Internet Firewall Dataset

Projekat za kurs Istraživanje podataka 1

Veljko Prodan

SADRŽAJ

1 UVOD

Ovaj izveštaj se bavi analizom skupa podataka pod nazivom "Internet firewall data", koji je prikupljen iz zapisa o internet saobraćaju na univerzitetskom firewall-u. Prikazana je primena algoritama klasifikacije (Stabla odlučivanja, K-najbližih suseda), klasterovanja (Algoritam K-sredina, Algoritam hijerarhijskog sakupljajućeg klasterovanja) i pravila pridruživanja u SPSS-u (Apriori algoritam).

2 ANALIZA SKUPA PODATAKA

Ovaj skup podataka je kompletan, bez ikakvih nedostajućih vrednosti. Sadrži 65532 instance sa sledećim atributima:

- Source Port
- Destination Port
- NAT Source Port
- NAT Destination Port
- Action
- Bytes
- Bytes Sent
- Bytes Received
- Packets
- Elapsed Time (sec)
- Packets Sent
- Packets Received

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65532 entries, 0 to 65531
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Source Port           65532 non-null  int64
 1   Destination Port      65532 non-null  int64
 2   NAT Source Port       65532 non-null  int64
 3   NAT Destination Port  65532 non-null  int64
 4   Action                65532 non-null  object
 5   Bytes                 65532 non-null  int64
 6   Bytes Sent            65532 non-null  int64
 7   Bytes Received        65532 non-null  int64
 8   Packets               65532 non-null  int64
 9   Elapsed Time (sec)    65532 non-null  int64
 10  pkts_sent             65532 non-null  int64
 11  pkts_received         65532 non-null  int64
dtypes: int64(11), object(1)
memory usage: 6.0+ MB
```

Figure 1: data.info()

Prilikom klasifikacije cilj je da predvidimo atribut ***action***, koji može imati vrednosti *allow*, *deny*, *drop* i *reset-both*.

Možemo da primetimo da je skup podataka nebalansiran. Instance u klasi *reset-both* čine svega 0.1% ukupnog broja instanci, pa će biti veoma teško predvideti tu klasu pri klasifikaciji.

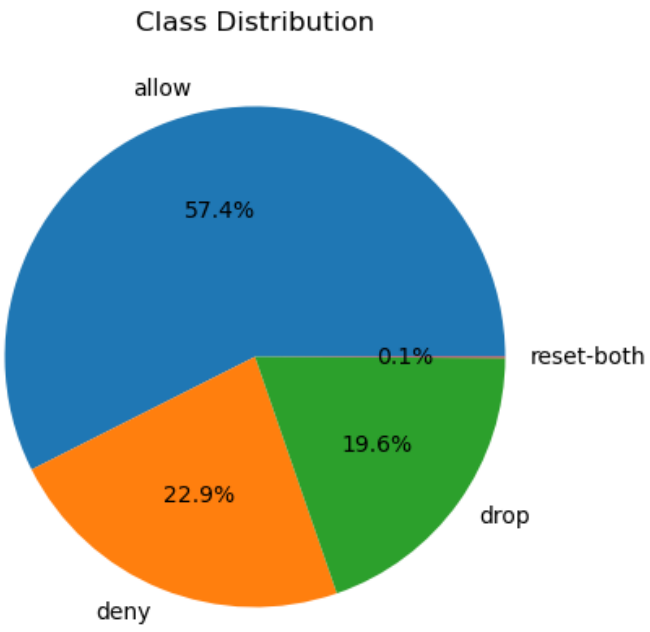
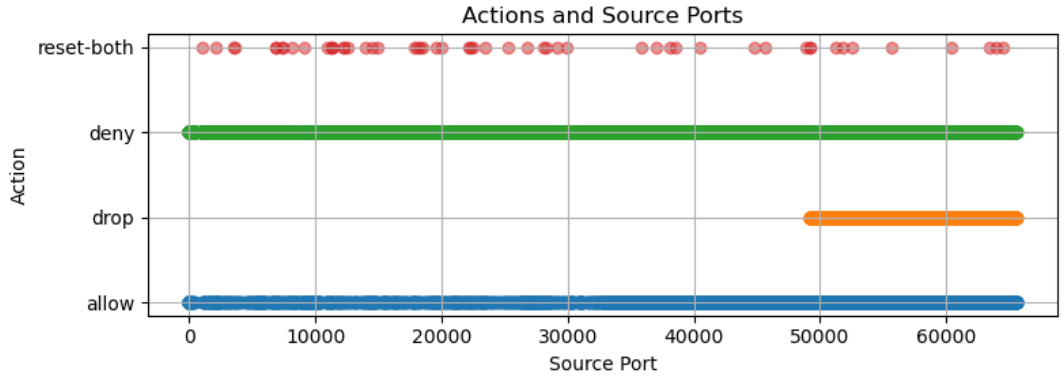
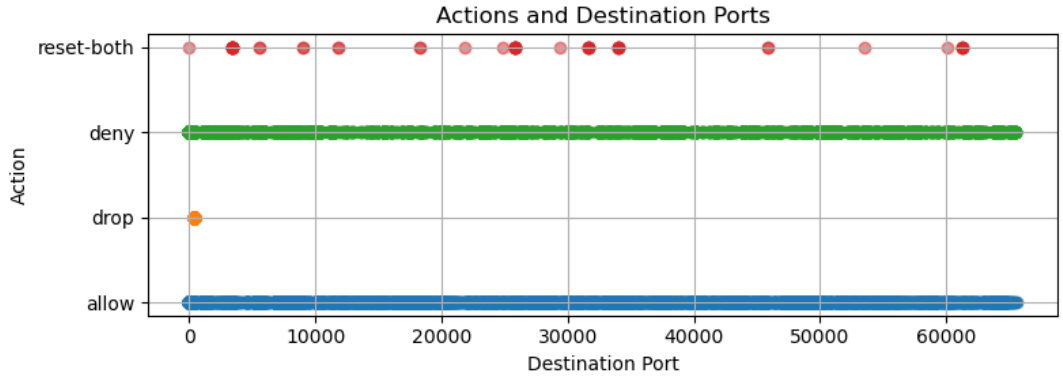
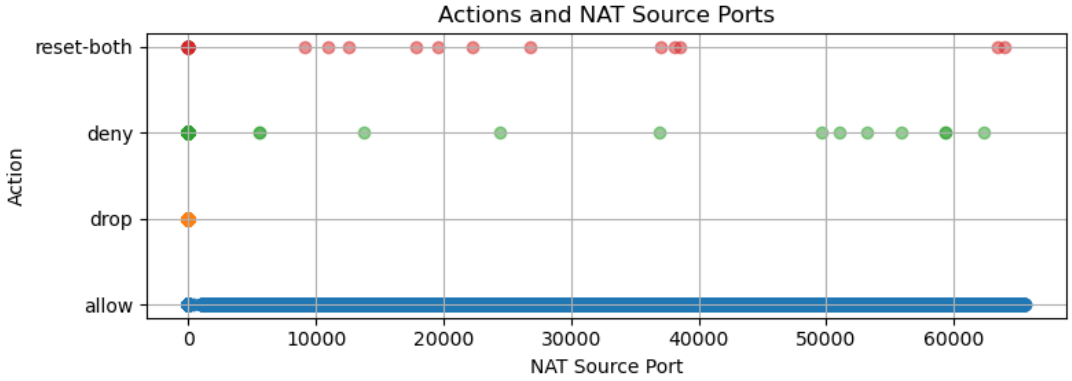
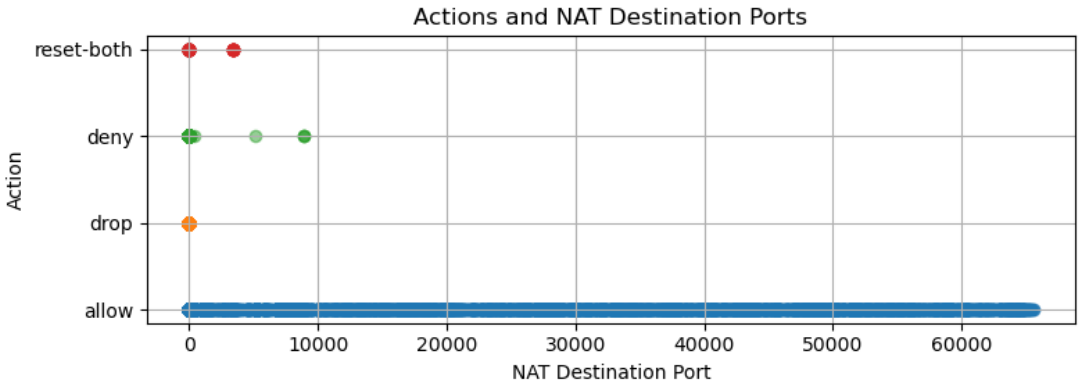


Figure 2: Distribucija klasa


Actions and Source Ports
reset-both
deny
drop
allow
Action
0
10000
20000
30000
40000
50000
60000
Source Port


Actions and Destination Ports
reset-both
deny
drop
allow
Action
0
10000
20000
30000
40000
50000
60000
Destination Port


Actions and NAT Source Ports
reset-both
deny
drop
allow
Action
0
10000
20000
30000
40000
50000
60000
NAT Source Port


Actions and NAT Destination Ports
reset-both
deny
drop
allow
Action
0
10000
20000
30000
40000
50000
60000
NAT Destination Port

3 PRETPROCESIRANJE

Na Figure 3 možemo videti da su vrednosti atributa *Bytes* skoro uvek zbir vrednosti atributa *Bytes Sent* i *Bytes Received*. Zbog ove gotovo konstantne veze, odlučeno je da se ukloni kolona *Bytes* kako bi se izbegla suvišnost u podacima. Slično je urađeno za atribut *Packets* zbog korelacije sa *pkts_sent* i *pkts_received*.

```
In [10]: print(data['Bytes'].corr(data['Bytes Sent'] + data['Bytes Received']))
         print(data['Packets'].corr(data['pkts_sent'] + data['pkts_received']))

0.9999999999999999
1.0
```

Figure 3

U matrici korelacije se jasno uočava povezanost između atributa *pkts_sent* i *Bytes Sent*. S obzirom na ovu korelaciju, uklanjamo atribut *pkts_sent*. Slično, uklanjamo atribut *pkts_received* zbog uske povezanosti sa atributom *Bytes Received*. Ovakva korekcija će unaprediti interpretaciju podataka i omogućiti fokusiranje na ključne karakteristike.

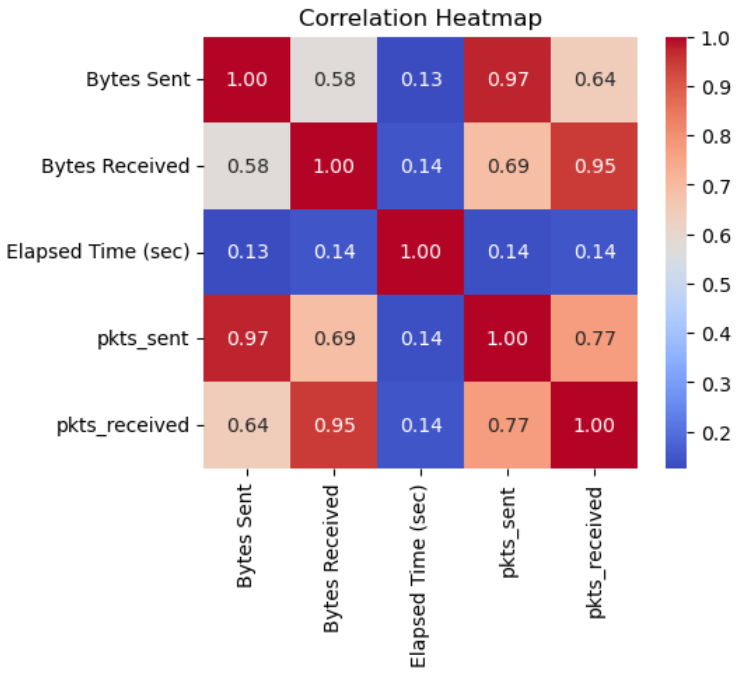


Figure 4: Matrica korelacije numeričkih atributa

S obzirom na to da atributi *Source Port, Destination Port, NAT Source Port* i *NAT Destination Port* imaju i numeričke i kategoričke karakteristike, čuvamo dve verzije skupa podataka.

U prvoj verziji, portovi su zadržani kao celobrojne vrednosti, kako bi se sačuvala njihova numerička priroda.

U drugoj verziji, iskoristili smo činjenicu da se portovi mogu kategorizovati u tri osnovne grupe:

1. Dobro poznati portovi (Well-known Ports): Ovi portovi imaju vrednosti u opsegu od 0 do 1023 i rezervisani su za najpoznatije i najčešće korišćene servise, kao što su HTTP (port 80) i HTTPS (port 443).
2. Registrovani portovi (Registered Ports): Ovi portovi imaju vrednosti u opsegu od 1024 do 49151 i koriste se za specifične servise koji nisu deo dobro poznate kategorije.
3. Dinamički ili privatni portovi (Dynamic or Private Ports): Ovi portovi imaju vrednosti u opsegu od 49152 do 65535 i namenjeni su privremenim ili dinamičkim vezama koje uspostavljaju klijentske aplikacije.

NAT (Network Address Translation) portovi ne mogu biti jasno svrstani u prethodno pomenute kategorije, pa ćemo izbaciti atribute *NAT Source Port* i *NAT Destination Port* u verziji dataset-a u kojoj smo portove kategorizovali.

	Source Port	Destination Port	Action	Bytes Sent	Bytes Received	Elapsed Time (sec)
0	Private	Well known	allow	94	83	30
1	Private	Registered	allow	1600	3168	17
2	Registered	Private	allow	118	120	1199
3	Private	Registered	allow	1438	1889	17
4	Private	Well known	allow	6778	18580	16

Figure 5: data.head() nakon izbacivanja NAT Port atributa

Verziju sa kategorizovanim portovima smo transformisali primenom *one-hot encoding* tehnike. Ovim omogućavamo efikasno reprezentovanje kategoričkih podataka kao binarne vektore čime olakšavamo dalju analizu i obradu podataka.

	Action	Bytes Sent	Bytes Received	Elapsed Time (sec)	Source Port_Registered	Source Port_Well known	Destination Port_Registered	Destination Port_Well known
0	allow	94	83	30	0	0	0	1
1	allow	1600	3168	17	0	0	1	0
2	allow	118	120	1199	1	0	0	0
3	allow	1438	1889	17	0	0	1	0
4	allow	6778	18580	16	0	0	0	1

Figure 6: data.head() nakon one-hot encoding-a

Za svrhe klasterovanja, primenjujemo **Analizu glavnih komponenti** (PCA) nad skupom podataka. Ova tehnika nam omogućava da redukujemo dimenzionalnost podataka, zadržavajući ključne informacije relevantne za analizu i grupisanje. Pre primene PCA, primenili smo postupak standardizacije podataka korišćenjem *Standard Scaler*-a kako bismo obezbedili da atributi imaju srednju vrednost 0 i standardnu devijaciju 1.

Nakon standardizacije peimenjujemo PCA kako bismo izdvojili tri ključne komponente koje zajedno obuhvataju značajan deo varijabilnosti u početnim podacima. Ove tri komponente zahvataju 67% varijanse skupa podataka,

čime smo zadržali ključne informacije za dalju analizu i proces klasterovanja. Ovakav pristup nam omogućava da radimo sa smanjenim brojem dimenzija, čime se olakšava interpretacija rezultata klasterovanja i omogućava efikasnija analiza podataka.

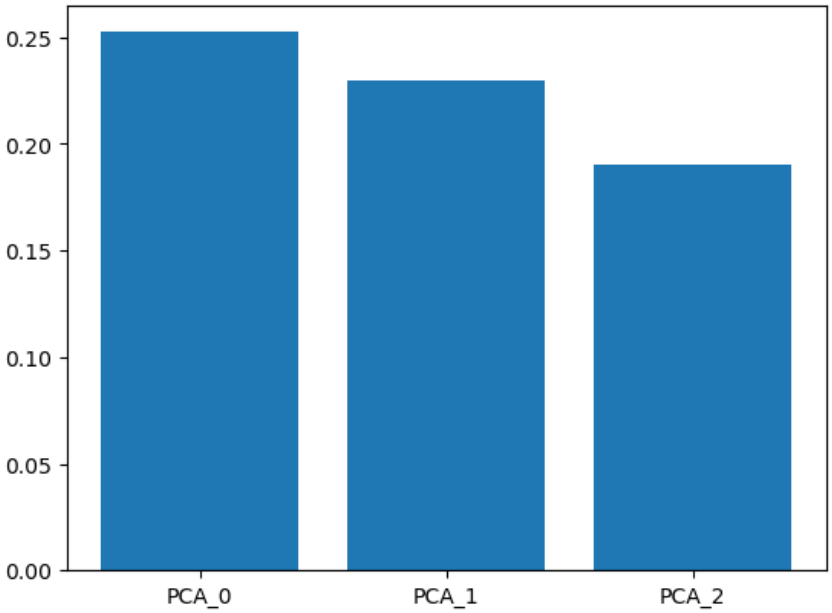


Figure 7: Udeo objašnjene varijense komponenti pojedinačno

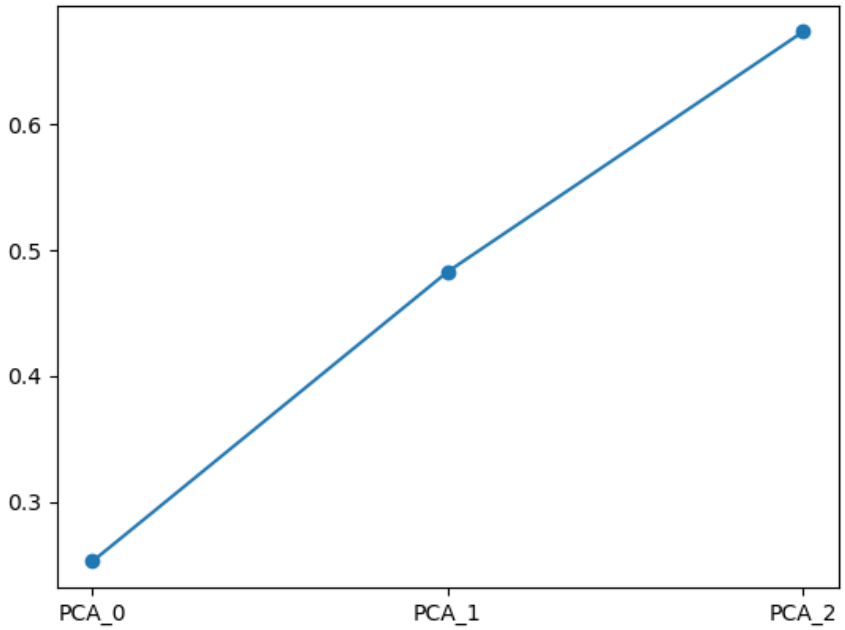


Figure 8: Kumulativna suma

4 KLASIFIKACIJA

U ovom poglavlju fokus je na algoritmima **Stabla odlučivanja** i **K-najbližih suseda**.

4.1 STABLA ODLUČIVANJA

Iz dataset-a izdvajamo ciljni atribut ***action***, i delimo ga na trening i test skup. Za test ćemo koristiti 25% podataka. Prvo isprobavamo stablo odlučivanja bez ikakvog nameštanja hiperparametara.

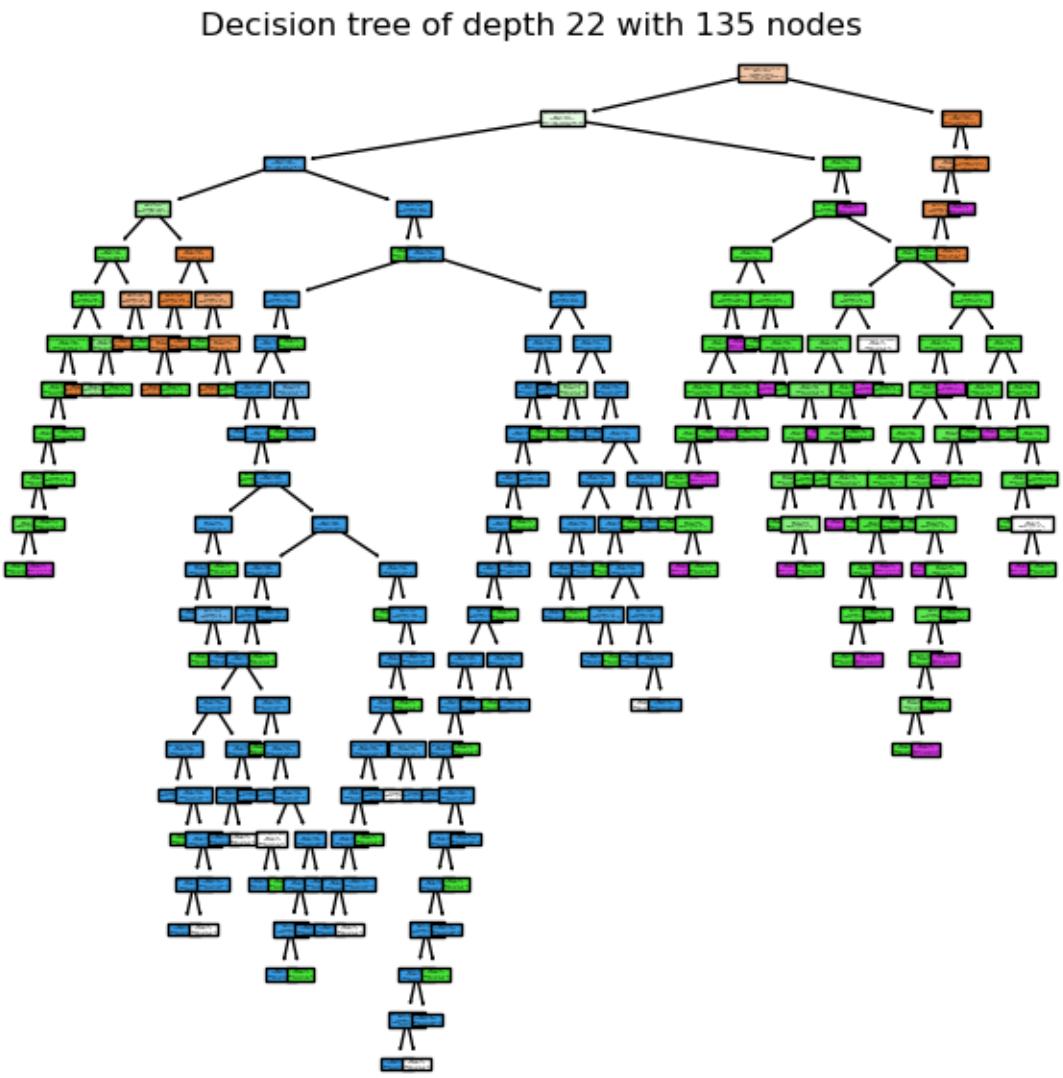


Figure 9: Stablo odlučivanja bez nameštanja hiperparametara

Na Figure 10 možemo videti metrike i matricu konfuzije za evaluaciju nad trening skupom.

```
Classification report for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------
              precision    recall  f1-score   support

       allow       1.00      1.00      1.00     28227
        deny       1.00      1.00      1.00     11240
        drop       1.00      1.00      1.00      9638
  reset-both       1.00      1.00      1.00        41

    accuracy                           1.00     49146
   macro avg       1.00      1.00      1.00     49146
weighted avg       1.00      1.00      1.00     49146

--------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------
            allow   drop  deny  reset-both
allow       28220      7     0           0
drop            0  11240     0           0
deny            0      6  9632           0
reset-both      0      0     0          41
--------------------------------------------------------------------------
```

Figure 10: Metrike i matrica konfuzije za evaluaciju nad trening skupom

Na Figure 11 možemo videti metrike i matricu konfuzije za evaluaciju nad test skupom. Definitivno se može reći da se model preprilagodio trening podacima, ali su rezultati zadovoljavajući. Za klasu *reset-both* su očekivano lošiji rezultati nego za ostale klase.

```
Classification report for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------
              precision    recall  f1-score   support

       allow       1.00      1.00      1.00      9410
        deny       0.99      1.00      1.00      3747
        drop       1.00      1.00      1.00      3213
  reset-both       0.70      0.54      0.61        13

    accuracy                           1.00     16383
   macro avg       0.92      0.88      0.90     16383
weighted avg       1.00      1.00      1.00     16383

--------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------
            allow  drop  deny  reset-both
allow        9406     3     0           1
drop            1  3737     7           2
deny            0    12  3201           0
reset-both      0     6     0           7
--------------------------------------------------------------------------
```

Figure 11: Metrike i matrica konfuzije za evaluaciju nad test skupom

4.1.1 GridSearch

Za pretragu kombinacija hiperparametara koristimo ***GridSearch***. Koristimo sledeće parametre:

1. **Kriterijum podele (criterion):**
    - **Gini**: mera nečistoće čvora. Ovaj kriterijum teži da minimizuje broj pogrešno klasifikovanih instanci
    - **Entropy**: Entropija je drugi kriterijum nečistoće. Koristi se kako bi se postigla ravnoteža između grananja i dubine stabla, i može rezultovati u bolje uravnoteženim stablima.
2. **Maksimalna dubina (max_depth):**
    - Ograničava maksimalnu dubinu stabla. Postavljanjem ove vrednosti, možete kontrolisati koliko grananja će se dešavati u modelu. Ovo je važan parametar za kontrolu kompleksnosti modela i prevenciju prenaučenosti (overfitting).
3. **Minimalni broj instanci za grananje (min_samples_split):**
    - Ovaj parametar postavlja minimalni broj instanci koje su neophodne kako bi se vršila dalja podela čvora. Ukoliko je broj instanci u čvoru manji od ove vrednosti, dalje grananje se neće izvršiti.
4. **Minimalni broj instanci u listu (min_samples_leaf):**
5. **Strategija podeljene tačke (splitter):**
    - **Best**: Ova opcija bira najbolju podeonu tačku koristeći kriterijum podele.
    - **Random**: Nasumično bira podeljenu tačku. Ova strategija može doneti raznolikost u stablu i pomoći u prevenciji preprilagođavanja.

```
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [12,15,18,20,None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'splitter': ['best', 'random']
}
```

Figure 12: GridSearch parametri

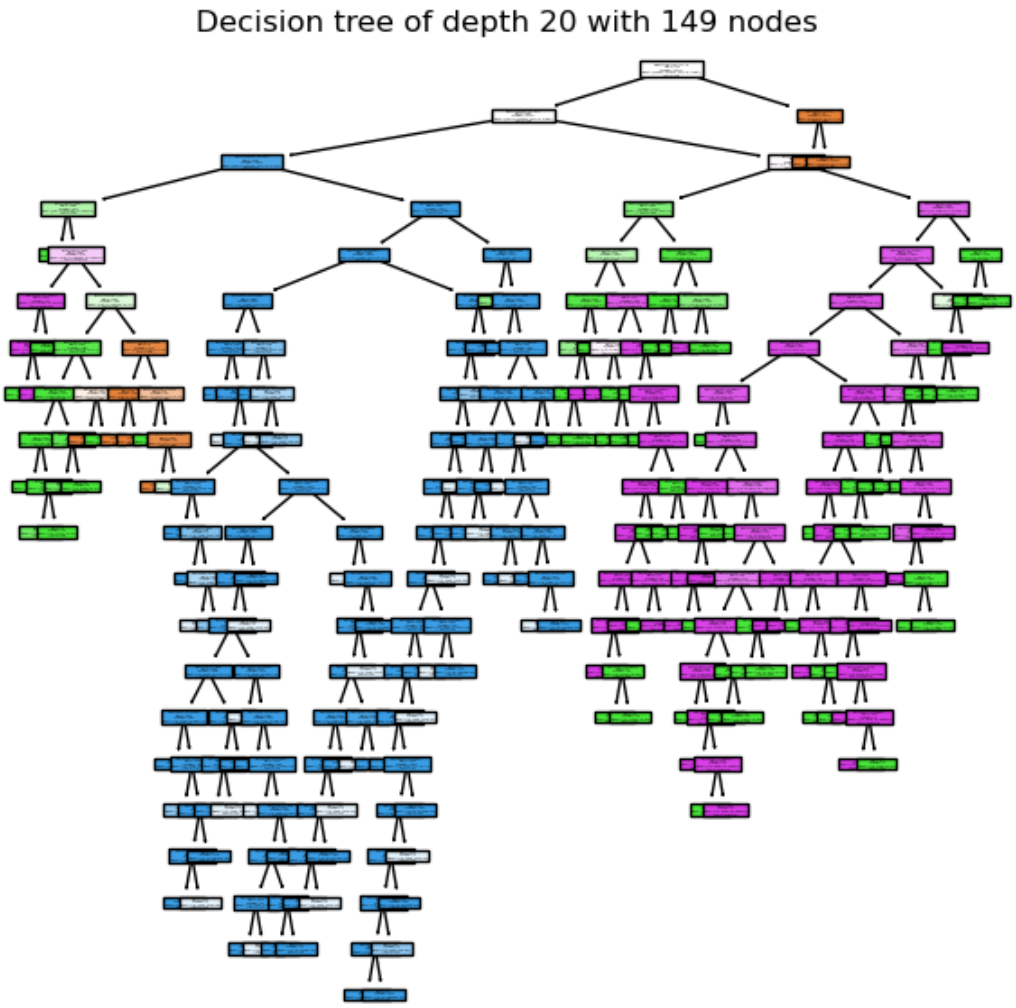


Figure 13: Stablo odlučivanja nakon GridSearch-a

```
Parameters of model DecisionTreeClassifier
ccp_alpha 0.0
class_weight balanced
criterion gini
max_depth 20
max_features None
max_leaf_nodes None
min_impurity_decrease 0.0
min_samples_leaf 2
min_samples_split 2
min_weight_fraction_leaf 0.0
random_state None
splitter best
```

Figure 14: Parametri nakon GridSearch-a

Na Figure 15 možemo videti metrike i matricu konfuzije nad trening skupom nakon GridSearch-a.

Classification report for model DecisionTreeClassifier on training data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	28227
deny	1.00	1.00	1.00	11240
drop	1.00	1.00	1.00	9638
reset-both	0.72	1.00	0.84	41
accuracy			1.00	49146
macro avg	0.93	1.00	0.96	49146
weighted avg	1.00	1.00	1.00	49146

Confusion matrix for model DecisionTreeClassifier on training data

	allow	drop	deny	reset-both
allow	28217	10	0	0
drop	0	11194	30	16
deny	0	1	9637	0
reset-both	0	0	0	41

Figure 15: Metrike i matrica konfuzije za evaluaciju nad trening skupom nakon GridSearch-a

Na Figure 16 možemo videti metrike i matricu konfuzije nad test skupom nakon GridSearch-a. Rezultati su skoro pa identični kao pre GridSearch-a.

Classification report for model DecisionTreeClassifier on test data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	9410
deny	1.00	1.00	1.00	3747
drop	1.00	1.00	1.00	3213
reset-both	0.35	0.46	0.40	13
accuracy			1.00	16383
macro avg	0.84	0.86	0.85	16383
weighted avg	1.00	1.00	1.00	16383

Confusion matrix for model DecisionTreeClassifier on test data

	allow	drop	deny	reset-both
allow	9406	3	0	1
drop	0	3729	8	10
deny	0	0	3213	0
reset-both	0	7	0	6

Figure 16: Metrike i matrica konfuzije nad test skupom nakon GridSearch-a.

4.1.2 Random forest classifier

Random forest classifier je algoritam koji se zasniva na ansambliranju stabala odlučivanja. On kombinuje predikcije više stabala kako bi doneo konačnu odluku o klasifikaciji. Ovaj pristup često vodi preciznijim predviđanjima u poređenju sa pojedinačnim stablima odlučivanja. Random Forest je takođe otporniji na preprilagođavanje i ima visoku sposobnost generalizacije na različite setove podataka.

Po rezulatatima na Figure 17 i Figure 18, možemo da zaključimo da se model ipak preprilagodio trening podacima, ali su rezultati nad test skupom ipak zadovoljavajuci, osim za klasu *reset-both* kao što je i očekivano.

Classification report for model RandomForestClassifier on training data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	28227
deny	1.00	1.00	1.00	11240
drop	1.00	1.00	1.00	9638
reset-both	1.00	1.00	1.00	41
accuracy			1.00	49146
macro avg	1.00	1.00	1.00	49146
weighted avg	1.00	1.00	1.00	49146

Confusion matrix for model RandomForestClassifier on training data

	allow	drop	deny	reset-both
allow	28220	7	0	0
drop	0	11234	6	0
deny	0	0	9638	0
reset-both	0	0	0	41

Figure 17: Metrike i matrica konfuzije za evaluaciju nad trening skupom

Classification report for model RandomForestClassifier on test data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	9410
deny	0.99	1.00	1.00	3747
drop	1.00	1.00	1.00	3213
reset-both	0.75	0.23	0.35	13
accuracy			1.00	16383
macro avg	0.94	0.81	0.84	16383
weighted avg	1.00	1.00	1.00	16383

Confusion matrix for model RandomForestClassifier on test data

	allow	drop	deny	reset-both
allow	9406	3	0	1
drop	0	3740	7	0
deny	0	11	3202	0
reset-both	0	10	0	3

Figure 18: Metrike i matrica konfuzije za evaluaciju nad test skupom

4.2 ALGORITAM K-NAJBLIŽIH SUSEDA

K-najbližih suseda (KNN) je jednostavan i intuitivan algoritam koji se koristi za klasifikaciju i regresiju. Osnovna ideja iza KNN algoritma je da se novi, nepoznati podaci klasifikuju ili predviđaju na osnovu sličnosti sa "k" najbližih podataka iz trening skupa.

Za primenu KNN-a ćemo koristiti dva skupa podataka. Prvi je skup gde su sačuvane numeričke vrednosti portova, a drugi je skup u kom smo podelili portove u tri različite kategorije i transformisali *one-hot encoding* tehnikom.

4.2.1 Prvi skup podataka

Iz skupa podataka izdvajamo ***action*** kolonu i delimo ga na trening i test skup u razmeri 3:1. Kako sve preostale atribute možemo da gledamo kao numeričke, podatke skaliramo *MinMaxScaler*-om, tako da su minimalne vrednosti atributa 0, a maksimalne 1. Skaliranje je ključno radi osiguravanja ravnomernog doprinosa svakog atributa u procesu donošenja odluke. Bez skaliranja, atributi sa većim vrednostima mogu dominirati u izračunavanjima udaljenosti, što može dovesti do netačnih rezultata.

Prvo primenjujemo KNN bez nameštanja hiperparametara. Na Figure 19 i Figure 20 možemo videti metrike i matrice konfuzije na trening i test skupu.

```
Classification report for model KNeighborsClassifier on training data
------------------------------------------------------------------------
              precision    recall  f1-score   support

       allow       1.00      1.00      1.00     28227
        deny       0.99      0.99      0.99     11240
        drop       1.00      1.00      1.00      9638
  reset-both       1.00      0.05      0.09        41

    accuracy                           1.00     49146
   macro avg       1.00      0.76      0.77     49146
weighted avg       1.00      1.00      1.00     49146

------------------------------------------------------------------------
Confusion matrix for model KNeighborsClassifier on training data
------------------------------------------------------------------------
            allow   drop  deny  reset-both
allow       28135     92     0           0
drop           20  11183    37           0
deny            0      3  9635           0
reset-both      8     31     0           2
------------------------------------------------------------------------
```

Figure 19: Metrike i matrica konfuzije za evaluaciju nad trening skupom

```
Classification report for model KNeighborsClassifier on test data
----------------------------------------------------------------------
              precision    recall  f1-score   support

       allow       1.00      1.00      1.00      9410
        deny       0.99      0.99      0.99      3747
        drop       1.00      1.00      1.00      3213
  reset-both       1.00      0.00      0.00        13

    accuracy                           1.00     16383
   macro avg       1.00      0.75      0.75     16383
weighted avg       1.00      1.00      1.00     16383

----------------------------------------------------------------------
Confusion matrix for model KNeighborsClassifier on test data
----------------------------------------------------------------------
            allow  drop  deny  reset-both
allow        9380    29     1           0
drop           12  3726     9           0
deny            0     0  3213           0
reset-both      3    10     0           0
----------------------------------------------------------------------
```

Figure 20: Metrike i matrica konfuzije za evaluaciju nad test skupom

4.2.1.1 GridSearch

Za podešavanje parametara ponovo koristimo ***GridSearch***:

1. **Broj suseda (n_neighbors):**
    - Ovaj parametar predstavlja broj najbližih suseda koje algoritam uzima u obzir
2. **Težine (weights):**
    - Uniform: svi susedi imaju istu težinu i doprinose jednakom delu u odluci.
    - Distance: težine su obrnuto proporcionalne udaljenosti od tačke koja se klasifikuje. Bliski susedi imaju veći uticaj na odluku.
3. **Mera udaljenosti (p):**
    - Parametar p određuje koju meru udaljenosti će KNN koristiti. Obično se koristi 1 (*Manhattan* udaljenost) ili 2 (Euklidska udaljenost).
4. **Algoritam za pronalaženje suseda (algorithm):**
    - auto: Algoritam će automatski odabrati najbolji algoritam na osnovu unetih podataka.
    - ball_tree, kd_tree: Koristi se struktura drveta za pretragu suseda.
    - brute: Koristi se brute-force pretraga

```
{'algorithm': 'auto', 'n_neighbors': 5, 'p': 1, 'weights': 'distance'}
```

Figure 21: Parametri odabrani GridSearch-om

```
Classification report for model KNeighborsClassifier on training data
--------------------------------------------------------------------------------
              precision    recall  f1-score   support

       allow       1.00      1.00      1.00     28227
        deny       1.00      1.00      1.00     11240
        drop       1.00      1.00      1.00      9638
  reset-both       1.00      1.00      1.00        41

    accuracy                           1.00     49146
   macro avg       1.00      1.00      1.00     49146
weighted avg       1.00      1.00      1.00     49146

--------------------------------------------------------------------------------
Confusion matrix for model KNeighborsClassifier on training data
--------------------------------------------------------------------------------
            allow   drop  deny  reset-both
allow       28220      7     0           0
drop            0  11240     0           0
deny            0      6  9632           0
reset-both      0      0     0          41
--------------------------------------------------------------------------------
```

Figure 22: Metrike i matrica konfuzije za evaluaciju nad trening skupom nakon GridSearch-a

```
Classification report for model KNeighborsClassifier on test data
--------------------------------------------------------------------------------
              precision    recall  f1-score   support

       allow       1.00      1.00      1.00      9410
        deny       0.99      1.00      0.99      3747
        drop       1.00      1.00      1.00      3213
  reset-both       0.50      0.08      0.13        13

    accuracy                           1.00     16383
   macro avg       0.87      0.77      0.78     16383
weighted avg       1.00      1.00      1.00     16383

--------------------------------------------------------------------------------
Confusion matrix for model KNeighborsClassifier on test data
--------------------------------------------------------------------------------
            allow  drop  deny  reset-both
allow        9386    22     1           1
drop            4  3733    10           0
deny            0     3  3210           0
reset-both      3     9     0           1
--------------------------------------------------------------------------------
```

Figure 23: Metrike i matrica konfuzije za evaluaciju nad test skupom nakon GridSearch-a

Na Figure 22 i Figure 23 možemo videti da smo dobili veoma slične rezultate kao pre GridSearch-a.

4.2.1.2 Bagging Classifier

Bagging Classifier za KNN je ansambl tehnika mašinskog učenja koja kombinuje više modela KNN kako bi poboljšala tačnost i stabilnost predviđanja. Ovaj pristup se oslanja na ideju da konsenzus više modela može doneti bolje rezultate od pojedinačnih modela. Konkretno, BaggingClassifier primenjuje KNN na različite podskupove trening podataka, omogućavajući svakom modelu da nauči specifične uzorke i odstupanja u podacima. Kombinovanjem ovih modela, BaggingClassifier postiže bolje generalizacije i smanjuje rizik od preprilagođavanja.

Classification report for model BaggingClassifier on training data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	28227
deny	0.99	1.00	0.99	11240
drop	1.00	1.00	1.00	9638
reset-both	1.00	0.05	0.09	41
accuracy			1.00	49146
macro avg	1.00	0.76	0.77	49146
weighted avg	1.00	1.00	1.00	49146

Confusion matrix for model BaggingClassifier on training data

	allow	drop	deny	reset-both
allow	28137	90	0	0
drop	17	11186	37	0
deny	0	1	9637	0
reset-both	7	32	0	2

Figure 24: Metrike i matrica konfuzije za evaluaciju nad trening skupom

Classification report for model BaggingClassifier on test data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	9410
deny	0.99	0.99	0.99	3747
drop	1.00	1.00	1.00	3213
reset-both	1.00	0.00	0.00	13
accuracy			1.00	16383
macro avg	1.00	0.75	0.75	16383
weighted avg	1.00	1.00	1.00	16383

Confusion matrix for model BaggingClassifier on test data

	allow	drop	deny	reset-both
allow	9378	31	1	0
drop	12	3725	10	0
deny	0	0	3213	0
reset-both	3	10	0	0

Figure 25: Metrike i matrica konfuzije za evaluaciju nad test skupom

4.2.2 Drugi skup podataka

Za skup podataka sa portovima kao kategoričkim atributima je korišćen isti proces nalaska optimalnih hiperparametara. Nema drastične razlike u rezultatima u odnosu na prvi skup podataka.

Classification report for model KNeighborsClassifier on training data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	28227
deny	1.00	1.00	1.00	11240
drop	1.00	1.00	1.00	9638
reset-both	0.86	0.15	0.25	41
accuracy			1.00	49146
macro avg	0.96	0.79	0.81	49146
weighted avg	1.00	1.00	1.00	49146

Confusion matrix for model KNeighborsClassifier on training data

	allow	drop	deny	reset-both
allow	28208	18	0	1
drop	5	11195	40	0
deny	0	0	9638	0
reset-both	0	35	0	6

Figure 26: Metrike i matrica konfuzije za evaluaciju nad trening skupom pre nameštanja parametara

Classification report for model KNeighborsClassifier on training data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	9410
deny	0.99	1.00	1.00	3747
drop	1.00	1.00	1.00	3213
reset-both	1.00	0.08	0.14	13
accuracy			1.00	16383
macro avg	1.00	0.77	0.78	16383
weighted avg	1.00	1.00	1.00	16383

Confusion matrix for model KNeighborsClassifier on training data

	allow	drop	deny	reset-both
allow	9403	7	0	0
drop	5	3730	12	0
deny	0	0	3213	0
reset-both	0	12	0	1

Figure 27: Metrike i matrica konfuzije za evaluaciju nad test skupom pre nameštanja parametara

Classification report for model KNeighborsClassifier on training data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	28227
deny	1.00	1.00	1.00	11240
drop	1.00	1.00	1.00	9638
reset-both	1.00	0.22	0.36	41
accuracy			1.00	49146
macro avg	1.00	0.80	0.84	49146
weighted avg	1.00	1.00	1.00	49146

Confusion matrix for model KNeighborsClassifier on training data

	allow	drop	deny	reset-both
allow	28220	7	0	0
drop	1	11199	40	0
deny	0	0	9638	0
reset-both	0	32	0	9

Figure 28: Metrike i matrica konfuzije za evaluaciju nad trening skupom nakon GridSearch-a

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	9410
deny	1.00	1.00	1.00	3747
drop	1.00	1.00	1.00	3213
reset-both	1.00	0.23	0.38	13
accuracy			1.00	16383
macro avg	1.00	0.81	0.84	16383
weighted avg	1.00	1.00	1.00	16383

Confusion matrix for model KNeighborsClassifier on test data

	allow	drop	deny	reset-both
allow	9403	7	0	0
drop	4	3731	12	0
deny	0	0	3213	0
reset-both	0	10	0	3

Figure 29: Metrike i matrica konfuzije za evaluaciju nad test skupom nakon GridSearch-a

Classification report for model BaggingClassifier on training data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	28227
deny	1.00	1.00	1.00	11240
drop	1.00	1.00	1.00	9638
reset-both	0.86	0.15	0.25	41
accuracy			1.00	49146
macro avg	0.96	0.79	0.81	49146
weighted avg	1.00	1.00	1.00	49146

Confusion matrix for model BaggingClassifier on training data

	allow	drop	deny	reset-both
allow	28206	20	0	1
drop	5	11195	40	0
deny	0	0	9638	0
reset-both	0	35	0	6

Figure 30: Metrike i matrica konfuzije za evaluaciju nad trening skupom nakon BaggingClassifier-a

Classification report for model BaggingClassifier on test data

	precision	recall	f1-score	support
allow	1.00	1.00	1.00	9410
deny	0.99	1.00	1.00	3747
drop	1.00	1.00	1.00	3213
reset-both	1.00	0.08	0.14	13
accuracy			1.00	16383
macro avg	1.00	0.77	0.78	16383
weighted avg	1.00	1.00	1.00	16383

Confusion matrix for model BaggingClassifier on test data

	allow	drop	deny	reset-both
allow	9403	7	0	0
drop	5	3730	12	0
deny	0	0	3213	0
reset-both	0	12	0	1

Figure 31:: Metrike i matrica konfuzije za evaluaciju nad test skupom nakon BeggingClassifier-a

5 KLASTEROVANJE

Klasterovanje je metoda nenadgledanog učenja koja ima za cilj grupisanje sličnih instanci u diskretne grupe, nazvane klasteri, na osnovu zajedničkih karakteristika. Ovaj proces omogućava otkrivanje prirodnih struktura u podacima i razumevanje odnosa među podacima.

Primenićemo dva algoritma klasterovanja: **Algoritam K-sredina** i **Algoritam sakupljajućeg hijerarhijskog klasterovanja**.

Za implementaciju ovih algoritama, koristimo dataset koji je prethodno podvrgnut analizi glavnih komponenti (PCA). Ovaj dataset sada poseduje tri ključna atributa, što će nam omogućiti dobru vizuelizaciju klasterovanja.

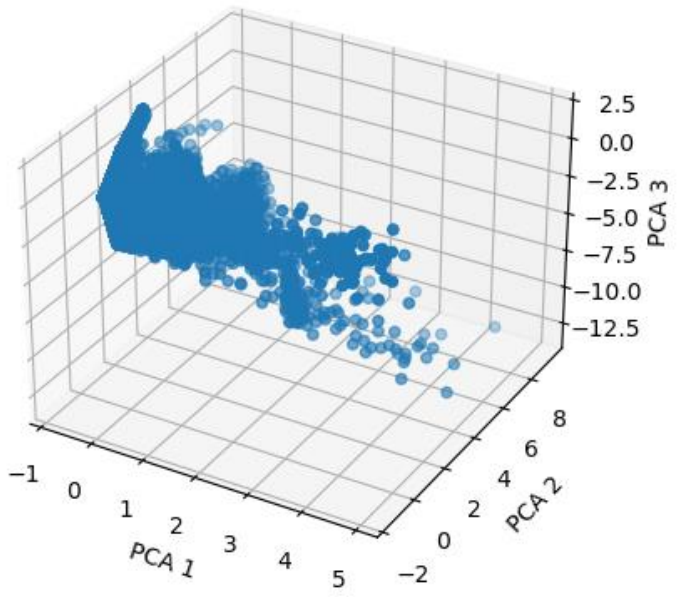
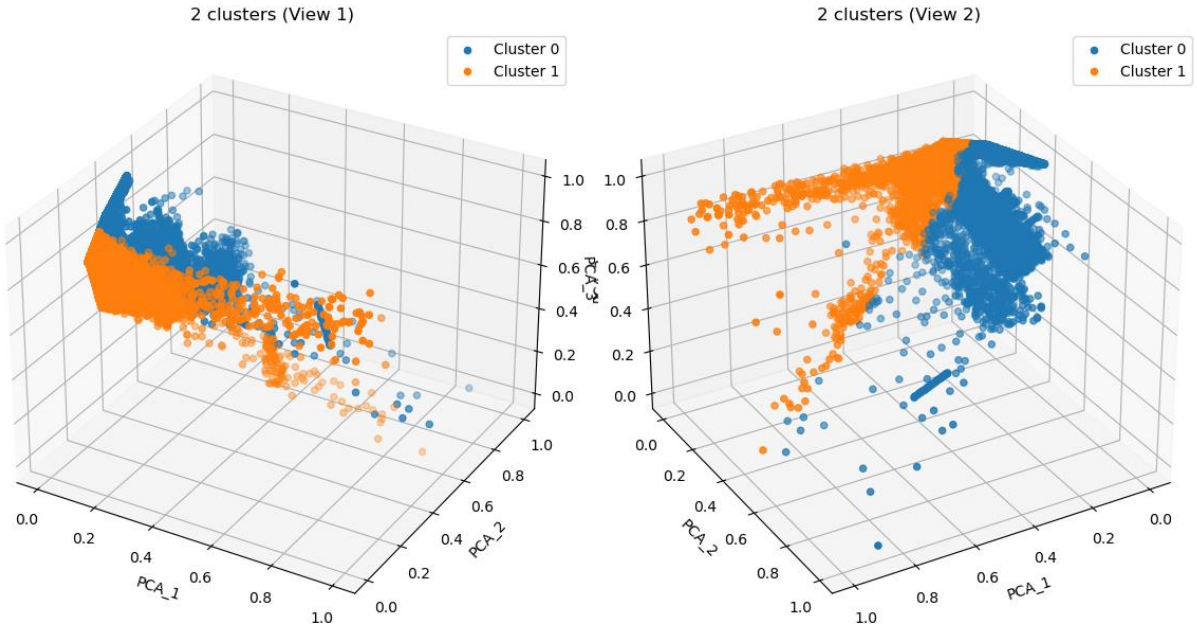


Figure 32: Vizuelizacija podataka pre klasterovanja

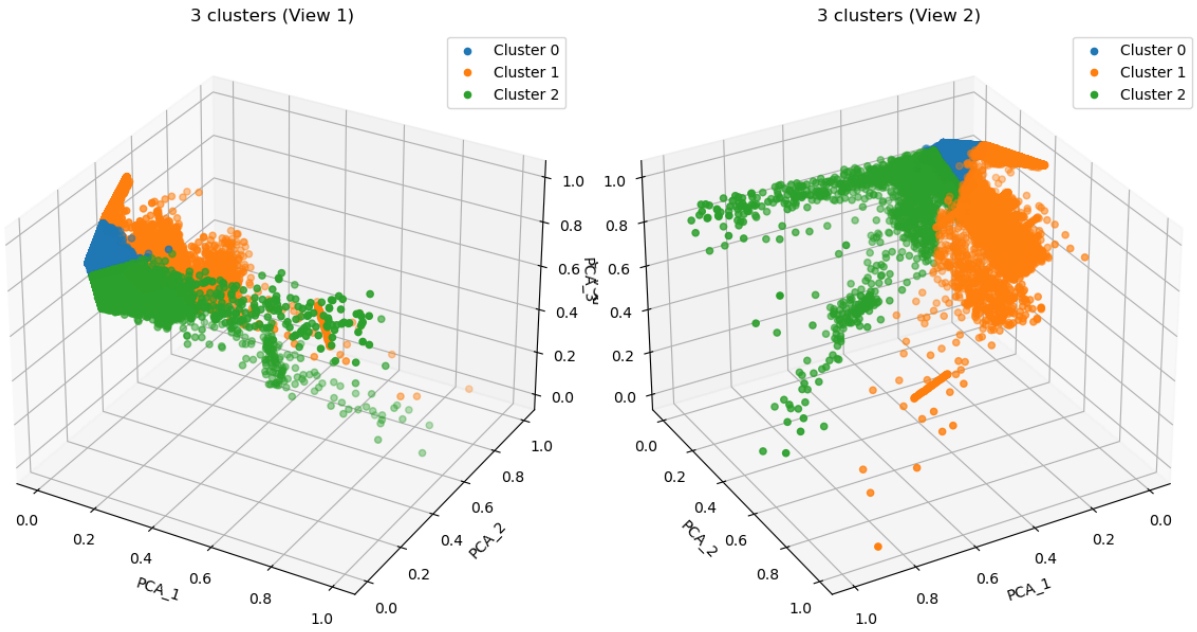
5.1 ALGORITAM K-SREDINA

K-means je iterativni algoritam klasterovanja koji počinje sa nasumičnim postavljanjem centroida klastera u prostoru podataka. Za svaku instancu, dodeljuje je klasteru čiji je centar najbliži, koristeći meru udaljenosti poput Euklidske. Centri klastera se zatim ažuriraju na osnovu novih grupacija, postavljajući ih u središta svih instanci dodeljenih tom klasteru. Ovaj proces se ponavlja dok se centri klastera ne stabilizuju, što ukazuje na konvergenciju algoritma.

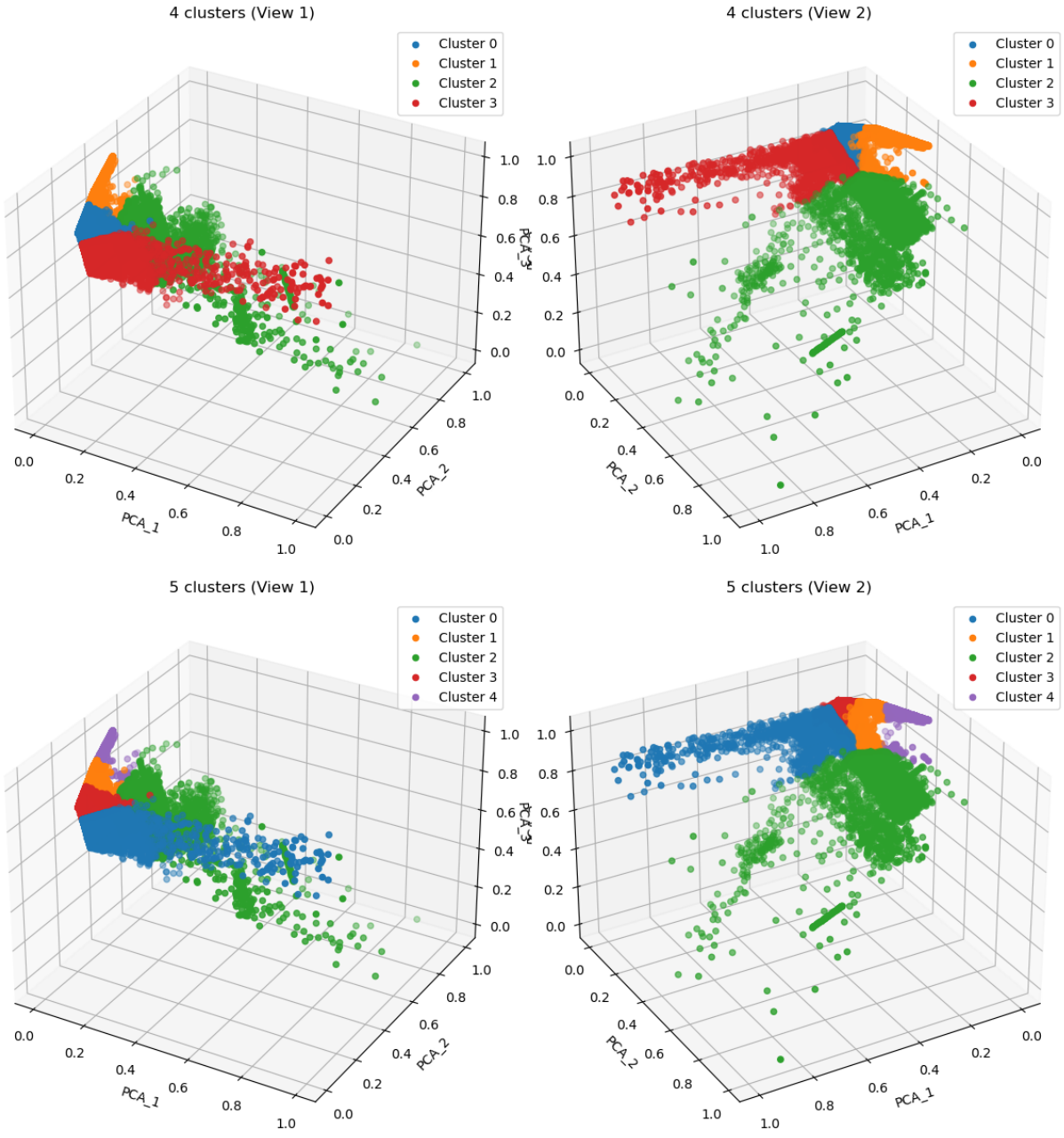
Na narednim graficima prikazano je isprobavanje različitih vrednosti parametra k.


2 clusters (View 1)
Cluster 0
Cluster 1
PCA_1
PCA_2
PCA_3

2 clusters (View 2)
Cluster 0
Cluster 1
PCA_1
PCA_2


3 clusters (View 1)
Cluster 0
Cluster 1
Cluster 2
PCA_1
PCA_2
PCA_3

3 clusters (View 2)
Cluster 0
Cluster 1
Cluster 2
PCA_1
PCA_2



Dolazimo do zaključka da je najbolji sa parametrom k=2 jer je u tom slučaju ***silhouette score*** najveći.

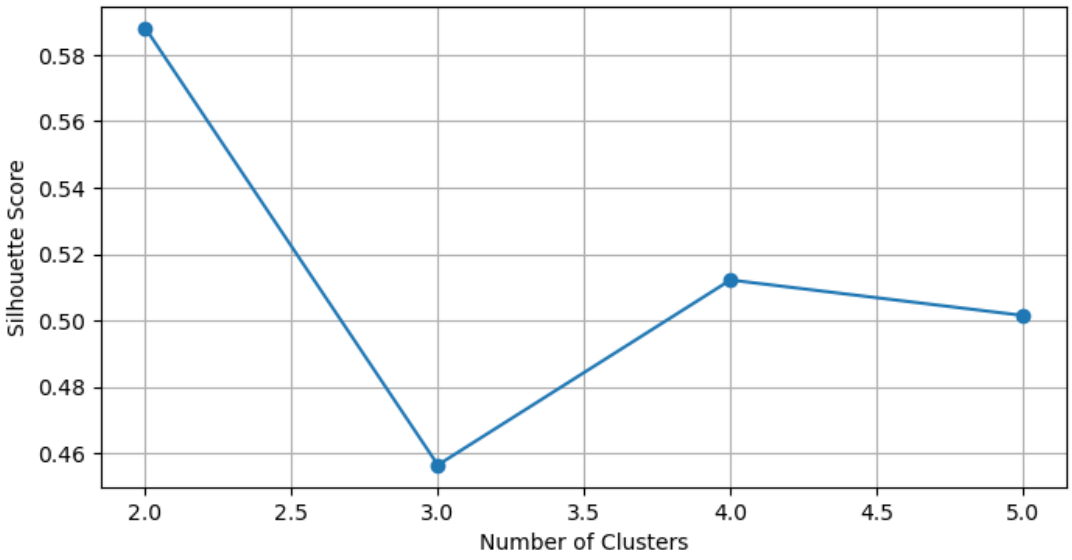
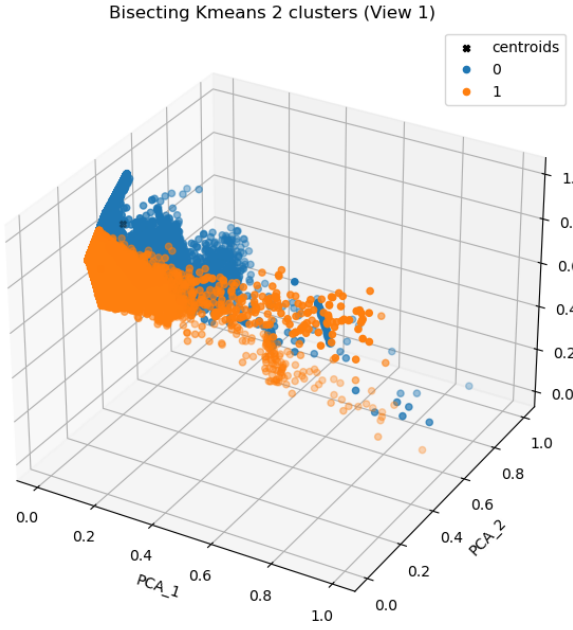


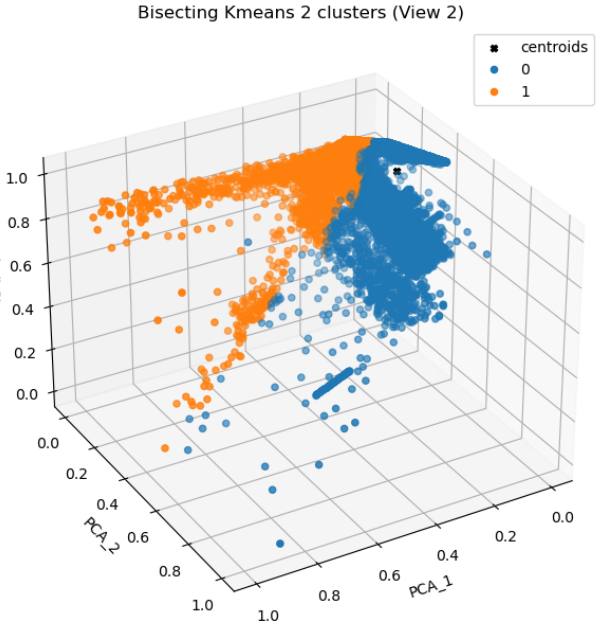
Figure 33: Silhouette score u odnosu na broj klastera

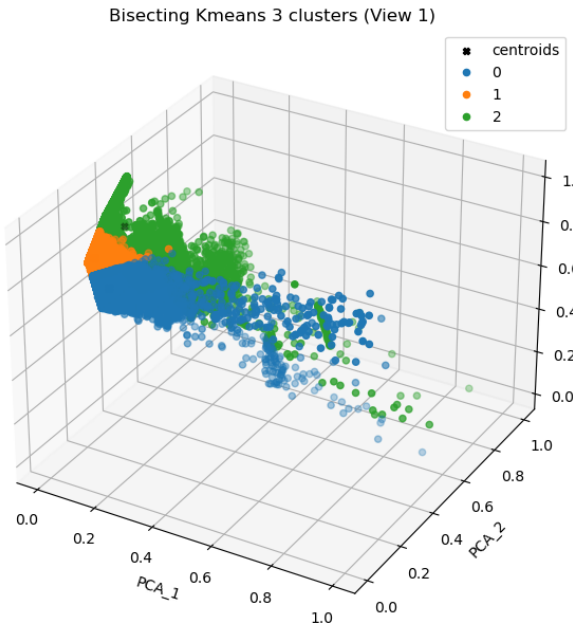
5.1.1 Bisecting K-means

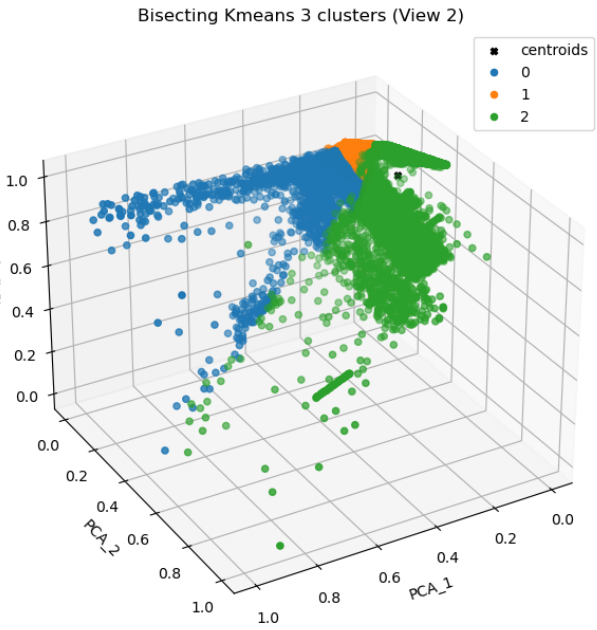
Bisecting K-means je algoritam za klasterovanje gde umesto da se počne sa jednim klasterom i podeli ga u više manjih, počinje se sa celim skupom podataka kao jednim klasterom. Zatim, iterativno deli taj klaster na dva manja klastera. Zatim se bira klaster sa najvećom inercijom koji će biti sledeći podeljen, sve dok se ne dođe do željenog broja klastera.

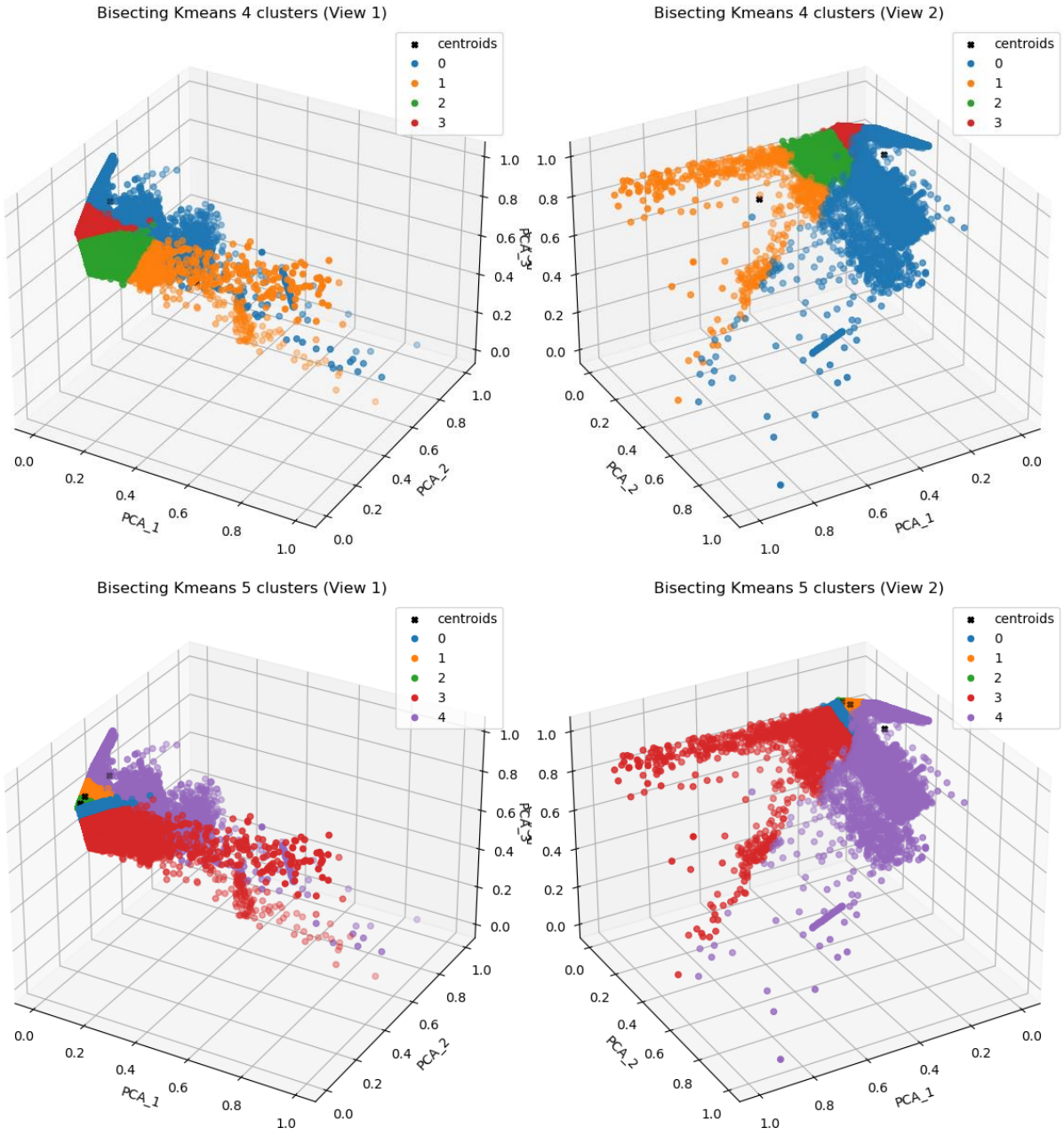
Na narednim graficima je prikazano isprobavanje različitih parametara k.


Bisecting Kmeans 2 clusters (View 1)
centroids
0
1
PCA_1
PCA_2
PCA_3


Bisecting Kmeans 2 clusters (View 2)
centroids
0
1
PCA_1
PCA_2
PCA_3


Bisecting Kmeans 3 clusters (View 1)
centroids
0
1
2
PCA_1
PCA_2
PCA_3


Bisecting Kmeans 3 clusters (View 2)
centroids
0
1
2
PCA_1
PCA_2
PCA_3



5.2 ALGORITAM SAKUPLJAJUĆEG HIJERARHIJSKOG KLASTEROVANJA

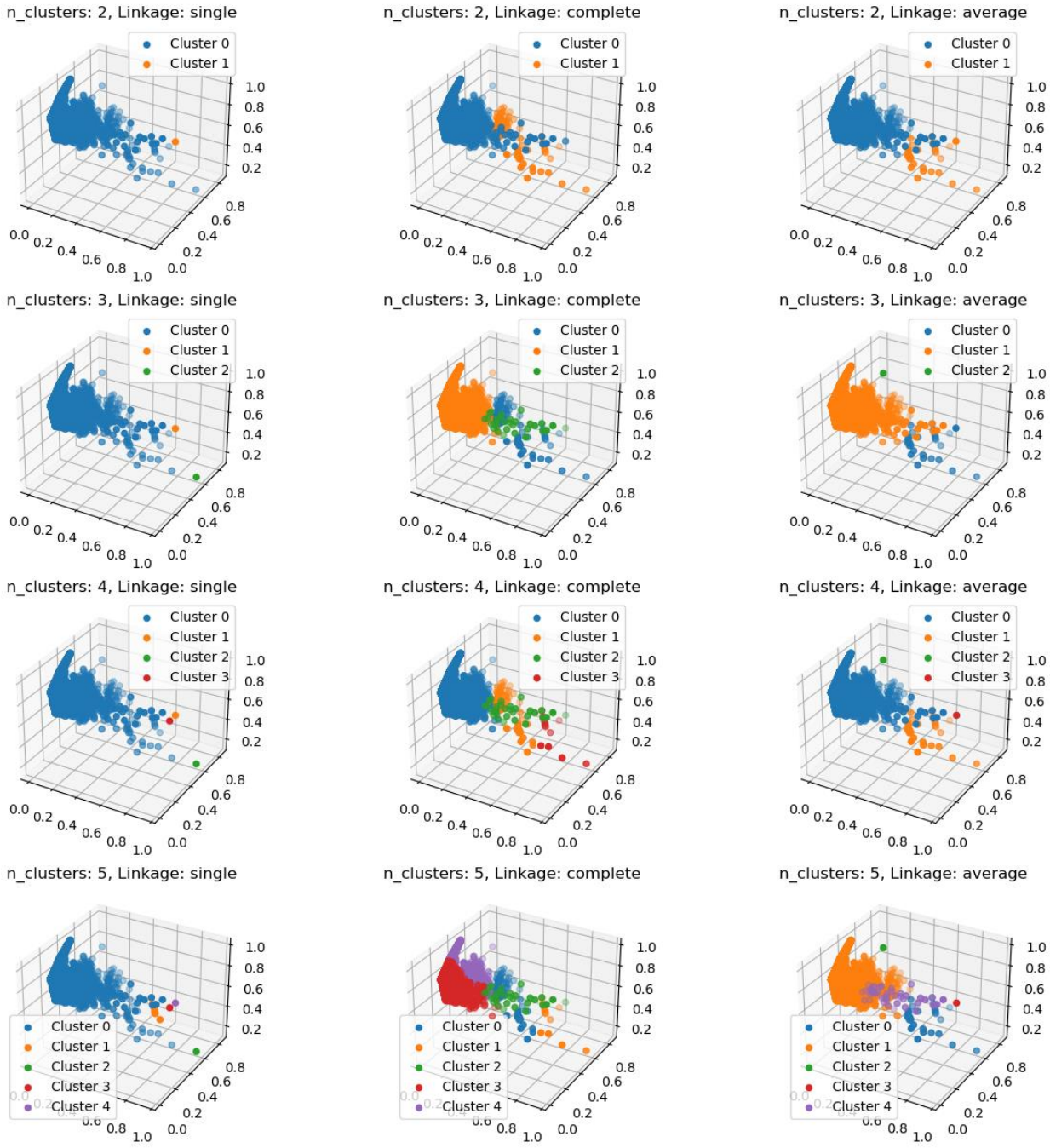
Agglomerative clustering je hijerarhijski algoritam klasterovanja koji počinje sa svakom instancom kao zasebnim klasterom, a zatim iterativno spaja najbliže klastera dok se ne formira jedan konačan klaster koji obuhvata sve podatke. Ovaj proces stvara drvo klastera, poznato kao dendrogram, koji pruža vizualnu reprezentaciju hijerarhijske strukture podataka.

Pored parametra k, testiraćemo 3 različite vrednosti parametra ***linkage***:

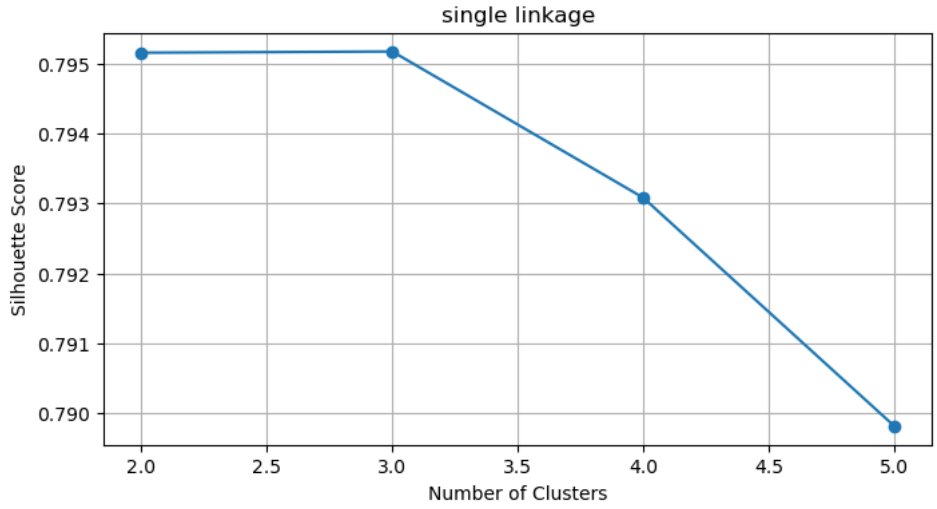
1. **single (minimalna udaljenost):** ova metoda koristi minimalnu udaljenost između svih tačaka u dva klastera kao meru udaljenosti između tih klastera. To znači da se uzima najmanja udaljenost između bilo koje dve tačke u različitim klasterima.

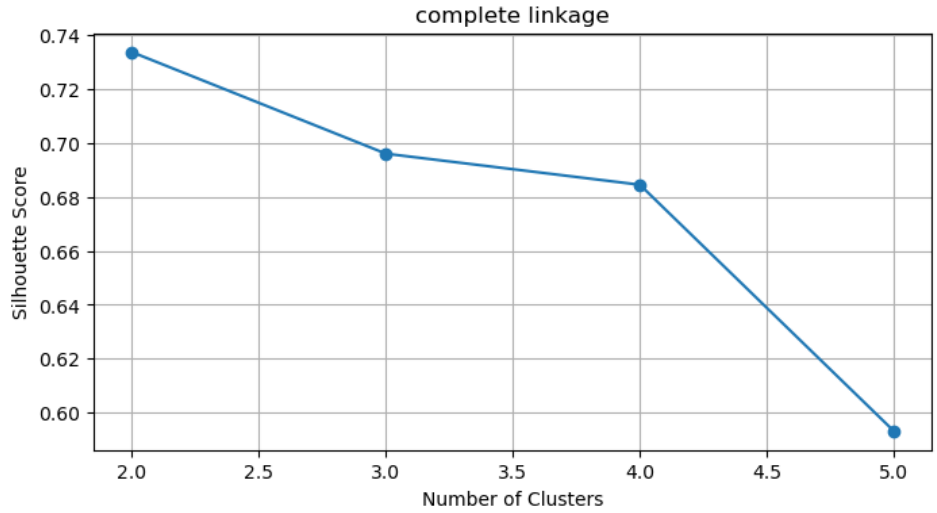
2. **complete (maksimalna udaljenost):** koristi maksimalnu udaljenost između svih tačaka u dva klastera kao meru udaljenosti između tih klastera. To znači da se uzima najveća udaljenost između bilo koje dve tačke u različitim klasterima.

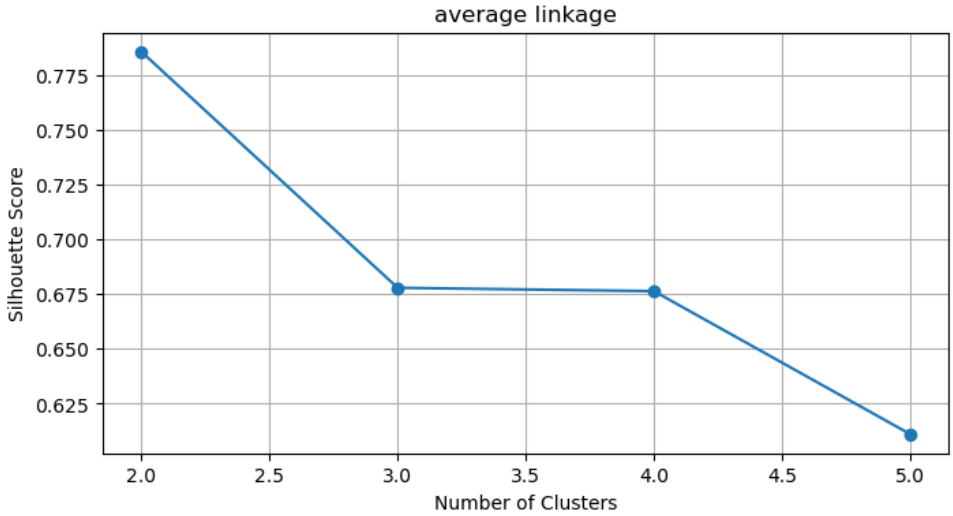
3. **average (prosečna udaljenost):** ova metoda koristi prosečnu udaljenost između svih tačaka u dva klastera kao meru udaljenosti između tih klastera. To znači da se uzima srednja vrednost udaljenosti između svih parova tačaka u različitim klasterima.



Na narednim graficima možemo videti vrednosti ***silhouette score*** u zavisnosti od k i *linkage* parametra







Na Figure 34 možemo videti tri dendrograma za tri različita atributa (PCA_1, PCA_2, PCA_3).

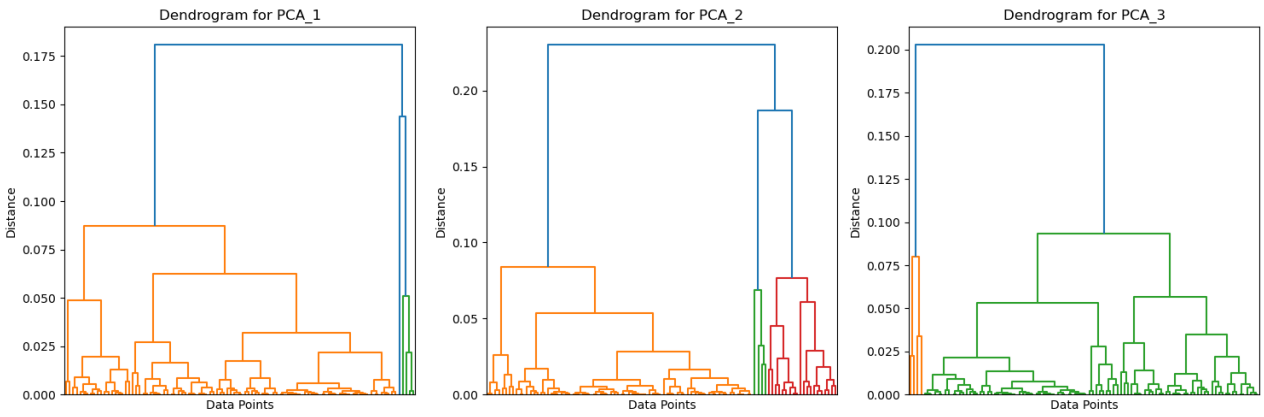


Figure 34

6 PRAVILA PRIDRUŽIVANJA

Korišćen je **Apriori algoritam** nad skupom podataka sa kategorizovanim *Source Port* i *Destination Port* atributima. Za sve atribute postavljena je uloga *both*, tako da se svaki atribut može biti i uzrok i posledica pravila.

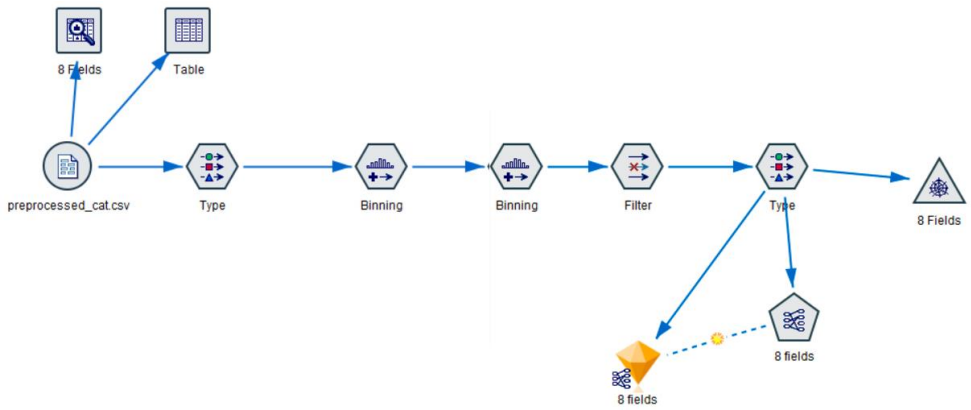


Figure 35: SPSS dijagram

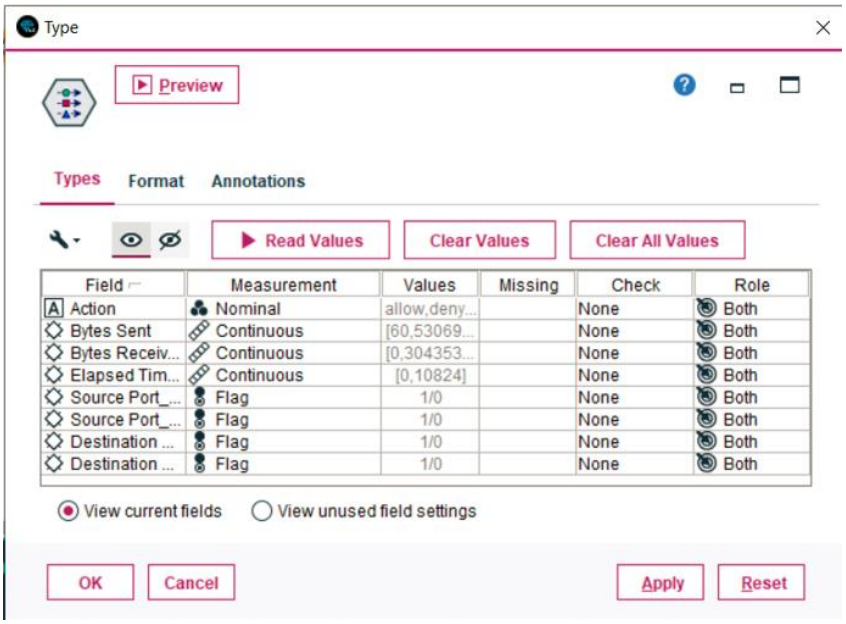


Figure 36

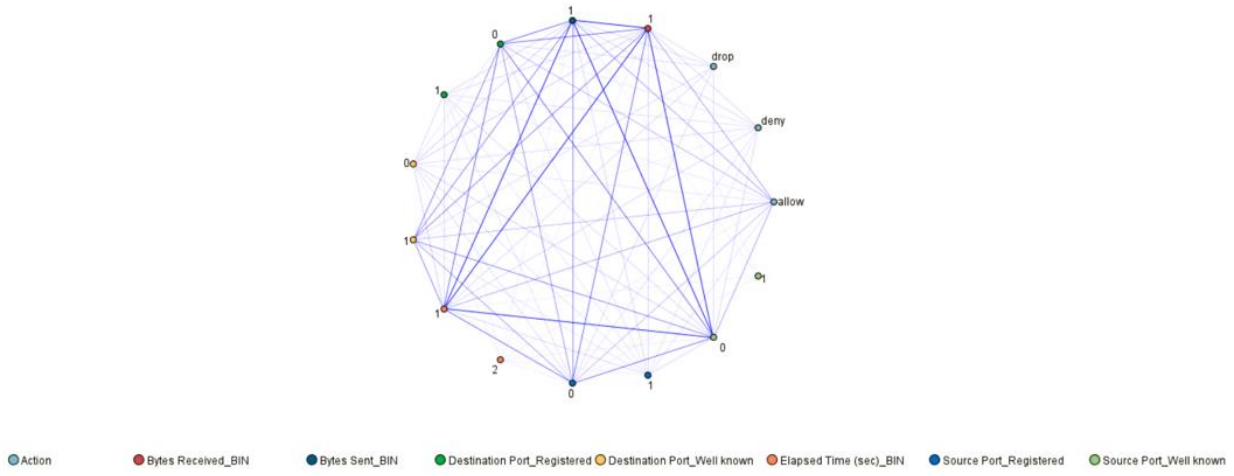
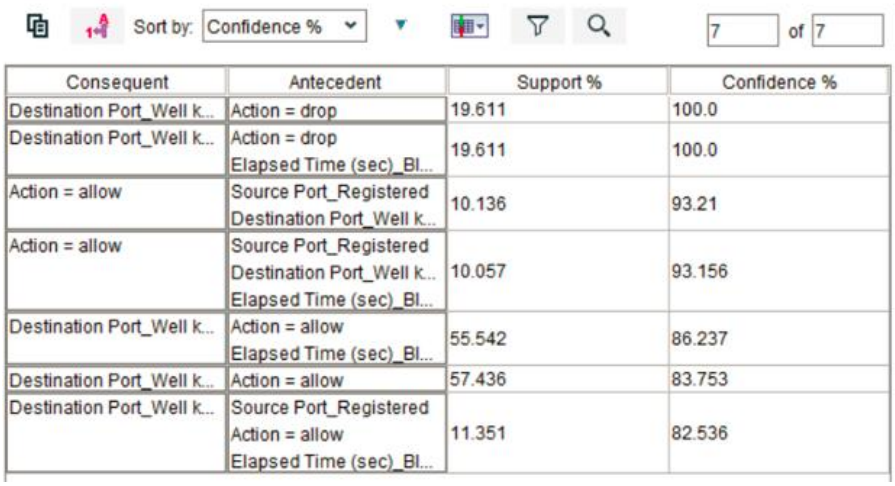


Figure 37

Na Figure 38 možemo videti tabelu podrške i pouzdanosti. *Support* se odnosi na učestalost pojavljivanja kombinacije atributa u skupu podataka, dok *confidence* predstavlja verovatnoću da će se ciljna klasa pojaviti kada su prisutni određeni atributi.



Consequent	Antecedent	Support %	Confidence %
Destination Port_Well k...	Action = drop	19.611	100.0
Destination Port_Well k...	Action = drop Elapsed Time (sec)_Bl...	19.611	100.0
Action = allow	Source Port_Registered Destination Port_Well k...	10.136	93.21
Action = allow	Source Port_Registered Destination Port_Well k... Elapsed Time (sec)_Bl...	10.057	93.156
Destination Port_Well k...	Action = allow Elapsed Time (sec)_Bl...	55.542	86.237
Destination Port_Well k...	Action = allow	57.436	83.753
Destination Port_Well k...	Source Port_Registered Action = allow Elapsed Time (sec)_Bl...	11.351	82.536

Figure 38