NBA_Rookies_Dataset

Projekat iz Istraživanja podataka 1 [R274]

Pavle Dušanić 287/2019

Septembar 2023

Sadržaj

1 Uvod

U ovom radu bavimo se analizom baze podataka "NBA_Rookies _Dataset" . Ova baza podataka sadrži statistike za 1340 igrača , bivše i trenutne, za njihovu prvu godinu igranja u američkoj nacionalnoj košarkaškoj ligi ("NBA"). Nad ovim podacima izvršićemo razne algoritme klasifikacije i klasterovanja radi grupisanja igrača koji imaju budućnost u ovoj ligi barem u narednih 5 godina, a i takođe ćemo koristiti softver SPSS radi dubljeg razumevanja podataka i nekih statističkih analiza. Predviđanja za košarku, kao i za bilo koji drugi sport, su teska da se precizno odrade. Postoji previse promenljivih koji se trenutno ne mogu kvantifikovati na pravi nacin da bi se dobio model koji moze da pravi realne i precizne predikcije svaki put.

2 Podaci sa kojim radimo

Podataka sa kojma radimo nema puno. Ova baza podataka ima 1340 instanci i 21 atribut.

Atributi naše baze:
1. Name : Imena igrača
2. GP : Broj odigranih igara
3. MIN : Prosek odigranih minuta
4. PTS: Prosecan broj poena
5. FGM: Broj uspesnih koseva (dvojki)
6. FGA: Broj pokusanih koseva (dvojki)
7. FG%: Odnos uspesnih i pokusanih koseva (dvojki)
8. 3P Made: Broj uspesnih koseva (trojki)
9. 3PA: Broj pokusanih koseva (trojki)
10. 3P%: Odnos uspesnih i pokusanih koseva (trojki)
11. FTM: Broj uspesnih slobodnih bacanja
12. FTA: Broj pokusaja slobodnih bacanja
13. FT%: Odnos uspesnih i pokusanih slobodnih bacanja
14. OREB: Broj ofanzivnih skokova
15. DREB: Broj defanzivnih skokova
16. REB: Ukupan broj skokova
17. AST: Broj asistencija
18. STL: Broj ukradenih lopti
19. BLK: Broj blokova
20. TOV: Broj preokreta
21. TARGET_5Yrs: Da li je igraču trajala karijera barem 5 godina

Skoro svi atributi su numerički, osim TARGET_5Yrs koji je FLAG i Name koji je ID.

	Description
Name	Name
GP	Games Played
MIN	MinutesPlayed
PTS	PointsPerGame
FGM	FieldGoalsMade
FGA	FieldGoalAttempts
FG%	FieldGoalPercent
3P Made	3PointMade
3PA	3PointAttempts
3P%	3PointAttempts
FTM	FreeThrowMade
FTA	FreeThrowAttempts
FT%	FreeThrowPercent
OREB	OffensiveRebounds
DREB	DefensiveRebounds
REB	Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
TARGET_5Yrs	Outcome: 1 if career length >= 5 yrs, 0 if < 5...

Tabela atributa baze podataka.

3 Priprema podataka i eksplorativna analiza

3.1 Nedostajuće vrednosti

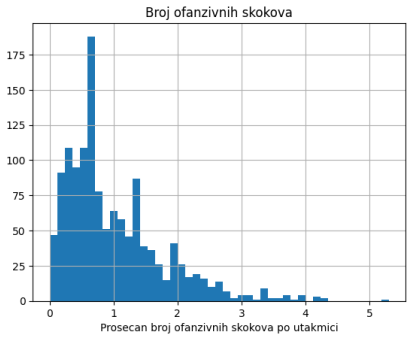
```
name           0
gp             0
min            0
pts            0
fgm            0
fga            0
fg             0
3p_made        0
3pa            0
3p            11
ftm            0
fta            0
ft             0
oreb           0
dreb           0
reb            0
ast            0
stl            0
blk            0
tov            0
target_5yrs    0
dtype: int64
```

Output komande data.isna().sum().

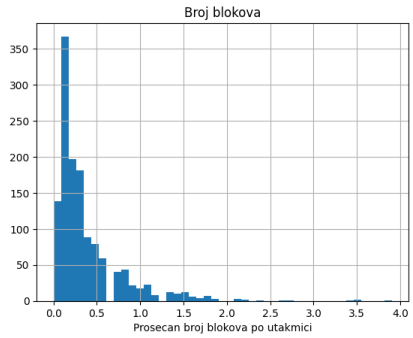
Jedine nedostajuce vrednosti u citavoj bazi se nalaze u koloni 3P. Te nedostajuce vrednosti su uglavnom posledica deljenja nulom posto postoje igraci koji nisu nijednom pokusali da daju trojku u svojoj "rookie"godini. Njih smo popunili nulama, komandom **data.fillna(0.0)** jer necemo da dropujemo čitave redove zbog manjka podataka.

3.2 Rad sa outlier-ima

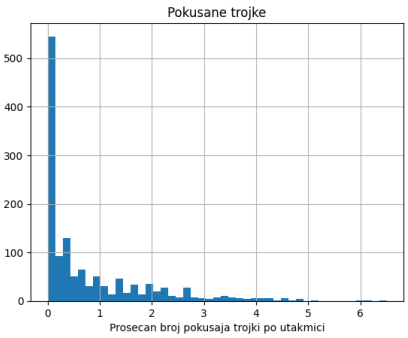
Primeri atributa sa outlier-ima



OREB



BLK



3PA

	name	gp	min	pts	fgm	fga	fg	3p_made	3pa	3p	...	fta	ft	oreb	dreb	reb	ast	stl	blk	tov	target_5yrs
45	Pete Chilcutt	69	11.8	3.6	1.6	3.6	45.2	0.0	0.0	100.0	...	0.4	82.1	1.1	1.6	2.7	0.6	0.5	0.3	0.6	1.0
609	Darius Miller	52	13.3	2.3	0.8	2.1	40.7	0.4	1.1	39.3	...	0.2	100.0	0.2	1.3	1.5	0.8	0.3	0.2	0.4	0.0
625	Jeremy Lamb	23	6.4	3.1	1.0	3.0	35.3	0.4	1.3	30.0	...	0.6	100.0	0.2	0.6	0.8	0.2	0.1	0.1	0.3	1.0
663	E'Twaun Moore	38	8.7	2.9	1.1	3.0	38.1	0.4	1.2	37.8	...	0.2	100.0	0.1	0.8	0.9	0.9	0.3	0.1	0.6	1.0
851	Steve Novak	35	5.5	1.5	0.5	1.4	36.0	0.4	1.1	33.3	...	0.1	100.0	0.1	0.6	0.7	0.2	0.1	0.0	0.1	1.0
1053	Mark Madsen	70	9.2	2.0	0.8	1.6	48.7	0.0	0.0	100.0	...	0.5	70.3	1.1	1.1	2.2	0.3	0.1	0.1	0.4	1.0
1108	Bryce Drew	34	13.0	3.5	1.4	3.8	36.4	0.5	1.4	32.7	...	0.2	100.0	0.1	0.9	0.9	1.5	0.3	0.1	0.9	1.0
1174	Erick Dampier	72	14.6	5.1	1.8	4.7	39.0	0.0	0.0	100.0	...	2.3	63.7	1.3	2.8	4.1	0.6	0.3	1.0	1.2	1.0
1253	Eric Mobley	46	12.8	3.9	1.7	2.9	59.1	0.0	0.0	100.0	...	1.0	48.9	1.2	2.1	3.3	0.5	0.2	0.6	0.5	0.0

Ekstremne vrednosti na atributima 3p i ft

	ftm	fta	ft
609	0.2	0.2	100.0
625	0.6	0.6	100.0
663	0.2	0.2	100.0
851	0.1	0.1	100.0
1108	0.2	0.2	100.0

Ekstremne ft vred

U većini slučajeva, skoro pola atributa ovih podataka sadrži neki statistički outlier. U zavisnosti od naše planirane upotrebe, ovi outlier-i se mogu izbaciti, a mogu i da se zadrže. Ako pravimo model koji će dobro da predviđa da li prosecan igrač ima buduću karijeru u ligi, onda slobodno mozemo da ih izbacimo, a pak ako želimo da napravimo precizan model za opšti slučaj u kome postoje igrači koji imaju karijeru samo zato što su odlični u nekom određenom aspektu igre, a prosečni ili čak ispod proseka u drugim, onda mozemo da ih zadržimo. Kako je prolazilo vreme, taktike u košarci su se drastično menjale i samim time potražnja za određene tipove igrača se menjala. U današnjem vremenu se uglavno traze visoki igrači i oni koji umeju da šutiraju trojke, za razliku od prethodnih era gde je snaga bila ključni faktor zbog agresivnije igre. Takođe jedno od problema je mala količina podataka i nedostatak pojedinih vaznih atributa igrača. Pošto radimo samo sa ovim podacima, uglavnom nam je bolje da zadržimo podatke pošto nema prevelikih outlier-a. Za ekstremne vrednosti na atributu 3p gde lazno stoji 100% za igrace koji nisu sutirali trojke, smo stavili 0%, a za atribut ft smo ostavili posto je tacan.

3.3 Rad sa duplikatima

	name	gp	min	pts	fgm	fga	fg	3p_made	3pa	3p	...	fta	ft	oreb	dreb	reb	ast	stl	blk	tov	target_5yrs
18	Larry Johnson	82	37.2	19.2	7.5	15.3	49.0	0.1	0.3	22.7	...	5.0	82.9	3.9	7.0	11.0	3.6	1.0	0.6	1.9	0.0
19	Larry Johnson	82	37.2	19.2	7.5	15.3	49.0	0.1	0.3	22.7	...	5.0	82.9	3.9	7.0	11.0	3.6	1.0	0.6	1.9	1.0
71	Dee Brown	82	23.7	8.7	3.5	7.5	46.4	0.1	0.4	20.6	...	1.9	87.3	0.5	1.7	2.2	4.2	1.0	0.2	1.7	0.0
72	Dee Brown	82	23.7	8.7	3.5	7.5	46.4	0.1	0.4	20.6	...	1.9	87.3	0.5	1.7	2.2	4.2	1.0	0.2	1.7	1.0
73	Dee Brown	49	9.2	1.9	0.7	2.0	32.7	0.1	0.6	21.4	...	0.8	64.9	0.2	0.7	0.8	1.7	0.5	0.1	0.6	0.0
74	Dee Brown	49	9.2	1.9	0.7	2.0	32.7	0.1	0.6	21.4	...	0.8	64.9	0.2	0.7	0.8	1.7	0.5	0.1	0.6	1.0
120	Tim Hardaway	79	33.7	14.7	5.9	12.5	47.1	0.3	1.1	27.4	...	3.5	76.4	0.7	3.2	3.9	8.7	2.1	0.1	3.3	0.0
121	Tim Hardaway	79	33.7	14.7	5.9	12.5	47.1	0.3	1.1	27.4	...	3.5	76.4	0.7	3.2	3.9	8.7	2.1	0.1	3.3	1.0
126	Glen Rice	77	30.0	13.6	6.1	13.9	43.9	0.2	0.9	24.6	...	1.6	73.4	1.3	3.3	4.6	1.8	0.9	0.3	1.5	1.0
127	Glen Rice	77	30.0	13.6	6.1	13.9	43.9	0.2	0.9	24.6	...	1.6	73.4	1.3	3.3	4.6	1.8	0.9	0.3	1.5	0.0
156	Michael Smith	65	9.5	5.0	2.1	4.4	47.6	0.0	0.4	7.1	...	1.0	82.8	0.6	0.9	1.5	1.2	0.1	0.0	0.8	0.0
157	Michael Smith	65	9.5	5.0	2.1	4.4	47.6	0.0	0.4	7.1	...	1.0	82.8	0.6	0.9	1.5	1.2	0.1	0.0	0.8	1.0
158	Michael Smith	82	21.2	6.9	2.7	5.0	54.2	0.0	0.0	0.0	...	3.2	48.5	2.1	3.8	5.9	0.8	0.7	0.6	1.3	0.0
159	Michael Smith	82	21.2	6.9	2.7	5.0	54.2	0.0	0.0	0.0	...	3.2	48.5	2.1	3.8	5.9	0.8	0.7	0.6	1.3	1.0
162	Charles Smith	60	8.7	2.9	1.0	2.2	44.4	0.0	0.1	0.0	...	1.3	69.7	0.2	0.9	1.2	1.7	0.6	0.1	0.6	1.0
164	Charles Smith	60	8.7	2.9	1.0	2.2	44.4	0.0	0.1	0.0	...	1.3	69.7	0.2	0.9	1.2	1.7	0.6	0.1	0.6	0.0
165	Charles Smith	71	30.4	16.3	6.1	12.4	49.5	0.0	0.0	0.0	...	5.5	72.5	2.4	4.1	6.5	1.5	1.0	1.3	2.1	1.0
167	Charles Smith	71	30.4	16.3	6.1	12.4	49.5	0.0	0.0	0.0	...	5.5	72.5	2.4	4.1	6.5	1.5	1.0	1.3	2.1	0.0
168	Charles Smith	34	8.6	3.5	1.4	3.7	39.2	0.4	1.4	31.9	...	0.3	54.5	0.4	0.4	0.8	0.6	0.3	0.2	0.8	1.0
170	Charles Smith	34	8.6	3.5	1.4	3.7	39.2	0.4	1.4	31.9	...	0.3	54.5	0.4	0.4	0.8	0.6	0.3	0.2	0.8	0.0
223	Mark Davis	33	7.8	3.8	1.5	3.1	48.0	0.0	0.3	10.0	...	1.0	82.4	0.5	0.6	1.1	0.4	0.4	0.1	0.4	1.0
224	Mark Davis	33	7.8	3.8	1.5	3.1	48.0	0.0	0.3	10.0	...	1.0	82.4	0.5	0.6	1.1	0.4	0.4	0.1	0.4	0.0
225	Mark Davis	57	10.0	3.3	1.0	2.6	36.9	0.1	0.2	30.8	...	2.0	63.8	1.0	1.2	2.2	0.8	0.7	0.4	1.2	1.0
226	Mark Davis	57	10.0	3.3	1.0	2.6	36.9	0.1	0.2	30.8	...	2.0	63.8	1.0	1.2	2.2	0.8	0.7	0.4	1.2	0.0
364	Charles Jones	78	20.1	8.4	3.0	5.8	52.0	0.0	0.1	0.0	...	3.6	64.8	1.8	3.3	5.1	1.6	0.6	0.8	1.8	1.0
365	Charles Jones	78	20.1	8.4	3.0	5.8	52.0	0.0	0.1	0.0	...	3.6	64.8	1.8	3.3	5.1	1.6	0.6	0.8	1.8	0.0
367	Charles Jones	29	16.4	3.7	1.3	4.2	31.7	0.7	2.1	31.1	...	0.8	50.0	0.3	1.1	1.4	1.4	0.6	0.2	1.0	1.0
368	Charles Jones	29	16.4	3.7	1.3	4.2	31.7	0.7	2.1	31.1	...	0.8	50.0	0.3	1.1	1.4	1.4	0.6	0.2	1.0	0.0
472	Walker Russell	68	11.1	2.7	1.0	2.7	36.4	0.0	0.3	11.1	...	0.9	81.0	0.3	0.8	1.1	1.9	0.2	0.0	1.4	0.0
473	Walker Russell	68	11.1	2.7	1.0	2.7	36.4	0.0	0.3	11.1	...	0.9	81.0	0.3	0.8	1.1	1.9	0.2	0.0	1.4	1.0
553	Larry Drew	76	20.8	6.6	2.6	6.4	40.7	0.1	0.2	23.5	...	1.8	79.7	0.3	1.3	1.6	3.3	1.2	0.1	2.2	1.0
554	Larry Drew	76	20.8	6.6	2.6	6.4	40.7	0.1	0.2	23.5	...	1.8	79.7	0.3	1.3	1.6	3.3	1.2	0.1	2.2	0.0
848	Bobby Jones	44	7.6	2.5	1.0	2.1	46.2	0.0	0.2	11.1	...	0.9	56.1	0.5	0.8	1.3	0.4	0.3	0.0	0.4	1.0
849	Bobby Jones	44	7.6	2.5	1.0	2.1	46.2	0.0	0.2	11.1	...	0.9	56.1	0.5	0.8	1.3	0.4	0.3	0.0	0.4	0.0
870	David Lee	67	16.9	5.1	2.0	3.4	59.6	0.0	0.0	0.0	...	1.8	57.7	1.6	2.9	4.5	0.6	0.5	0.3	0.8	1.0
871	David Lee	67	16.9	5.1	2.0	3.4	59.6	0.0	0.0	0.0	...	1.8	57.7	1.6	2.9	4.5	0.6	0.5	0.3	0.8	0.0
1126	Cedric Henderson	82	30.8	10.1	4.2	8.8	48.0	0.0	0.0	0.0	...	2.3	71.6	0.9	3.1	4.0	2.0	1.2	0.6	2.0	1.0

Duplikati čija je jedina razlika u ciljnoj klasi.

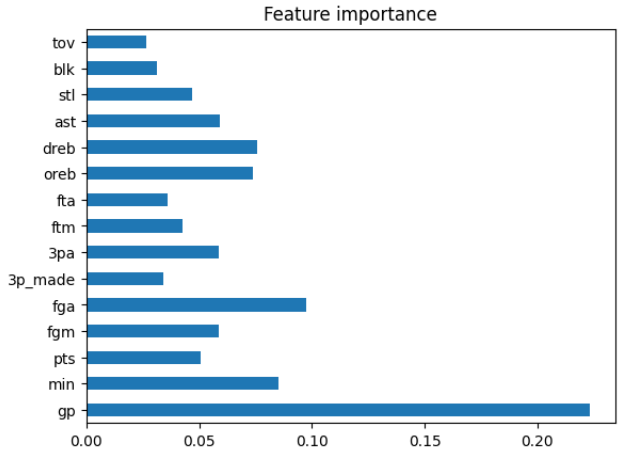
U podacima se nalaze duplikati ciji su svi atributi isti osim atributa **"target_5yrs"** . U Pandas-u postoji ugradjena metoda unutar klase DataFrame koja nalazi duplikate unutar jednog datafrejm-a i vraca ih. Kao što vidimo, vratio nam je i igrače sa istim imenom i prezimenom ali za njih vidimo da su zapravo druge osobe zbog njihovih statistika, a nema smisla da budu iste osobe pošto je ovo tabela prve godine igranja u NBA-ju za te igrače.

4 Klasifikacija

Za algoritme klasifikacije smo izabrali Stablo odlucivanja (DecisionTreeClassifier), ansambl stabla odlucivanja (RandomForest) i neuronsku mrezu (MLPClassifier). Unutar samih fajlova je dodat i jedna model zvani AutoGluon koji automatski vrsi svo predprocesiranje nad podacima i radi ansambl raznih algoritama da vidi koji ima najbolje rezultate. On je dodat kao provera, posto prethodno pomenutim algoritmima se ne postizu ohrabrujuci rezultati. Podatke smo podelili na trening i test preko funkcije train_test_split importovane iz sklearn.model_selection. Zato sto nam je ciljni atribut target_5years, ovo se svodi na problem binarne klasifikacije.

4.1 Stablo odlučivanja

Koristeći GridSearchCV iz sklearn.model_selection dobijamo sledeći skup parametara:
DecisionTreeClassifier(criterion='gini', max_depth=4) **Accuracy pre: 0.6062 Accuracy posle: 0.6815**



Feature importance za Decision Tree Classifier.

4.2 RandomForest

Kombinuje više stabala odlučivanja. Svako drvo se trenira na slučajno izabranom podskupu podataka, a klasifikacija se obavlja na osnovu glasanja svakog drveta. Random Forest ima sposobnost smanjenja preprilagođavanja i poboljšanja sposobnosti rada nad neviđenim podacima, u odnosu na pojedinačno stablo odlučivanja.

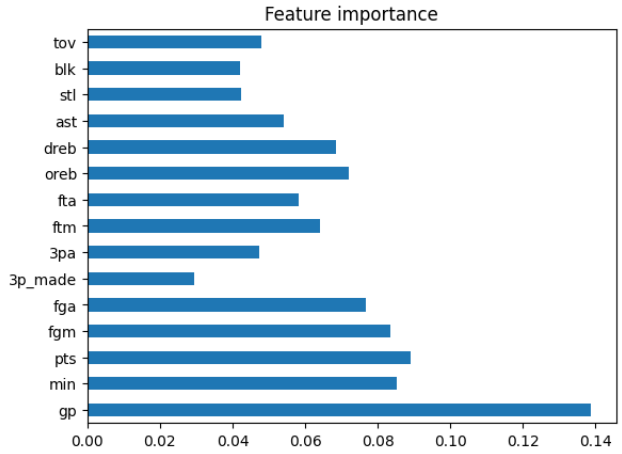
Koristeći GridSearchCV dobijamo:
RandomForestClassifier(max_depth=5, criterion= gini, n_estimators=400)
Accuracy pre: 0.7054 Accuracy posle: 0.7209

48	49
27	134

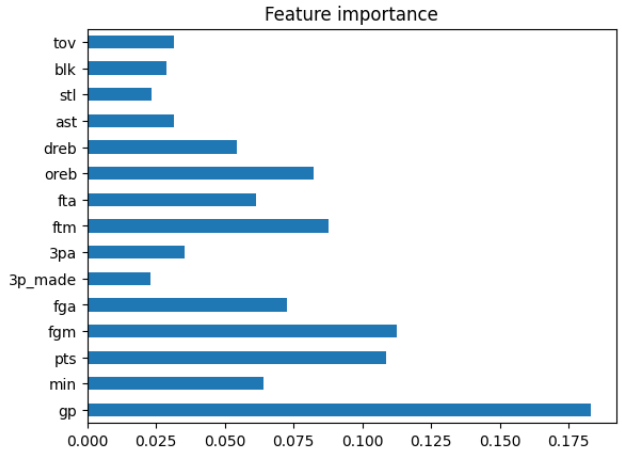
Matrica konfuzije za Random Forest pre GridSearch-a.

61	36
36	125

Matrica konfuzije za Random Forest posle GridSearch-a.



Feature importance za Random Forest Classifier pre GridSearch-a.



Feature importance za Random Forest Classifier posle GridSearch-a.

4.3 Neuronske mreze

Multi-layer Perceptron (MLP) je algoritam nadgledanog ucenja koji uci funkciju koja slika realne brojeve iz dimenzije m u realne brojeve dimenzije n, gde je m dimenzija ulaza, a n dimenzija izlaza. Sa datim atributima i ciljnim atributom y, on moze da radi klasifikaciju ili regresiju. Postoje skriveni slojevi izmedju ulaznog i izlaznog sloja koji ne moraju da budu nuzno linearne funkcije. Prednost mu je to sto moze da nauci linearne i ne linearne modele i pogodan je za velike kolicine podataka ali posto nasa baza ne sadrzi puno podataka, ne mozemo da iskoristimo ovaj algoritam maksimalno.

Koristeći GridSearchCV dobijamo:
MLPClassifier(learning_rate_init=0.001,learning_rate='adaptive',max_iter = 400,hidden_layer_sizes=200, tol=0.0004,activation='logistic',early_stopping='false') **Accuracy pre: 0.6899 Accuracy posle: 0.6744**

50	47
37	124

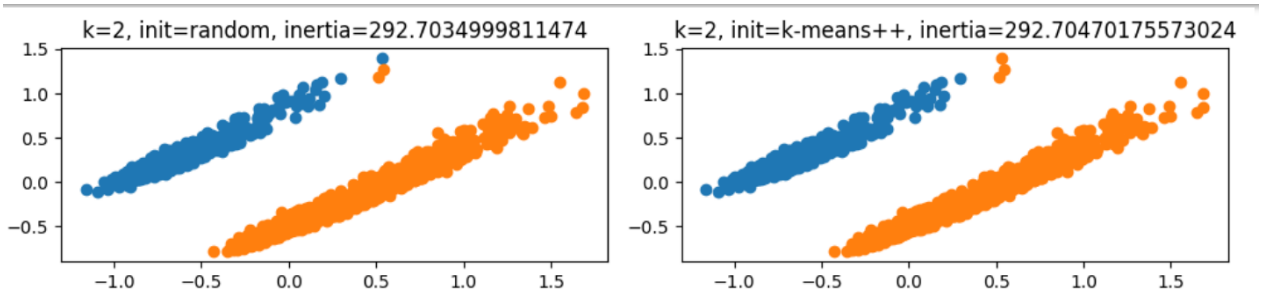
Matrica konfuzije za MLPClassifier.

5 Klasterovanje

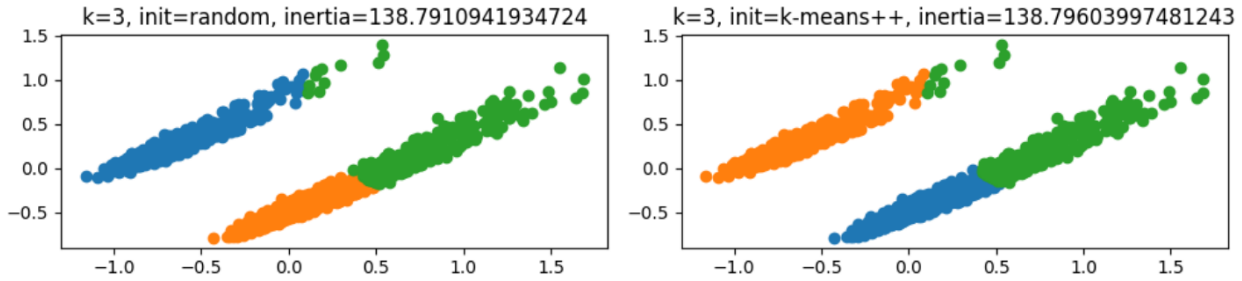
Klasterovanje se moze definisati kao problem prepoznavanje odredjenih grupa podataka koje su dovoljno slicne unutar jedne grupa, a istovremeno i dovoljno razlicite od ostalih. Za algoritme klasterovanja smo izabrali KMeans algoritam i DBSCAN algoritam. Nad podacima je izvrsen PCA algoritam da bi se smanjila dimenzionalnost podataka na dve dimenzija za prikazivanje na grafikonima, posto nasi podaci imaju veliki broj atributa. Koriscen je i MinMaxScaler za normalizaciju podataka.

5.1 KMeans

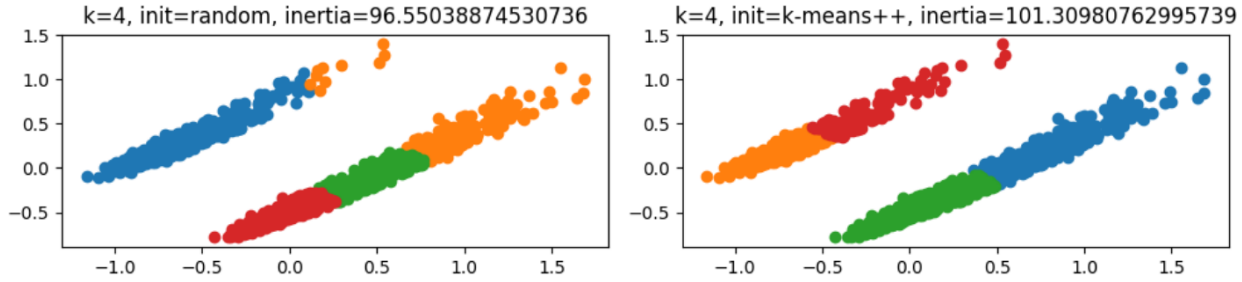
Algoritam K sredina. Za inicijalizaciju su izabrane random i k-means++ metode. K-means++ koristi statisticke metode koje su zasnovane na doprinosu svake tacke ukupnoj inerciji da bi izabrao pocetne centroide, dok random bira nasumicne centroide. Algoritam je radjen i za vise od cetiri klastera, ali je prikazano samo prvih par posto je najveci pad inercije upravo na ovom manjem broju klastera.



KMeans za k=2.

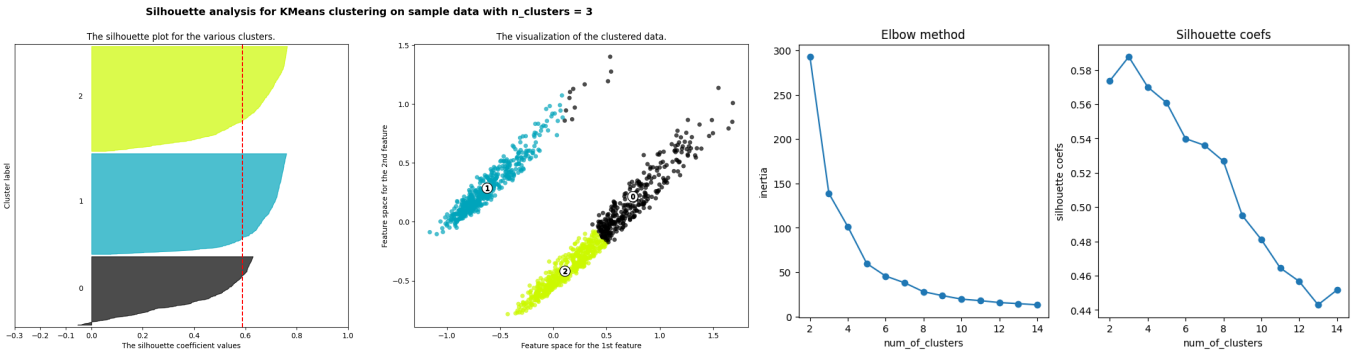


KMeans za k=3.



KMeans za k=4.

Pomocu metode lakta i koeficijenata siluete, zakljucujemo da je najbolji izbor za n=3 klastera cak iako nam izgleda da je najbolji izbor za n=2 klastera kada pogledamo vizualizaciju podataka.



Grafikon siluete za n=3 i vizualizacija podataka **Metod lakta i koeficijenti siluete**

5.2 DBSCAN - Density Based Spatial Clustering of Applications with Noise

Za razliku od prethodno navedenog algoritma K sredina, za ovaj algoritam se ne trazi zadavanje broja klastera kao parametra, vec se samo trazi epsilon okolina jednog klastera kao parametar. Ovaj algoritam je zasnovan na gustini podataka, u smislu da grupise podatke koji su vrlo gusto spakovani jedni uz druge u jedan klaster a vrlo bliske podatke koji su u redjem regionu i ne spadaju ni u jedan drugi obliznji klaster da oznaci kao sum. Algoritam je pusten da radi i nad obicnim visoko dimenzionalnim podacima , a i na redukovanim podacima PCA algoritmom.

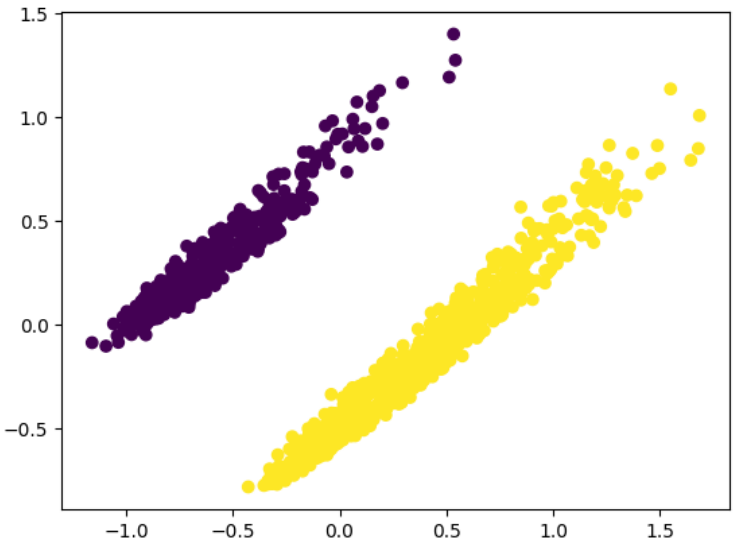
```
Estimated number of clusters: 2
Estimated number of noise points: 54
```

Sum i broj klastera pre PCA algoritma.

```
Estimated number of clusters: 2
Estimated number of noise points: 0
```

Sum i broj klastera posle PCA algoritma

U sklearn-ovoj implementaciji DBSCAN-a, automatski se stavljaju labele na sve tacke. To nam pomaze da vidimo koliko tacaka algoritam prepoznaje kao sum a koliko ne. Od svih podataka samo 54 tacke su oznacene kao sum, ali to mozemo da smanjimo. Vidimo da je nakon primene, PCA algoritam eliminisao sum koji je prepoznao DBSCAN u originalnim podacima.



Prikaz rezultata DBSCAN-a.

6 Pravila pridruživanja

Primenjeni algoritam u SPSS-u.

6.1 Association rules

Posto su nasi podaci neprekidnog tipa, jedini algoritam koji mozemo da koristimo jeste Association rules iz SPSS-a. Stavili smo atribute tako da mogu da budu i posledicni i uzrocni (to jest i sa desne i sa leve strane pravila). Za pravila smo izabrali da moze da bude najvise jedan posledicni i najvise pet uzrocnih. Izabran je Apriori algoritam za pravljenje pravila. Ukljucena je i lift statistika. Kriterijum za pravila je stavljen na pouzdanost, i da se boduju jedino kada predvidjanja nisu u ulazu.

Build Settings

Maximum Number of Rules	1.000
Minimum Condition Support	0,05
Minimum Confidence	0,10
Minimum Rule Support	0,05
Minimum Lift	2,00
Maximum Number of Items in a Rule	10
Maximum Number of Items in a Condition	5
Maximum number of Items in a Prediction	1
Use only True Value for Flag Fields	False
Allow Rules without Conditions	False
Evaluation Measure Sorting the Rules	Confidence

Podesavanja za Association rules u SPSS-u.

Most Interesting Rules by Confidence

Rank	Rule ID	Condition	Prediction	Sorted By Confidence(%)	Other Evaluation Statistics			
					Condition Support (%)	Rule Support (%)	Lift	Deployability (%)
1	1	6,200 ≤ pts < 11,700 4,600 ≤ fga < 8,400 blk ≤ 0,780 target_5yrs = 1	2,280 ≤ fgm < 4,260	100,00	13,28	13,28	3,19	0,00
2	2	gp > 68 6,200 ≤ pts < 11,700 4,600 ≤ fga < 8,400 blk ≤ 0,780	2,280 ≤ fgm < 4,260	100,00	11,61	11,61	3,19	0,00
3	3	6,200 ≤ pts < 11,700 4,600 ≤ fga < 8,400 ftm ≤ 1,540 blk ≤ 0,780	2,280 ≤ fgm < 4,260	100,00	11,38	11,38	3,19	0,00
4	4	39 ≤ gp < 54 min ≤ 10,660 fgm ≤ 2,280	fga ≤ 4,600	100,00	10,17	10,17	2,11	0,00
5	5	39 ≤ gp < 54 min ≤ 10,660 pts ≤ 6,200 fgm ≤ 2,280	fga ≤ 4,600	100,00	10,17	10,17	2,11	0,00
6	6	39 ≤ gp < 54 min ≤ 10,660 fgm ≤ 2,280 ftm ≤ 1,540	fga ≤ 4,600	100,00	10,17	10,17	2,11	0,00
7	7	39 ≤ gp < 54 min ≤ 10,660 fgm ≤ 2,280 reb ≤ 3,020	fga ≤ 4,600	100,00	10,17	10,17	2,11	0,00

Neka od dobijenih pravila koja se nalaze u segmentu onih sortiranih po pouzdanosti.

Most Interesting Rules by Rule Support

Rank	Rule ID	Condition	Prediction	Sorted By Rule Support(%)	Other Evaluation Statistics			
					Condition Support (%)	Confidence (%)	Lift	Deployability (%)
1	1	6,200 ≤ pts < 11,700 4,600 ≤ fga < 8,400 blk ≤ 0,780 target_5yrs = 1	2,280 ≤ fgm < 4,260	13,28	13,28	100,00	3,19	0,00
2	2	gp > 68 6,200 ≤ pts < 11,700 4,600 ≤ fga < 8,400 blk ≤ 0,780	2,280 ≤ fgm < 4,260	11,61	11,61	100,00	3,19	0,00
3	3	6,200 ≤ pts < 11,700 4,600 ≤ fga < 8,400 ftm ≤ 1,540 blk ≤ 0,780	2,280 ≤ fgm < 4,260	11,38	11,38	100,00	3,19	0,00
4	4	39 ≤ gp < 54 min ≤ 10,660 fgm ≤ 2,280	fga ≤ 4,600	10,17	10,17	100,00	2,11	0,00
5	5	39 ≤ gp < 54 min ≤ 10,660 pts ≤ 6,200 fgm ≤ 2,280	fga ≤ 4,600	10,17	10,17	100,00	2,11	0,00
6	6	39 ≤ gp < 54 min ≤ 10,660 fgm ≤ 2,280 ftm ≤ 1,540	fga ≤ 4,600	10,17	10,17	100,00	2,11	0,00
7	7	39 ≤ gp < 54 min ≤ 10,660 fgm ≤ 2,280 reb ≤ 3,020	fga ≤ 4,600	10,17	10,17	100,00	2,11	0,00

Neka od dobijenih pravila koja se nalaze u segmentu onih 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	469	60,000 ≤ ft < 80,000 1,060 ≤ oreb < 2,120 2,080 ≤ dreb < 3,960 0,960 ≤ tov < 1,820	3,020 ≤ reb < 5,740	3,45	7,44	100,00	7,44	0,00
2	572	43,760 ≤ fg < 53,740 60,000 ≤ ft < 80,000 1,060 ≤ oreb < 2,120 2,080 ≤ dreb < 3,960 target_5yrs = 1	3,020 ≤ reb < 5,740	3,45	7,28	100,00	7,28	0,00
3	573	10,660 ≤ min < 18,220 1,060 ≤ oreb < 2,120 2,080 ≤ dreb < 3,960	3,020 ≤ reb < 5,740	3,45	7,21	100,00	7,21	0,00
4	578	10,660 ≤ min < 18,220 1,060 ≤ oreb < 2,120 2,080 ≤ dreb < 3,960 ast ≤ 2,120	3,020 ≤ reb < 5,740	3,45	7,21	100,00	7,21	0,00
5	675	10,660 ≤ min < 18,220 3p_made ≤ 0,460 1,060 ≤ oreb < 2,120 2,080 ≤ dreb < 3,960	3,020 ≤ reb < 5,740	3,45	7,06	100,00	7,06	0,00
6	676	10,660 ≤ min < 18,220 3pa ≤ 1,300 1,060 ≤ oreb < 2,120 2,080 ≤ dreb < 3,960	3,020 ≤ reb < 5,740	3,45	7,06	100,00	7,06	0,00
7	720	10,660 ≤ min < 18,220 3p_made ≤ 0,460 3pa ≤ 1,300 1,060 ≤ oreb < 2,120 2,080 ≤ dreb < 3,960	3,020 ≤ reb < 5,740	3,45	7,06	100,00	7,06	0,00

Neka od dobijenih pravila koja se nalaze u segmentu onih sortiranih po liftu.

Information for Most Frequent Items

Item name	Records (%)	Conditions (%)	Predictions (%)
blk ≤ 0,780	86,72	19,70	0,00
3p_made ≤ 0,460	79,51	18,80	0,00
3pa ≤ 1,300	77,54	24,40	0,00
ast ≤ 2,120	77,39	16,00	0,00
ftm ≤ 1,540	72,31	16,50	0,00
60,000 ≤ ft < 80,000	68,44	22,00	0,00
fta ≤ 2,040	68,21	16,50	0,00
oreb ≤ 1,060	62,82	15,60	0,00
target_5yrs = 1	62,14	7,70	0,00
reb ≤ 3,020	60,93	16,50	0,00
dreb ≤ 2,080	60,77	15,70	0,00
pts ≤ 6,200	56,83	14,20	0,00
fgm ≤ 2,280	53,34	22,90	0,00
stl ≤ 0,500	52,28	13,20	0,00
20,000 ≤ 3p < 40,000	48,79	1,20	0,00
fga ≤ 4,600	47,34	0,00	80,70
tov ≤ 0,960	46,74	13,10	0,00
3p ≤ 20,000	46,28	19,20	0,00
43,760 ≤ fg < 53,740	45,68	45,40	0,00
33,780 ≤ fg < 43,760	44,01	0,10	0,00
gp > 68	41,96	2,70	0,00
target_5yrs = 0	37,86	9,20	0,00
0,960 ≤ tov < 1,820	37,10	2,00	0,00
10,660 ≤ min < 18,220	34,52	1,10	0,00
0,500 ≤ stl < 1,000	33,99	3,50	0,00
4,600 ≤ fga < 8,400	32,70	12,30	0,00
2,280 ≤ fgm < 4,260	31,34	6,70	10,70
2,080 ≤ dreb < 3,960	29,89	1,80	0,00
6,200 ≤ pts < 11,700	29,82	10,70	6,80
3,020 ≤ reb < 5,740	28,98	0,00	1,80
1,060 ≤ oreb < 2,120	27,92	1,80	0,00
2,040 ≤ fta < 4,080	25,27	3,20	0,00
min ≤ 10,660	24,20	77,30	0,00
18,220 ≤ min < 25,780	22,84	0,50	0,00
1,540 ≤ ftm < 3,080	21,85	6,70	0,00
39 ≤ gp < 54	21,70	14,60	0,00

Procenat pojavljivanja pravila, kao i koliko cesto se nalaze sa leve i sa desne strane pravila.

7 Zaključak

Kao i sa svakom bazom podataka, vrlo je bitna priprema i razumevanje podataka sa kojima radimo. Ako ne razumemo podatke ne mozemo da budemo sigurni koje algoritme mozemo da iskoristimo, kako treba da ih pripremimo, sta su zapravo dobre vrednosti a sta greske itd. Kolicina gresaka u ovoj bazi je bila minimalna, ali su postojale neke anomalije koje bi uticale na preciznost naseg modela iako bi to bio zanemarljiv uticaj. Nakon primene raznih tehnika, preciznost za klasifikaciju nije mogla da predje preko 72% sto samo pokazuje tezinu rada sa ovom bazom. Klasterovanje se uglavnom dobro pokazalo, posebno sa DBSCAN-om. Algoritmom Association rules smo otkrili neka pravila koja su zanimljiva iz pogleda kosarkaske statistike.

8 Literatura

DataWorld: NBA_Rookies_Dataset
Github repozitorijum kursa Istraživanje podataka 1: Link ka githubu
AutoGluon:Link ka dokumentacji
Scikit-learn Machine Learning in Python: Link ka sajtu