ostvari željeni broj klastera.

Algoritam DBSCAN klastere pronalazi na osnovu gustine instanci, pri čemu se ne navodi željeni broj klastera. Prednost algoritma je što može naći klastere proizvoljnog oblika. Prikazan je rezlutat primene DBSSCAN algoritma za različite kombinacije vrednosti njegovih parametara:

- Eps - prag za rastojanja suseda. Dve instance su susedne ako je njihovo medjusobno rastojanje manje ili jednako Eps
- MinPts - prag za broj suseda instanci

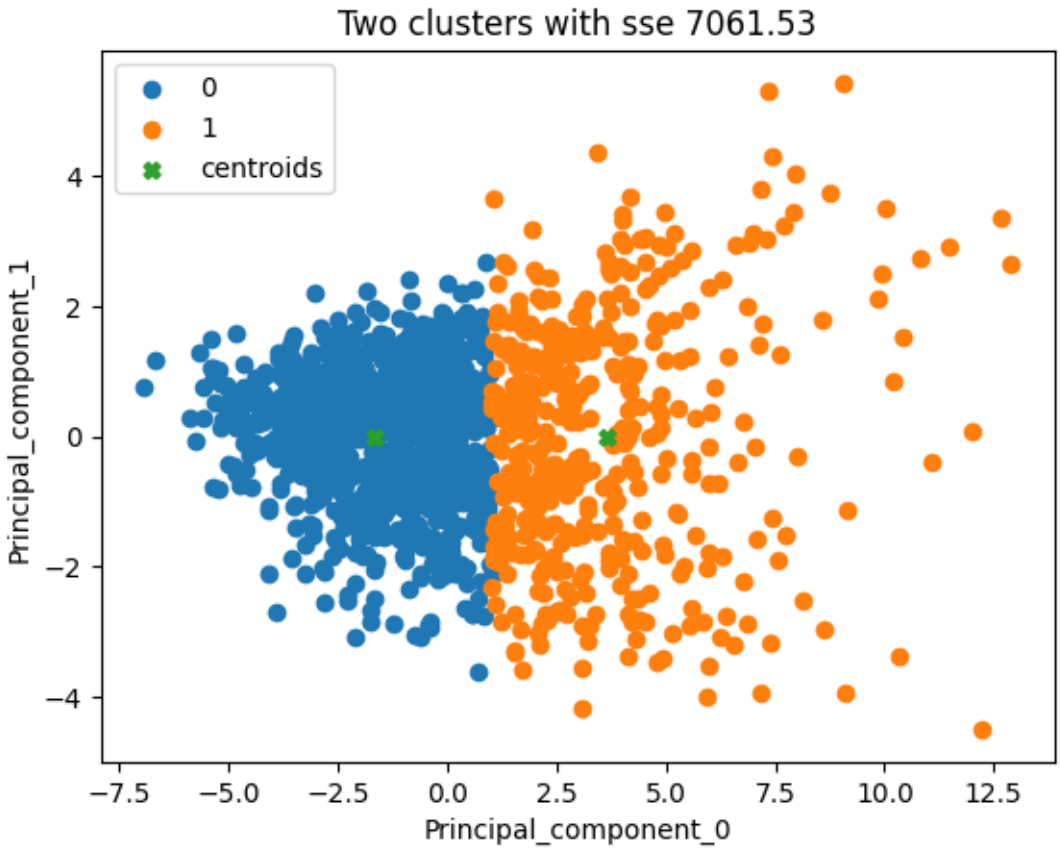


Figure 3.1: Kmeans za k = 2

Sa grafičkog prikaza primene hijerarhijskog klasterovanja kao i na osnovu vrednosti koeficijenta siluete, možemo se uveriti da hijerarhijsko klasterovanje postiže bolje rezultate nego k-means algoritam i bolje realizuje našu pretpostavku o potencijalnim klasterima.

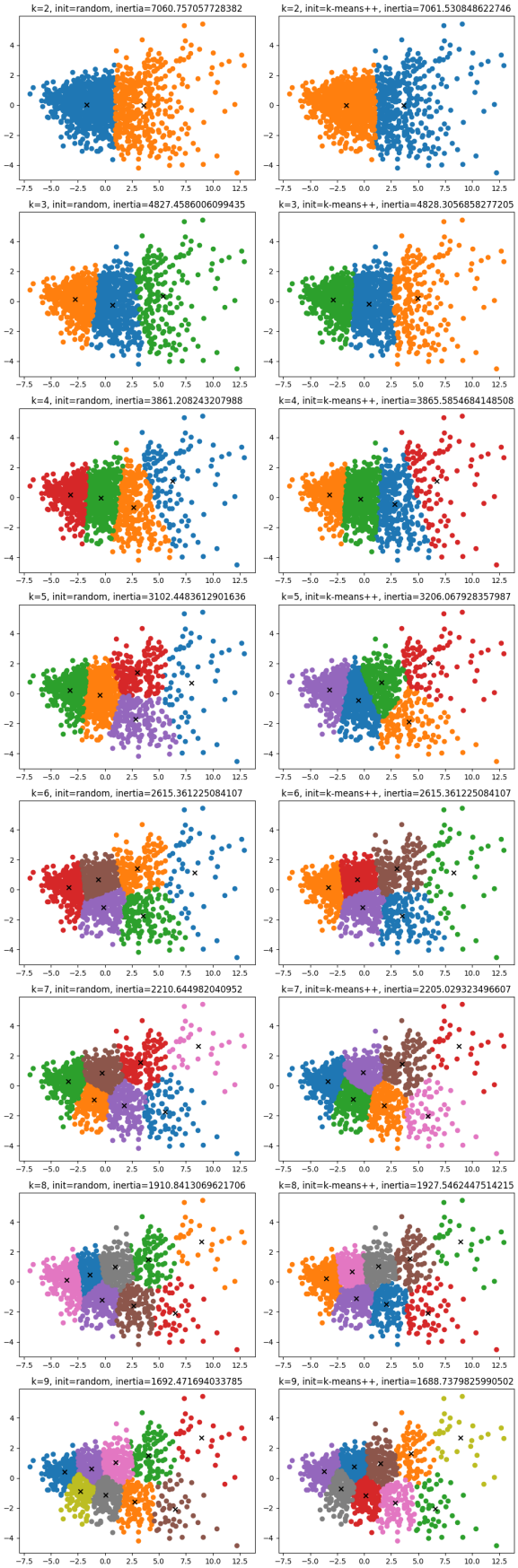


Figure 3.2: Prikaz rezlutata algoritma k means za različito k

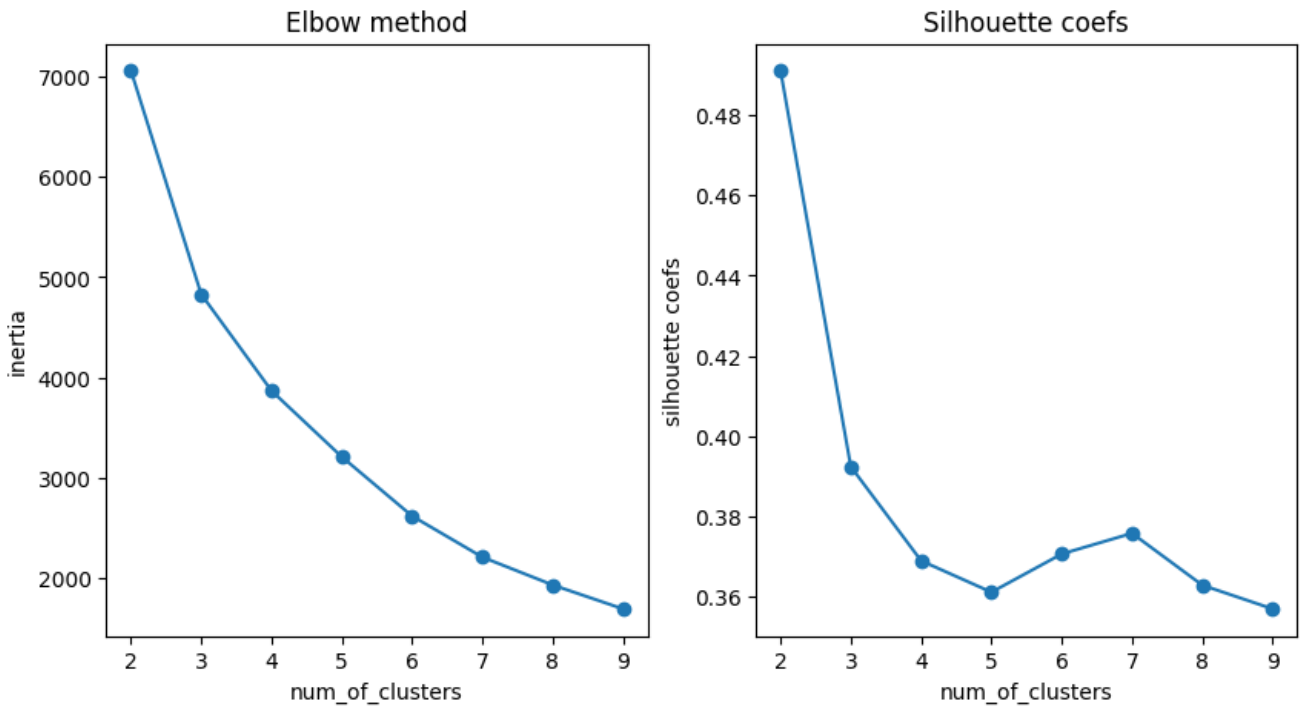


Figure 3.3: Elbow method i Silhouette coefs

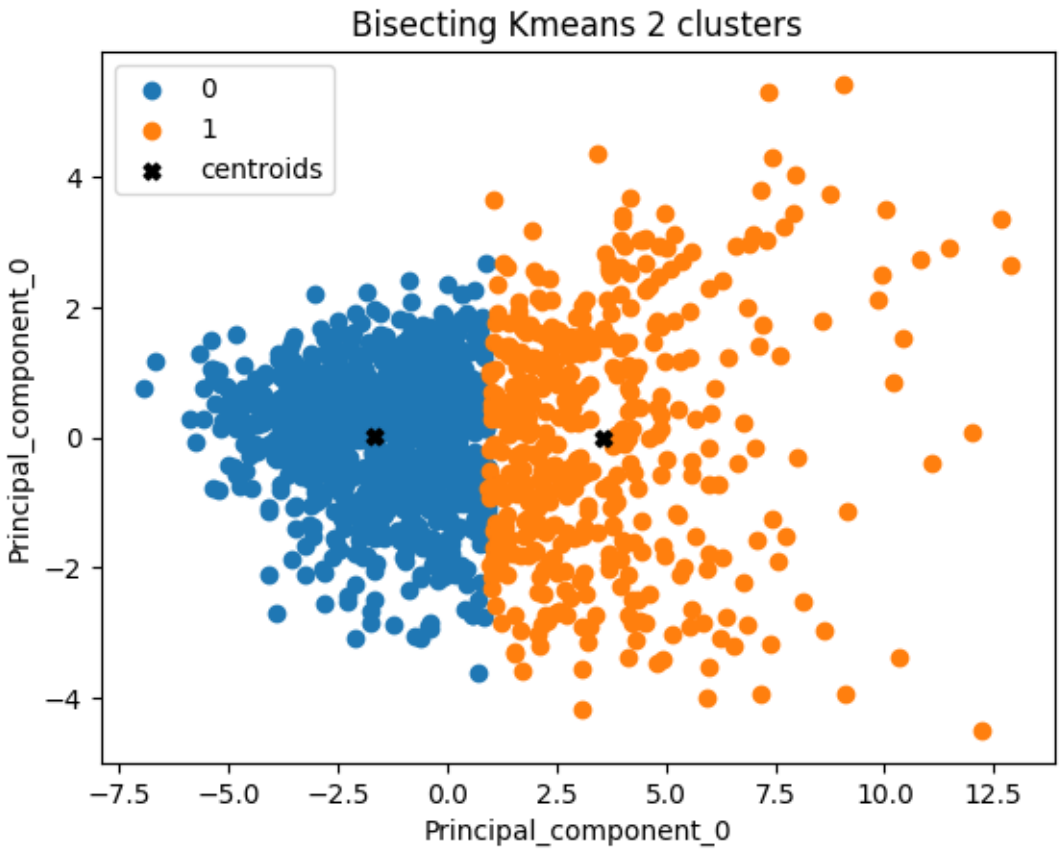


Figure 3.4: Bisecting K-means

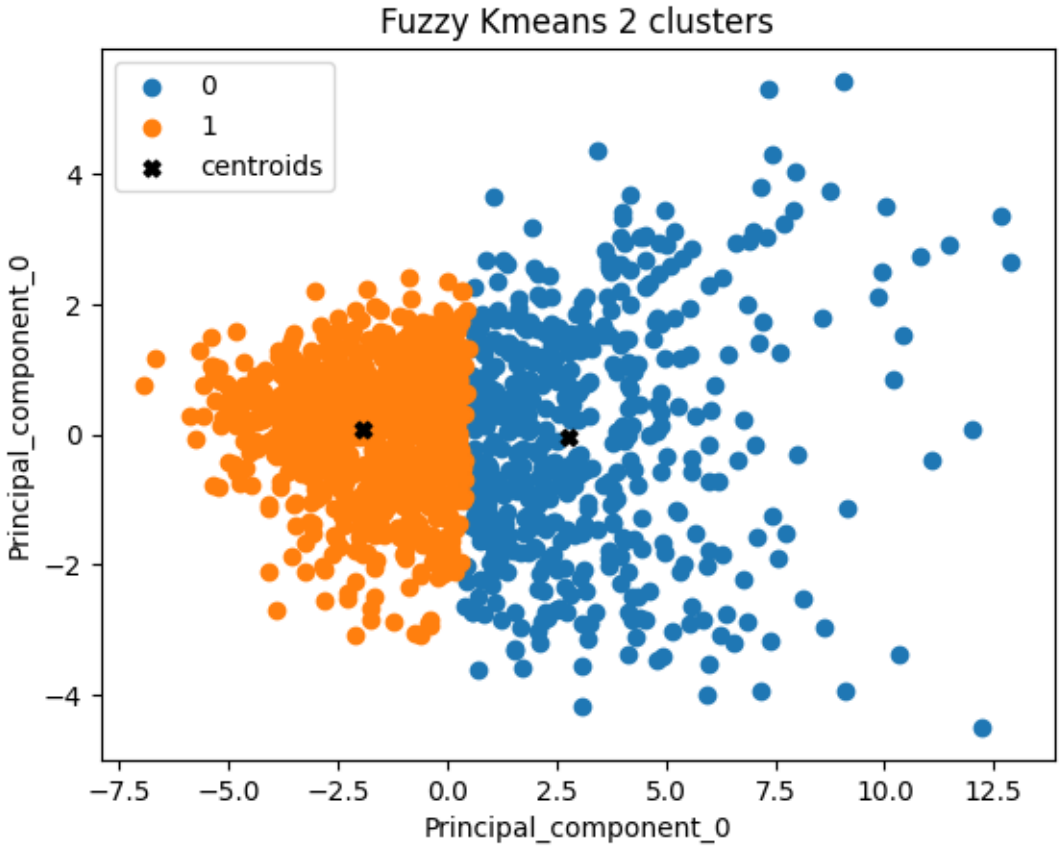


Figure 3.5: Fuzzy C-means

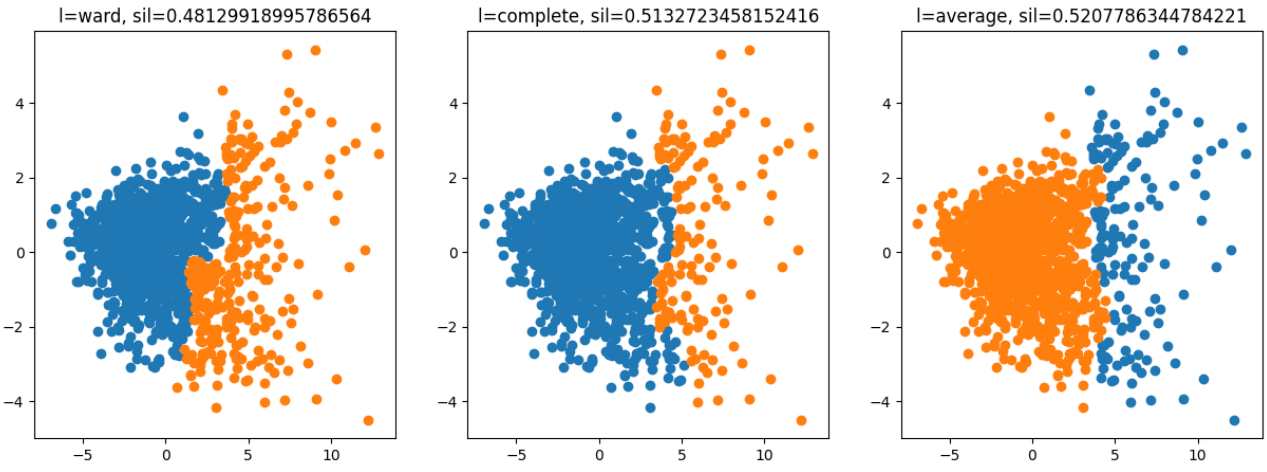


Figure 3.6: Agglomerative clustering

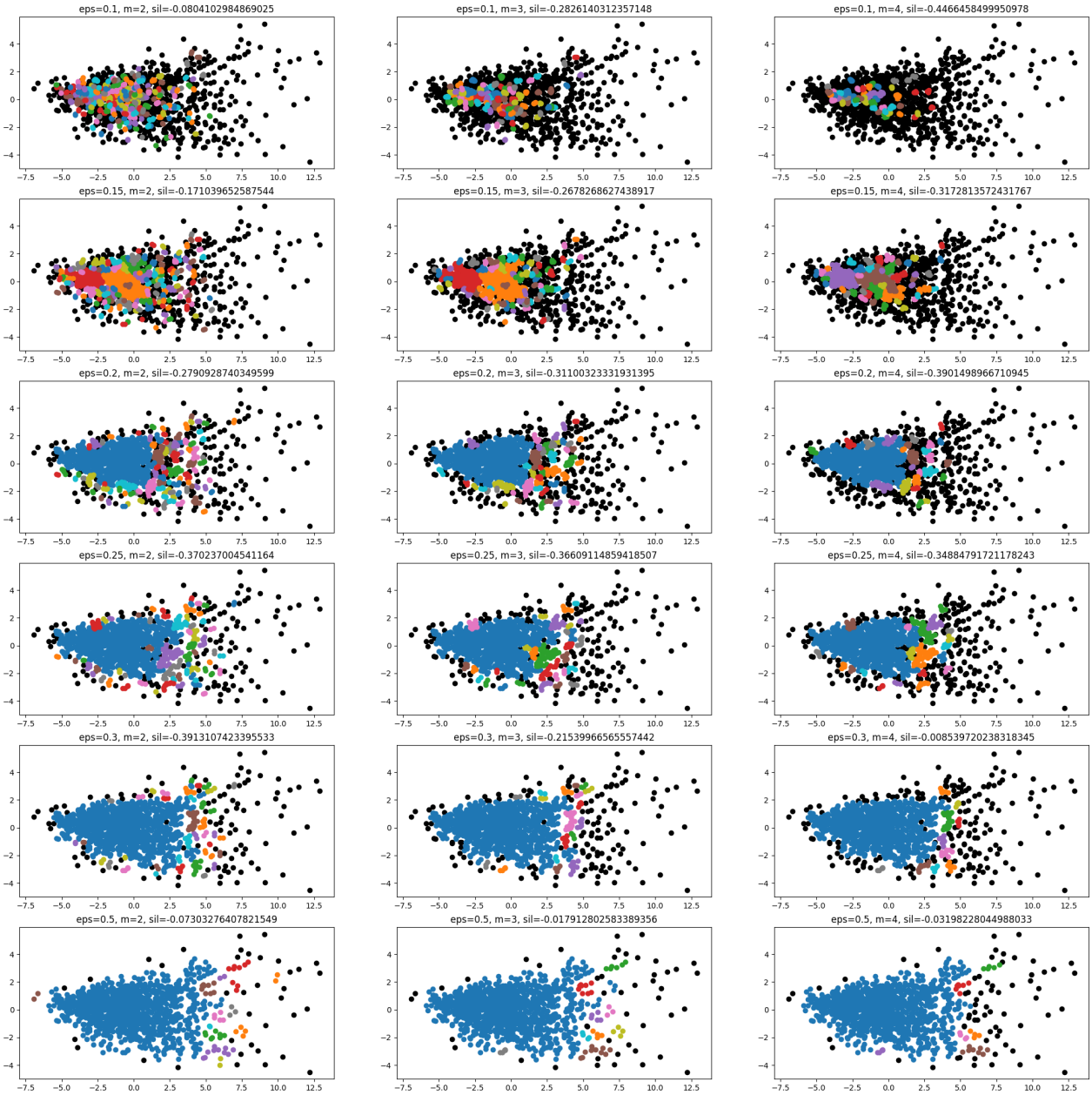


Figure 3.7: DBSCAN algoritam

Pravila pridruživanja

4.1 Čvor Association Rules u SPSS Modeleru

Za izdvajanje pravila pridruživanja u IBM SPSS Modeleru koristimo čvor Association Rules. Ovaj čvor ne radi sa podacima u transakcionom obliku i atributi u tabeli koja se koristi mogu biti različitih tipova. Svaka vrednost u kategoričkom atributu se posmatra kao jedna stavka, a nad numeričkim atributima se vrši diskretizacija i svaka nastala grupa se posmatra kao jedna stavka.

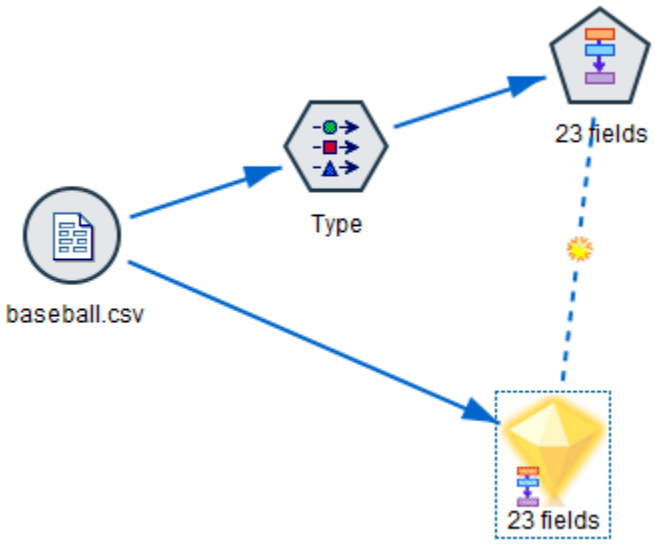


Figure 4.1: Association Rules u SPSS Modeleru

Most Interesting Rules by Confidence

Rank	Rule ID	Condition	Prediction	Other Evaluation Statistics				
				Sorted By Confidence(%)	Condition Support (%)	Rule Support (%)	Lift	Deployability (%)
1	1	Hits ≤ 889.600 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.90	20.90	2.04	0.00
2	2	Hits ≤ 889.600 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.90	20.90	2.04	0.00
3	3	Hits ≤ 889.600 Triples ≤ 61.800 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.82	20.82	2.04	0.00
4	4	Hits ≤ 889.600 Triples ≤ 61.800 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.82	20.82	2.04	0.00
5	5	Hits ≤ 889.600 Doubles ≤ 163.200 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.60	20.60	2.04	0.00
6	6	Hits ≤ 889.600 Doubles ≤ 163.200 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.60	20.60	2.04	0.00
7	7	Hits ≤ 889.600 Doubles ≤ 163.200 Triples ≤ 61.800 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.52	20.52	2.04	0.00
8	8	Hits ≤ 889.600 Doubles ≤ 163.200 Triples ≤ 61.800 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.52	20.52	2.04	0.00
9	9	Runs ≤ 465.200 Hits ≤ 889.600 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.45	20.45	2.04	0.00
10	10	Runs ≤ 465.200 Hits ≤ 889.600 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	100.00	20.45	20.45	2.04	0.00

Figure 4.2: Prvih 10 pravila sortiranih po pouzdanosti

Most Interesting Rules by Rule Support

					Other Evaluation Statistics			
Rank	Rule ID	Condition	Prediction	Sorted By Rule Support(%)	Condition Support (%)	Confidence (%)	Lift	Deployability (%)
1	1	Hits ≤ 889.600 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.90	20.90	100.00	2.04	0.00
2	2	Hits ≤ 889.600 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.90	20.90	100.00	2.04	0.00
3	3	Hits ≤ 889.600 Triples ≤ 61.800 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.82	20.82	100.00	2.04	0.00
4	4	Hits ≤ 889.600 Triples ≤ 61.800 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.82	20.82	100.00	2.04	0.00
5	5	Hits ≤ 889.600 Doubles ≤ 163.200 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.60	20.60	100.00	2.04	0.00
6	6	Hits ≤ 889.600 Doubles ≤ 163.200 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.60	20.60	100.00	2.04	0.00
7	7	Hits ≤ 889.600 Doubles ≤ 163.200 Triples ≤ 61.800 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.52	20.52	100.00	2.04	0.00
8	8	Hits ≤ 889.600 Doubles ≤ 163.200 Triples ≤ 61.800 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.52	20.52	100.00	2.04	0.00
9	9	Runs ≤ 465.200 Hits ≤ 889.600 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.45	20.45	100.00	2.04	0.00
10	10	Runs ≤ 465.200 Hits ≤ 889.600 Home_runs ≤ 151.000 0.299 ≤ Slugging_pct < 0.397	RBIs ≤ 476.200	20.45	20.45	100.00	2.04	0.00

Figure 4.3: Prvih 10 pravila sortiranih po podršci

Most Interesting Rules by Lift

Rank	Rule ID	Condition	Prediction	Sorted By Lift	Other Evaluation Statistics			
					Condition Support (%)	Confidence (%)	Rule Support (%)	Deployability (%)
1	180	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600	At_bats ≤ 3,012.200	3.87	14.85	100.00	14.85	0.00
2	181	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Triples ≤ 61.800	At_bats ≤ 3,012.200	3.87	14.85	100.00	14.85	0.00
3	182	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Home_runs ≤ 151.000	At_bats ≤ 3,012.200	3.87	14.85	100.00	14.85	0.00
4	183	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Walks ≤ 424.800	At_bats ≤ 3,012.200	3.87	14.85	100.00	14.85	0.00
5	189	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Triples ≤ 61.800 Home_runs ≤ 151.000	At_bats ≤ 3,012.200	3.87	14.85	100.00	14.85	0.00
6	190	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Triples ≤ 61.800 Walks ≤ 424.800	At_bats ≤ 3,012.200	3.87	14.85	100.00	14.85	0.00
7	191	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Home_runs ≤ 151.000 Walks ≤ 424.800	At_bats ≤ 3,012.200	3.87	14.85	100.00	14.85	0.00
8	207	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 RBIs ≤ 476.200	At_bats ≤ 3,012.200	3.87	14.78	100.00	14.78	0.00
9	208	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Triples ≤ 61.800 RBIs ≤ 476.200	At_bats ≤ 3,012.200	3.87	14.78	100.00	14.78	0.00
10	209	Number_seasons ≤ 13.200 Games_played ≤ 824.400 Hits ≤ 889.600 Home_runs ≤ 151.000 RBIs ≤ 476.200	At_bats ≤ 3,012.200	3.87	14.78	100.00	14.78	0.00

Figure 4.4: Prvih 10 pravila sortiranih po liftu

Zakljucak

Primenom opisanig algoritama 'Istraživanja podataka' nad analizirajućim skupom podataka iz predmetne baze bejzbol igrača, uočeno je da su najprecizniji rezultati postignuti pri rešavanju problema klasifikacije podataka, što potvrdjuje činjenicu da je i sam skup podataka predvidjen za isti. Svi algoritmi su ostvarili zadovoljavajuće performanse, s tim da je potredno izdvojiti sledeća dva:

- Algoritam slučajnih šuma
- Algoritam K najbližih suseda

Njihov uspeh, u predmetnoj analizi, leži u tehnikama i posebnim strukturama koje primenjuju.

U svetu u kojem je znanje moć, sposobnost pretvaranja sirovih podataka u informacije je od nreprocenjive vrednosti. Zato je važno prepoznati pravi značaj istraživanja podataka, a dalji razvoj računarske tehnologije ovaj pristup sve više uvodiće u svakodnevni život.