Olimpijske igre

Marko Paunović

Matematički fakultet
Univerzitet u Beogradu

15.9.2023.

Pregled

Uvod

	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: Izgled skupa podataka

Izbacivanje nedostajućih vrednosti

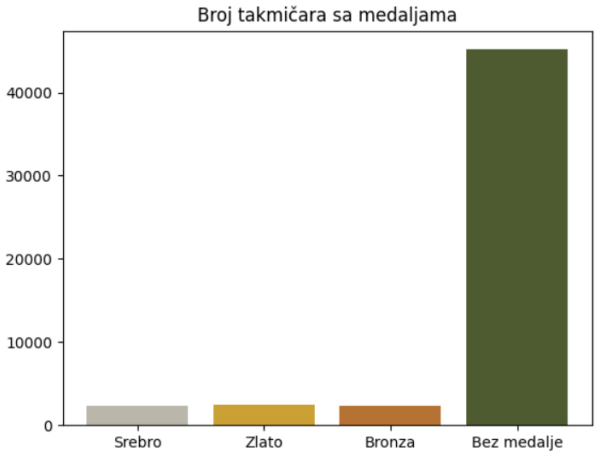
```
ID          0
Name        0
Sex         0
Age      2732
Height  16254
Weight  17101
Team        0
NOC         0
Games       0
Year        0
Season      0
City        0
Sport       0
Event       0
Medal   60310
dtype: int64
```

Slika: Broj nedostajućih vrednosti pre

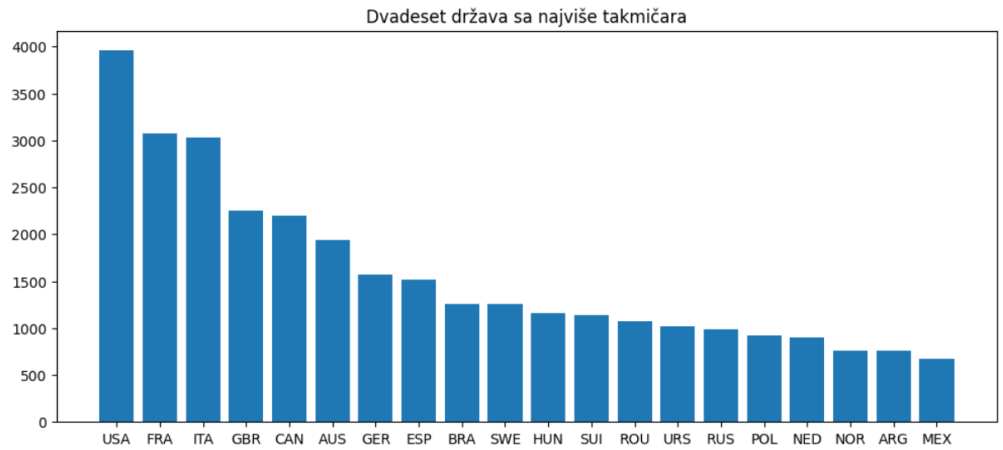
```
ID          0
Name        0
Sex         0
Age         0
Height      0
Weight      0
Team        0
NOC         0
Games       0
Year        0
Season      0
City        0
Sport       0
Event       0
Medal   45165
dtype: int64
```

Slika: i posle izbacivanja

Uvod



Slika: Broj takmičara sa medaljama



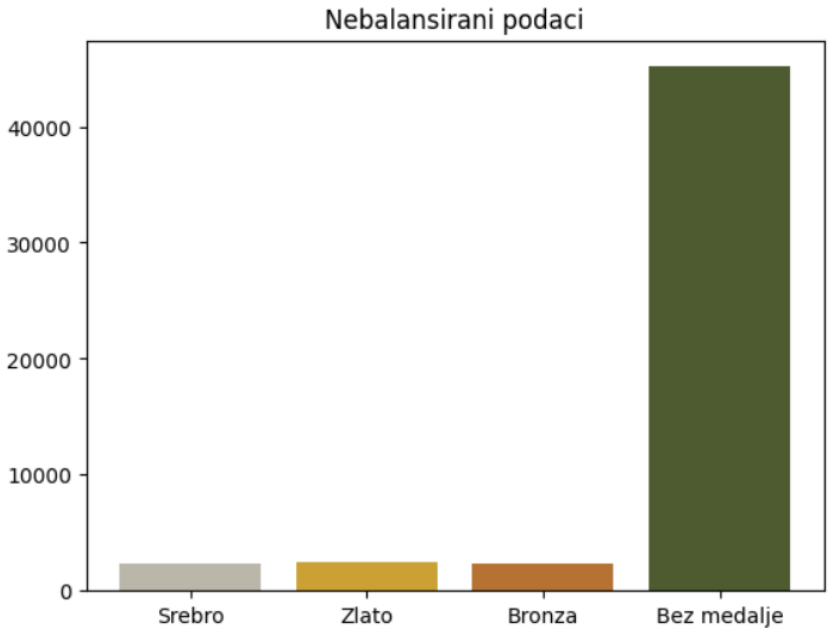
Slika: 20 država sa najviše takmičara

Kodiranje kategoričkih atributa

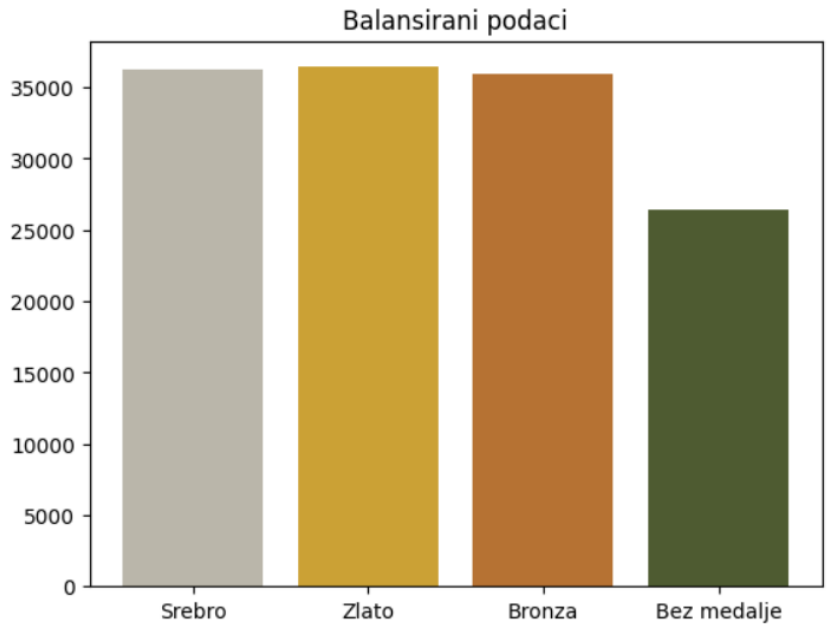
	Sex	Age	Height	Weight	NOC	Games	Sport
0	0	24.0	180.0	80.0	40	0	6
1	0	23.0	170.0	60.0	40	0	26
2	1	21.0	185.0	82.0	140	1	42
3	1	21.0	185.0	82.0	140	1	42
4	1	25.0	185.0	82.0	140	1	42

Slika: Kodirani kategorički atributi

Balansiranje podataka



Slika: Nebalansirani podaci



Slika: Balansirani podaci

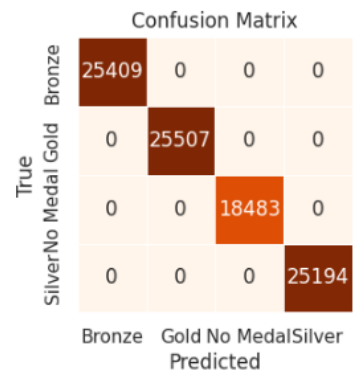
Standardizacija podataka

	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: Skalirani podaci za klasterovanje

Stabla odlučivanja

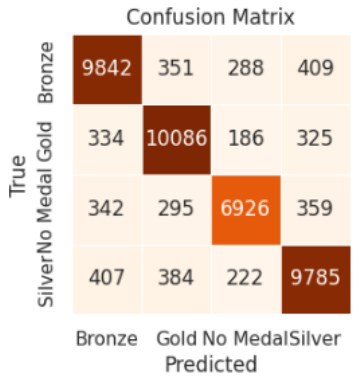
Train data:




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

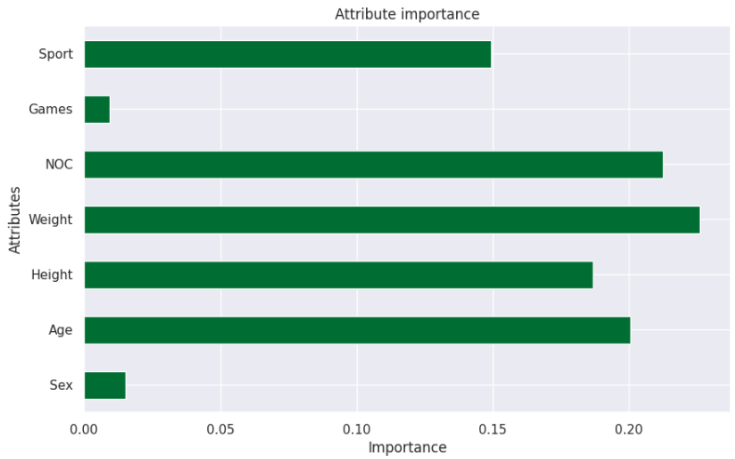
Slika: Matrica konfuzije

Test data:



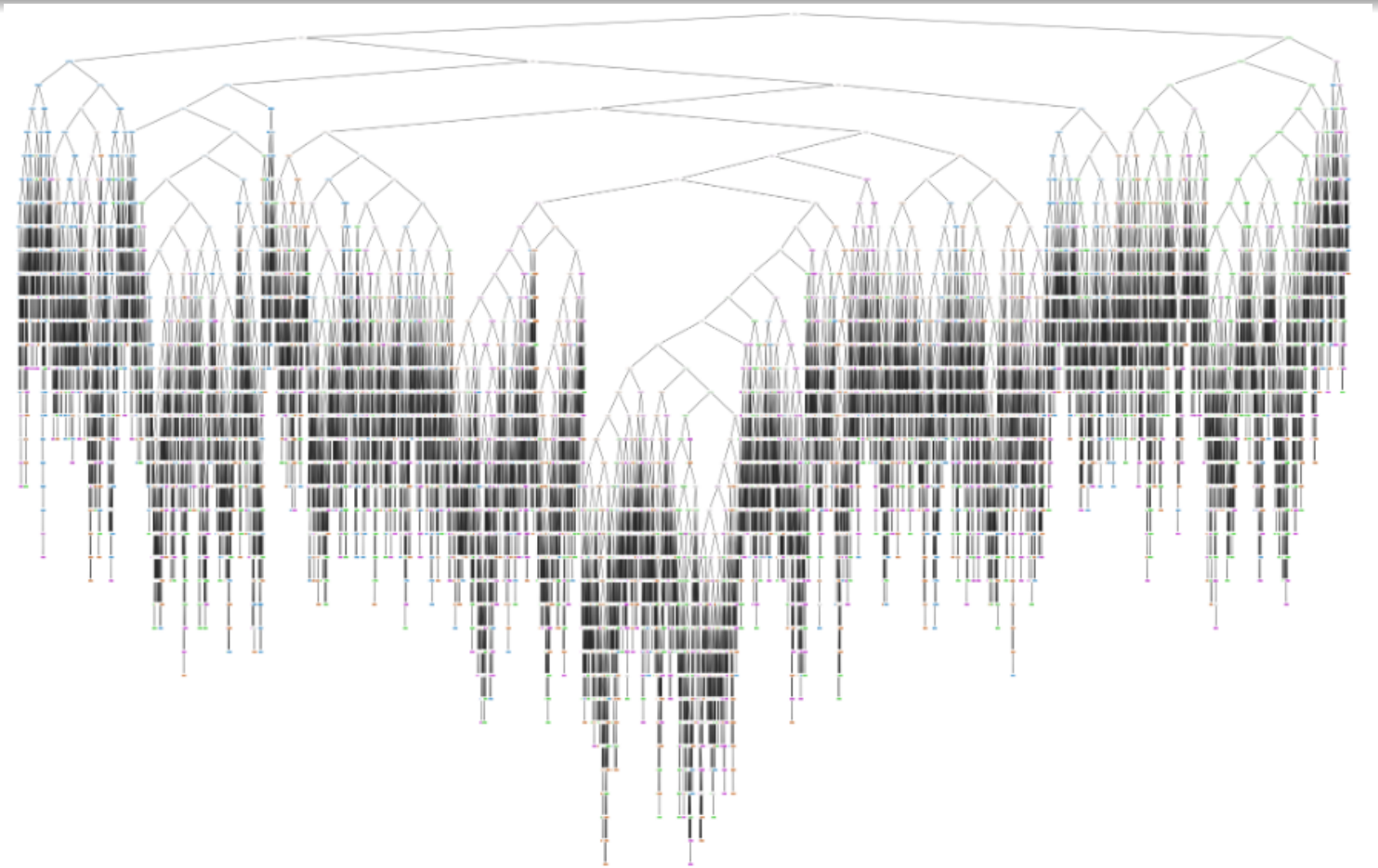

```
Accuracy score:  0.9037517574800819
Precision score:  0.9041044226606616
Recall score:  0.901730585740576
F1 score:  0.902814724740074
```

Slika: Matrica konfuzije



Slika: Uticaj atributa

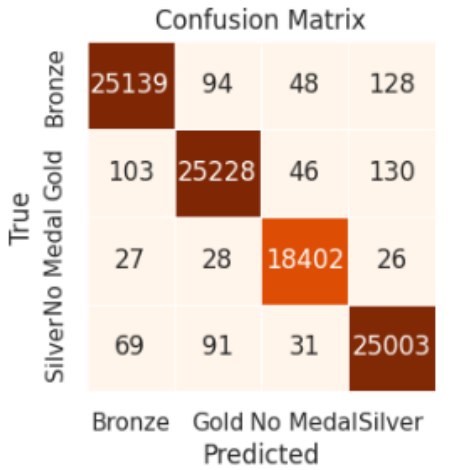
Stabla odlučivanja



Slika: Izgled stabla odlučivanja

Optimizovana stabla odlučivanja

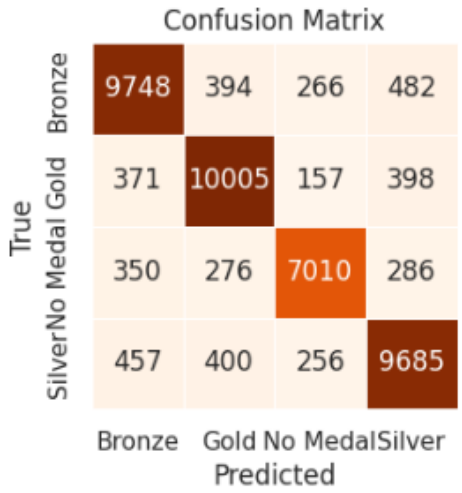
Train data:




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

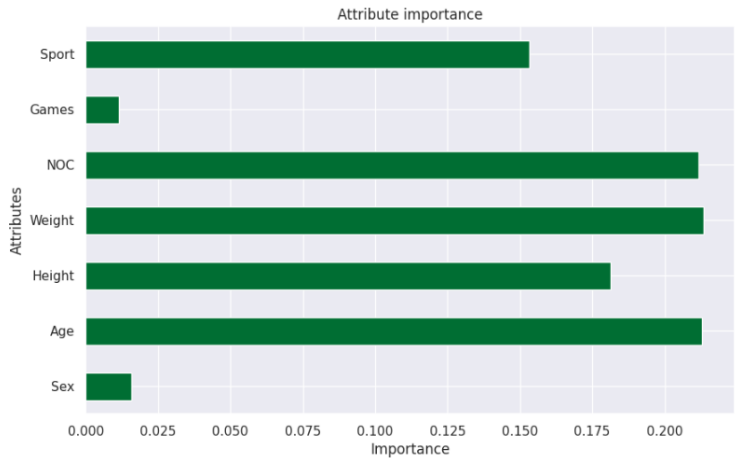
Slika: Matrica konfuzije

Test data:




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

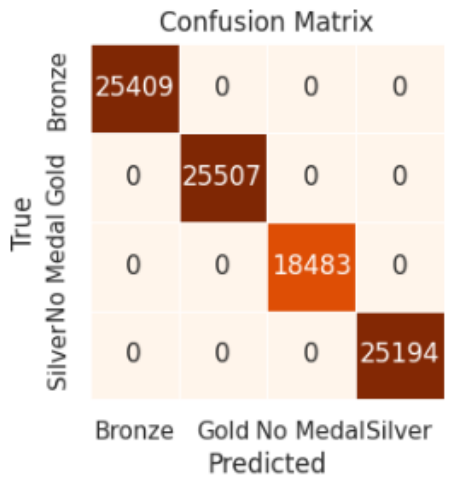
Slika: Matrica konfuzije



Slika: Uticaj atributa

Slučajne šume

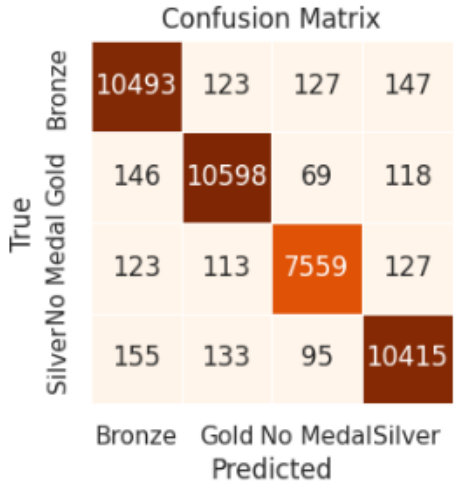
Train data:




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

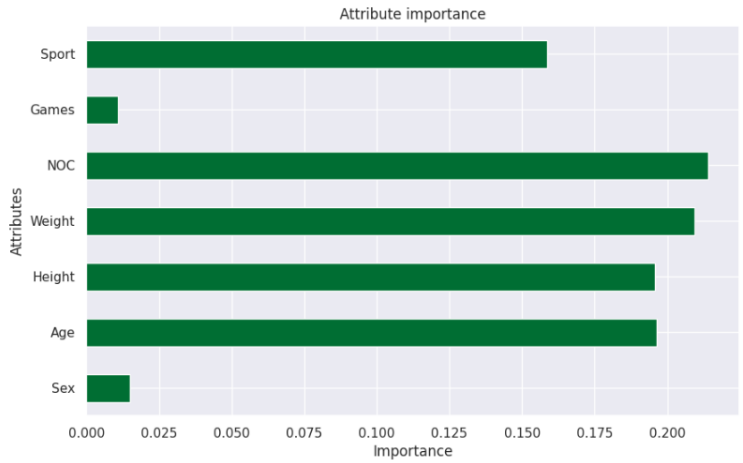
Slika: Matrica konfuzije

Test data:




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

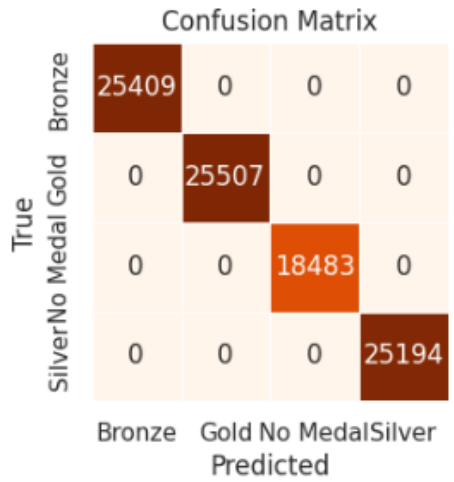
Slika: Matrica konfuzije



Slika: Uticaj atributa

Optimizovane slučajne šume

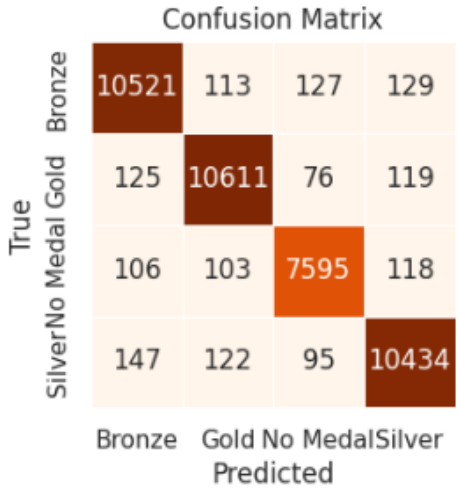
Train data:




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

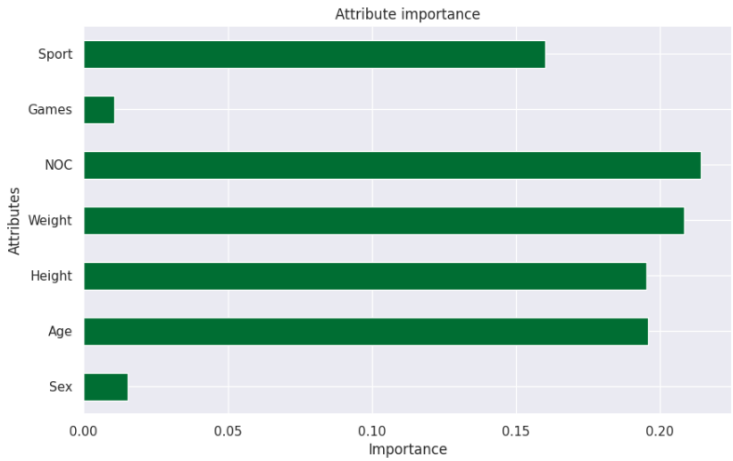
Slika: Matrica konfuzije

Test data:



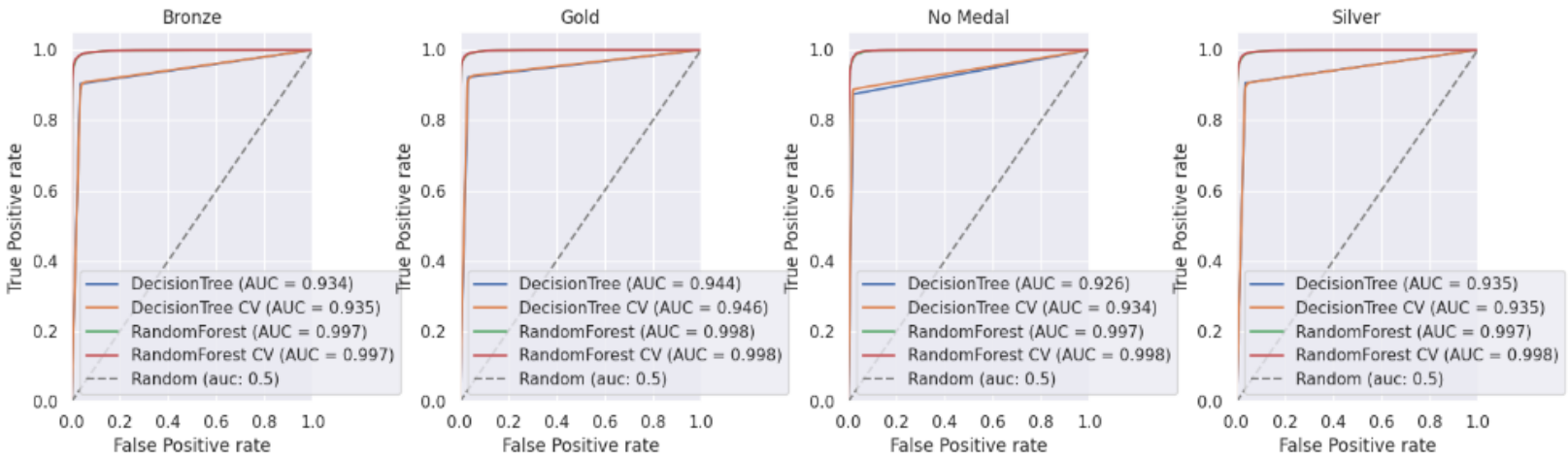

```
Accuracy score:  0.9659603857822945
Precision score:  0.9657009145755936
Recall score:  0.9654634401766679
F1 score:  0.9655811626081118
```

Slika: Matrica konfuzije



Slika: Uticaj atributa

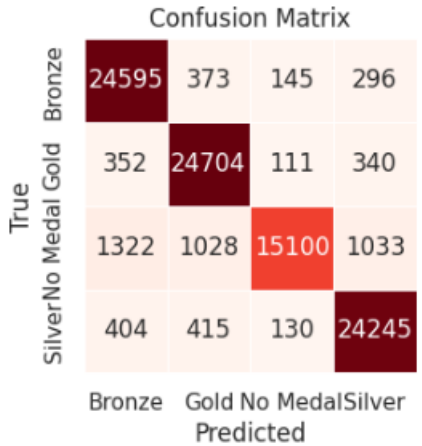
ROC kriva



Slika: ROC kriva za višeklasnu klasifikaciju

K najbližih suseda i optimizacaija

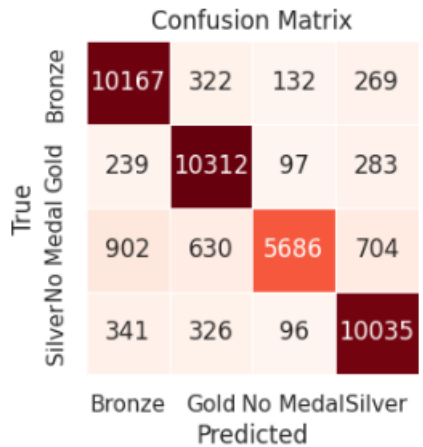
Train data:




```
Accuracy score:    0.9371095112746186
Precision score:   0.9410714503138327
Recall score:      0.9289454492651009
F1 score:          0.9329884922948736
```

Slika: Matrica konfuzije

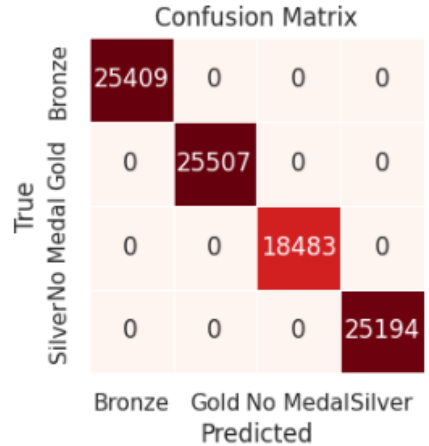
Test data:




```
Accuracy score:    0.8929232135369133
Precision score:   0.8993011811385817
Recall score:      0.8810169215998018
F1 score:          0.8856814731992828
```

Slika: Matrica konfuzije

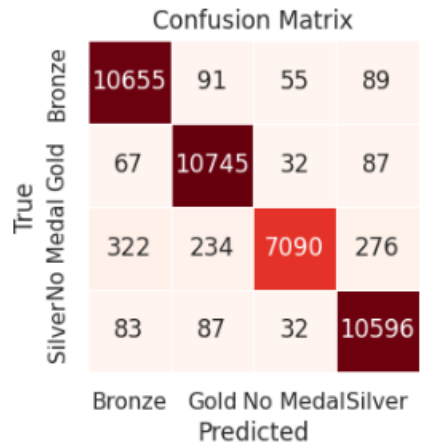
Train data:




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

Slika: Matrica konfuzije

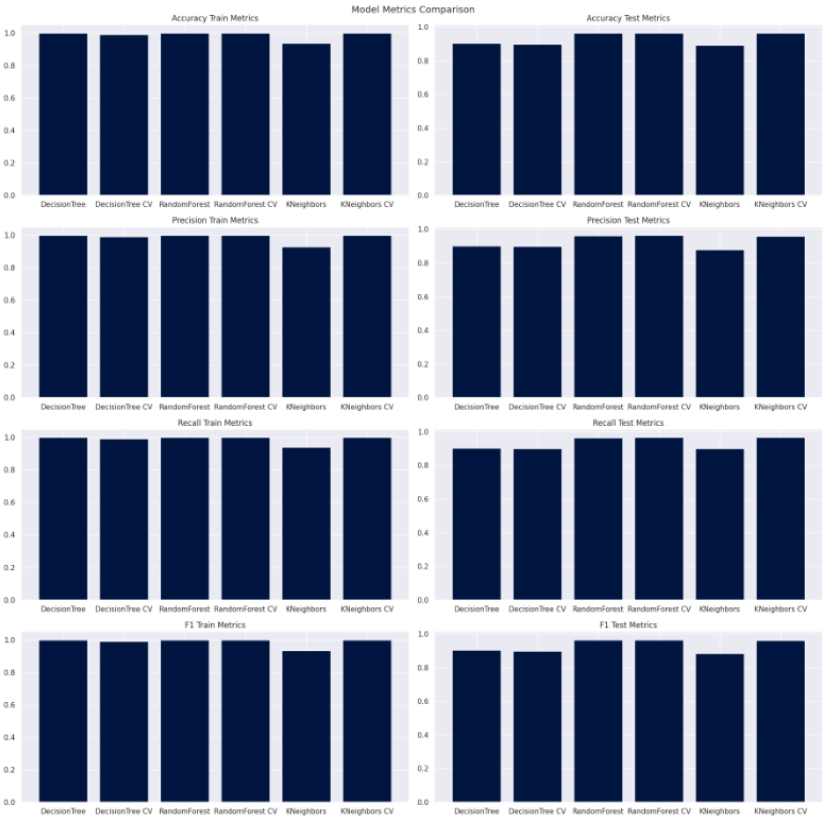
Test data:

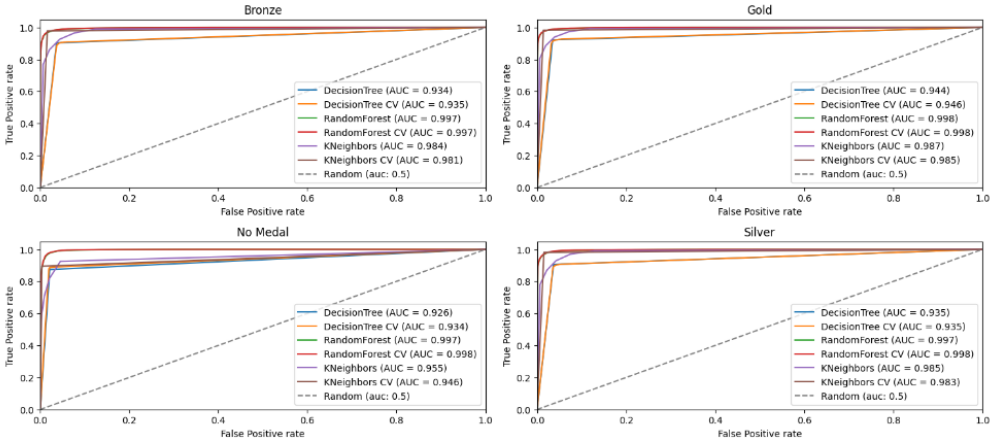



```
Accuracy score:    0.9641104067487235
Precision score:   0.965808411002146
Recall score:      0.9594183977361936
F1 score:          0.9620067487044386
```

Slika: Matrica konfuzije

Poređenje modela

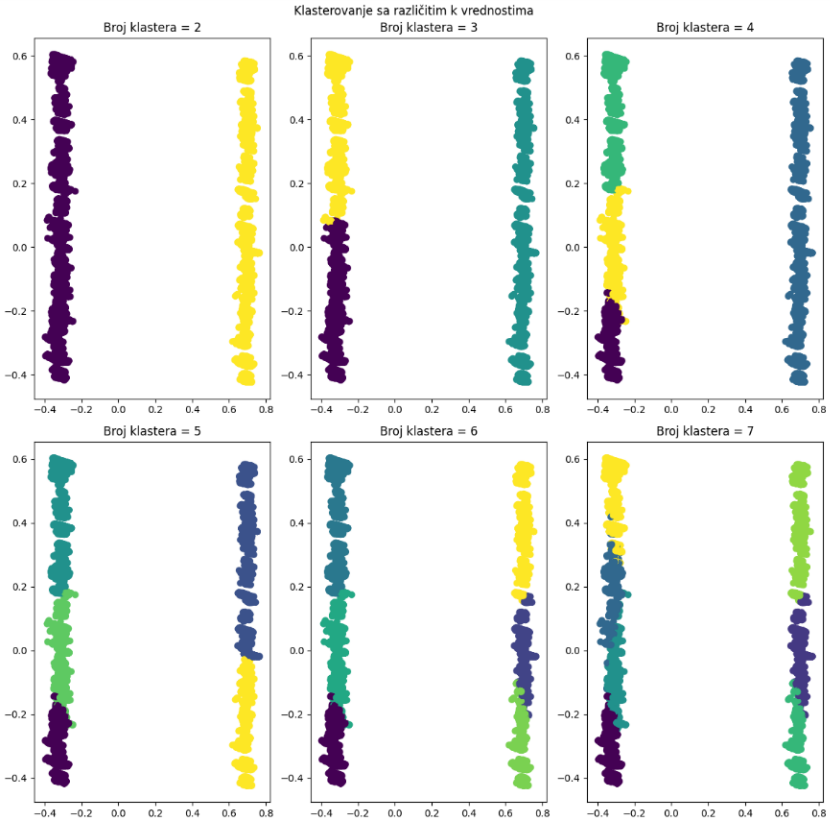


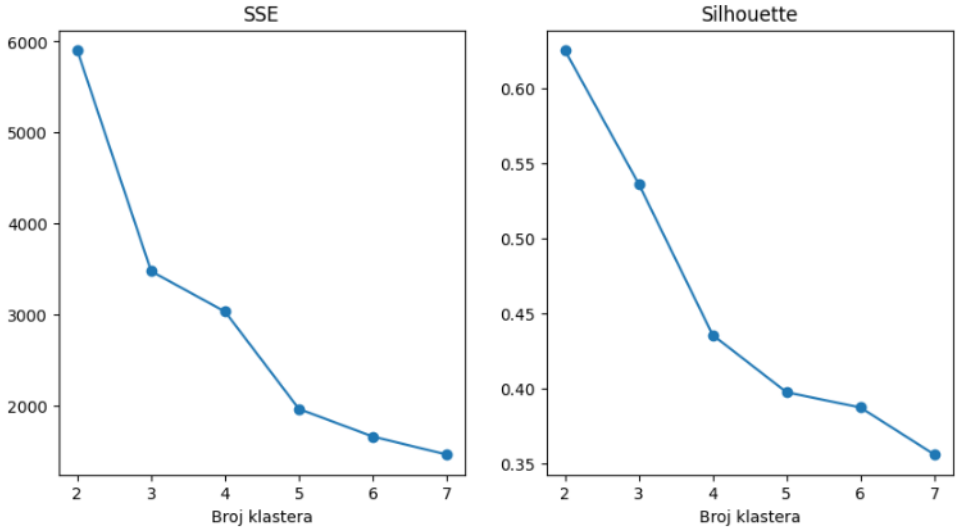


K-sredina

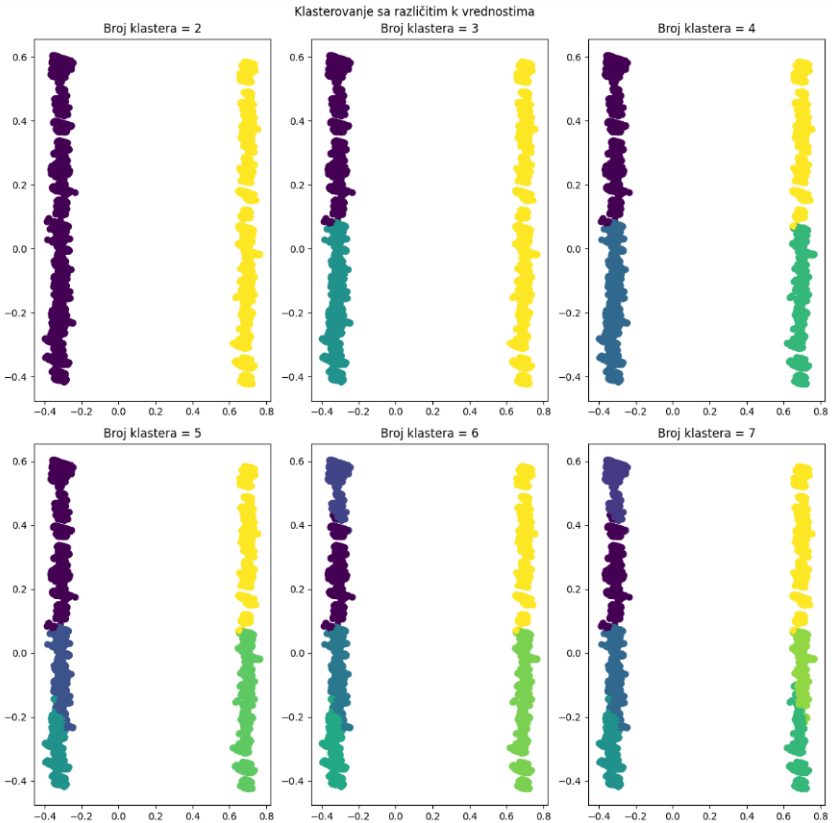

Matrica Korelacije
Sex
Age
Height
Weight
NOC
Games
Sport
1.00 -0.16 -0.49 -0.51 -0.01 0.01 0.01
-0.16 1.00 0.13 0.21 -0.00 0.00 -0.05
-0.49 0.13 1.00 0.78 0.01 -0.03 0.03
-0.51 0.21 0.78 1.00 0.02 0.00 0.09
-0.01 -0.00 0.01 0.02 1.00 0.06 -0.01
0.01 0.00 -0.03 0.00 0.06 1.00 -0.12
0.01 -0.05 0.03 0.09 -0.01 -0.12 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

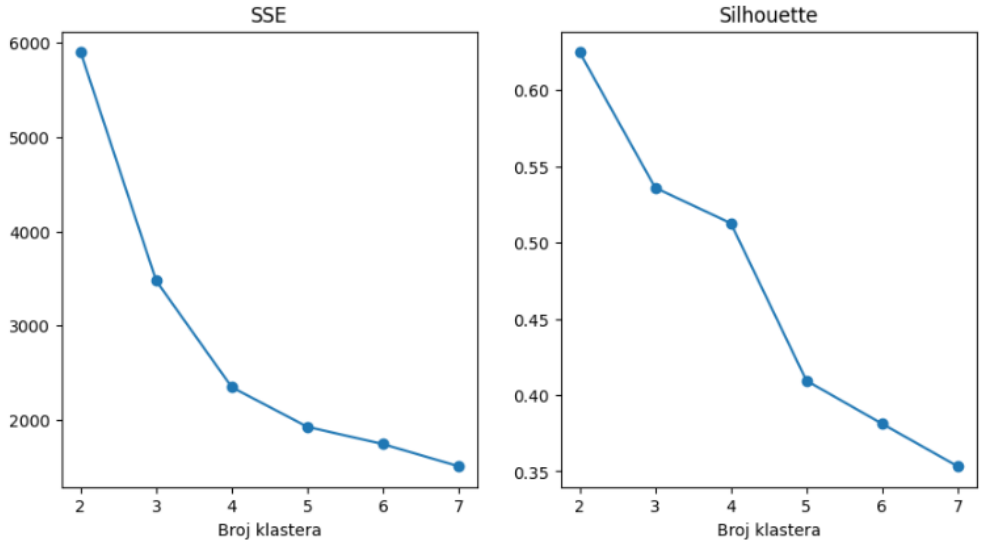
K-sredina


Klasterovanje sa različitim k vrednostima
Broj klastera = 2
Broj klastera = 3
Broj klastera = 4
Broj klastera = 5
Broj klastera = 6
Broj klastera = 7

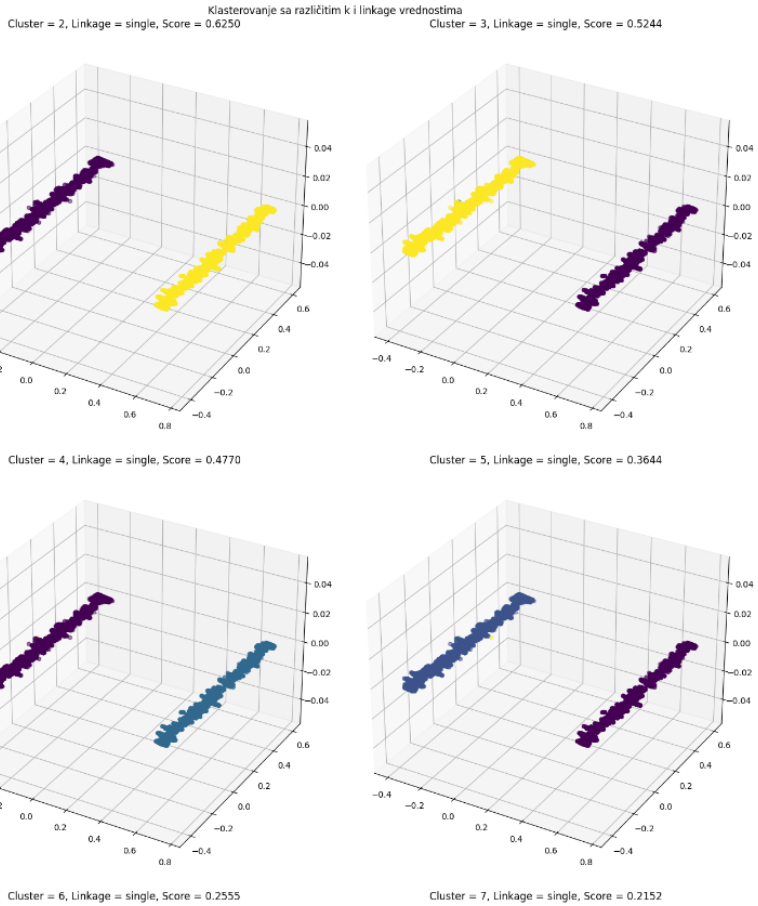

SSE
Broj klastera
Silhouette
Broj klastera

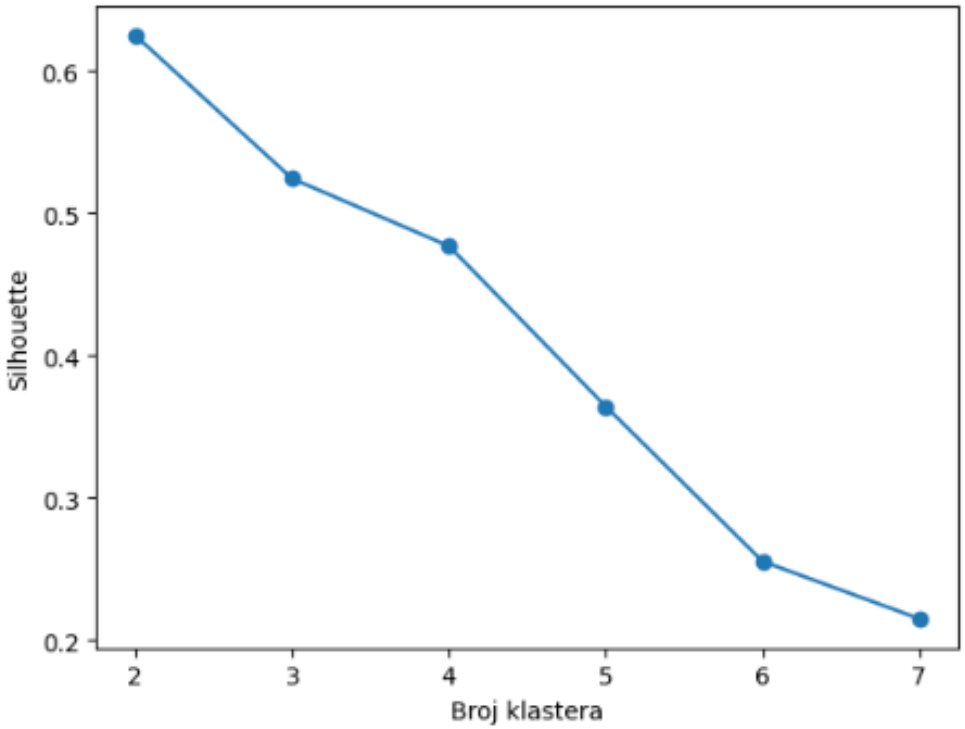
K-sredina sa podelom



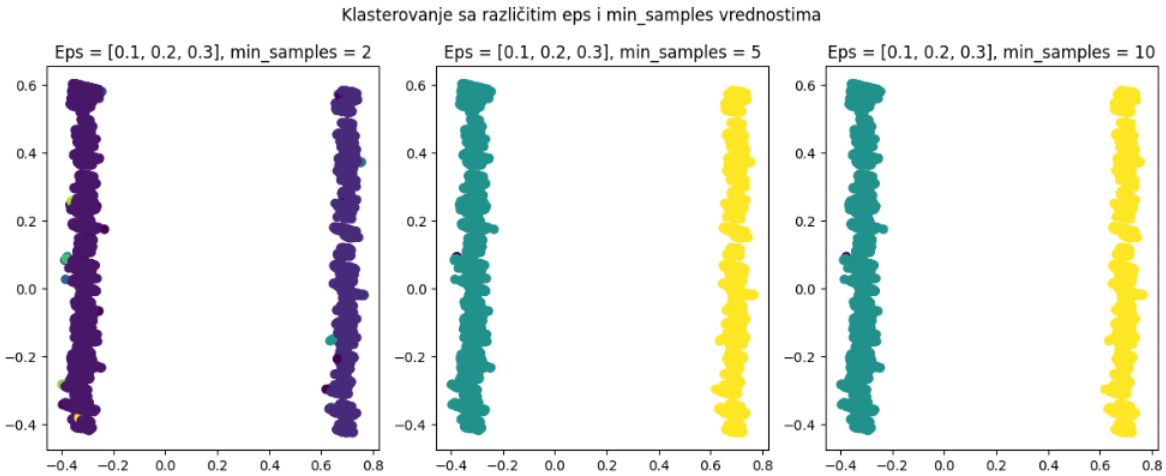


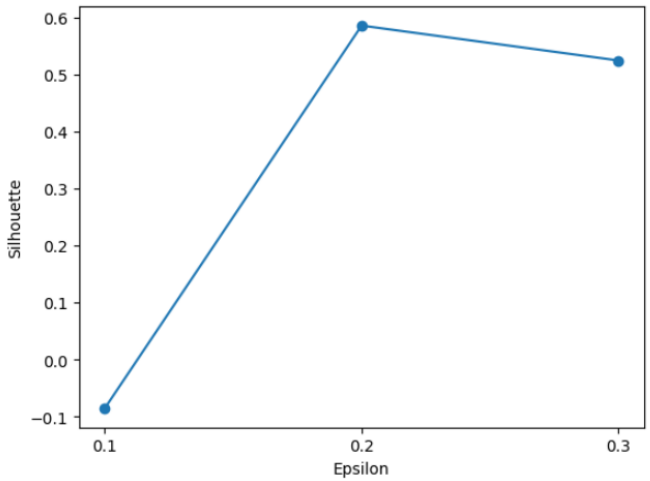
Aglomerativno klasterovanje





DBSCAN


Klasterovanje sa različitim eps i min_samples vrednostima
Eps = [0.1, 0.2, 0.3], min_samples = 2
Eps = [0.1, 0.2, 0.3], min_samples = 5
Eps = [0.1, 0.2, 0.3], min_samples = 10

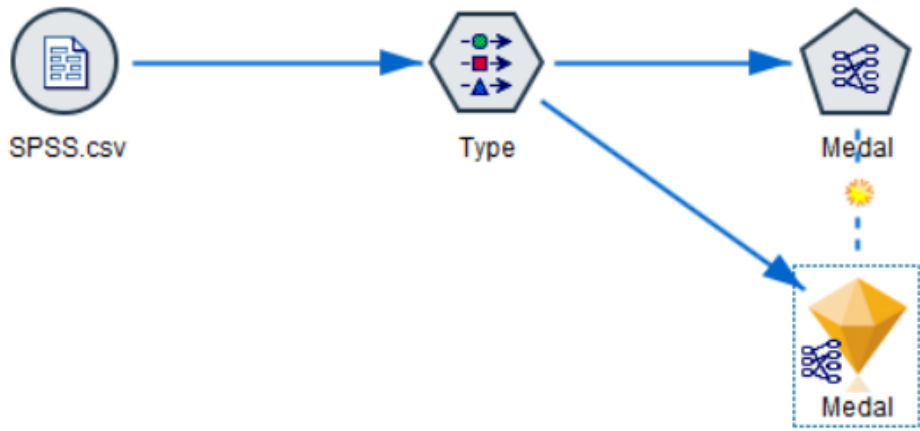

Silhouette
Epsilon

Pravila pridruživanja - Apriori

Sort by: Lift 12 of 12

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: Pravila pridruživanja



Slika: SPSS

Hvala na pažnji!