



Travel

REVIEW DATASET

ISTRAŽIVANJE
PODATAKA 1

MATEMATIČKI FAKULTET

JELENA MAKSIMOVIĆ 193/2018

PROFESOR: NENAD MITIĆ
ASISTENT: STEFAN KAPUNAC

ANALIZA PODATAKA

SKUP PODATAKA "TRAVEL REVIEW RATINGS" SASTOJI SE OD 5456 INSTANCI I 25 ATRIBUTA.

- CATEGORY 1: JEDINSTVENI KORISNIČKI IDENTIFIKATOR
- CATEGORY 2: PROSEČNA OCENA KORISNIKA ZA CRKVE
- CATEGORY 3: PROSEČNA OCENA KORISNIKA ZA ODMARALIŠTA
- CATEGORY 4: PROSEČNA OCENA KORISNIKA ZA PLAŽE
- CATEGORY 5: PROSEČNA OCENA KORISNIKA ZA PARKOVE
- CATEGORY 6: PROSEČNA OCENA KORISNIKA ZA POZORIŠTA
- CATEGORY 7: PROSEČNA OCENA KORISNIKA ZA MUZEJE
- CATEGORY 8: PROSEČNA OCENA KORISNIKA ZA TRŽNE CENTRE
- CATEGORY 9: PROSEČNA OCENA KORISNIKA ZA ZOO VRTOVE
- CATEGORY 10: PROSEČNA OCENA KORISNIKA ZA RESTORANE
- CATEGORY 11: PROSEČNA OCENA KORISNIKA ZA KAFIĆE/PUB-OVE
- CATEGORY 12: PROSEČNA OCENA KORISNIKA ZA LOKALNE USLUGE
- CATEGORY 13: PROSEČNA OCENA KORISNIKA ZA BRZE HRANE/PICERIJE
- CATEGORY 14: PROSEČNA OCENA KORISNIKA ZA HOTELE I DRUGE SMEŠTAJNE OBJEKTE
- CATEGORY 15: PROSEČNA OCENA KORISNIKA ZA BAROVE SA SOKOVIMA
- CATEGORY 16: PROSEČNA OCENA KORISNIKA ZA GALERIJE UMETNOSTI

- CATEGORY 17: PROSEČNA OCENA KORISNIKA ZA NOĆNE KLUBOVE
- CATEGORY 18: PROSEČNA OCENA KORISNIKA ZA BAZENE
- CATEGORY 19: PROSEČNA OCENA KORISNIKA ZA TERETANE
- CATEGORY 20: PROSEČNA OCENA KORISNIKA ZA PEKARE
- CATEGORY 21: PROSEČNA OCENA KORISNIKA ZA SALONE LEPOTE I SPA CENTRE
- CATEGORY 22: PROSEČNA OCENA KORISNIKA ZA KAFIĆE
- CATEGORY 23: PROSEČNA OCENA KORISNIKA ZA VIDIKOVCE
- CATEGORY 24: PROSEČNA OCENA KORISNIKA ZA SPOMENIKE
- CATEGORY 25: PROSEČNA OCENA KORISNIKA ZA BAŠTE

	User	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8	Category 9	...	Category 16	Category 17	Category 18	Category 19	Category 20
0	User 1	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33	...	0.59	0.50	0.00	0.50	0.00
1	User 2	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00	0.50	0.00
2	User 3	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00	0.50	0.00
3	User 4	0.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33	...	0.59	0.50	0.00	0.50	0.00
4	User 5	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00	0.50	0.00
...
5451	User 5452	0.91	5.00	4.00	2.79	2.77	2.57	2.43	1.09	1.77	...	0.66	0.65	0.66	0.69	5.00
5452	User 5453	0.93	5.00	4.02	2.79	2.78	2.57	1.77	1.07	1.76	...	0.65	0.64	0.65	1.59	1.00
5453	User 5454	0.94	5.00	4.03	2.80	2.78	2.57	1.75	1.05	1.75	...	0.65	0.63	0.64	0.74	5.00
5454	User 5455	0.95	4.05	4.05	2.81	2.79	2.44	1.76	1.03	1.74	...	0.64	0.63	0.64	0.75	5.00
5455	User 5456	0.95	4.07	5.00	2.82	2.80	2.57	2.42	1.02	1.74	...	0.64	0.62	0.63	0.78	5.00

5456 rows × 26 columns

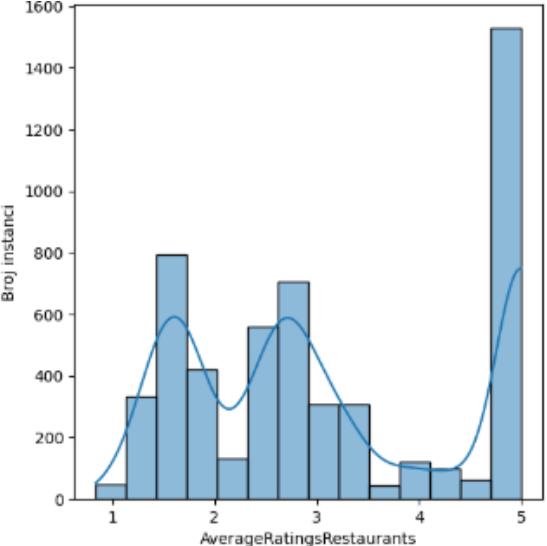
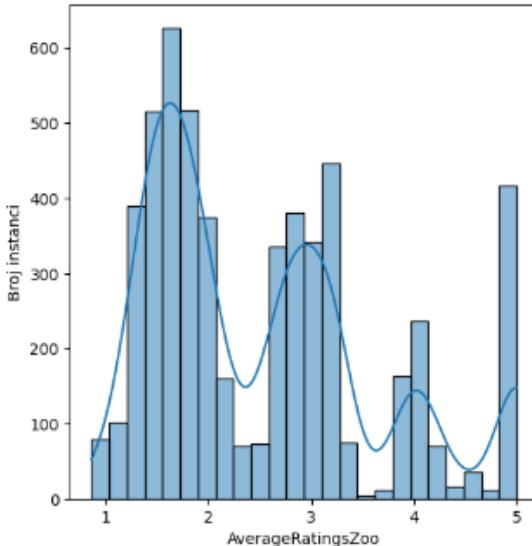
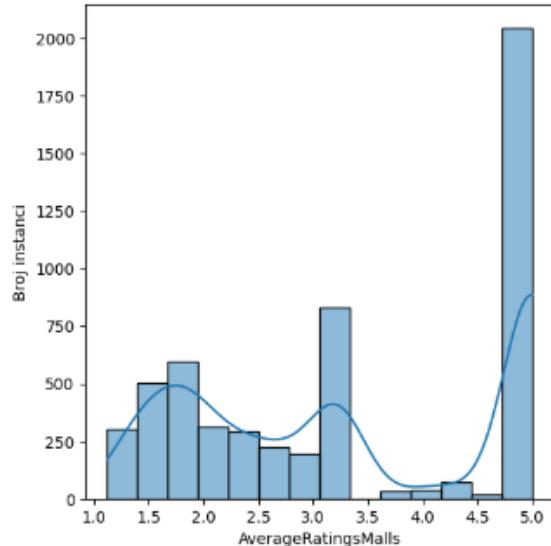
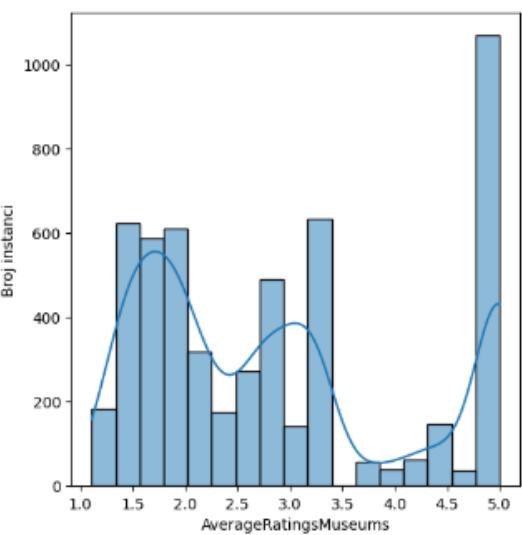
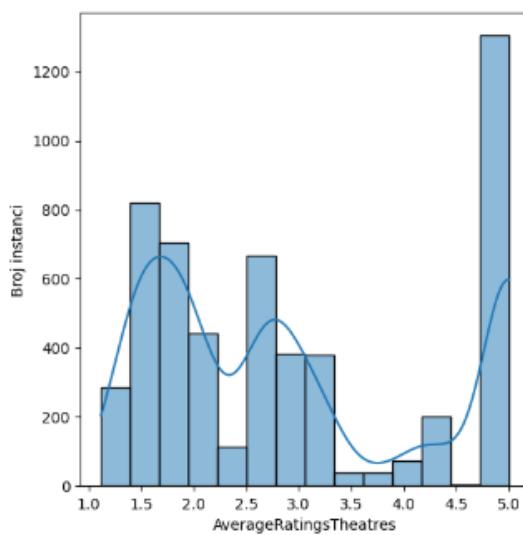
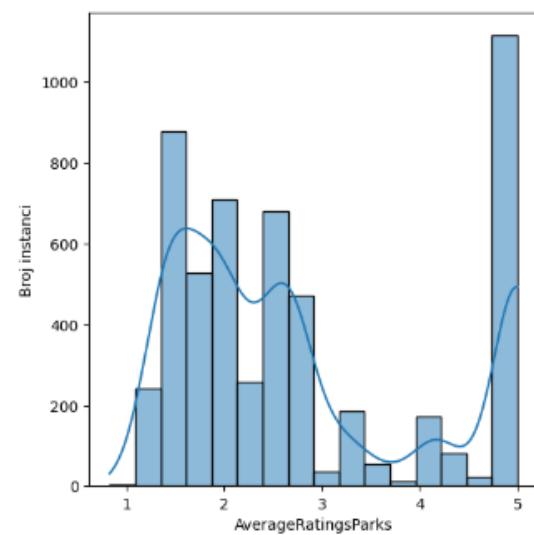
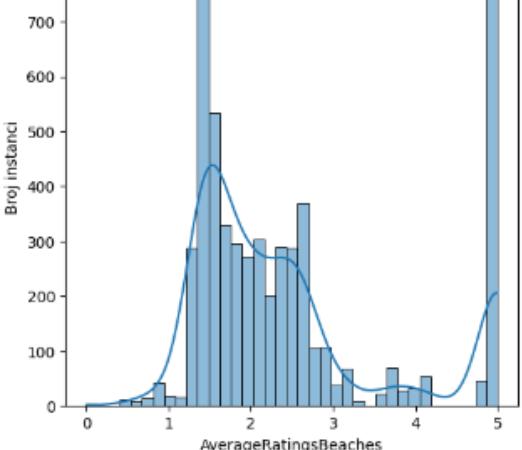
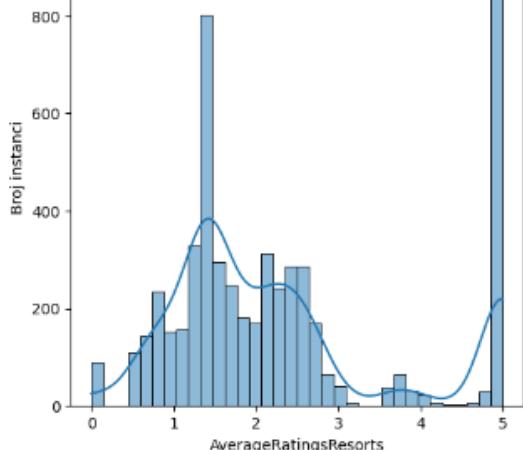
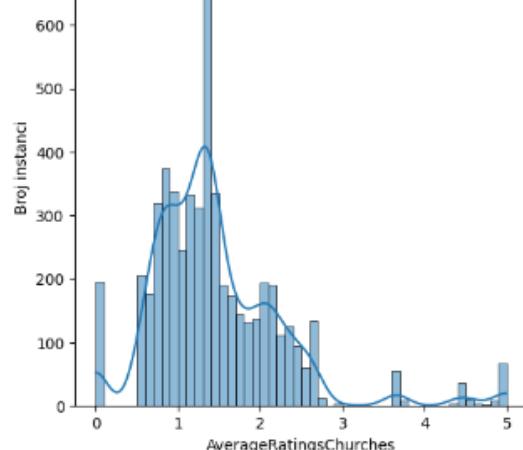
PRETPROCESIRANJE

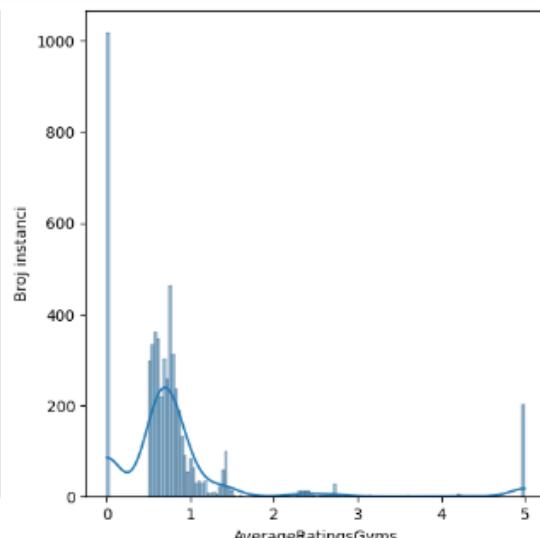
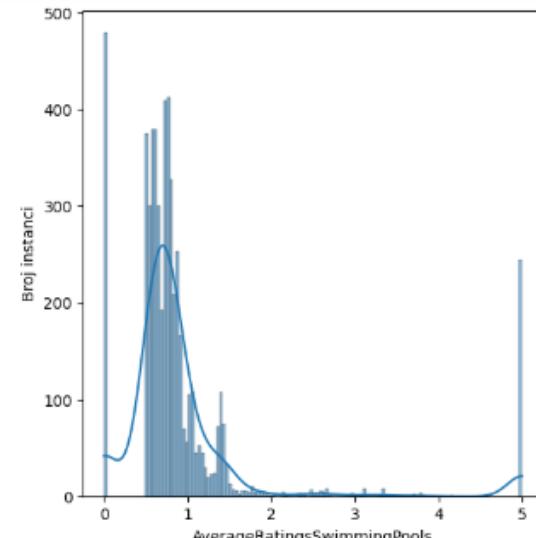
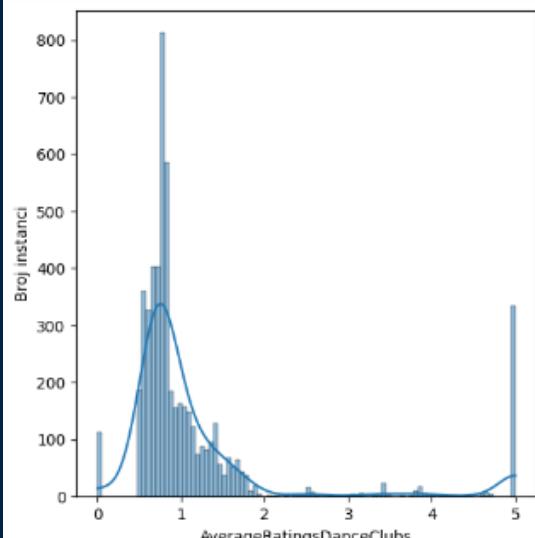
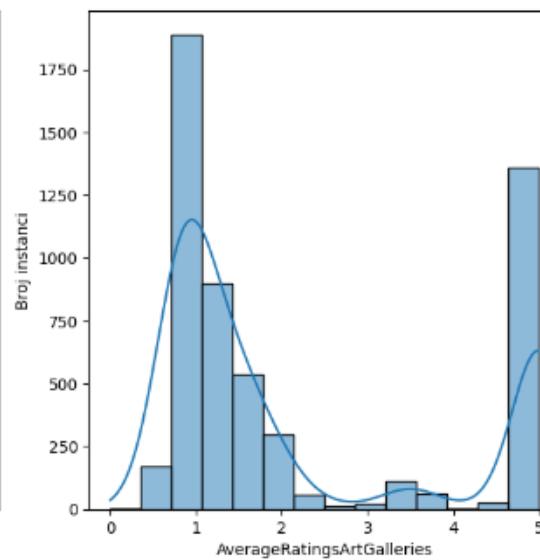
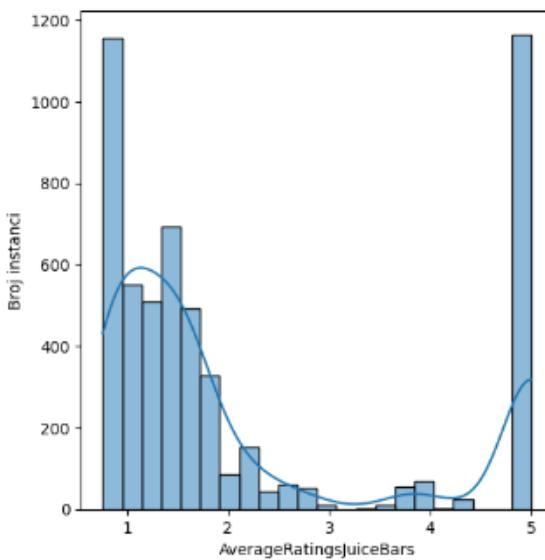
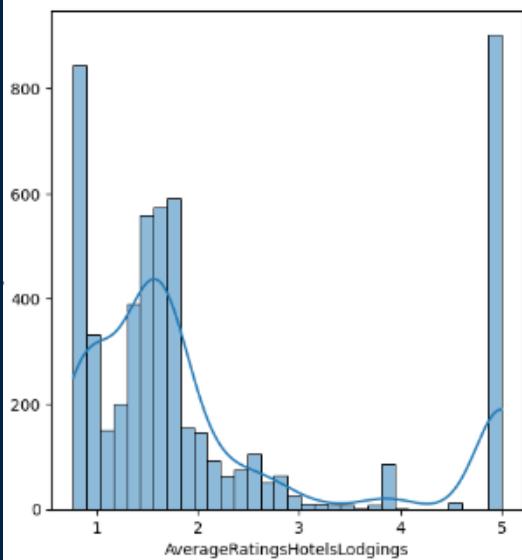
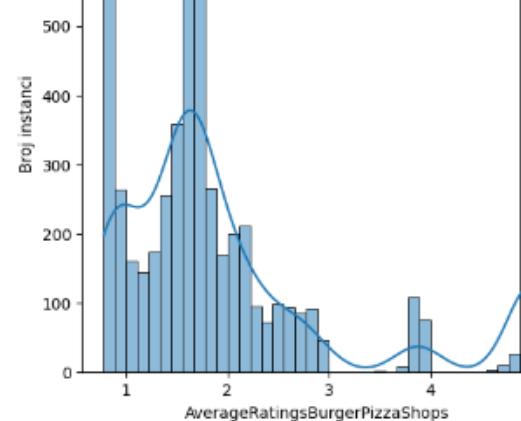
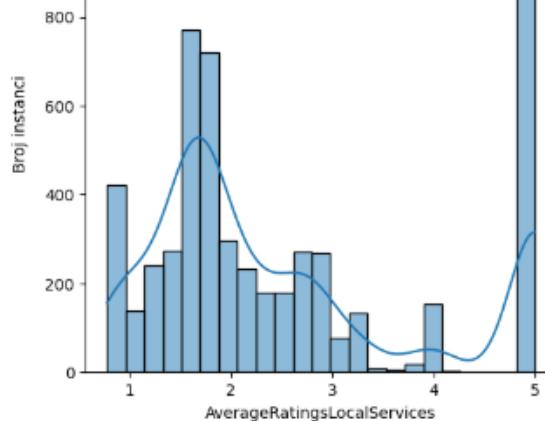
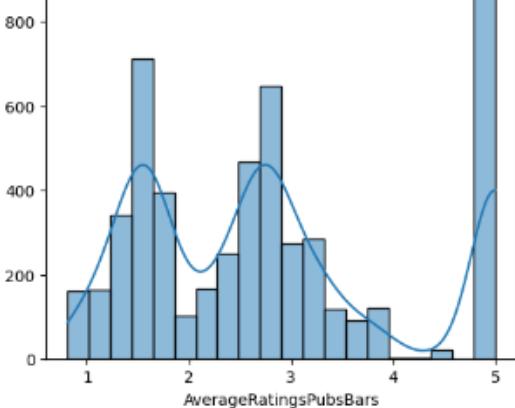
- KONVERZIJA OCENA U NUMERIČKE VREDNOSTI
- PROVERA I UKLANJANJE NEPOTREBNIH KOLONA ('UNNAMED: 25', 'USER')
- PREIMENOVANJE KOLONA
- STATISTIČKA ANALIZA NUMERIČKIH KOLONA
- RUKOVANJE SA NAN VREDNOSTIMA
- PROVERA AUTLAJERA I EKSTREMNIH VREDNOSTI
- VIZUALIZACIJA NUMERIČKIH KOLONA

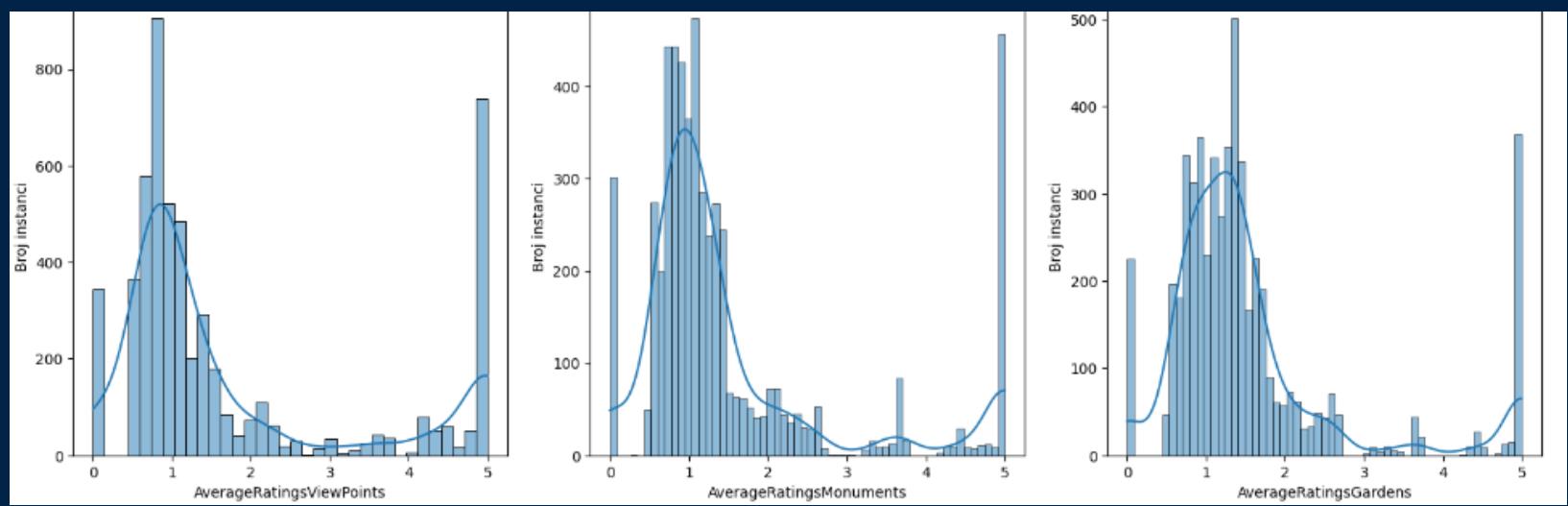
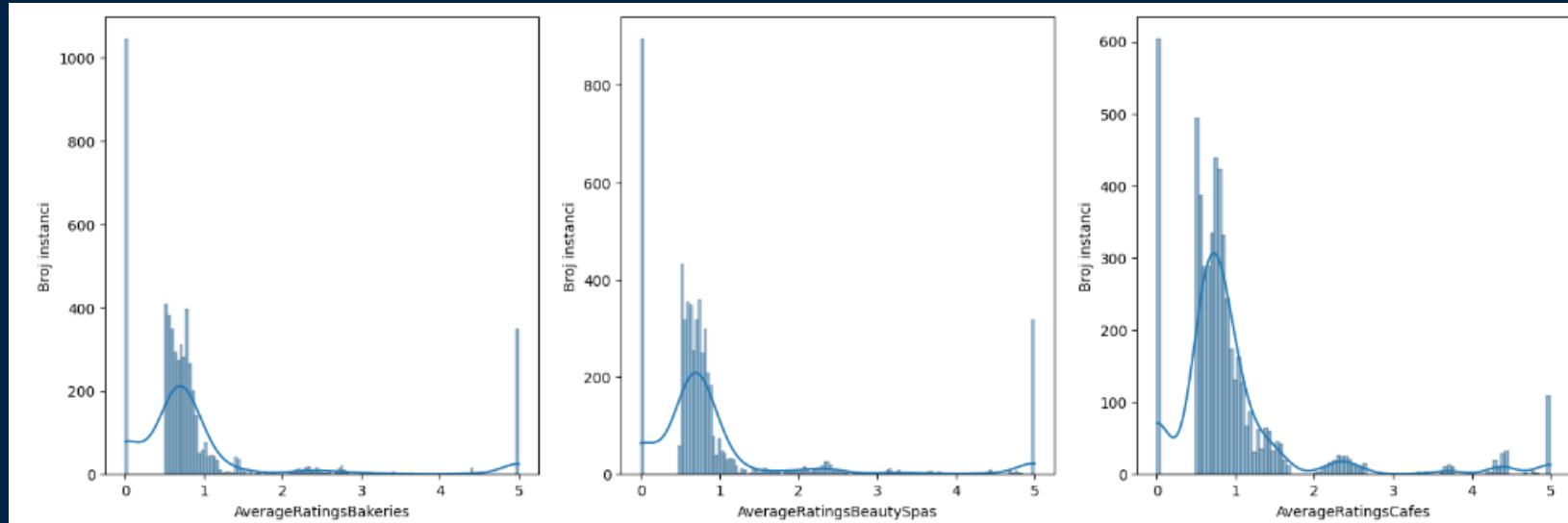


	AverageRatingsChurches	AverageRatingsResorts	AverageRatingsBeaches	AverageRatingsParks	AverageRatingsTheatres	AverageRatingsMuseums	AverageRatingsArtGalleries
count	5456.000000	5456.000000	5456.000000	5456.000000	5456.000000	5456.000000	5456.000000
mean	1.455720	2.319707	2.489331	2.796886	2.958941	2.89349	3.000000
std	0.827604	1.421438	1.247815	1.309159	1.339056	1.28240	1.300000
min	0.000000	0.000000	0.000000	0.830000	1.120000	1.11000	1.100000
25%	0.920000	1.360000	1.540000	1.730000	1.770000	1.79000	1.800000
50%	1.340000	1.905000	2.060000	2.460000	2.670000	2.68000	2.700000
75%	1.810000	2.682500	2.740000	4.092500	4.312500	3.84000	4.000000
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

8 rows × 24 columns









MATRICA KORELACIJE

KORISTI SE ZA VIZUALIZACIJU VEZA IZMEĐU ATRIBUTA

Matrica korelaciјe

AverageRatingsChurches	1.00	0.25	0.15	0.07	0.04	-0.09	-0.26	-0.18	-0.29	-0.27	-0.21	-0.26	-0.18	-0.24	-0.13	0.07	0.13	0.16	0.18	0.20	0.31	0.36	0.41	0.49
AverageRatingsResorts	0.25	1.00	0.33	0.17	0.15	0.05	-0.05	-0.00	-0.05	-0.09	-0.22	-0.16	-0.21	-0.13	-0.07	-0.03	-0.08	-0.03	0.03	0.10	0.09	0.02	0.08	0.13
AverageRatingsBeaches	0.15	0.33	1.00	0.40	0.33	0.16	-0.07	-0.19	-0.22	-0.18	-0.16	-0.24	-0.18	-0.16	-0.13	-0.02	-0.08	-0.12	-0.08	-0.01	0.00	0.13	0.12	0.08
AverageRatingsParks	0.07	0.17	0.40	1.00	0.63	0.32	-0.07	-0.13	-0.17	-0.12	-0.12	-0.17	-0.15	-0.31	-0.27	0.02	-0.13	-0.18	-0.19	-0.09	-0.06	0.28	0.17	0.09
AverageRatingsTheatres	0.04	0.15	0.33	0.63	1.00	0.49	0.08	0.00	0.17	0.10	0.12	0.10	0.09	0.28	0.32	-0.06	-0.18	0.24	-0.26	-0.19	0.13	0.12	0.13	0.10
AverageRatingsMuseums	-0.09	0.05	0.18	0.32	0.49	1.00	0.36	0.20	0.11	-0.02	-0.15	-0.16	-0.14	-0.15	-0.19	-0.15	-0.23	-0.27	-0.27	-0.23	-0.20	-0.09	-0.08	-0.07
AverageRatingsMalls	-0.26	-0.05	-0.07	-0.07	0.08	0.39	1.00	0.41	0.43	0.26	0.10	0.03	0.03	0.09	0.09	-0.14	-0.21	-0.23	-0.27	-0.23	-0.26	-0.36	-0.22	-0.25
AverageRatingsZoo	-0.18	-0.00	-0.19	-0.13	-0.00	0.20	0.41	1.00	0.54	0.55	0.29	0.00	-0.01	-0.02	-0.06	-0.12	-0.20	-0.24	-0.26	-0.25	-0.27	-0.26	-0.17	-0.14
AverageRatingsRestaurants	-0.29	-0.05	-0.22	-0.17	-0.17	0.11	0.43	0.54	1.00	0.56	0.26	-0.01	0.02	0.03	0.13	-0.12	-0.23	-0.27	-0.27	-0.16	-0.19	-0.27	-0.27	-0.33
AverageRatingsPubsBars	-0.27	-0.09	-0.18	-0.12	-0.10	-0.02	0.26	0.55	0.56	1.00	0.47	0.13	0.07	-0.00	0.04	-0.03	-0.21	-0.27	-0.32	-0.25	0.23	-0.18	-0.21	-0.26
AverageRatingsLocalServices	-0.21	0.22	0.16	0.12	0.12	0.15	0.10	0.29	0.26	0.47	1.00	0.32	0.26	0.06	0.03	0.00	-0.05	0.11	0.19	0.29	0.33	0.10	0.13	0.18
AverageRatingsBurgerPizzaShops	-0.26	-0.16	-0.24	-0.17	-0.10	-0.16	0.03	0.00	-0.01	0.13	0.32	1.00	0.47	0.35	0.15	-0.05	0.03	0.06	0.02	-0.13	-0.28	-0.32	-0.21	-0.18
AverageRatingsHotelsLodgings	-0.18	-0.21	-0.18	-0.15	-0.09	-0.14	0.03	-0.01	0.02	0.07	0.26	0.47	1.00	0.51	0.20	-0.05	0.02	0.08	0.06	-0.06	-0.21	-0.20	-0.16	-0.13
AverageRatingsJuiceBars	-0.24	-0.13	-0.16	-0.31	-0.28	-0.15	0.09	-0.02	0.03	-0.00	0.06	0.35	0.51	1.00	0.37	-0.01	0.08	0.10	0.12	0.03	-0.10	-0.29	-0.29	-0.21
AverageRatingsArtGalleries	-0.13	-0.07	-0.13	-0.27	-0.32	-0.19	0.09	-0.06	0.13	0.04	-0.03	0.15	0.20	0.37	1.00	0.10	0.06	0.08	0.07	0.07	0.06	-0.18	-0.16	-0.21
AverageRatingsDanceClubs	-0.07	-0.03	-0.02	0.02	0.06	-0.15	-0.14	-0.12	-0.12	-0.03	-0.00	-0.05	-0.05	-0.01	0.10	1.00	0.37	0.22	0.02	0.06	0.16	0.10	0.05	0.03
AverageRatingsSwimmingPools	-0.13	0.08	0.08	0.13	0.18	0.23	0.21	0.20	0.23	0.21	0.05	0.03	0.02	0.08	0.06	0.37	1.00	0.51	0.28	0.08	0.17	0.11	0.13	0.15
AverageRatingsGyms	-0.16	-0.03	-0.12	-0.18	-0.24	-0.27	-0.23	-0.24	-0.27	-0.27	-0.11	0.06	0.08	0.10	0.08	0.22	0.51	1.00	0.43	0.20	0.19	0.09	0.14	0.17
AverageRatingsBakeries	-0.18	0.03	-0.08	-0.19	-0.26	-0.27	-0.27	-0.28	-0.27	-0.32	-0.19	0.02	0.06	0.12	0.07	0.02	0.28	0.43	1.00	0.32	0.19	0.07	0.09	0.13
AverageRatingsBeautySpas	-0.20	0.10	-0.01	-0.09	-0.19	-0.23	-0.23	-0.25	-0.16	-0.25	-0.29	-0.13	-0.06	0.03	0.07	0.06	0.08	0.20	0.32	1.00	0.30	0.19	0.15	0.12
AverageRatingsCafes	-0.31	0.09	0.00	-0.06	-0.13	-0.20	-0.26	-0.27	-0.19	-0.23	-0.33	-0.28	-0.21	-0.10	0.06	0.16	0.17	0.19	0.19	0.30	1.00	0.37	0.35	0.30
AverageRatingsViewPoints	-0.36	0.02	0.13	0.28	0.12	0.09	-0.36	-0.26	-0.27	-0.18	0.10	0.32	0.20	0.29	-0.18	0.10	0.11	0.09	0.07	0.19	0.37	1.00	0.47	0.32
AverageRatingsMonuments	-0.41	0.08	0.12	0.17	0.13	0.08	0.22	0.17	0.27	0.21	0.13	0.21	0.16	0.29	-0.16	0.05	0.13	0.14	0.09	0.15	0.35	0.47	1.00	0.46
AverageRatingsGardens	-0.49	0.13	0.09	0.09	0.10	-0.07	-0.25	-0.14	-0.33	-0.26	-0.18	-0.18	-0.13	-0.21	-0.21	0.03	0.15	0.17	0.13	0.12	0.30	0.32	0.46	1.00

KLASIFIKACIJA

KORIŠĆENA SU TRI RAZLIČITA MODELA ZA KLASIFIKACIJU:

- **RANDOM FOREST**
- **LOGISTIC REGRESSION**
- **SUPPORT VECTOR MACHINE (SVM)**

- KREIRANA JE NOVA CILJNA PROMENLJIVA KORISTEĆI PROSEČNE OCENE I PRAGOVE ZA KLASIFIKACIJU NA TRI KATEGORIJE:
"NISKA," "SREDNJA," I "VISOKA"
- PODACI SU PODELJENI NA TRENING, VALIDACIONI I TEST SKUP U ODНОСУ 70-15-15.

LOGISTIČKA REGRESIJA

GRID SEARCH: Najbolji parametri: {'C': 0.1, 'solver': 'newton-cg'}

- TAČNOST MODELA NA VALIDACIONOM SKUPU: 0.733

- Izveštaj o klasifikaciji za sve tri klase (niska, srednja, visoka):

Izveštaj o klasifikaciji:				
	precision	recall	f1-score	support
niska	0.92	0.75	0.83	580
srednja	0.63	0.66	0.64	210
visoka	0.22	0.96	0.36	28
accuracy			0.73	818
macro avg	0.59	0.79	0.61	818
weighted avg	0.82	0.73	0.76	818

Ovaj izveštaj pruža detaljne informacije o performansama modela za svaku od klase. Na primer, za klasu "niska" (low) preciznost iznosi 0.92, što znači da je model tačno klasifikovao 92% instanci koje pripadaju toj klasi. Recall (odziv) za istu klasu iznosi 0.75, što ukazuje na to da je model identifikovao 75% instanci te klase. F1-skor je harmonijska sredina preciznosti i odziva i iznosi 0.83 za "niska" klasu. Izveštaj pruža informacije za sve tri klase.

- MATRICA KONFUZIJE:

```
Matrica konfuzije:  
[[435  82  63]  
 [ 38 138  34]  
 [  1   0  27]]
```

Matrica konfuzije prikazuje broj tačnih i netačnih klasifikacija za svaku od klase. Iz matrice možemo videti da model dobro radi u klasifikaciji "niska" i "srednja" klase, dok ima manje uspeha sa "visoka" klase.

RANDOM FOREST



GRID SEARCH: Najbolji parametri: {class_weight='balanced', n_estimators=200, random_state=42}

- **TAČNOST MODELNA NA VALIDACIONOM SKUPU: 0.943**
- Izveštaj o klasifikaciji za sve tri klase (niska, srednja, visoka):

Izveštaj o Klasifikaciji:

	precision	recall	f1-score	support
niska	0.94	0.98	0.96	580
srednja	0.94	0.88	0.91	210
visoka	0.90	0.64	0.75	28
accuracy			0.94	818
macro avg	0.93	0.83	0.87	818
weighted avg	0.94	0.94	0.94	818

Iz izveštaja o klasifikaciji možemo videti preciznost, odziv (recall) i F1-skor za svaku od klasa (niska, srednja, visoka).

Visoka preciznost i odziv za nisku i srednju klasu ukazuju na dobro modeliranje ovih klasa, dok niži odziv i preciznost za visoku klasu sugeriraju da model ima veće poteškoće u klasifikaciji ove klase.

F1-skor je mera balansa između preciznosti i odziva. Za nisku klasu, F1-skor iznosi 0.96, za srednju klasu 0.91, a za visoku klasu 0.75.

- **MATRICA KONFUZIJE:**

Matrica konfuzije:

```
[[569  10   1]
 [ 25 184   1]
 [  9   1 18]]
```

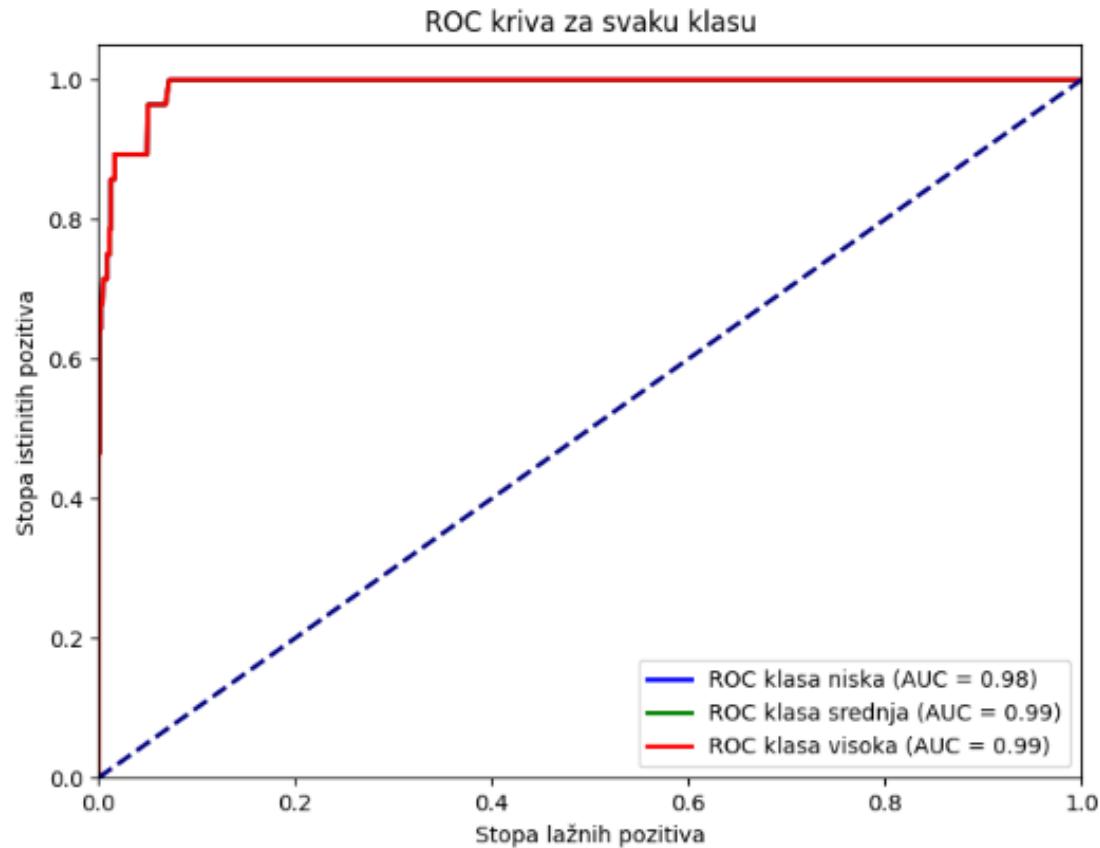
Iz matrice konfuzije možemo videti da postoji relativno mali broj netačnih klasifikacija

RANDOM FOREST



- AUC VREDNOSTI ZA SVAKU KLASU:

AUC za klasu niska: 0.9847906403940886
AUC za klasu srednja: 0.9876958020050125
AUC za klasu visoka: 0.9912296564195299



Visoke AUC vrednosti (blizu 1) za svaku od klase ukazuju na dobru sposobnost modela da razlikuje između tih klasa. Što je AUC bliže 1, to je bolje.

SVM

(Support Vector Machine):

GRID SEARCH: Najbolji parametri: {'C': 10, 'class_weight': None, 'kernel': 'rbf'}

- **TAČNOST MODELA NA VALIDACIONOM SKUPU: 0.929**
- Izveštaj o klasifikaciji za sve tri klase (niska, srednja, visoka):

Izveštaj o klasifikaciji:

	precision	recall	f1-score	support
niska	0.93	0.98	0.95	580
srednja	0.93	0.84	0.88	210
visoka	0.89	0.57	0.70	28
accuracy			0.93	818
macro avg	0.92	0.80	0.84	818
weighted avg	0.93	0.93	0.93	818

- **MATRICA KONFUZIJE:**

Matrica konfuzije za SVM model:

```
[[567 12 1]
 [ 32 177 1]
 [ 11   1 16]]
```

=> ZAKLJUČAK JE DA JE RANDOM FOREST MODEL NAJBOLJI IZBOR ZA DALJU ANALIZU

Evaluacija na test skupu:

- TAČNOST NA TEST SKUPU: 0.9474969474969475

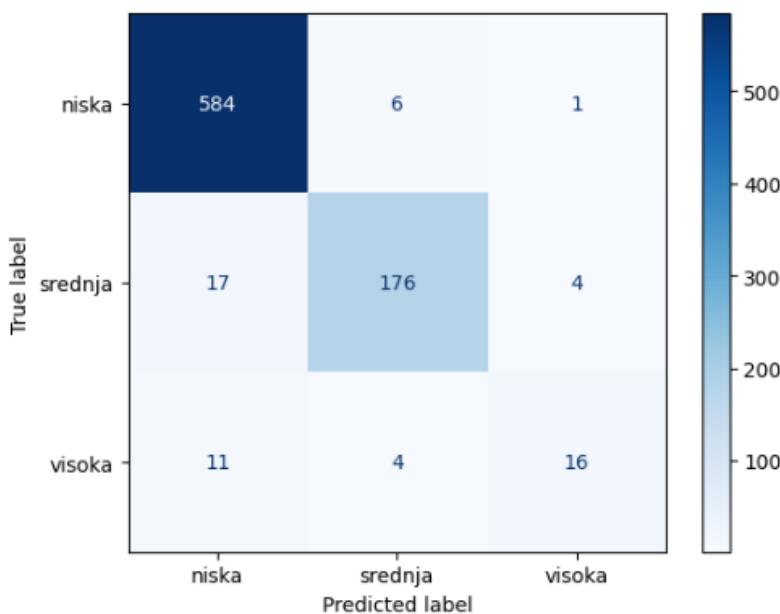
Izveštaj o klasifikaciji na test skupu:

	precision	recall	f1-score	support
niska	0.95	0.99	0.97	591
srednja	0.95	0.89	0.92	197
visoka	0.76	0.52	0.62	31
accuracy			0.95	819
macro avg	0.89	0.80	0.84	819
weighted avg	0.95	0.95	0.94	819

- MATRICA KONFUZIJE:

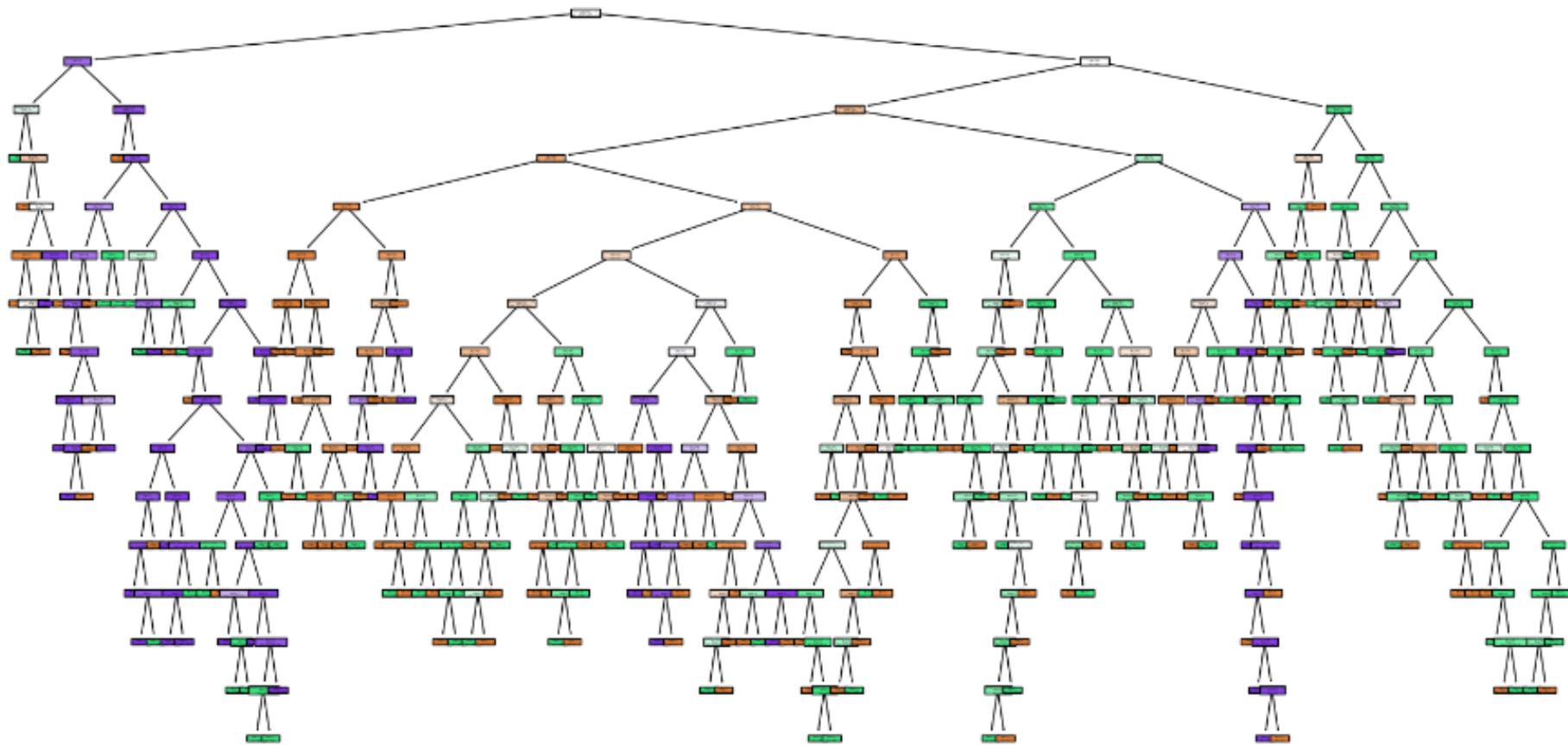
Matrica konfuzije:

```
[[435  82  63]
 [ 38 138  34]
 [  1   0  27]]
```



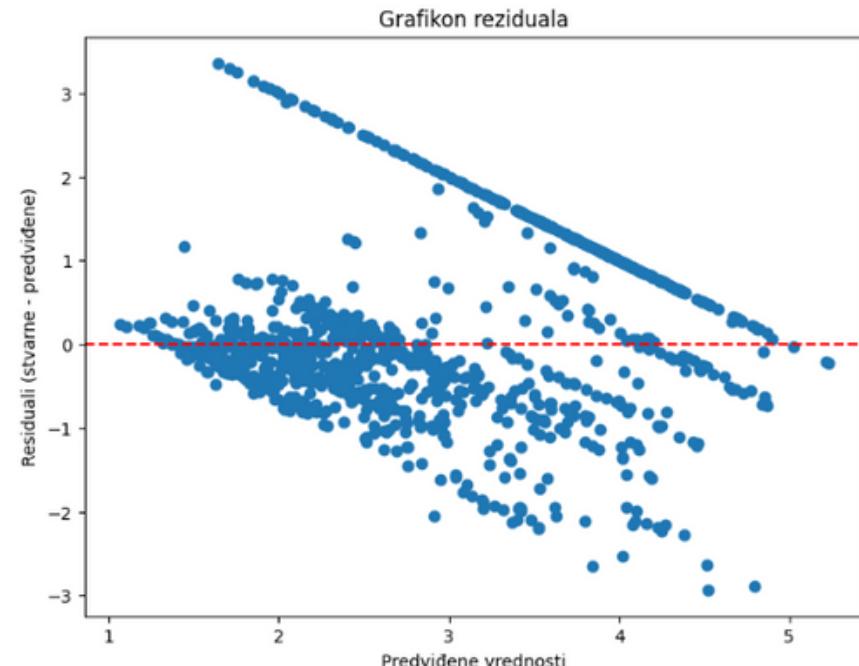
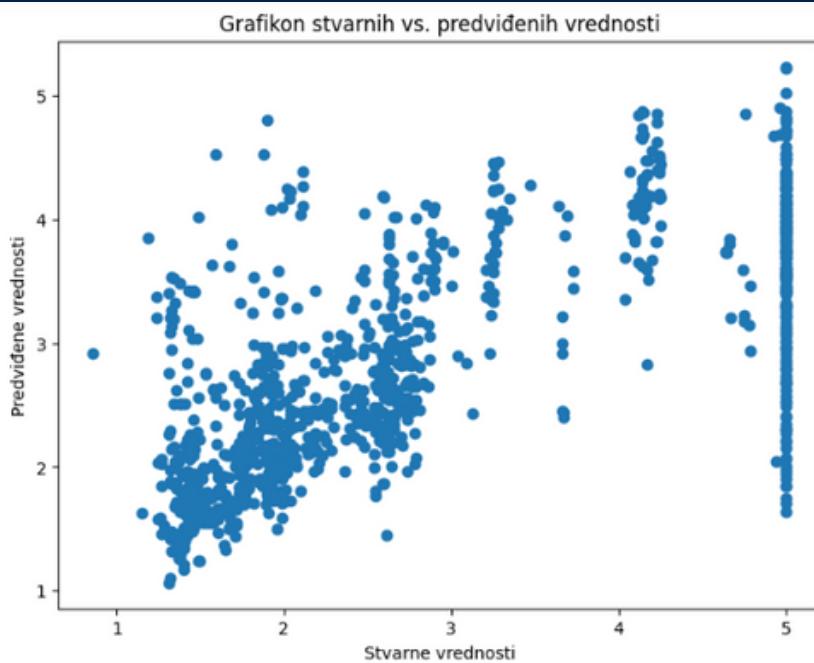
**REZULTATI UKAZUJU NA TO DA RANDOM FOREST MODEL DOBRO RADI
NA TEST SKUPU I IMA VISOKE PERFORMANSE U KLASIFIKACIJI.**

PRIKAZ STABLA ODLUČIVANJA



LINEARNA REGRESIJA

- PRIMENA: UPOTREBA PYTHON BIBLIOTEKE SCIKIT-LEARN.
- EVALUIRANJE MODELA KORŠIĆENJEM SREDNJE KVADRATNE GREŠKE (MSE) I R-KVADRAT KAO METRIKE ZA MERENJE TAČNOSTI
- PRIMENA LINEARNE REGRESIJE NA DVA RAZLIČITA CILJNA ATRIBUTA: "AVERAGERATINGSCOURSES" I "AVERAGERATINGSPARKS".

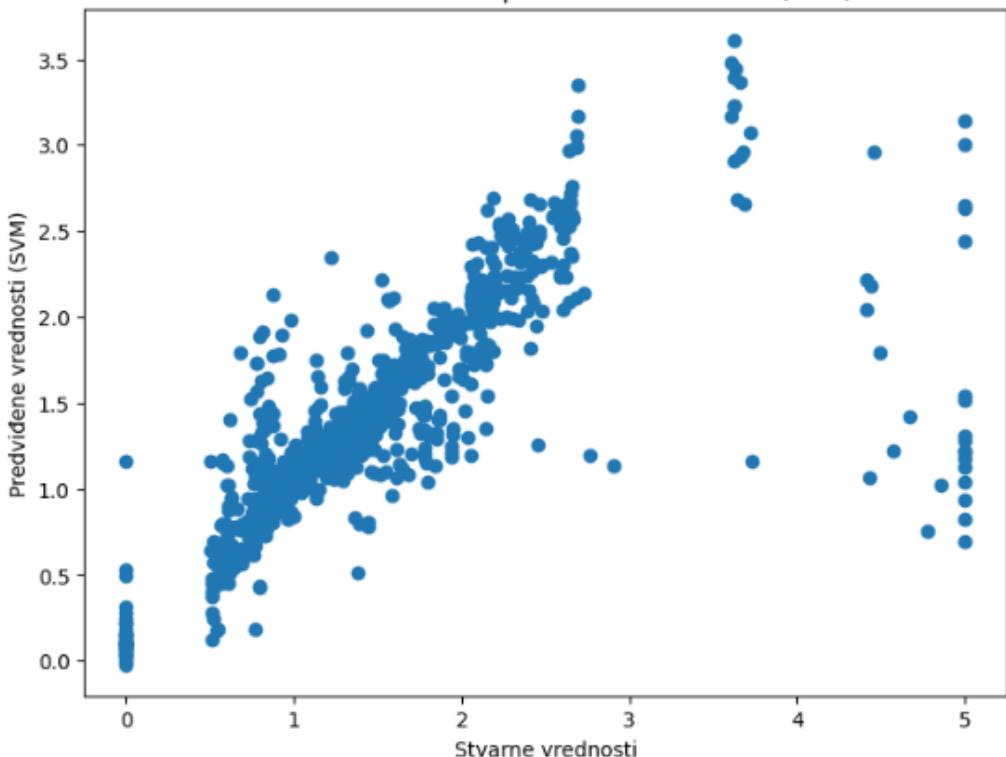


Support Vector Machine (SVM) za Regresiju

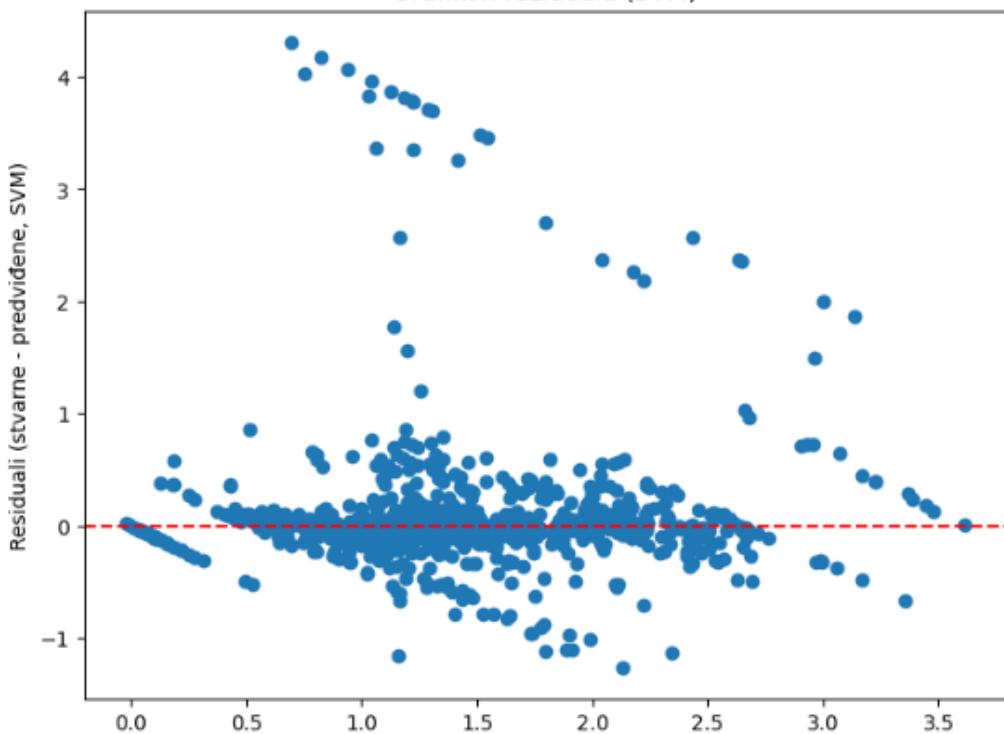
SUPPORT VECTOR REGRESSOR (SVR)

- Kreiramo i treniramo SVM model za regresiju (SVR) sa jezgrom 'rbf' na trening skupu podataka.
- Srednja kvadratna greška (MSE) od 0.339 ukazuje na prosečnu kvadratnu razliku između stvarnih vrednosti i predviđenih vrednosti. Manja vrednost MSE sugerije na bolju preciznost modela.
- R-kvadrat od 0.514 znači da ovaj SVM model objašnjava oko 51.4% varijanse u prosečnim ocenama crkava. Vrednost R-kvadrata od 1 označava savršeno objašnjenje varijanse, dok vrednost od 0 ukazuje na to da model ne objašnjava ništa.
=> **SVM regresija (koristeći jezgro 'rbf')** daje bolje performanse za predviđanje određenih ciljnih atributa u poređenju sa linearnom regresijom.

Grafikon stvarnih vs. predviđenih vrednosti (SVM)



Grafikon reziduala (SVM)

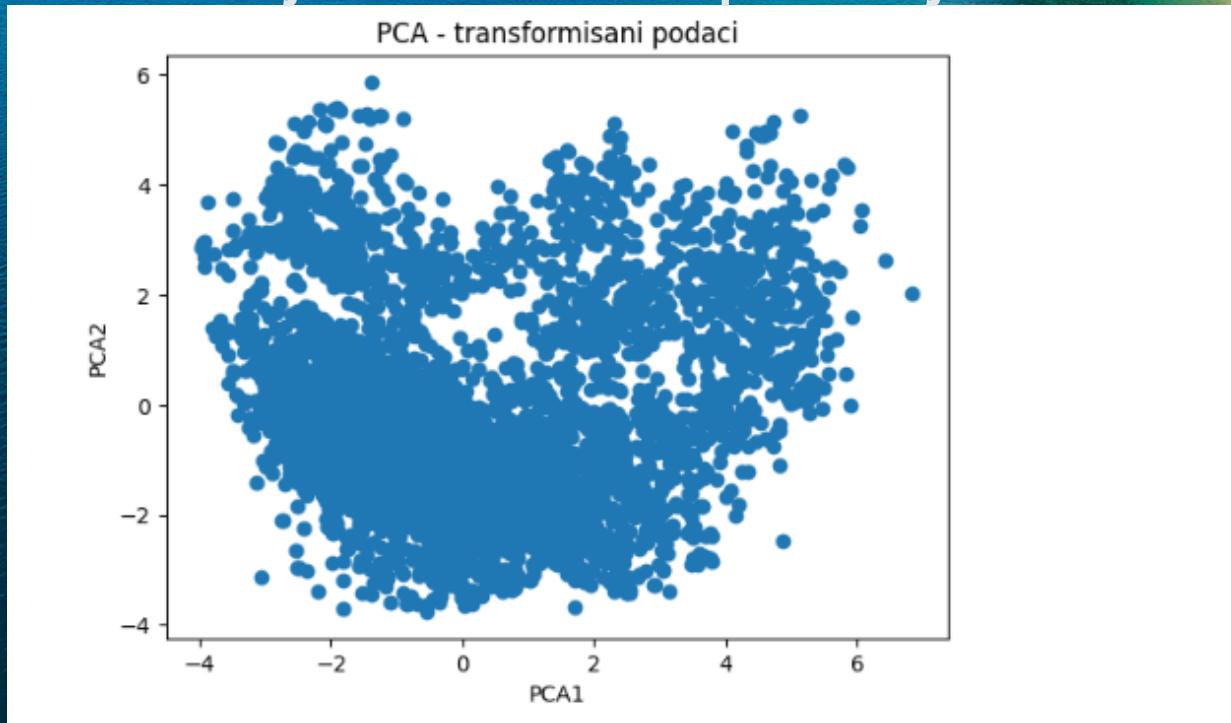


KLASTEROVANJE

Identifikacija sličnih grupa korisnika koji imaju slične preference u vezi sa atrakcijama (grupe korisnika koji vole muzeje itd..)

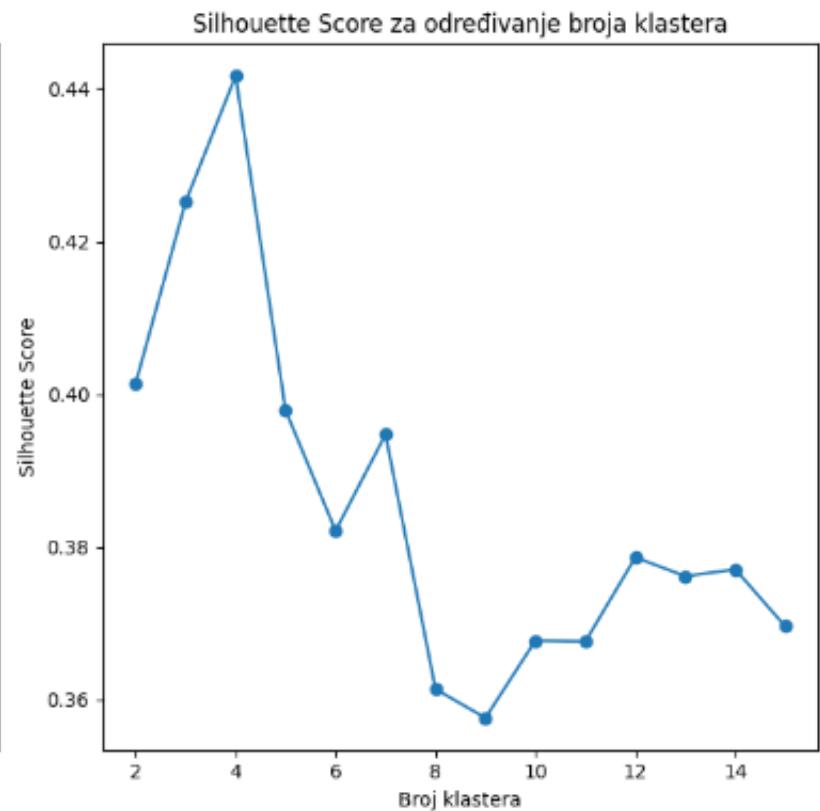
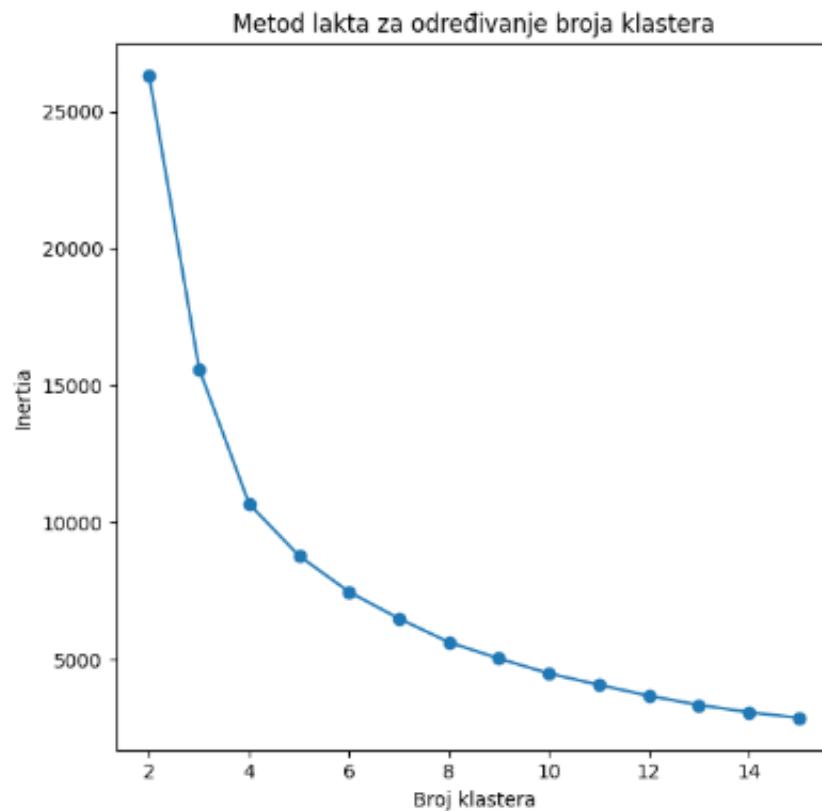
PCA

- Kada kao cilj postavimo da želimo da se zadrži 80% ukupne varijanse
->PCA bira 13 komponenti da bi zadovoljio uslov
- Instanciranje PCA modela sa 2 komponente za vizualizaciju
-> sačuvano je oko 34.2% ukupne varijanse



K-MEANS ALGORITAM

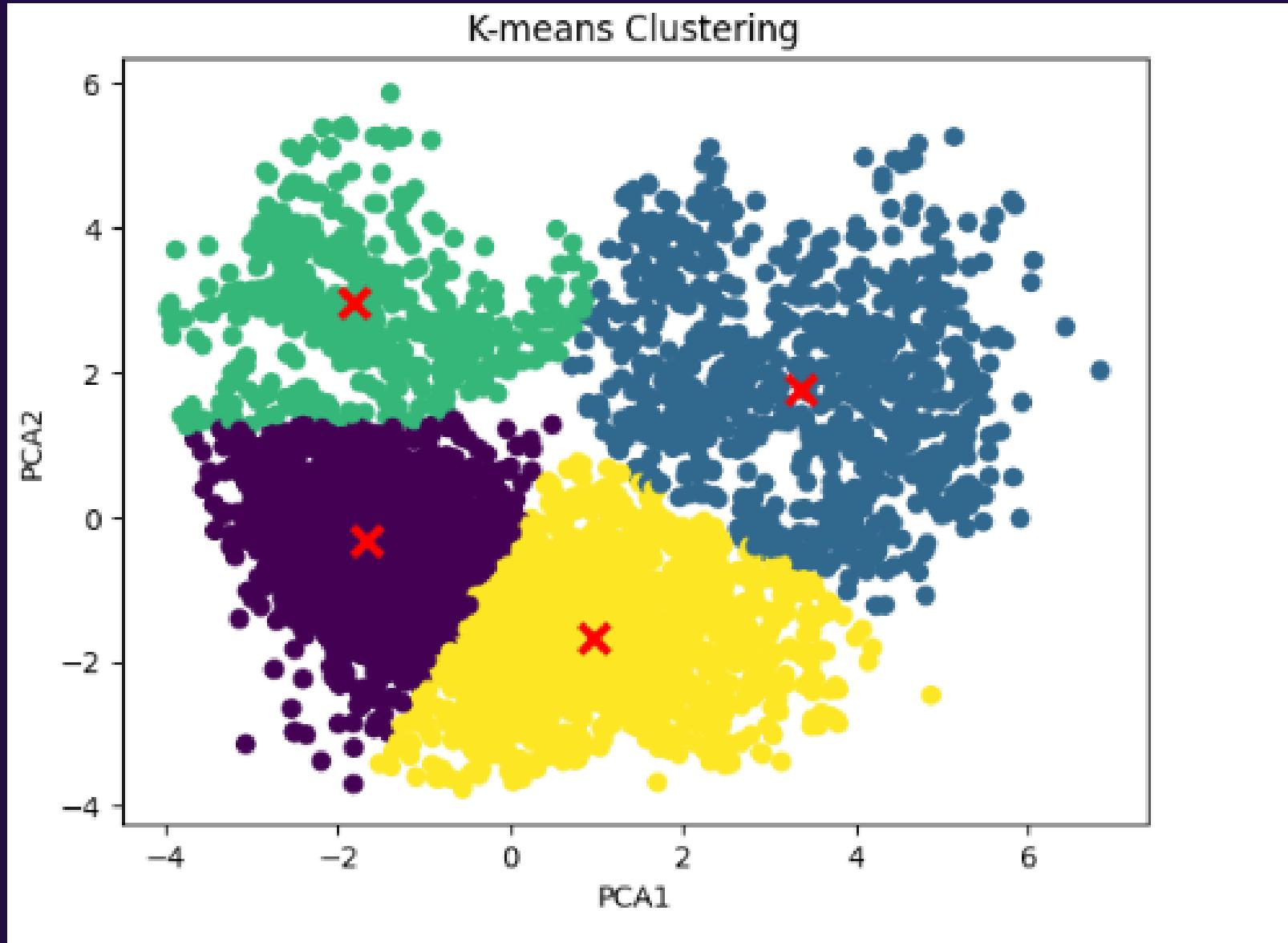
- Analiza broja klastera u rasponu od 2 do 15 klastera



: #sa grafika se vidi da je lakat za 4 klastera, a i silueta skor je najbolji takodje u 4

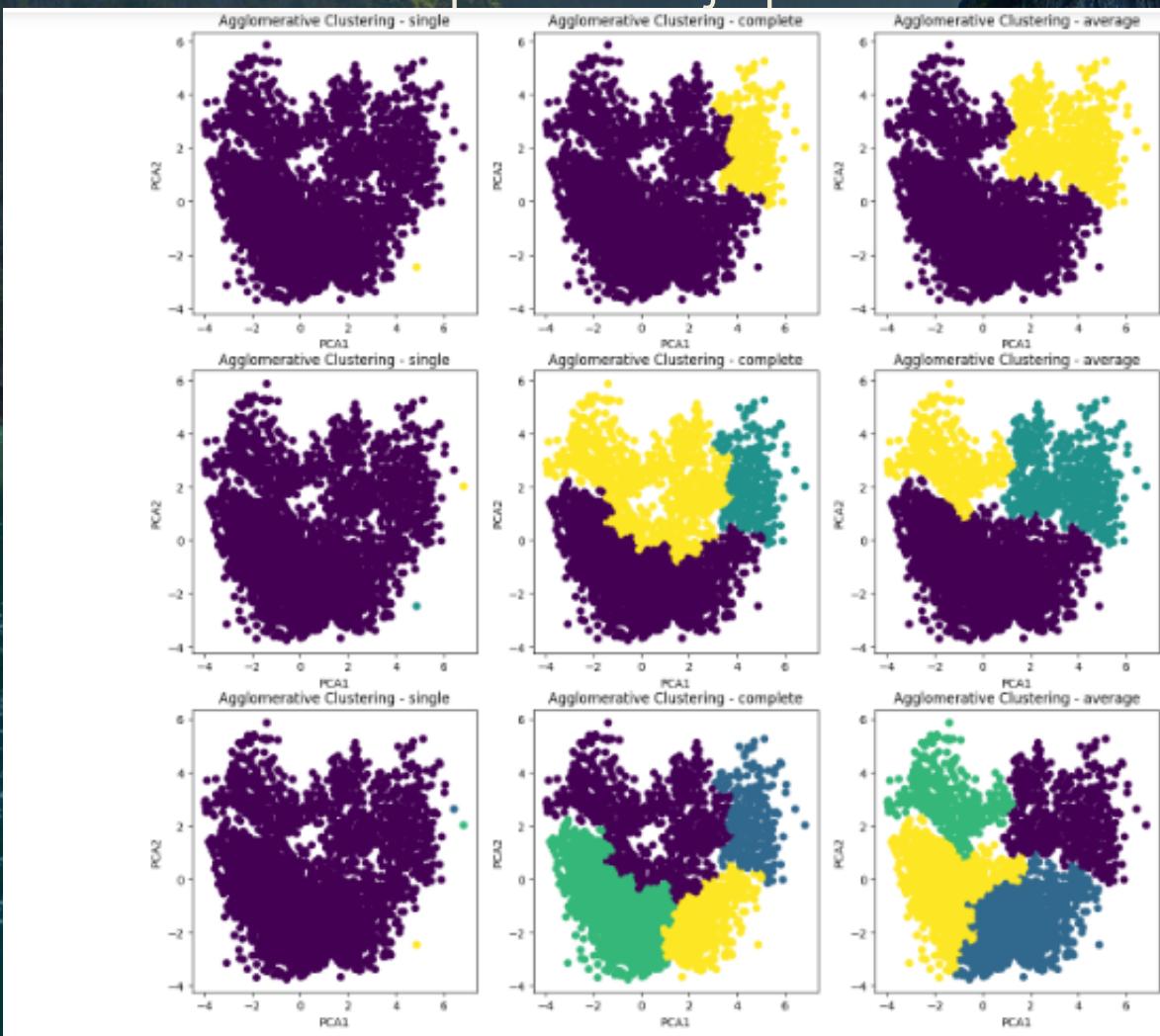
**K-MEANS JE ALGORITAM KOJI GRUPIŠE PODATKE NA OSNOVU NJIHOVE
SLIČNOSTI I MINIMIZIRA VARIJANSU UNUTAR KLASTERA.**

(modelovanje podataka sa 4 klastera)



HIJERARHIJSKO KLASTEROVANJE

- Vizualizacija klasterovanja za različite brojeve klastera (2, 3 i 4) i različite načine povezivanja podataka.



Silhouette skor:

Upoređivanje kvaliteta klasterovanja između K-means i hijerarhijskog klasterovanja:

```
kmeans_score = silhouette_score(data, kmeans.labels_)
```

```
kmeans_score
```

```
0.4416851319331852
```

```
agg_score = silhouette_score(data, agg.labels_)
```

```
agg_score
```

```
0.4287924465463875
```

K-means i hijerarhijsko klasterovanje daju relativno slične Silhouette skorove, pri čemu je K-means imao nešto bolji rezultat.

Oba algoritma imaju sličan kvalitet klasterovanja za ovaj skup podataka, ali K-means može biti efikasniji za veće skupove podataka, dok hijerarhijsko klasterovanje omogućava dublje razumevanje strukture klastera.

An aerial photograph of a tropical beach. The sand is a light tan color, and the ocean water is a vibrant turquoise. Waves are crashing onto the shore, creating white foam. Several small figures of people are scattered across the beach, appearing as tiny dots due to the wide-angle shot.

HVALA NA PAŽNJI!