

Olimpijske Igre

Marko Paunović
mi20104@alas.matf.bg.ac.rs

15.9.2023.

Sadržaj

1	Uvod	2
1.1	Analiza skupa podataka	2
1.2	Nedostajuće vrednosti	3
1.3	Medalje	4
2	Preprocesiranje	5
2.1	Izbacivanje nepotrebnih kolona i kodiranje kategorickih atributa	5
2.2	Balansiranje podataka	5
2.3	Standardizacija podataka	6
3	Opis modela	6
3.1	Klasifikacija	6
3.1.1	Stabla odlučivanja	7
3.1.2	Slučajne šume	10
3.1.3	K najbližih suseda	13
3.1.4	Poređenje modela	15
3.2	Klasterovanje	17
3.2.1	K-sredina	17
3.2.2	K-sredina sa podelom	19
3.2.3	Aglomerativno klasterovanje	20
3.2.4	DBSCAN-a	21
3.3	Pravila pridruživanja	22
3.3.1	Apriori	22
4	Zaključak	23
5	Reference	23

1 Uvod

Olimpijske igre su jedna od najznačajnijih sportskih manifestacija na svetu, događaj koji se održava na svake 4 godine i u kojem učestvuju hiljade sportista širom sveta takmičući u različitim sportskim disciplinama.

Cilj ovog projekta je da istražimo obrasce i trendove u istorijskim rezultatima Olimpijskih igara kako bismo bolje razumeli faktore koji su uticali na uspeh sportista tokom vremena.

Skup podataka koji ćemo razmatrati je skup Olympic Data koji možete naći na linku: <https://www.kaggle.com/datasets/bhanupratapbiswas/olympic-data>

1.1 Analiza skupa podataka

Naš skup podataka se sastoji od velikog broja takmičara koji su učestvovali na Olimpijskim igrama zadnjih 120 godina. Skup ukupno ima 70 000 takmičara(instanci) koji su opisani sa 15 atributa:

- ID - redni broj takmičara
- Name - ime i prezime
- Sex - pol
- Height - visina
- Weight - težina
- Team - država za koju se takmiči
- NOC - Nacionalni olimpijski komitet (oznaka države)
- Games - godina održavanja i da li su letnje ili zimske igre
- Year - godina održavanja igara
- Season - da li su letnje ili zimske igre
- City - grad gde se igre održavaju
- Sport - sport u kome se takmiči
- Event - disciplina u kojoj se takmiči
- Medal - osvojena medalja

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

Slika 1: Izgled skupa podataka

	ID	Age	Height	Weight	Year
count	70000.000000	67268.000000	53746.000000	52899.000000	70000.000000
mean	18081.846986	25.644645	175.505303	70.900216	1977.766457
std	10235.613253	6.485239	10.384203	14.217489	30.103306
min	1.000000	11.000000	127.000000	25.000000	1896.000000
25%	9325.750000	21.000000	168.000000	61.000000	1960.000000
50%	18032.000000	25.000000	175.000000	70.000000	1984.000000
75%	26978.000000	28.000000	183.000000	79.000000	2002.000000
max	35658.000000	88.000000	223.000000	214.000000	2016.000000

Slika 2: Opis skupa podataka

1.2 Nedostajuće vrednosti

Ova oblast spada u preprocesiranje ali izbacićemo sad nedostajuće vrednosti zato što ih nećemo uopšte koristiti i da bi nam statistika i opis skupa podataka bila preciznija. Izbacujemo samo redove gde nedostaju vrednosti kolona Age, Height i Weight, a ne izbacujemo nedostajuće vrednosti kolone Medal zato što nam ona ukazuje na to koliko takmičara nije osvojilo medalju.

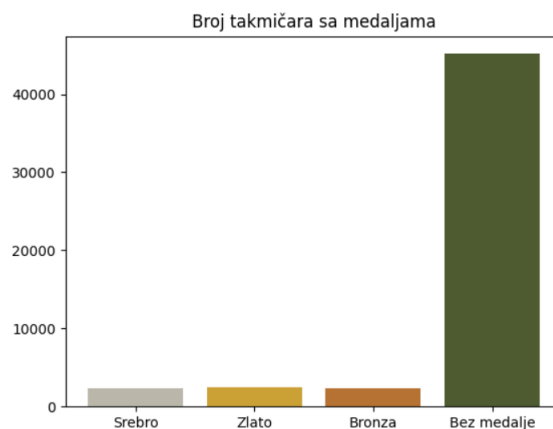
ID	0	ID	0
Name	0	Name	0
Sex	0	Sex	0
Age	2732	Age	0
Height	16254	Height	0
Weight	17101	Weight	0
Team	0	Team	0
NOC	0	NOC	0
Games	0	Games	0
Year	0	Year	0
Season	0	Season	0
City	0	City	0
Sport	0	Sport	0
Event	0	Event	0
Medal	60310	Medal	45165
dtype: int64		dtype: int64	

Slika 3: Broj nedostajućih vrednosti pre

Slika 4: i posle izbacivanja

1.3 Medalje

Vidimo da imamo sličan broj zlatnih, srebrnih i bronzanih medalja što nam ukazuje da je naš skup podataka u tom pogledu balansiran. Međutim broj takmičara bez medalja je mnogo puta veći pa ćemo morati da balansiramo podatke kako bi modeli bili precizniji.



Slika 5: Broj takmičara sa medaljama

2 Preprocesiranje

Preprocesiranje je jedan od najvažnijih koraka u analizi podataka. Ona podrazumeva proces čišćenja, transformacije i pripreme sirovih podataka kako bi se dobili kvalitetni i upotrebljivi podaci za analizu.

2.1 Izbacivanje nepotrebnih kolona i kodiranje kategoričkih atributa

Izbacujemo kolone koje nam nisu od značaja poput ID, Name, Team, Year, Season, City, Event.

Kategorički atributi su atributi koji sadrže kategoričke vrednosti i oni nisu pogodni za algoritme mašinskog učenja. Zato moramo da ih pretvorimo u numeričke vrednosti koristeći razna kodiranja.

Koristićemo Label encoding koje mapira svaku kategoriju u jedinstvenu numeričku vrednost.

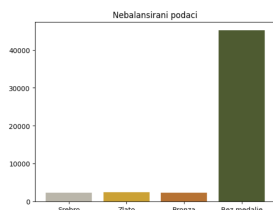
	Sex	Age	Height	Weight	NOC	Games	Sport
0	0	24.0	180.0	80.0	40	0	6
1	0	23.0	170.0	60.0	40	0	26
2	1	21.0	185.0	82.0	140	1	42
3	1	21.0	185.0	82.0	140	1	42
4	1	25.0	185.0	82.0	140	1	42

Slika 6: Kodiran i očišćen skup

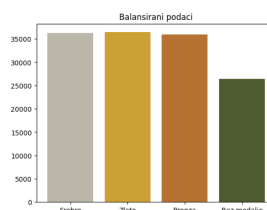
2.2 Balansiranje podataka

Potrebno je da balansiramo podatke zato što imamo mnogo više takmičara bez medalje (45000) nego takmičara sa medaljom (3*2000).

Za balansiranje koristimo SMOTEEN (Synthetic Minority Over-sampling Technique for Uneven Labels) tehniku koja kombinuje dve popularne tehnike: SMOTE (Synthetic Minority Over-sampling Technique) i ENN (Edited Nearest Neighbors). Koristimo je kako bismo generisali nove takmičare koji su slični takmičarima koji su dobili medalju. SMOTE generiše nove, dok ENN uklanja potencijalno štetne takmičare.



Slika 7: Nebalansirani podaci



Slika 8: Balansirani podaci

2.3 Standardizacija podataka

Standardizaciju radimo kako bi sve naše vrednosti bile između 0 i 1 da bi naši modeli pridavali podjednaku važnost svim atributima.

	Sex	Age	Height	Weight	NOC	Games	Sport
0	1.0	0.547170	0.406250	0.164021	0.466063	0.0	0.666667
1	0.0	0.415141	0.467244	0.315969	0.647059	0.0	0.444444
2	1.0	0.094340	0.520833	0.185185	0.013575	0.0	0.944444
3	0.0	0.323428	0.566956	0.266828	0.325792	0.0	0.444444
4	1.0	0.169811	0.556946	0.238919	0.303167	0.0	0.796296
...
94588	0.0	0.213329	0.580982	0.289384	0.303167	0.0	0.296296
94589	0.0	0.200528	0.453771	0.196772	0.407240	0.0	0.444444
94590	1.0	0.343895	0.426079	0.172041	0.167421	0.0	0.666667
94591	0.0	0.264151	0.427083	0.222222	0.307692	0.0	0.407407
94592	1.0	0.188679	0.354167	0.164021	0.303167	1.0	0.222222

Slika 9: Skaliran X train

Ovi podaci nam služe za klasifikaciju. Podatke za klasterovanje nismo delili na Xtrain i Xtest pa standardizovali nego smo standardizovali ceo X skup.

	Sex	Age	Height	Weight	NOC	Games	Sport
0	0.0	0.226415	0.552083	0.291005	0.180995	0.0	0.111111
1	0.0	0.207547	0.447917	0.185185	0.180995	0.0	0.481481
2	1.0	0.169811	0.604167	0.301587	0.633484	1.0	0.777778
3	1.0	0.169811	0.604167	0.301587	0.633484	1.0	0.777778
4	1.0	0.245283	0.604167	0.301587	0.633484	1.0	0.777778
...
52261	0.0	0.150943	0.343750	0.164021	0.420814	0.0	0.185185
52262	0.0	0.150943	0.677083	0.359788	0.054299	0.0	0.203704
52263	0.0	0.264151	0.447917	0.211640	0.054299	0.0	0.055556
52264	0.0	0.339623	0.447917	0.211640	0.054299	0.0	0.055556
52265	0.0	0.207547	0.552083	0.227513	0.235294	1.0	0.555556

Slika 10: Skalirani podaci za pravila pridruživanja

3 Opis modela

3.1 Klasifikacija

Klasifikacija je tehnika koja razvrstava podatke na osnovu atributa i dodeljuje im klase. Cilj je naučiti model da predviđa pripadnost klasama novih i nepoznatih instanci.

3.1.1 Stabla odlučivanja

Kao prvi model klasifikacije koristimo stabla odlučivanja. On radi na principu konstrukcije stabla sa čvorovima koji predstavljaju odluke i granama koje predstavljaju moguće ishode tih odluka. Tako se formira drvolika struktura. Počevši od korenskog čvora donosi se odluka u odnosu na jedan atribut...

Train data:

		Confusion Matrix			
True	Bronze	25409	0	0	0
	Gold	0	25507	0	0
	No Medal	0	0	18483	0
	Silver	0	0	0	25194
		Predicted			
		Bronze	Gold	No Medal	Silver

Accuracy score: 1.0
Precision score: 1.0
Recall score: 1.0
F1 score: 1.0

Test data:

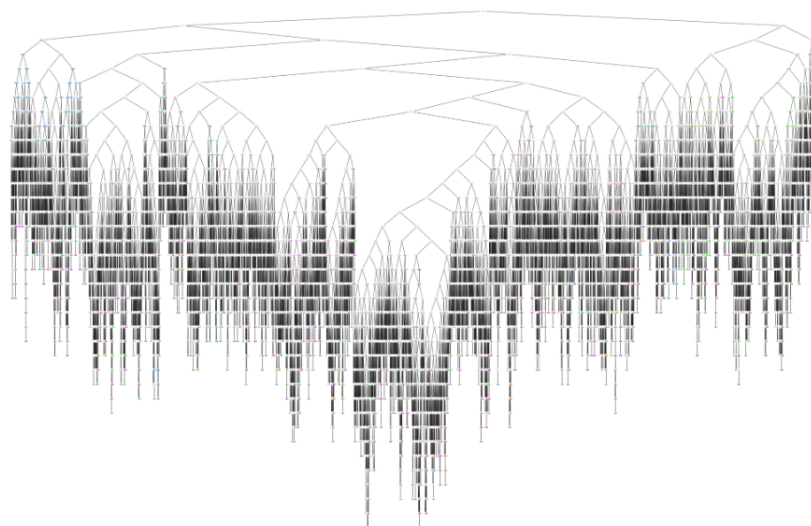
		Confusion Matrix			
True	Bronze	9842	351	288	409
	Gold	334	10086	186	325
	No Medal	342	295	6926	359
	Silver	407	384	222	9785
		Predicted			
		Bronze	Gold	No Medal	Silver

Accuracy score: 0.9037517574800819
Precision score: 0.9041044226606616
Recall score: 0.901730585740576
F1 score: 0.902814724740074

Slika 11: Matrica konfuzije treninga

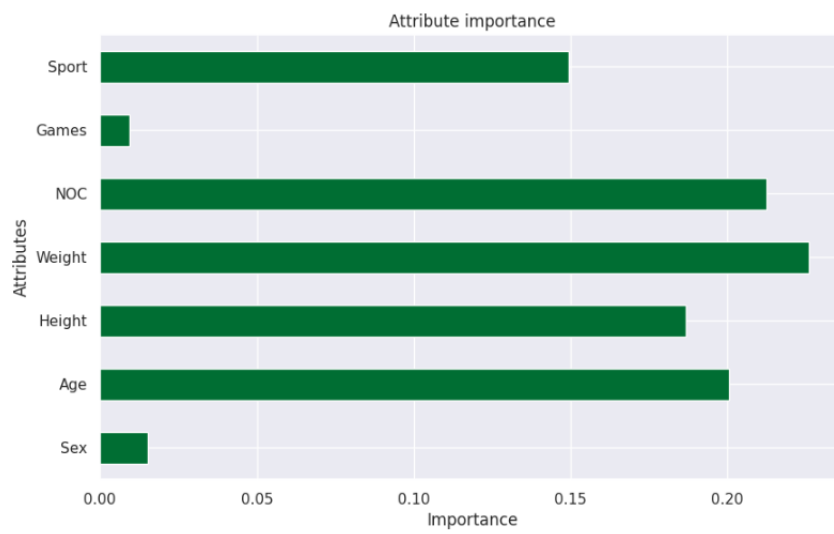
Slika 12: Matrica konfuzije testa

Vidimo da je model na treningu u potpunosti precizan što može da ukaže na to da se prilagodio, međutim i na testu postiže dobre rezultate, te se nije prilagodio.



Broj čvorova: 22835, dubina: 36

Slika 13: Izgled stabla odlučivanja



Slika 14: Uticaj atributa na odluku

Vidimo da pol, godina i to da li su igre letnje ili zimske ne utiče mnogo na odluku.

Optimizaciju ovog modela vršimo pomoću GridSearchCV koji kombinuje zadate parametre i pronalazi model koji ima najbolju tačnost (accuracy). GridSearchCV kombinuje hiper-parametre modela i tako traži najtačniji model.

Train data:

		Confusion Matrix			
True	Bronze	25139	94	48	128
	Gold	103	25228	46	130
	No Medal	27	28	18402	26
	Silver	69	91	31	25003
		Predicted			
		Bronze	Gold	No Medal	Silver

Accuracy score: 0.9913207108348397
Precision score: 0.9914489734816345
Recall score: 0.9916180236263527
F1 score: 0.9915314063170824

Slika 15: Matrica konfuzije treninga

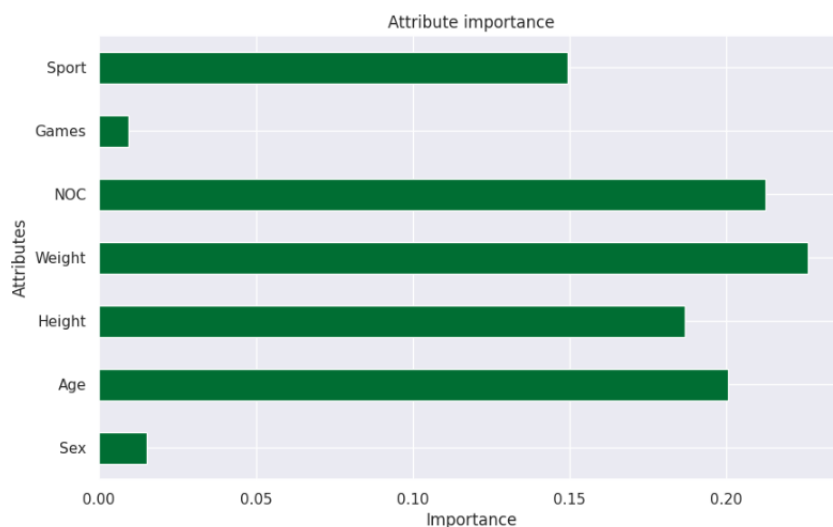
Test data:

		Confusion Matrix			
True	Bronze	9748	394	266	482
	Gold	371	10005	157	398
	No Medal	350	276	7010	286
	Silver	457	400	256	9685
		Predicted			
		Bronze	Gold	No Medal	Silver

Accuracy score: 0.8990404775412545
Precision score: 0.8999515699598092
Recall score: 0.8980557153526952
F1 score: 0.8989419322958763

Slika 16: Matrica konfuzije testa

Vidimo da se optimizovani model ponaša gore od običnog. Potrebno je još malo podešavati hiper-parametre kako bi model bio bolji.



Slika 17: Uticaj atributa na odluku

3.1.2 Slučajne šume

Slučajne šume (Random Forest) je ansambl algoritam koji se sastoji od više stabala odlučivanja i kombinovanjem njihovih rezultata dobija konačan rezultat. Svako stablo kao ulazne vrednosti ima nasumičan podskup ukupnih atributa.

Train data:

Confusion Matrix				
True	Bronze	Gold	No Medal	Silver
Bronze	25409	0	0	0
Gold	0	25507	0	0
No Medal	0	0	18483	0
Silver	0	0	0	25194
Predicted				

Accuracy score: 1.0
Precision score: 1.0
Recall score: 1.0
F1 score: 1.0

Slika 18: Matrica konfuzije treninga

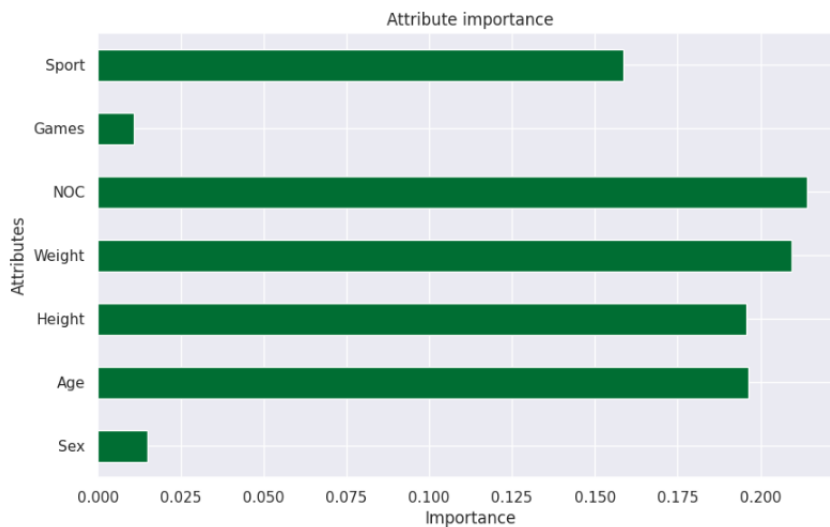
Test data:

Confusion Matrix				
True	Bronze	Gold	No Medal	Silver
Bronze	10493	123	127	147
Gold	146	10598	69	118
No Medal	123	113	7559	127
Silver	155	133	95	10415
Predicted				

Accuracy score: 0.9635924126193236
Precision score: 0.9635430619185383
Recall score: 0.9629473560494971
F1 score: 0.9632391503328528

Slika 19: Matrica konfuzije testa

Ovde takođe vidimo da je model na treningu u potpunosti precizan što može da ukaže na to da se prilagodio, međutim i na testu postiže odlične rezultate, te se nije prilagodio.



Slika 20: Uticaj atributa na odluku

Optimizaciju radimo isto kao i gore, preko GridSearchCV.

Train data:

Confusion Matrix				
True	Bronze	Gold	No Medal	Silver
	25409	0	0	0
	0	25507	0	0
	0	0	18483	0
	0	0	0	25194
Predicted				

Accuracy score: 1.0
Precision score: 1.0
Recall score: 1.0
F1 score: 1.0

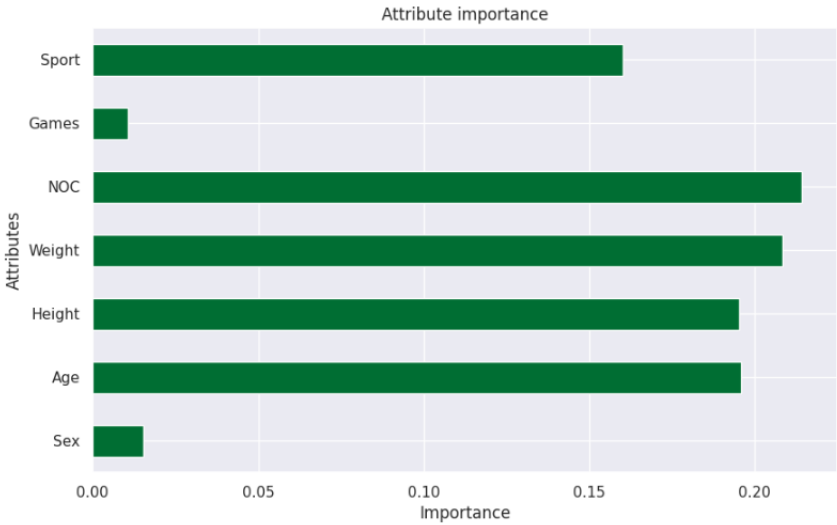
Slika 21: Matrica konfuzije treninga

Test data:

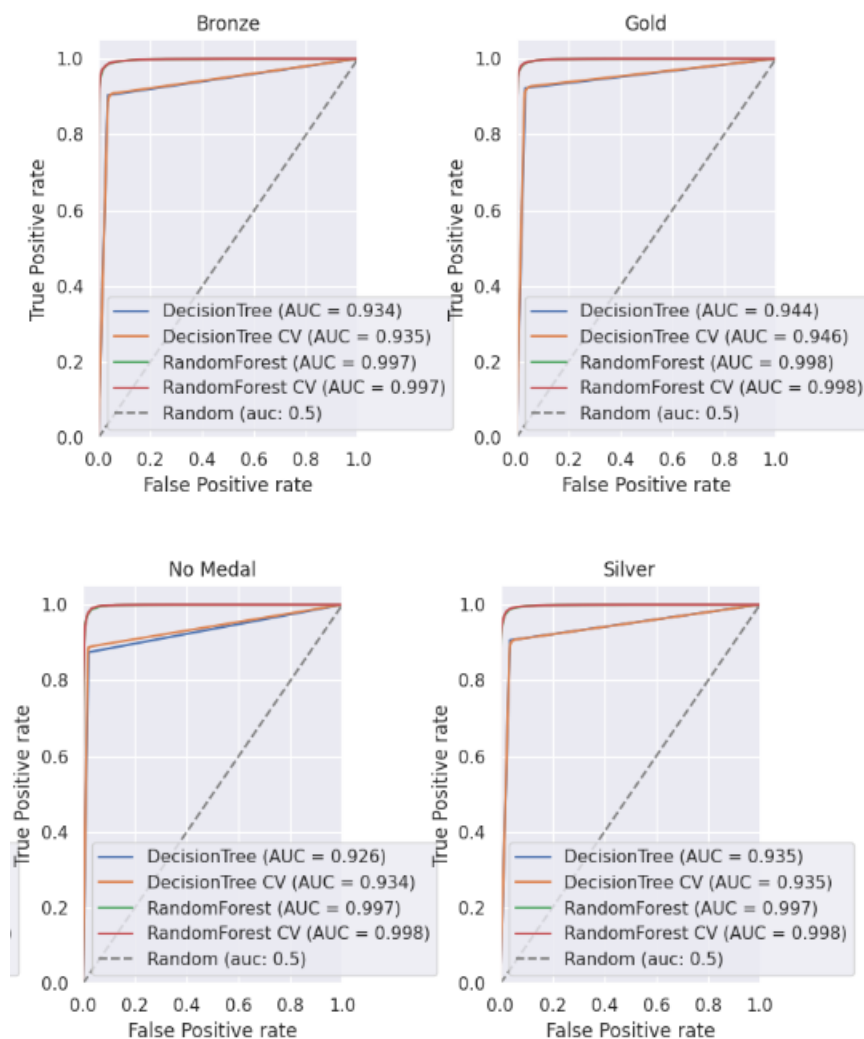
Confusion Matrix				
True	Bronze	Gold	No Medal	Silver
	10521	113	127	129
	125	10611	76	119
	106	103	7595	118
	147	122	95	10434
Predicted				

Accuracy score: 0.9659603857822945
Precision score: 0.9657009145755936
Recall score: 0.9654634401766679
F1 score: 0.9655811626081118

Slika 22: Matrica konfuzije testa



Slika 23: Uticaj atributa na odluku



Slika 24: ROC kriva za višeklasnu klasifikaciju

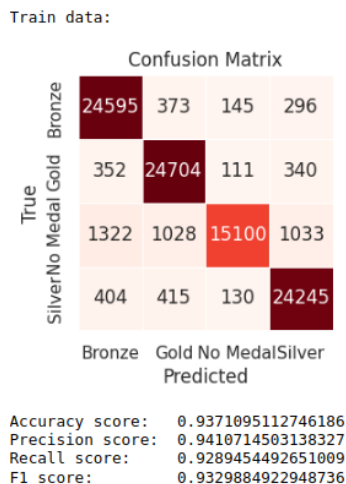
ROC kriva je grafički prikaz performansi klasifikacionog modela, posebno za binarnu klasifikaciju. Zato samo i ispisivali za svaku vrednost ciljne kolone po jedan grafik koji ukazuje na to da li je tačno klasifikovano ili nije (binarna klasifikacija).

ROC kriva se konstruiše tako što se prate dva ključna metrička pokazatelja: stopa lažno pozitivnih (False Positive Rate - FPR) i stopa tačno pozitivnih (True Positive Rate - TPR). Na dijagonali se nalazi trivijalni model koji nasumično donosi odluke.

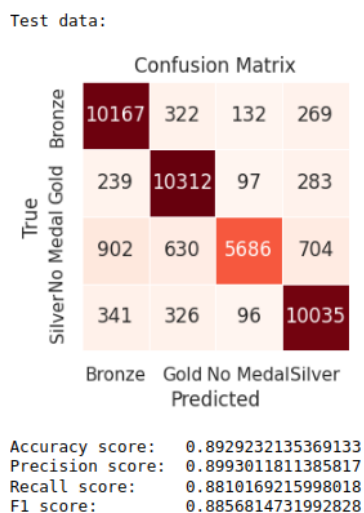
Oдавде vidimo da je najtačniji model RandomForestCV.

3.1.3 K najbližih suseda

K najbližih suseda radi tako što se prvo odabere broj najbližih suseda k , zatim se za svaku instancu odredi udaljenost od trenutne instance, odabere k najbližih i na osnovu klasa najbližih odredi klasu trenutne instance.

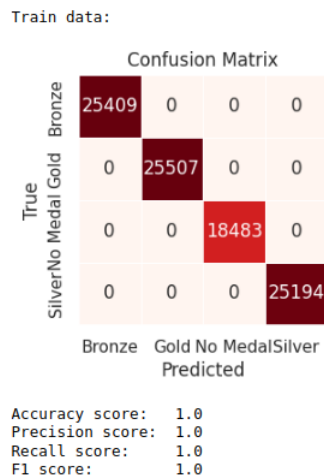


Slika 25: Matrica konfuzije treninga

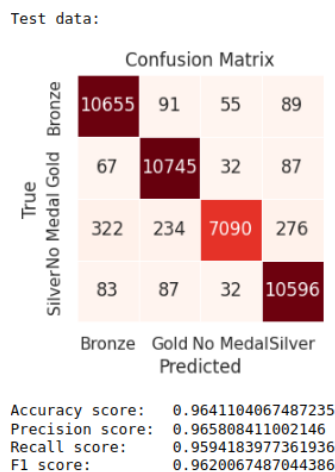


Slika 26: Matrica konfuzije testa

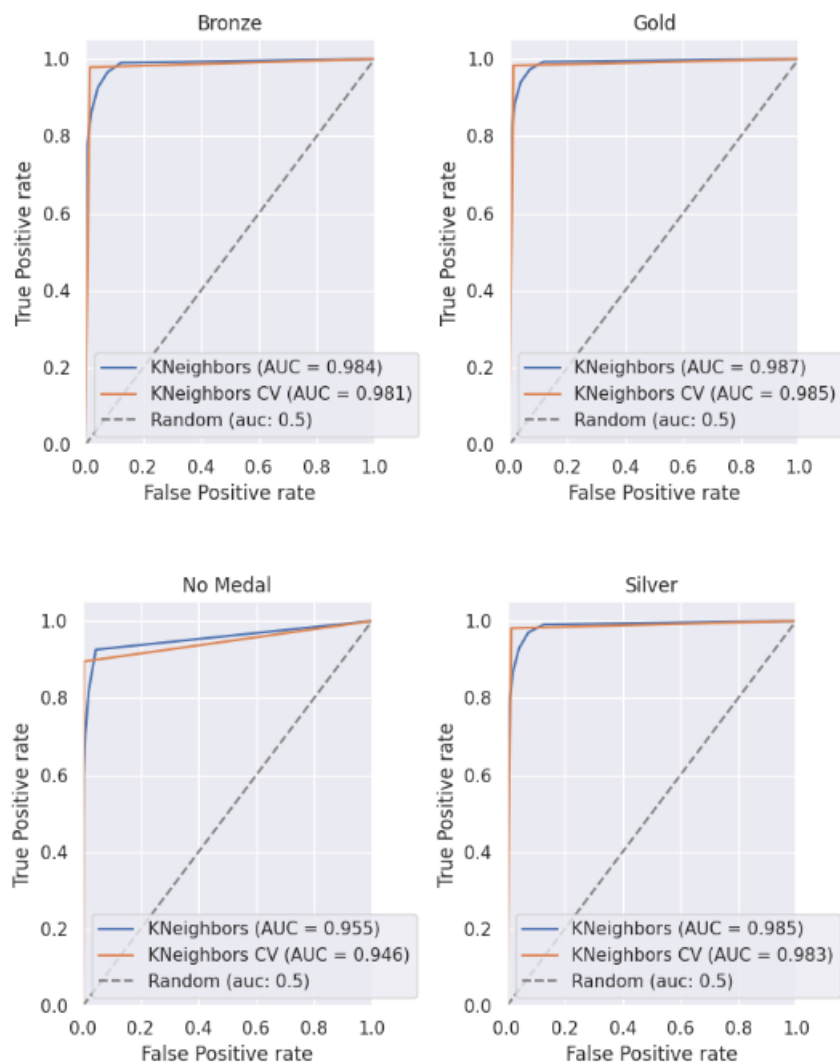
Optimizaciju ovog modela vršimo pomoću GridSearchCV koji podešava hiperparametre modela i izdvađa model sa najboljom tačnošću



Slika 27: Matrica konfuzije treninga



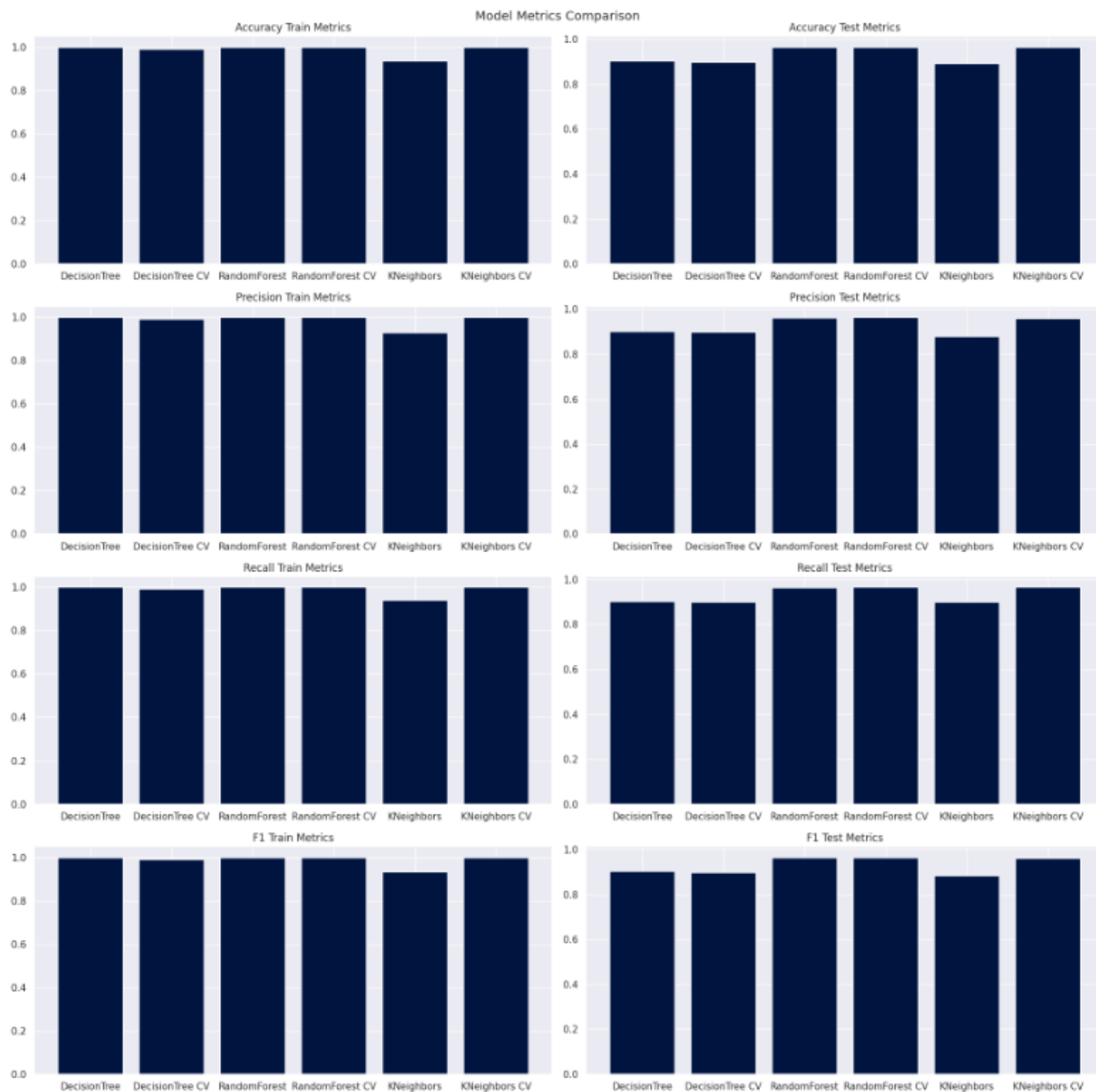
Slika 28: Matrica konfuzije testa



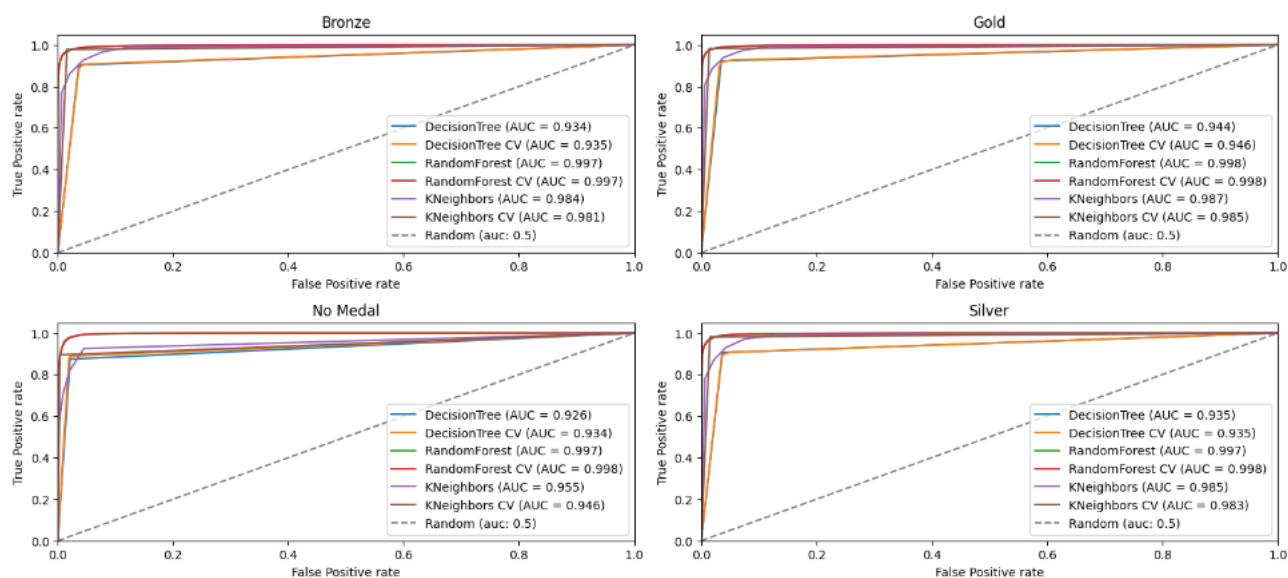
Slika 29: ROC kriva za višeklasnu klasifikaciju

Vidimo da se oba modela muče sa klasifikaciom takmičara bez medalja, ali model bez podešavanja hiper-parametara je objektivno bolji.

3.1.4 Poređenje modela



Slika 30: Poređenje metrika svih modela



Slika 31: ROC kriva za višeklasnu klasifikaciju za sve modele

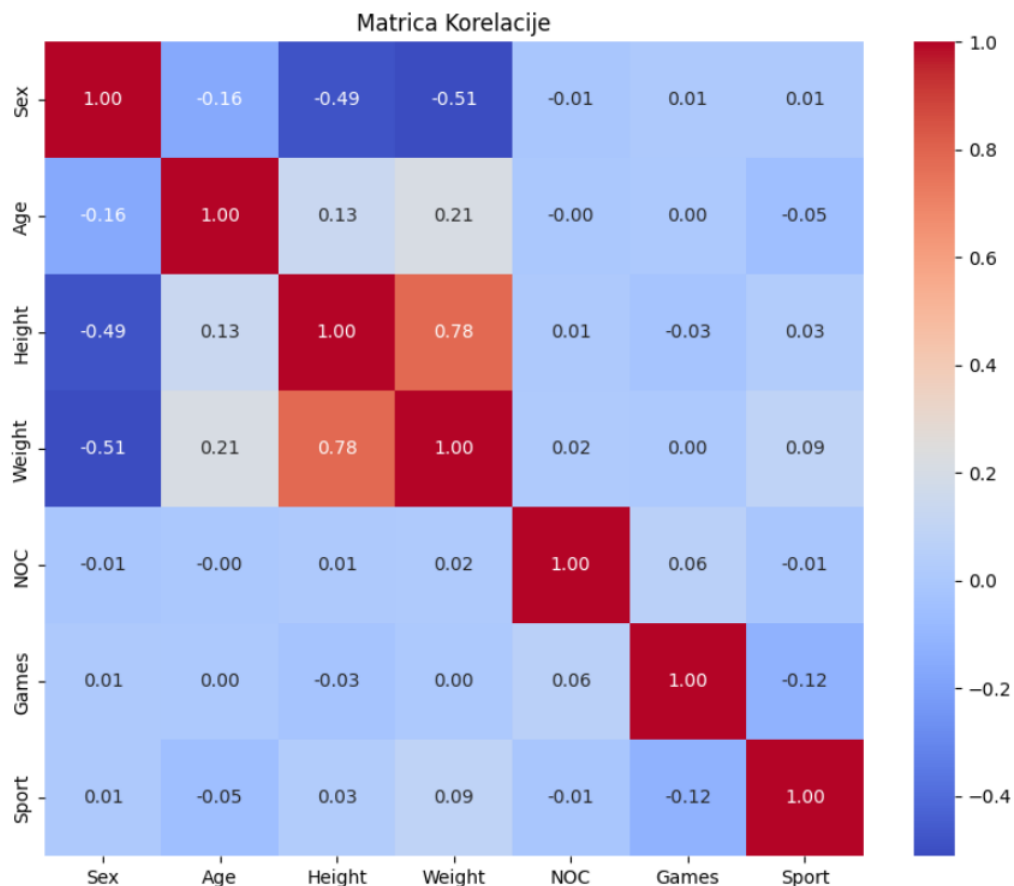
Sa ove 2 slike vidimo da najbolje metrike imaju modeli slučajnih šuma i optimizovan model K najbližih suseda.

3.2 Klasterovanje

Klasterovanje je tehnika u analizi podataka koja se koristi za grupisanje sličnih podataka zajedno u klastere (grupe), tako da podaci unutar jednog klastera budu sličniji jedni drugima nego podaci iz drugih klastera. Algoritmi klasterovanja pokušavaju da nađu nekakvu zakonitost među podacima i grupišu ih po tome.

3.2.1 K-sredina

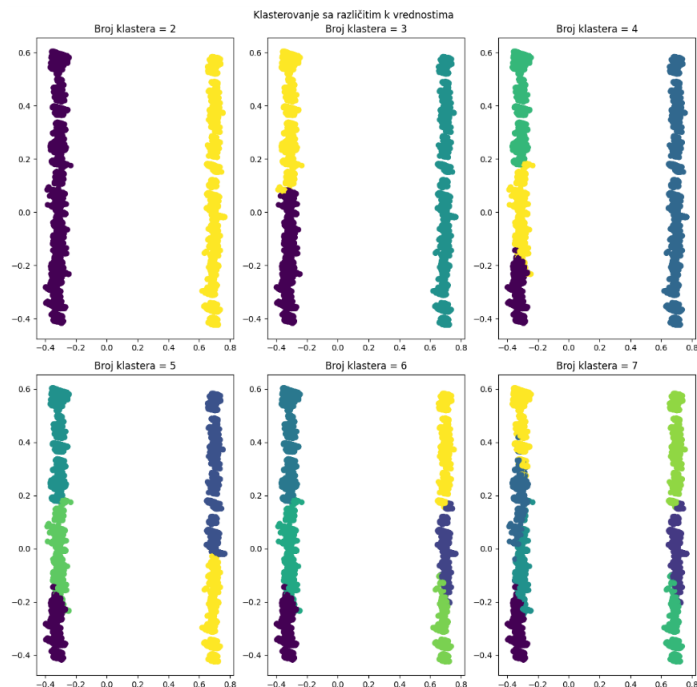
K-sredina je algoritam koji klasteruje podatke u unapred određen broj klastera. Algoritam radi tako što se, obično nasumice, odaberu početni centri klastera, zatim se izračuna udaljenost svake instance od centrova i dodeljuje joj se klaster čijem je centru najbliža. Nakon toga se centri ažuriraju tako da budu u sredini instanci koje joj pripadaju. Ovo se ponavlja dok centri ne počnu da se pomeraju zanemarljivo malo.



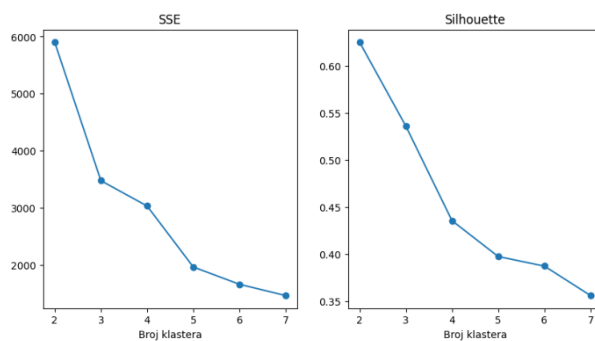
Slika 32: Matrica korelacije

Vidimo iz matrice korelacije da su kolone NOC i Games slabo zavisne od ostalih kolona i obrnuto. Zato možemo da ih izbacimo i smanjimo vreme treni-

ranja modela uz potencijalno malo smanjenja tačnosti algoritma.



Slika 33: Klasterovanje pomoću K-sredina



Slika 34: SSE i Silueta

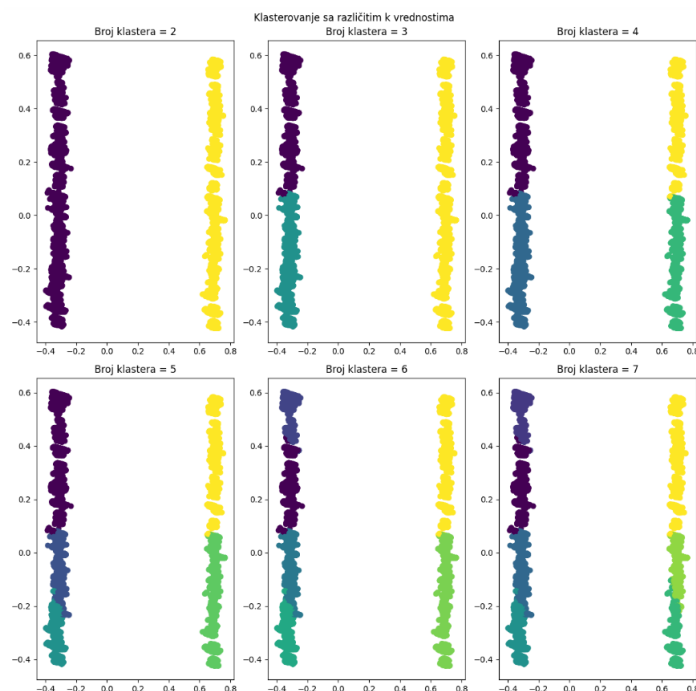
SSE (Sum of Squared Errors) je metrika koja meri koliko su tačke unutar istog klastera blizu centroide tog klastera. Što je niža vrednost SSE, to su klasteri kompaktniji.

Silhouette je metrika koja meri koliko su tačke dobro povezane sa svojim sopstvenim klasterom u poređenju sa drugim klasterima. Silhouette koeficijent se izračunava kao $(b - a) / \max(a, b)$, gde je a prosečna udaljenost od tačaka istog klastera, a b prosečna udaljenost od tačaka najbližeg klastera. Vrednost Silhouette koeficijenta varira od -1 (loše grupisanje) do 1 (dobro grupisanje).

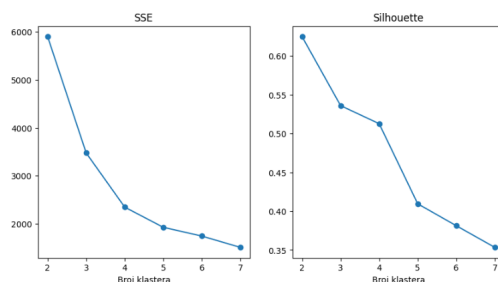
U našem primeru, ako klasterujemo u 2 klastera, vidimo da imamo klastere čije su tačke daleko od svojih centroida, ali da su dobro razdvojeni tj. da nema poljapanja. Ova model najtačnije klasteruje podatke u poredjenju sa modelima koji klasteruju u više klastera.

3.2.2 K-sredina sa podelom

Ovo je varijanta algoritma k-sredina, razlika je u tome što ovaj algoritam radi iterativno i u svakom koraku deli jedan klaster na 2 dela dok ne dođe do traženog broja klastera.



Slika 35: Klasterovanje pomoću K-sredina sa podelom

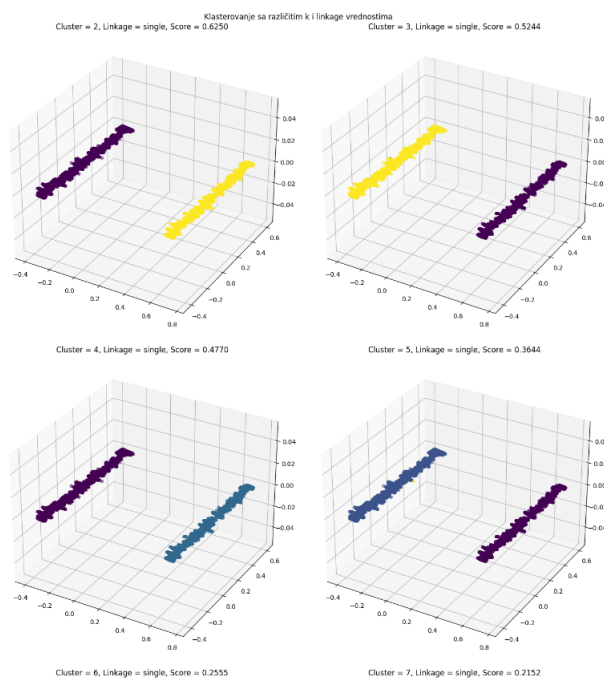


Slika 36: SSE i silueta

Isto kao i sa prošlim modelom, najbolji model je onaj koji klasteruje u 2 klastera.

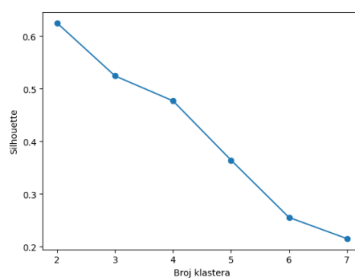
3.2.3 Aglomerativno klasterovanje

Ovaj algoritam se koristi za grupisanje u dendrograme. Svaku instancu tretira kao jedan klaster i u svakom koraku spaja je sa najbližim klasterom. To se ponavlja dok podaci ne postanu jedan klaster. Algoritam pamti korake spajanja i može da formira hijerarhijsku strukturu.



Slika 37: Klasterovanje pomoću Aglomerativnog klasterovanja

Linkage određuje kako se računa udaljenost između dva klastera i utiče na strukturu hijerarhijske dendrograme koja se generiše tokom ovog procesa.



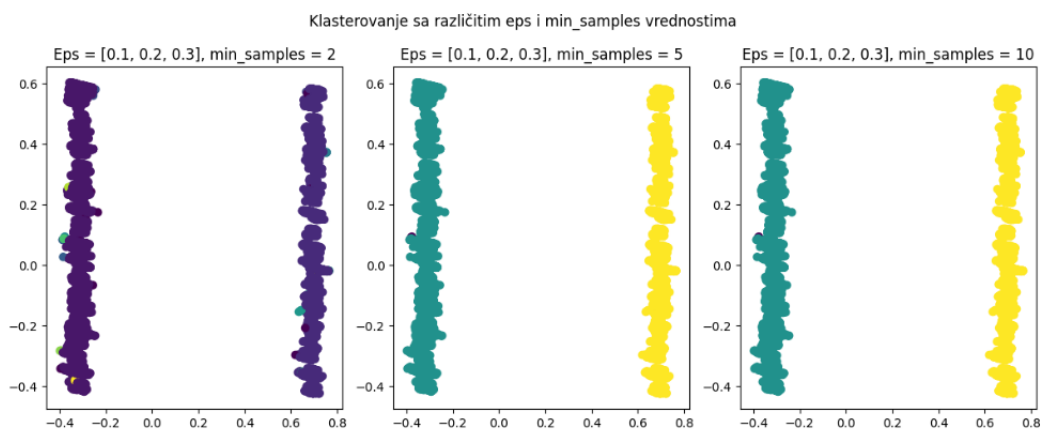
Slika 38: Silueta

I ovaj model, poput ostalih najbolje klasifikuje u 2 klastera.

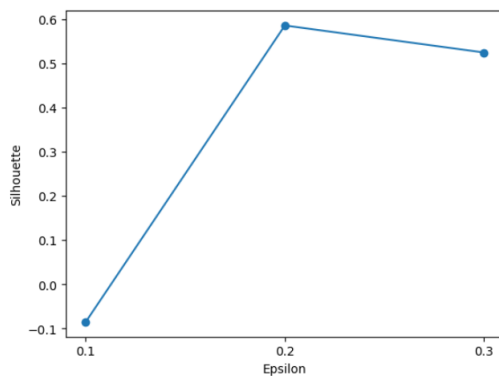
3.2.4 DBSCAN-a

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) je algoritam koji se posebno dobro nosi sa podacima koji mogu imati različite oblike i veličine klastera i sa prisustvom buke.

Radi tako što algoritam bira neobrađenu tačku iz skupa podataka i sve tačke koje su u njenoj epsilon-okolini. Ukoliko je broj tačaka u tom krugu veći od zadate min samples vrednosti sve te tačke pripadaju istom klasteru. Ovo se ponavlja za sve tačke iz skupa.



Slika 39: Klasterovanje pomoću DBSCAN



Slika 40: Zavisnost siluete od epsilon

Za epsilon = 0.2 i min samples = 5 model radi najbolje

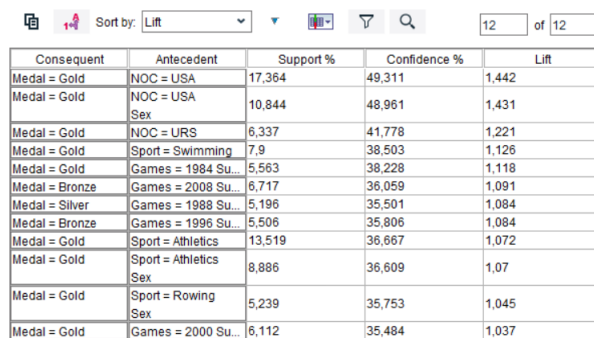
3.3 Pravila pridruživanja

Pravila pridruživanja su metoda u analizi podataka koja se koristi za otkrivanje zavisnosti i veza između različitih elemenata ili atributa u velikim skupovima podataka.

3.3.1 Apriori

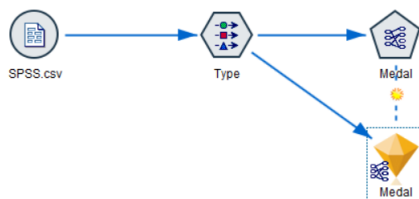
Ovaj algoritam pomaže u identifikaciji pravila pridruživanja između različitih atributa koji se često javljaju zajedno u transakcijama. Osnovna ideja Apriori algoritma je da koristi podršku (support) i pouzdanost (confidence) kako bi identifikovao česte kombinacije i pravila pridruživanja.

Algoritam prvo računa pouzdanost za sve proizvode, zatim za kombinaciju 2 takva proizvoda, pa 3, 4... Ukoliko je pouzdanost za neki od ovih manja od zadatog praga, ta kombinacija se odbacuje i njeni "potomci" se ne razmatraju. Ostali čvorovi se zovu česti čvorovi. Na osnovu čestih čvorova, algoritam generiše pravila pridruživanja koja se sastoje od antecedenta (prethodnika) i consequenta (posledice).



Consequent	Antecedent	Support %	Confidence %	Lift
Medal = Gold	NOC = USA	17,364	49,311	1,442
Medal = Gold	NOC = USA Sex	10,844	48,961	1,431
Medal = Gold	NOC = URS	6,337	41,778	1,221
Medal = Gold	Sport = Swimming	7,9	38,503	1,126
Medal = Gold	Games = 1984 Su...	5,563	38,228	1,118
Medal = Bronze	Games = 2008 Su...	6,717	36,059	1,091
Medal = Silver	Games = 1988 Su...	5,196	35,501	1,084
Medal = Bronze	Games = 1996 Su...	5,506	35,806	1,084
Medal = Gold	Sport = Athletics	13,519	36,667	1,072
Medal = Gold	Sport = Athletics Sex	8,886	36,609	1,07
Medal = Gold	Sport = Rowing Sex	5,239	35,753	1,045
Medal = Gold	Games = 2000 Su...	6,112	35,484	1,037

Slika 41: Pravila pridruživanja



Slika 42: Primena Apriori algoritma u SPSS-u

4 Zaključak

U okviru ovog projekta, detaljno smo istražili i analizirali različite modele u oblasti analize podataka i primenili ih na naš skup podataka Olimpijskih igara. Naš cilj je bio da istražimo obrasce i trendove u istorijskim rezultatima Olimpijskih igara kako bismo bolje razumeli faktore koji su uticali na uspeh sportista tokom vremena.

Primenili smo različite modele za predviđanje uspeha sportista na osnovu dostupnih atributa, neki su radili gore, a neki bolje nego očekivano i identifikovali smo one koji daju najbolje rezultate. Kroz analizu i vizualizaciju podataka, prikazali smo zanimljive obrasce u rezultatima i sada možemo sa predznanjem da pratimo Olimpijske igre u Parizu 2024 godine.

5 Reference

1. <https://github.com/MATF-istrazivanje-podataka-1/materijali-2022-2023>
2. <https://matplotlib.org/stable/index.html>