Travel Review Ratings

Autor: Jelena Maksimović

Projekat u okviru kursa Istraživanje podataka 1, Matematički fakultet u Beogradu
Profesor: Nenad Mitić
Asistent: Stefan Kapunac

Sadržaj

1 Uvod

1.1 Pregled zadatka

Cilj ovog istraživanja je analiza skupa podataka "Travel Review Ratings" koji sadrži Google recenzije atrakcija širom Evrope. Skup podataka se sastoji od ocena korisnika na 24 različite kategorije atrakcija, a ocene korisnika se kreću u opsegu od 1 do 5. Zadatak je da izvršimo eksplorativnu analizu ovog skupa podataka, primena različitih tehnika pretprocesiranja, kao i primena algoritama za klasifikaciju i klasterovanje radi dobijanja dubljih uvida u podatke i omogućavanja boljeg razumevanja korisničkih ocena atrakcija.

1.2 Motivacija

Analiza korisničkih ocena može pružiti korisne informacije o atrakcijama i njihovoj popularnosti. Razumevanje faktora koji utiču na ocene može pomoći u donošenju odluka u oblasti turizma, marketinga i unapređenja atrakcija. Takođe, klasifikacija i klasterovanje ovih podataka mogu doprineti personalizovanim preporukama za korisnike.

2 Eksplorativna analiza podataka

2.1 Skup podataka

Skup podataka "Travel Review Ratings" sastoji se od 5456 instanci sa 25 atributa. Atributi uključuju ocene korisnika za različite kategorije atrakcija, kao i jedinstvene identifikatore korisnika.

	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	Categ
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
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
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
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
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
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
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
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
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
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

5456 rows × 26 columns

Prikaz podataka

2.2 Atributi

- **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/pubar.
- **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 sok barove.
- **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 za lepoto 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.

3 Eksplorativna analiza podataka

3.1 Pretprocesiranje

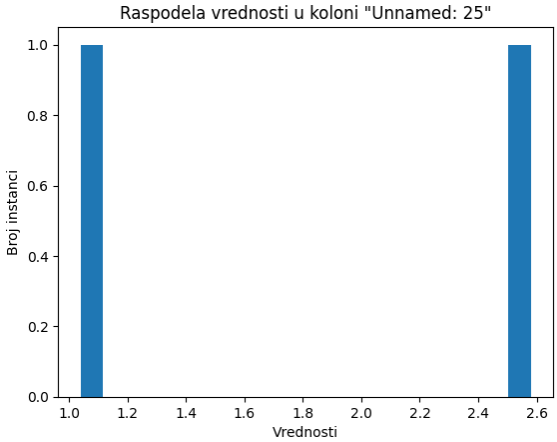
Priprema podataka za analizu.

1. Konverzija ocena u numeričke vrednosti:

   Prva važna faza pretprocesiranja podataka bila je konverzija ocena korisnika, koje su prvobitno bile sačuvane kao stringovi, u numeričke vrednosti. Ovo nam omogućava da radimo sa ocenama kao sa brojevima, što je neophodno za analizu i primenu algoritama.

2. Provera i uklanjanje nepotrebnih kolona:

   Nakon razmatranja svih kolona u skupu podataka radi utvrđivanja njihove korisnosti za našu analizu, može se primetiti da kolona "Unnamed : 25" sadrži većinom NaN vrednosti i samo dve različite stvarne vrednosti (1.04 i 2.58). Odlučili smo se za uklanjanje ove kolone kako bismo smanjili dimenzionalnost skupa podataka.



3. Izbacivanje kolone "User":

   Kolona "User" sadrži jedinstvene vrednosti za svakog korisnika i ne pruža značajne informacije za analizu podataka. Izbacujemo je kako bismo olakšali obradu podataka.

4. Preimenovanje kolona:

   Da bismo unapredili čitljivost i razumevanje podataka, preimenovali smo kolone kako bi imale smislene nazive.

```
new_column_names = {
    'Category 1': 'AverageRatingsChurches',
    'Category 2': 'AverageRatingsResorts',
    'Category 3': 'AverageRatingsBeaches',
    'Category 4': 'AverageRatingsParks',
    'Category 5': 'AverageRatingsTheatres',
    'Category 6': 'AverageRatingsMuseums',
    'Category 7': 'AverageRatingsMalls',
    'Category 8': 'AverageRatingsZoo',
    'Category 9': 'AverageRatingsRestaurants',
    'Category 10': 'AverageRatingsPubsBars',
    'Category 11': 'AverageRatingsLocalServices',
    'Category 12': 'AverageRatingsBurgerPizzaShops',
    'Category 13': 'AverageRatingsHotelsLodgings',
    'Category 14': 'AverageRatingsJuiceBars',
    'Category 15': 'AverageRatingsArtGalleries',
    'Category 16': 'AverageRatingsDanceClubs',
    'Category 17': 'AverageRatingsSwimmingPools',
    'Category 18': 'AverageRatingsGyms',
    'Category 19': 'AverageRatingsBakeries',
    'Category 20': 'AverageRatingsBeautySpas',
    'Category 21': 'AverageRatingsCafes',
    'Category 22': 'AverageRatingsViewPoints',
    'Category 23': 'AverageRatingsMonuments',
    'Category 24': 'AverageRatingsGardens'
}
```

5. Statistička analiza numeričkih kolona:

   Korišćenjem ‘data.describe()‘ funkcije, dobili smo osnovne statističke informacije o numeričkim atributima, uključujući srednju vrednost, standardnu devijaciju, minimum, maksimum i kvartile. Ovo nam pomaže da bolje razumemo raspodelu podataka i identifikujemo potencijalne outliere.

	AverageRatingsChurches	AverageRatingsResorts	AverageRatingsBeaches	AverageRatingsParks	AverageRatingsTheatres	AverageRatingsMuseums	Av
count	5456.000000	5456.000000	5456.000000	5456.000000	5456.000000	5456.00000	
mean	1.455720	2.319707	2.489331	2.796886	2.958941	2.89349	
std	0.827604	1.421438	1.247815	1.309159	1.339056	1.28240	
min	0.000000	0.000000	0.000000	0.830000	1.120000	1.11000	
25%	0.920000	1.360000	1.540000	1.730000	1.770000	1.79000	
50%	1.340000	1.905000	2.060000	2.460000	2.670000	2.68000	
75%	1.810000	2.682500	2.740000	4.092500	4.312500	3.84000	
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	

8 rows × 24 columns

6. Rukovanje sa NaN vrednostima:

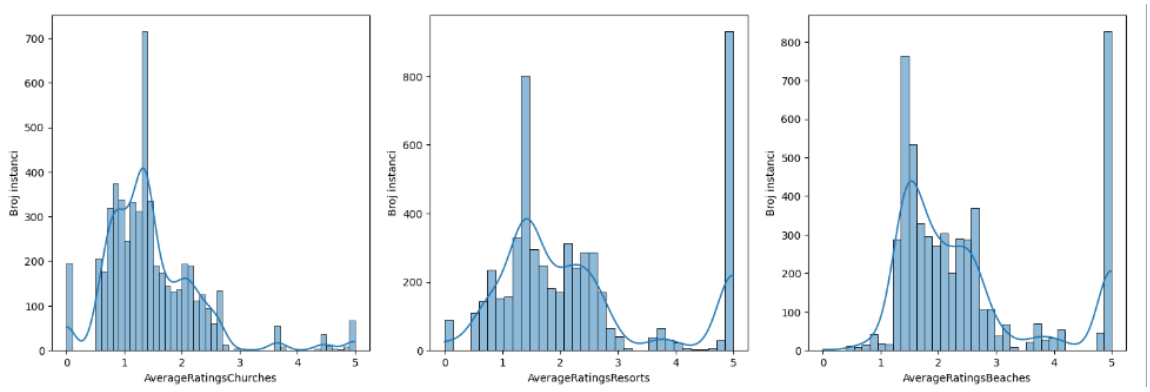
   Što se tiče provere i obrade Nan vrednosti u skupu podataka imamo prisustvo samo 3 NaN vrednosti, dok su sve ostale vrednosti u kolonama poznate. Odlučili smo se za zamenu ovih 3 NaN vrednosti prosečnim vrednostima po odgovarajućim kolonama, čime smo sačuvali podatke i sprečili gubitak instanci. Ova strategija je izabrana kako bismo očuvali što više informacija.

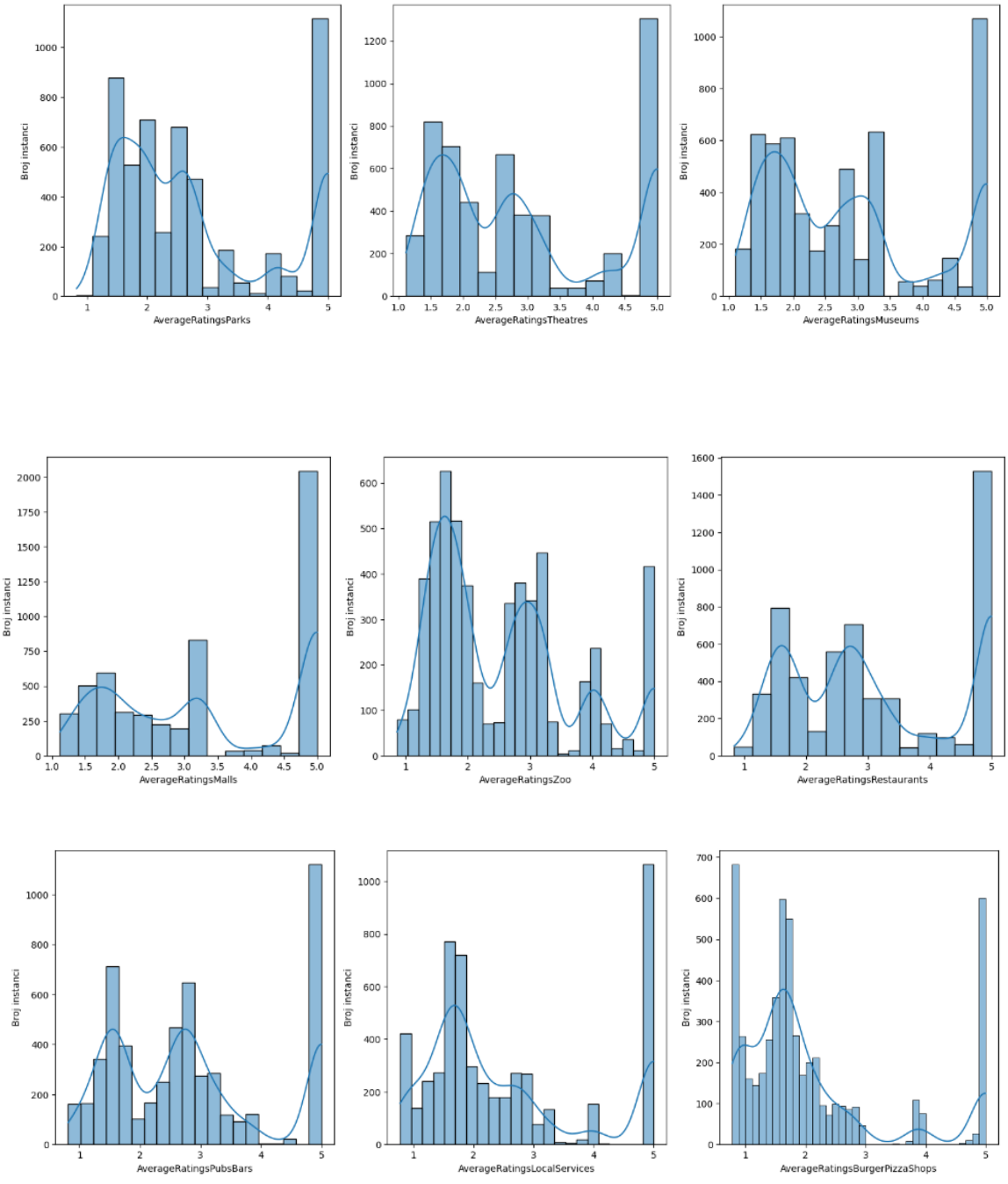
7. Provera autlajera i ekstremnih vrednosti:

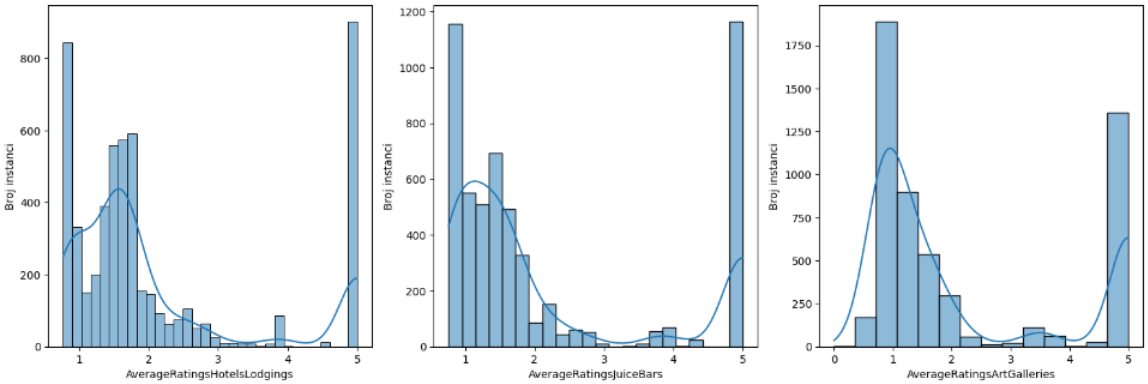
   U analizi skupa podataka nismo identifikovali autlajere ni ekstremne vrednosti. Sve ocene u našem skupu podataka nalaze se u intervalu od 1 do 5, što je očekivano za ocene korisnika. Iz tog razloga, nisu preduzeti koraci za izbacivanje ili tretiranje autlajera.

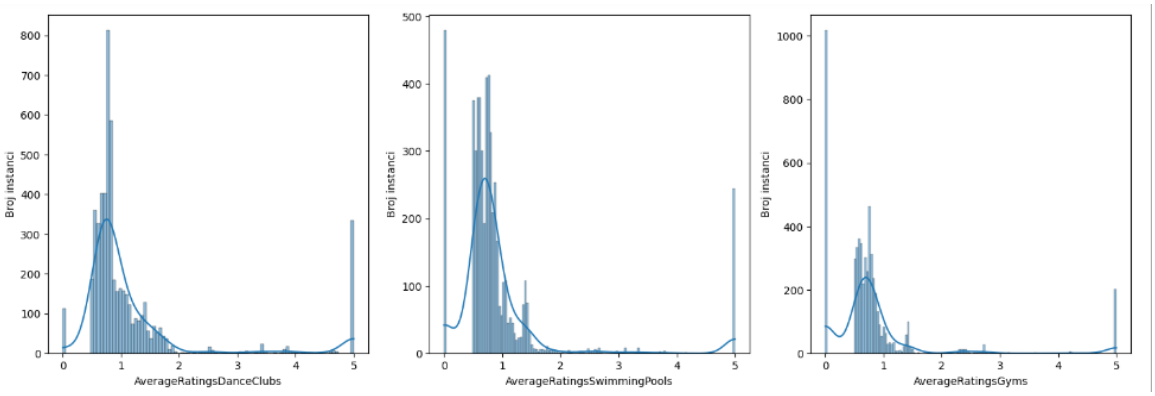
8. Vizualizacija numeričkih kolona:

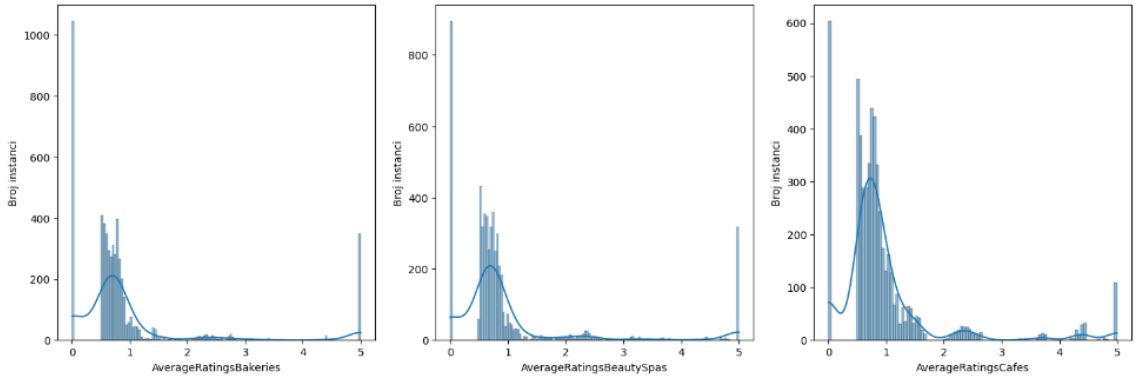
   Dodati su i grafovi i histogram sa podacima kako bi se omogućilo bolje razumevanje raspodele numeričkih atributa.

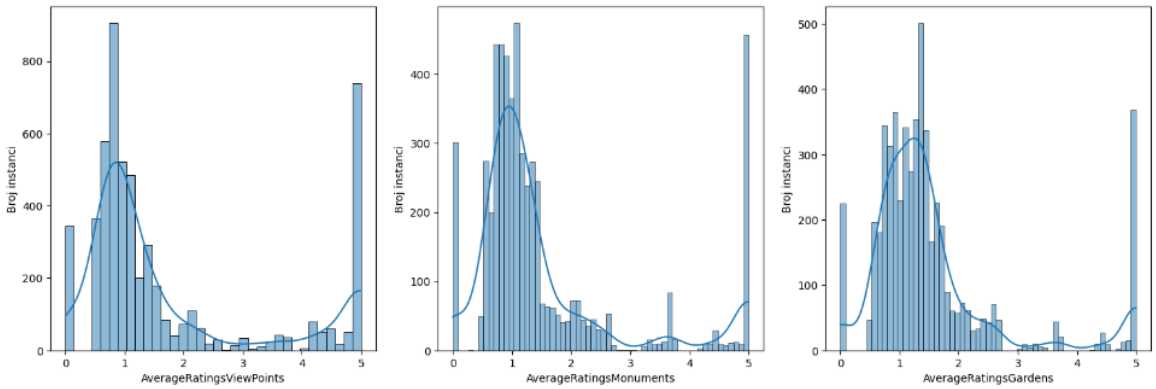



Broj instanci
AverageRatingsParks
AverageRatingsTheatres
AverageRatingsMuseums
AverageRatingsMalls
AverageRatingsZoo
AverageRatingsRestaurants
AverageRatingsPubsBars
AverageRatingsLocalServices
AverageRatingsBurgerPizzaShops

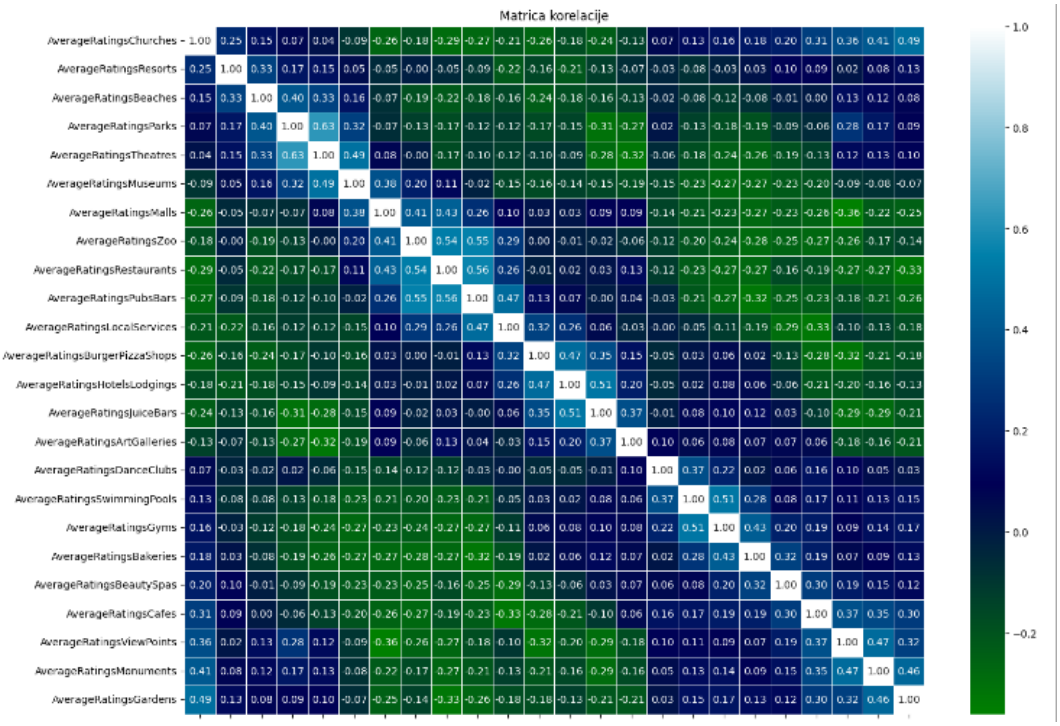

Broj instanci
AverageRatingsHotelsLodgings
Broj instanci
AverageRatingsJuiceBars
Broj instanci
AverageRatingsArtGalleries


Broj instanci
AverageRatingsDanceClubs
Broj instanci
AverageRatingsSwimmingPools
Broj instanci
AverageRatingsGyms


Broj instanci
AverageRatingsBakeries
Broj instanci
AverageRatingsBeautySpas
Broj instanci
AverageRatingsCafes



9. Matrica korelacije: Za identifikaciju linearne korelacije između numeričkih atributa, kreiramo matricu korelacije. Ova analiza će nam pomoći u identifikaciji potencijalnih veza između atributa, što može biti korisno za buduće analize, uključujući klasifikaciju i klasterovanje.



Matrica korelacije između atributa.

Slika prikazuje matricu korelacije između različitih atributa u našem skupu podataka. Matrica korelacije se koristi za vizualizaciju veza između atributa. Vrednosti blizu 1 ukazuju na jaku pozitivnu korelaciju, dok vrednosti blizu -1 ukazuju na jaku negativnu korelaciju. Vrednosti blizu 0 ukazuju na slabo ili nikakvo korelisanje između atributa.

4 Klasifikacija

U procesu klasifikacije nad našim skupom podataka bavili smo se procenom kvaliteta različitih atrakcija na osnovu ocena korisnika i to ima nekoliko koristi i potencijalnih benefita:

- **Segmentacija atrakcija:** Klasifikacija omogućava grupisanje atrakcija u različite kategorije (niska, srednja, visoka) na osnovu prosečnih ocena korisnika. Ovo može biti korisno za organizaciju, promociju i analizu atrakcija. Na primer, organizatori događaja ili turističkih tura mogu koristiti ovu klasifikaciju da identifikuju atrakcije koje bi bile najzanimljivije za određenu ciljanu grupu.
- **Personalizovane preporuke:** Na osnovu klasifikacije, može se razviti sistem preporuka koji korisnicima sugeriše atrakcije koje najbolje odgovaraju njihovim preferencijama. Korisnici mogu dobiti personalizovane preporuke za atrakcije koje su klasifikovane kao "visoke" ili "srednje" kvalitete, poboljšavajući njihovo iskustvo.
- **Efikasnija analiza podataka:** Klasifikacija može olakšati analizu i interpretaciju podataka. Na primer, istraživači i analitičari mogu koristiti klasifikaciju da identifikuju obrasce i trendove u ocenama za različite tipove atrakcija.
- **Donošenje odluka:** Organizacije ili poslovni entiteti koji se bave turizmom ili zabavom mogu koristiti ovu klasifikaciju kako bi doneli informisane odluke o tome na koje atrakcije treba usmeriti svoje resurse ili kako bi poboljšali kvalitet svojih ponuda.

Klasifikacija omogućava da se izvuče vrednost iz dataseta tako što se organizuju i analiziraju podaci na način koji je koristan za donošenje odluka i poboljšanje poslovanja ili korisničkog iskustva.

Za postizanje ovog cilja, korišćena su tri različita modela za klasifikaciju: Logistic Regression, Random Forest, i Support Vector Machine (SVM). Svaki model je prošao kroz sledeće korake kako bi se evaluirala njegova efikasnost i performanse.

4.1 Koraci

1. **Priprema podataka:**
   Prvo je kreirana nova ciljna promenljiva koristeći prosečne ocene i pragove za klasifikaciju na tri kategorije: "niska," "srednja," i "visoka."
2. **Podela podataka:**
   Podaci su podeljeni na trening, validacioni i test skup u odnosu 70-15-15. Ovo omogućava treniranje, optimizovanje i testiranje modela.
3. **Logistička Regresija:**
   Primena Logistic Regression modela i korišćenje Grid Search-a za optimizaciju hiperparametara.
   - Rezultati:
     - 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 klasa. 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.

     - AUC vrednosti za svaku klasu.

```
AUC za klasu niska: 0.8890683859750798
AUC za klasu srednja: 0.8728540100250628
AUC za klasu visoka: 0.9717902350813743
```

ROC (Receiver Operating Characteristic) kriva i AUC (Area Under the Curve) vrednosti koriste se za procenu performansi modela u pogledu procene verovatnoća za različite klase. AUC vrednosti se koriste za merenje koliko dobro model razdvaja klase.

Naš model ima dobre AUC vrednosti za sve tri klase: visoka (0.97), srednja (0.87) i niska (0.88). Visoka AUC vrednost ukazuje na dobro razdvajanje klasa.

– Matrica konfuzije.

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

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

4. **Random Forest:**
   Primenili smo Random Forest model i takođe koristili Grid Search za optimizaciju hiperparametara.

   - Rezultati:
     – Tačnost modela na validacionom skupu: 0.943
     – Izveštaj o klasifikaciji za sve tri klase.

```
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
```

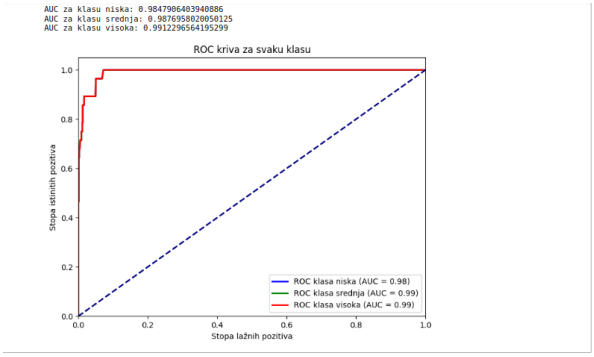
Iz izveštaja o klasifikaciji možemo videti preciznost, odziv (recall) i F1-skor za svaku od klasa (niska, srednja, visoka).

Visoka preciznost i odziv za nisku i srednju klasu ukazuju na dobro modeliranje ovih klasa, dok niži odziv i preciznost za visoku klasu sugerišu 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.

U ukupnom weighted avg F1-skor iznosi 0.94, što ukazuje na dobar balans između preciznosti i odziva na nivou svih klasa.

– AUC vrednosti za svaku klasu.



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

– 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

5. **SVM (Support Vector Machine):**
   Koristili smo SVM model i ponovo Grid Search za optimizaciju hiperparametara.

   - Rezultati:
     – Tačnost modela na validacionom skupu: 0.929
     – Izveštaj o klasifikaciji za sve tri klase.

```
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
```

     – AUC vrednosti za svaku klasu.

```
AUC za klasu niska: 0.9570486815415822
AUC za klasu srednja: 0.9694000626566416
AUC za klasu visoka: 0.9543851717902351
```

     – Matrica konfuzije.

```
Matrica konfuzije za SVM model:
 [[567  12   1]
 [ 32 177   1]
 [ 11   1  16]]
```

   S obzirom na sve prethodno prikazane rezultate, zaključujemo da je Random Forest model naš najbolji izbor za dalju analizu. Kako bismo dobili potpune informacije o performansama ovog modela u stvarnim uslovima, sledeći korak je primeniti ga na test skupu podataka.

6. **Evaluacija na Test Skupu**

   Tačnost na test skupu: 0.9474969474969475

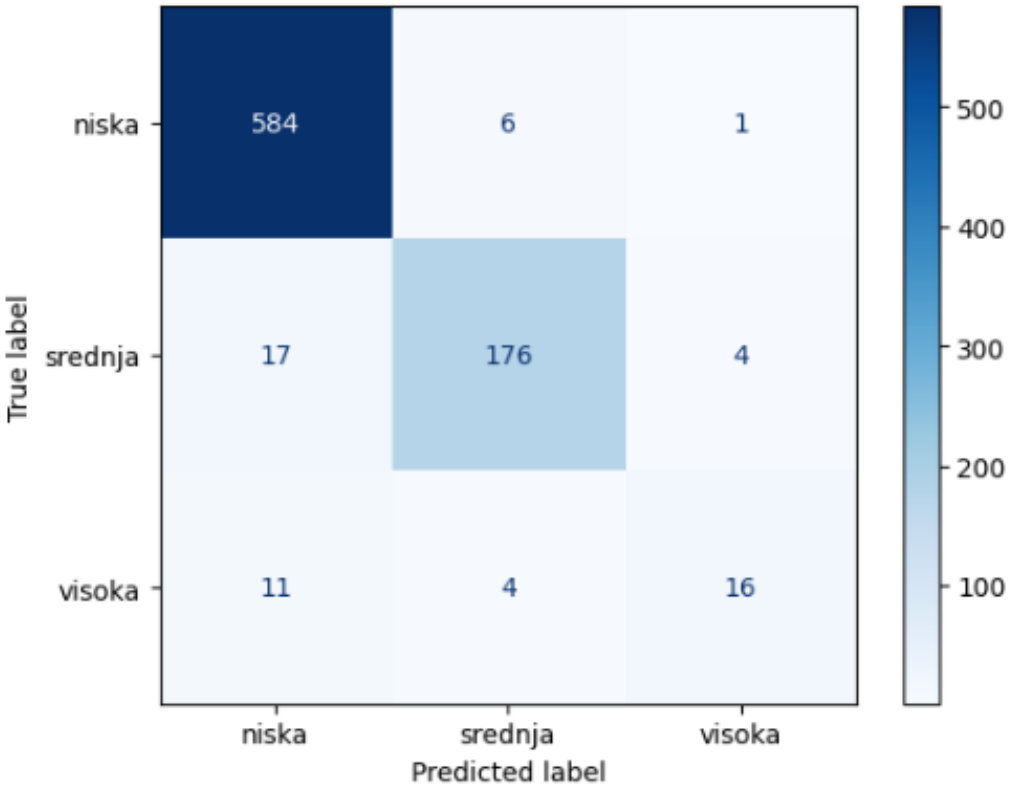
```
Izveštaj o klasifikaciji na test skupu:
              precision    recall  f1-score   support

       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]]
```



Ova matrica konfuzije prikazuje tačne i netačne klasifikacije za svaku od klasa na test skupu. Evo analize rezultata:

Za klasu "niska": Model je tačno klasifikovao većinu instanci u ovu klasu, što se vidi po visokim vrednostima na glavnoj dijagonali matrice konfuzije. Postoji mali broj lažnih pozitiva i lažnih negativa, što ukazuje na dobru preciznost i odziv za ovu klasu.

Za klasu "srednja": Model takođe pokazuje dobru tačnost za ovu klasu, s većinom tačnih klasifikacija. Preciznost i odziv su visoki, što ukazuje na dobar balans između tačnosti i odziva.

Za klasu "visoka": Model ima poteškoća u klasifikaciji ove klase, što se vidi po nižim vrednostima na glavnoj dijagonali matrice. Postoji veći broj lažnih negativa, što znači da model često klasifikuje instance iz ove klase kao druge klase.

Opšti zaključak je da Random Forest model dobro radi za klasifikaciju klasa "niska" i "srednja," ali može se unaprediti za klasu "visoka."

```python
accuracy = accuracy_score(y_test, y_test_pred_rf)
print(f"Tačnost: {accuracy:.2f}")

precision = precision_score(y_test, y_test_pred_rf, average='weighted')
print(f"Preciznost: {precision:.2f}")

recall = recall_score(y_test, y_test_pred_rf, average='weighted')
print(f"Odziv: {recall:.2f}")

f1 = f1_score(y_test, y_test_pred_rf, average='weighted')
print(f"F1-skor: {f1:.2f}")
```

```
Tačnost: 0.95
Preciznost: 0.95
Odziv: 0.95
F1-skor: 0.94
```

Tačnost (Accuracy): Model ima tačnost od 0.95, što znači da tačno klasifikuje 95% instanci na test skupu. Ovo ukazuje na dobar ukupan performans modela.

Preciznost (Precision): Preciznost je 0.95, što znači da je prosečna preciznost modela za sve klase visoka. To znači da je većina pozitivnih predviđanja tačna i da ima malo lažnih pozitiva.

Odziv (Recall): Odziv je takođe 0.95, što ukazuje na dobar odziv modela za sve klase. To znači da model dobro pronalazi većinu pozitivnih instanci i ima malo lažnih negativa.

F1-skor: F1-skor je 0.94, što je harmonična sredina između preciznosti i odziva. Ovo je dobar pokazatelj balansa između tačnosti i kompletnosti modela.

U suštini, ovi rezultati ukazuju na to da Random Forest model dobro radi na test skupu i ima visoke performanse u klasifikaciji.

Prikaz stabla odlučivanja:

5 Linearna Regresija

Na ovom dataset-u "Travel Review Ratings" primenićemo i linearnu regresiju iz nekoliko razloga:

- **Interpretabilnost:** Linearna regresija pruža jednostavno tumačenje modela. Koeficijenti atributa omogućavaju nam direktno razumevanje kako svaki atribut utiče na ciljni atribut, na primer, kako prosečna ocena restorana varira sa prosečnim ocenama drugih atrakcija.
- **Brza implementacija:** Linearna regresija je brza za implementaciju i izračun, što je korisno za brzu analizu i iteraciju sa različitim eksperimentima.
- **Očekivanje linearnih veza:** Pretpostavili smo da postoji linearan odnos između prosečnih ocena različitih atrakcija i ciljnog atributa (na primer, prosečnih ocena crkava i prosečnih ocena restorana).

5.1 Implementacija Linearne Regresije

Odrađeni su sledeći koraci:

- Učitavanje dataset-a i konvertovanje svih numeričkih vrednosti u odgovarajući format (float64).
- Izbacivanje nepotrebnih kolona, kao što su kolona sa korisničkim identifikatorima i neimenovana kolona sa nedostajućim vrednostima.
- Primena linearne regresije koristeći Python biblioteku scikit-learn.
- Evaluiranje modela koristeći srednju kvadratnu grešku (MSE) i R-kvadrat (R^2) kao metrike za merenje tačnosti i sposobnosti modela da objasni varijaciju ciljnog atributa.

5.1.1 Rezultati

Primenjujemo linearnu regresiju na dva različita ciljna atributa: "AverageRatingsChurches" (prosečne ocene crkava) i "AverageRatingsParks" (prosečne ocene parkova).

Za atribut "AverageRatingsChurches"

- Prvo definišemo ciljni atribut kao "AverageRatingsChurches" i biramo sve druge atribute kao ulazne promenljive.
- Podatke delimo na trening i test skup podataka koristeći 'train_test_split' funkciju. Ovo omogućava procenu tačnosti modela.
- Kreiramo i treniramo model linearne regresije na trening skupu podataka.
- Koristimo trenirani model da predviđamo vrednosti na test skupu.
- Evaluiramo model koristeći metrike srednje kvadratne greške (MSE) i R-kvadrata (R^2) kako bismo ocenili tačnost modela.

Rezultati za "AverageRatingsChurches"

- Srednja kvadratna greška (MSE): 0.438
- R-kvadrat (R^2): 0.371

Za atribut "AverageRatingsParks"

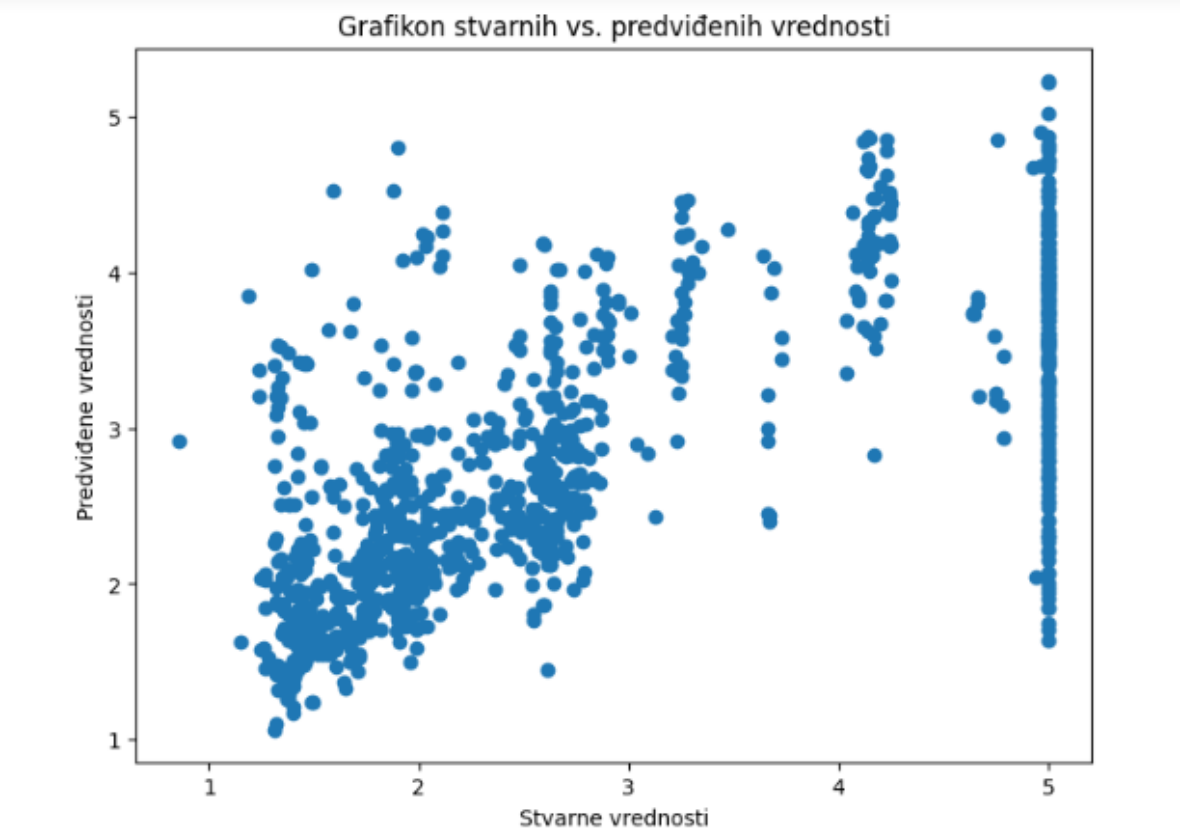
- Slično kao za crkve, definišemo ciljni atribut kao "AverageRatingsParks" i biramo sve druge atribute kao ulazne promenljive.
- Ponavljamo korake 2-5 za ovaj ciljni atribut.

Rezultati za "AverageRatingsParks"

- Srednja kvadratna greška (MSE): 0.879
- R-kvadrat (R^2): 0.498

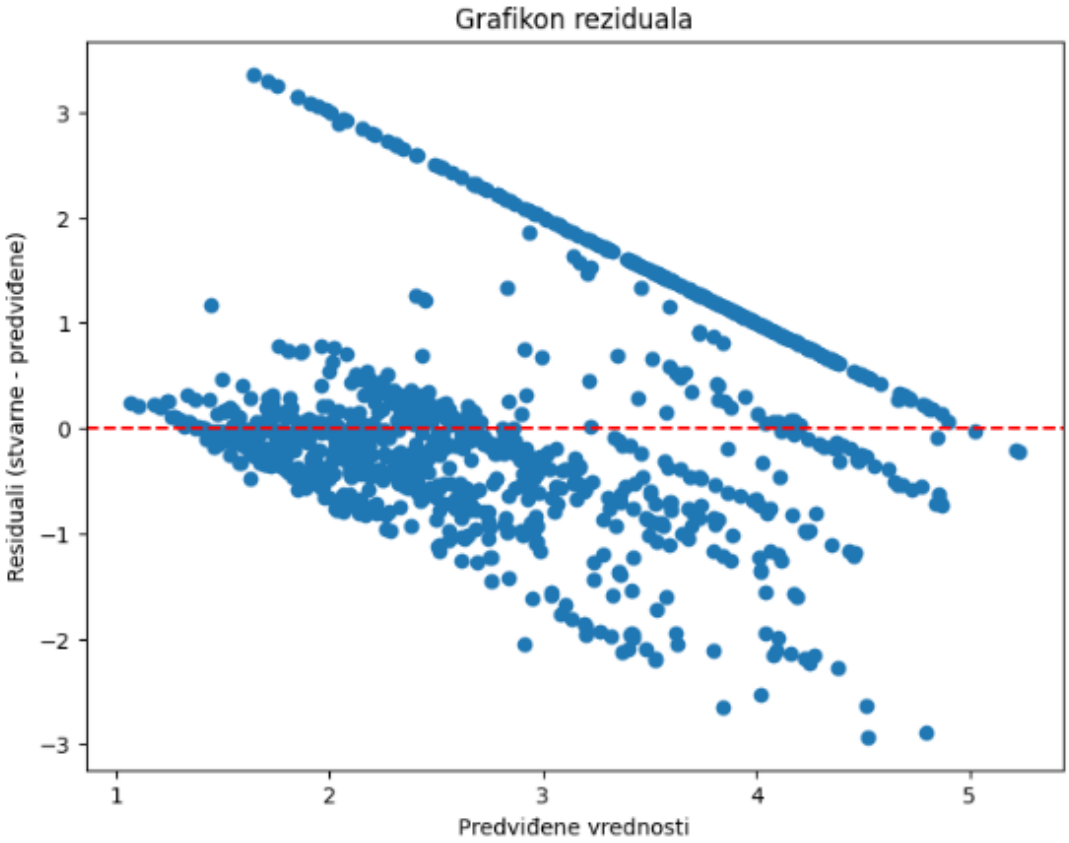
Nakon analize sa linearne regresije, vizualizujemo rezultate kako bismo bolje razumeli ponašanje modela. Grafikoni "stvarnih vs. predviđenih vrednosti" prikazuju koliko dobro model predviđa stvarne vrednosti, dok grafikoni reziduala pomažu u identifikaciji grešaka u modelu.

5.1.2 Grafikon stvarnih vs. predviđenih vrednosti



Ovaj grafikon prikazuje tačke u obliku rasipanja gde x-osa predstavlja stvarne vrednosti, a y-osa predviđene vrednosti. Idealno, tačke bi trebalo da se nalaze duž dijagonale, što znači da su stvarne i predviđene vrednosti skoro identične.

5.1.3 Računanje reziduala



Reziduali su razlika između stvarnih vrednosti i vrednosti koje je model predvideo, tj. to su greške koje model pravi u svojim predviđanjima. Drugi grafikon prikazuje rezidualne greške u odnosu na predviđene vrednosti. Linija na y=0 je horizontalna crvena linija koja predstavlja nulu reziduala. Ako reziduali leže oko horizontalne crvene linije na nuli (koja predstavlja nulu greške), to ukazuje na dobro kalibrisan model.

5.2 Primena Modela Support Vector Machine (SVM) za Regresiju

Sledeći korak u analizi podataka je primena modela Support Vector Machine (SVM) za regresiju, tačnije, Support Vector Regressor (SVR). SVM je još jedan algoritam koji se koristi za regresiju i ima svoje prednosti u odnosu na linearnu regresiju.

5.2.1 Za ciljni atribut "AverageRatingsChurches"

1. Definišemo ciljni atribut kao "AverageRatingsChurches" i biramo sve druge atribute kao ulazne promenljive.
2. Vršimo podelu podataka na trening i test skup koristeći 'train_test_split' funkciju.
3. Kreiramo i treniramo SVM model za regresiju (SVR) sa jezgrom 'rbf' na trening skupu podataka.
4. Koristimo trenirani SVM model da predviđamo vrednosti na test skupu.
5. Evaluiramo SVM model koristeći metrike srednje kvadratne greške (MSE) i R-kvadrata (R^2) kako bismo ocenili tačnost modela.

5.2.2 Rezultati za "AverageRatingsChurches"

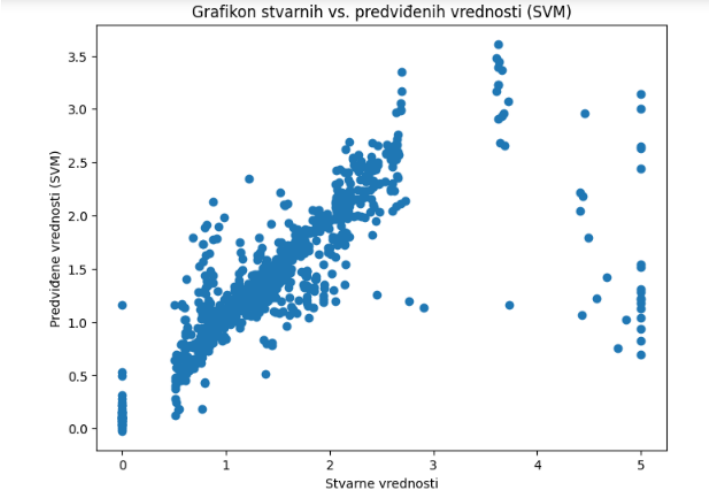
- Srednja kvadratna greška (MSE): 0.339
- R-kvadrat (R^2): 0.514

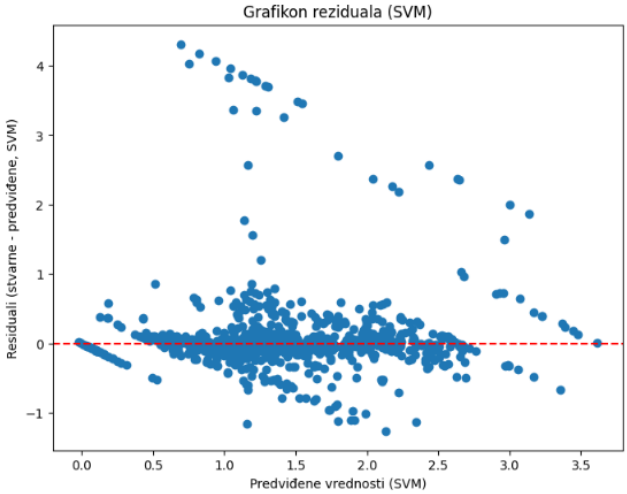
Srednja kvadratna greška (MSE) od 0.339 ukazuje na prosečnu kvadratnu razliku između stvarnih vrednosti i predviđenih vrednosti. Manja vrednost MSE sugeriše na bolju preciznost modela.

R-kvadrat (R^2) od 0.514 znači da ovaj SVM model objašnjava oko 51.4% varijacije u prosečnim ocenama crkava. Vrednost R-kvadrata od 1 označava savršeno objašnjenje varijacije, dok vrednost od 0 ukazuje na to da model ne objašnjava ništa. Ovaj rezultat sugeriše da ovaj SVM model dobro objašnjava varijaciju u ciljnom atributu.

Iz ovih rezultata zaključujemo da SVM model (sa jezgrom 'rbf') daje bolje performanse za predviđanje prosečnih ocena crkava u poređenju sa modelom linearne regresije za ovaj dataset i ciljni atribut.

5.2.3 Grafikoni stvarnih vs. predviđenih vrednosti i grafikoni reziduala za SVM model





Zaključak: SVM regresija (koristeći jezgro 'rbf') daje bolje performanse za predviđanje određenih ciljnih atributa u poređenju sa linearnom regresijom

6 Klasterovanje

Klasterovanje je tehnika mašinskog učenja koja se koristi za grupisanje podataka sličnih karakteristika zajedno u cilju identifikacije prirodnih grupa ili klastera. Ovaj proces omogućava da se podaci koji su međusobno slični grupišu u istom klasteru.

Razlozi za primenu klasterovanja nad ovim skupom podataka:

6.1 Segmentacija korisnika

Klasterovanje može pomoći da se identifikuju slične grupe korisnika koji imaju slične preference u vezi sa atrakcijama. Na primer, mogu se otkriti grupe korisnika koji vole muzeje, grupe koje više vole plaže, i tako dalje. Ovo može biti korisno za personalizaciju ponude ili marketinških strategija za svaku grupu.

6.2 Otkrivanje obrazaca

Klasterovanje može pomoći u otkrivanju skrivenih obrazaca ili grupisanju sličnih atrakcija zajedno. Na primer, može se primetiti da se muzeji i galerije često grupišu zajedno u isti klaster, što ukazuje na to da postoje slične karakteristike među njima koje privlače isti tip posetilaca.

6.3 Analiza tržišta

Razumevanje kako su atrakcije grupisane može biti korisno za analizu tržišta. Može se identifikovati koje kategorije atrakcija dominiraju u određenim regionima ili gradovima, što može pomoći u donošenju odluka o širenju ili prilagođavanju ponude.

6.4 Pretprocesiranje Podataka

6.4.1 Standardizacija podataka

Prvi korak u pretprocesiranju podataka je standardizacija numeričkih vrednosti u datasetu koristeći StandardScaler. Ovo je važno jer omogućava skaliranje podataka tako da imaju srednju vrednost 0 i standardnu devijaciju 1, što olakšava obradu podataka i analizu.

6.4.2 PCA (Principal Component Analysis) analiza za smanjenje dimenzionalnosti

Korak 1: Određivanje broja komponenti

Prvi korak u analizi PCA je određivanje broja komponenti koje čuvaju određeni procenat varijanse. Ovde smo postavili kao cilj da želimo da zadržimo 80% ukupne varijanse. **Dobijamo da je PCA odabrao 13 komponenata koje zajedno čuvaju 80% varijanse**

6.4.3 Korak 2: Instanciranje PCA modela

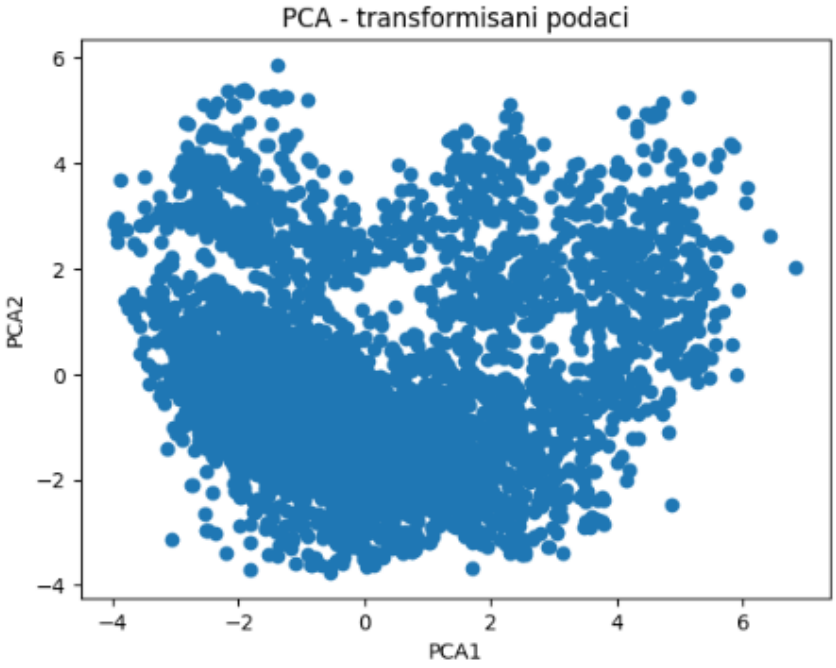
Nakon što odredimo broj komponenata, instanciramo PCA model sa željenim brojem komponenata (u ovom slučaju 2) za vizualizaciju. Takođe, prikazujemo koliko varijanse je sačuvano korišćenjem ovih komponenata i to je oko 34.2% ukupne varijacije.

Transformacija podataka

Nakon što odaberemo odgovarajući broj komponenata, transformišemo naše podatke koristeći PCA. Podatke ćemo predstaviti koristeći nove varijable (PCA1 i PCA2) umesto originalnih atributa.

Vizualizacija rezultata PCA

Prikazujemo rezultate PCA transformacije na grafiku kako bismo vizualno analizirali strukturu podataka.



6.4.4 Analiza broja klastera

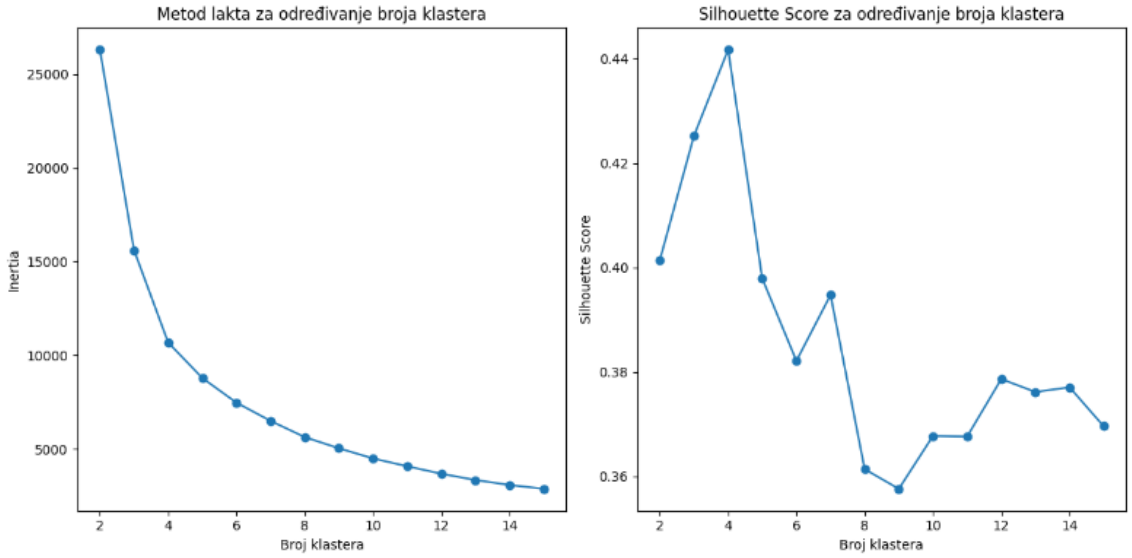
Korišćenje K-means algoritma

Koristimo K-means algoritam za analizu broja klastera u rasponu od 2 do 15 klastera. Za svaki broj klastera u rasponu, primenjujemo K-means algoritam na podatke.

Metod lakta (inertia) meri koliko su tačke u klasterima blizu svojih centara. Što je inertia niža, to je bolje, jer ukazuje na kompaktnije klasterovanje.

Takođe računamo **Silhouette skor** za svaki broj klastera. Silhouette skor meri koliko su tačke unutar istog klastera slične jedna drugoj i koliko su različite od tačaka u drugim klasterima. Viši Silhouette skorovi ukazuju na bolje klasterovanje.

Odabir broja klastera: Na osnovu rezultata metoda lakta i Silhouette skora, biramo broj klastera koji najbolje odgovara strukturi podataka. Obično biramo broj klastera gde se metoda lakta počinje "lomiti" ili gde Silhouette skor dostiže maksimum.

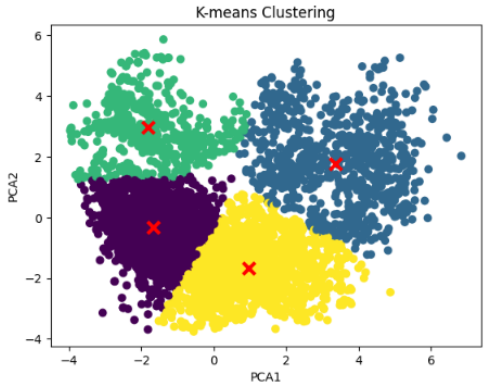



```
: #sa grafika se vidi da je lakat za 4 klastera, a i silueta skor je najbolji takodje u 4
```

6.5 K-means

Primenjujemo K-means algoritam za klasterovanje. K-means je algoritam koji grupiše podatke na osnovu njihove sličnosti i minimizira varijansu unutar klastera. Na osnovu prethodno dobijenog broja klastera biramo da modelujemo podatke sa 4 klastera.

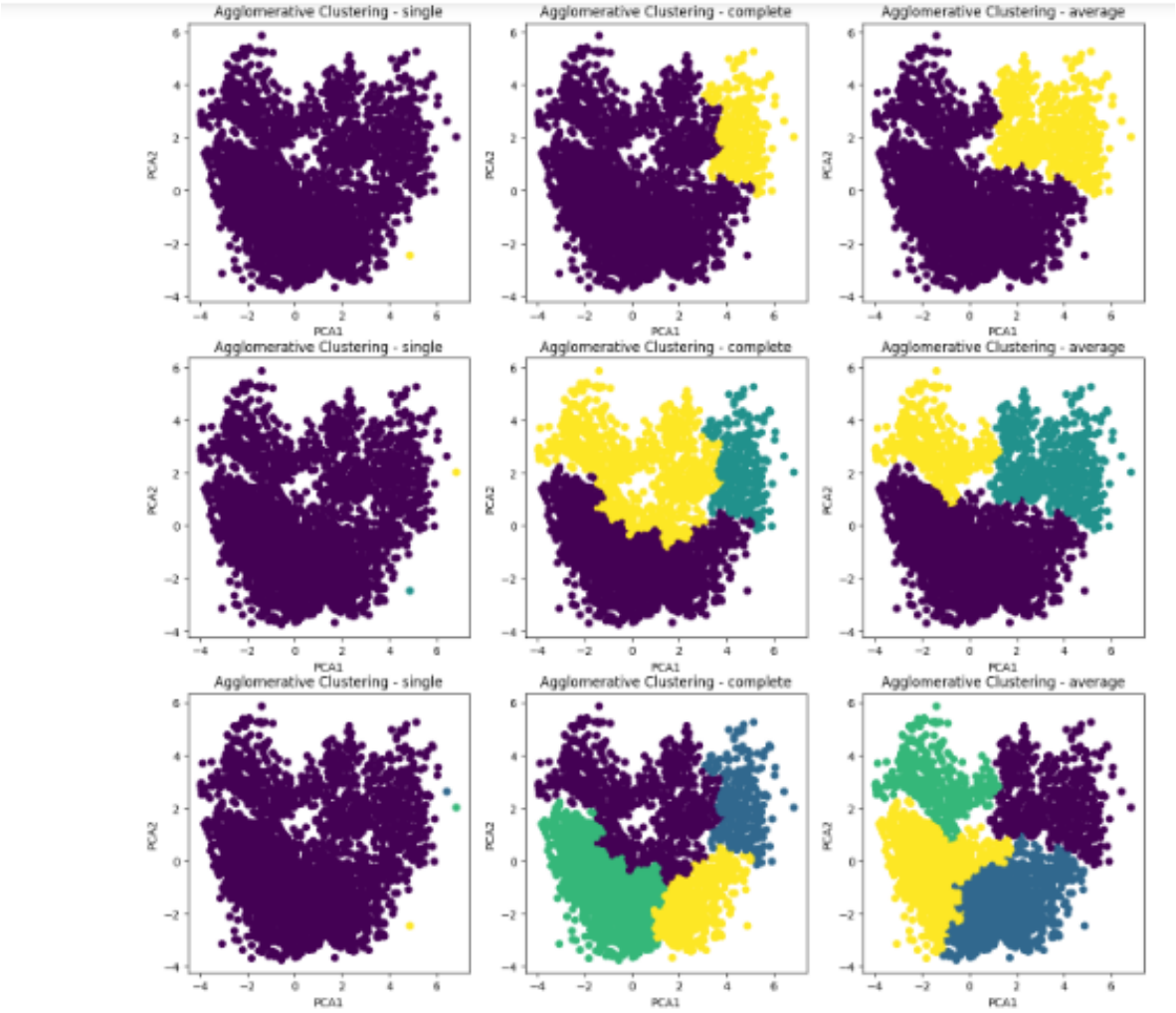
Nakon toga, vizualizujemo rezultate klasterovanja koristeći PCA komponente PCA1 i PCA2 kako bismo dobili uvid u strukturu klastera. Crveni "x" označava centralne tačke (centre) svakog klastera.



6.6 Hijerarhijsko klasterovanje:

Primenjujemo i hijerarhijsko klasterovanje koje formira hijerarhiju klastera i omogućava različite načine povezivanja podataka.

Pomoću hijerarhijskog klasterovanja, vizualizovali smo klasterovanje za različite brojeve klastera (2, 3 i 4) i različite načine povezivanja podataka.



Nakon ispitivanja odlučili smo da ćemo koristiti "average" vezu za povezivanje.

Silhouette skor:

Zbog upoređivanja kvaliteta klasterovanja između K-means i hijerarhijskog klasterovanja, koristimo Silhouette skor.

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

```
0.4416851319331852
```

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

```
0.4287924465463075
```

Rezultati pokazuju da su i K-means i hijerarhijsko klasterovanje daju relativno slične Silhouette skorove, pri čemu je K-means imao nešto bolji rezultat. Ovi rezultati sugerišu da oba algoritma imaju sličan kvalitet klasterovanja za ovaj skup podataka, ali da K-means može biti efikasniji za veće skupove podataka, dok hijerarhijsko klasterovanje omogućava dublje razumevanje strukture klastera.

Crtanje dendrograma za svaki klaster kako bi se vizualizovalo hijerarhijsko klasterovanje