Internet Firewall Data

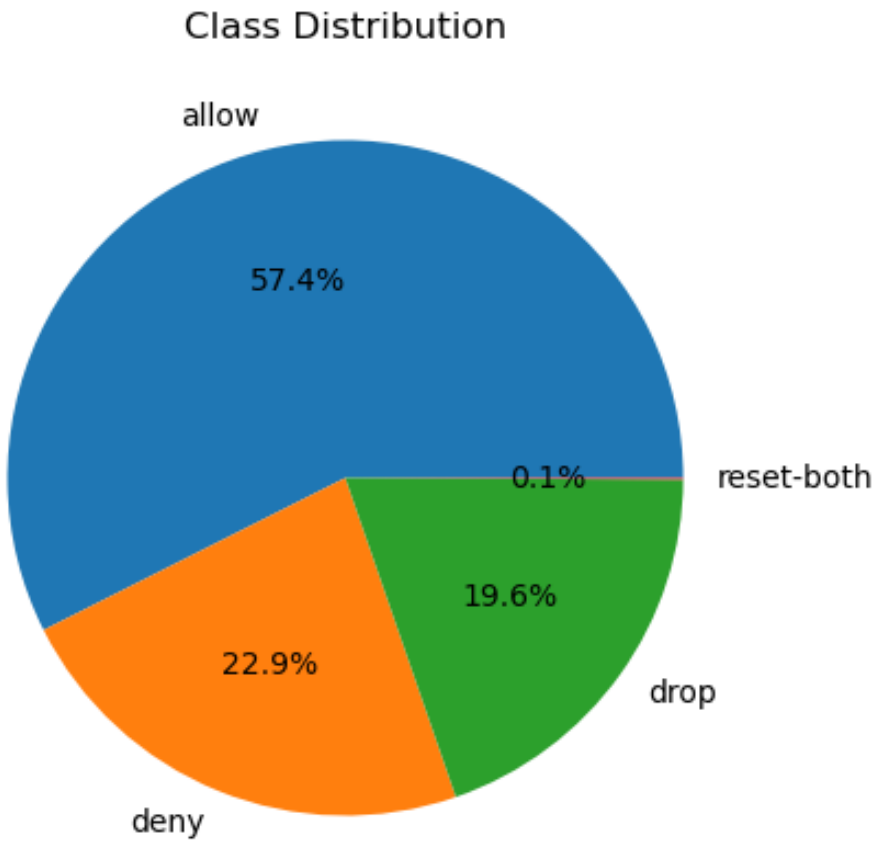
Veljko Prodan

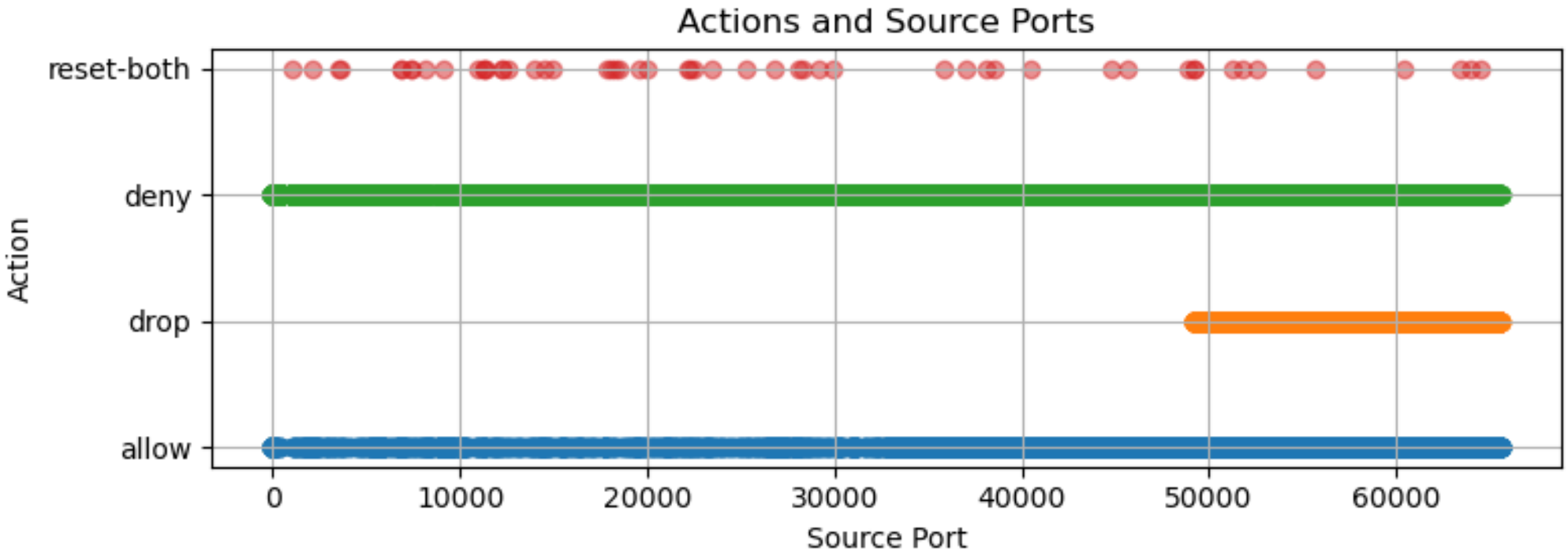
163/2019

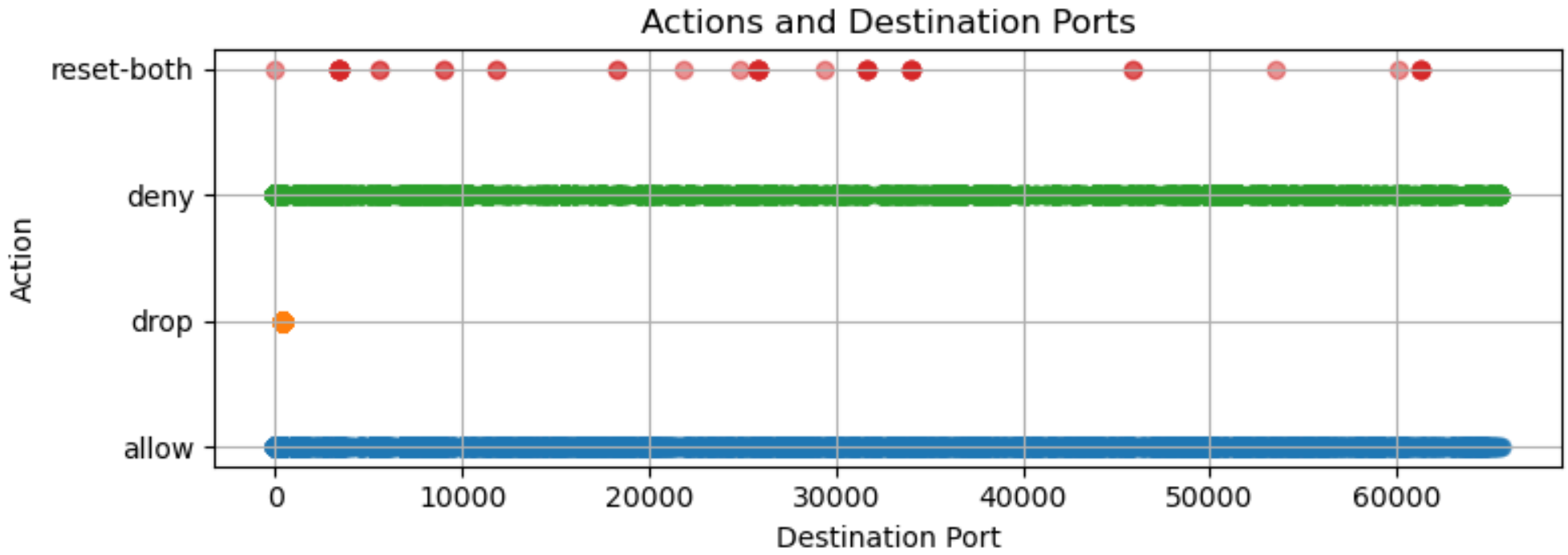
Analiza skupa podataka

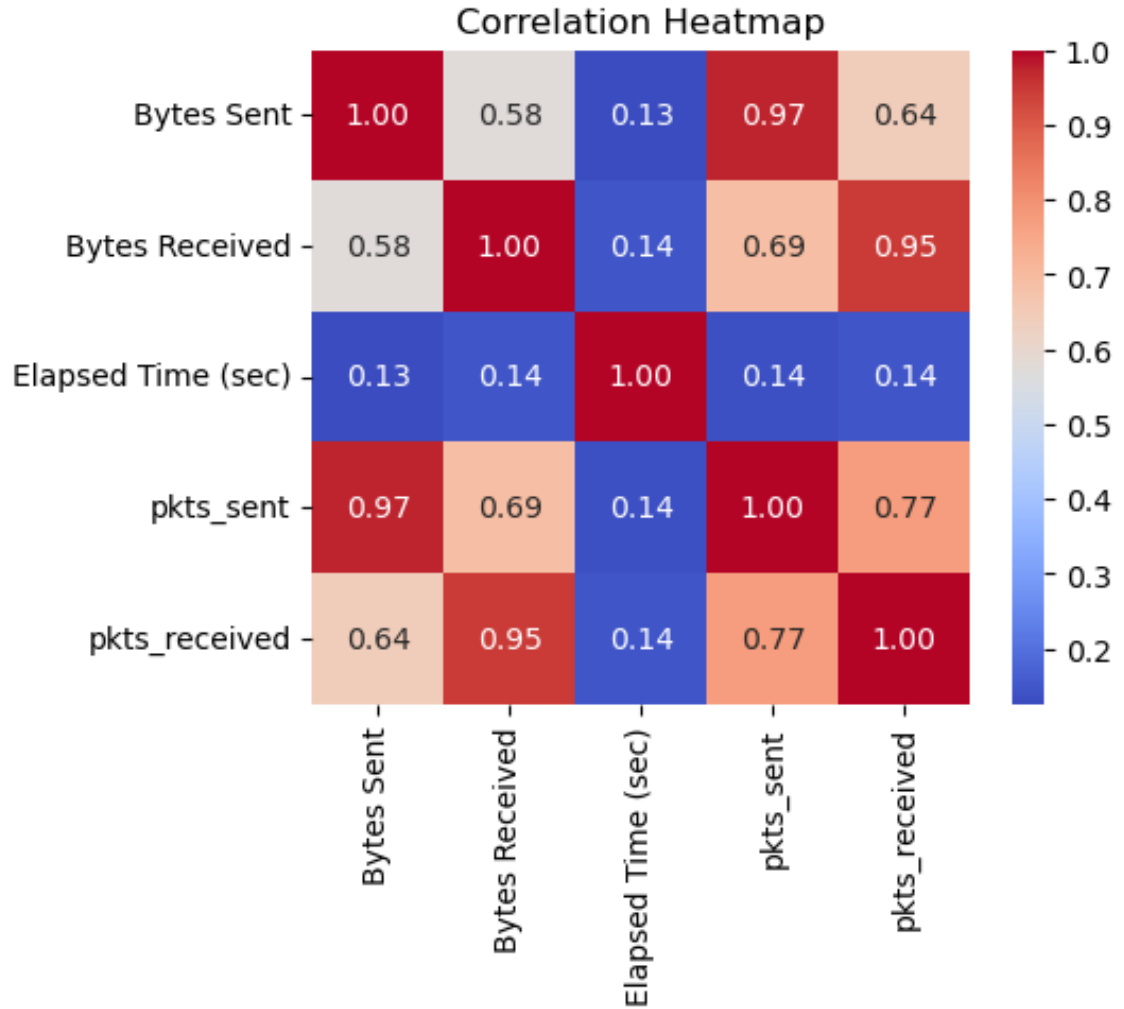
- 65532 instance
- 12 atributa

```
Source Port           65532 non-null  int64
Destination Port      65532 non-null  int64
NAT Source Port       65532 non-null  int64
NAT Destination Port  65532 non-null  int64
Action                65532 non-null  object
Bytes                 65532 non-null  int64
Bytes Sent            65532 non-null  int64
Bytes Received        65532 non-null  int64
Packets               65532 non-null  int64
Elapsed Time (sec)    65532 non-null  int64
pkts_sent             65532 non-null  int64
pkts_received         65532 non-null  int64
```




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

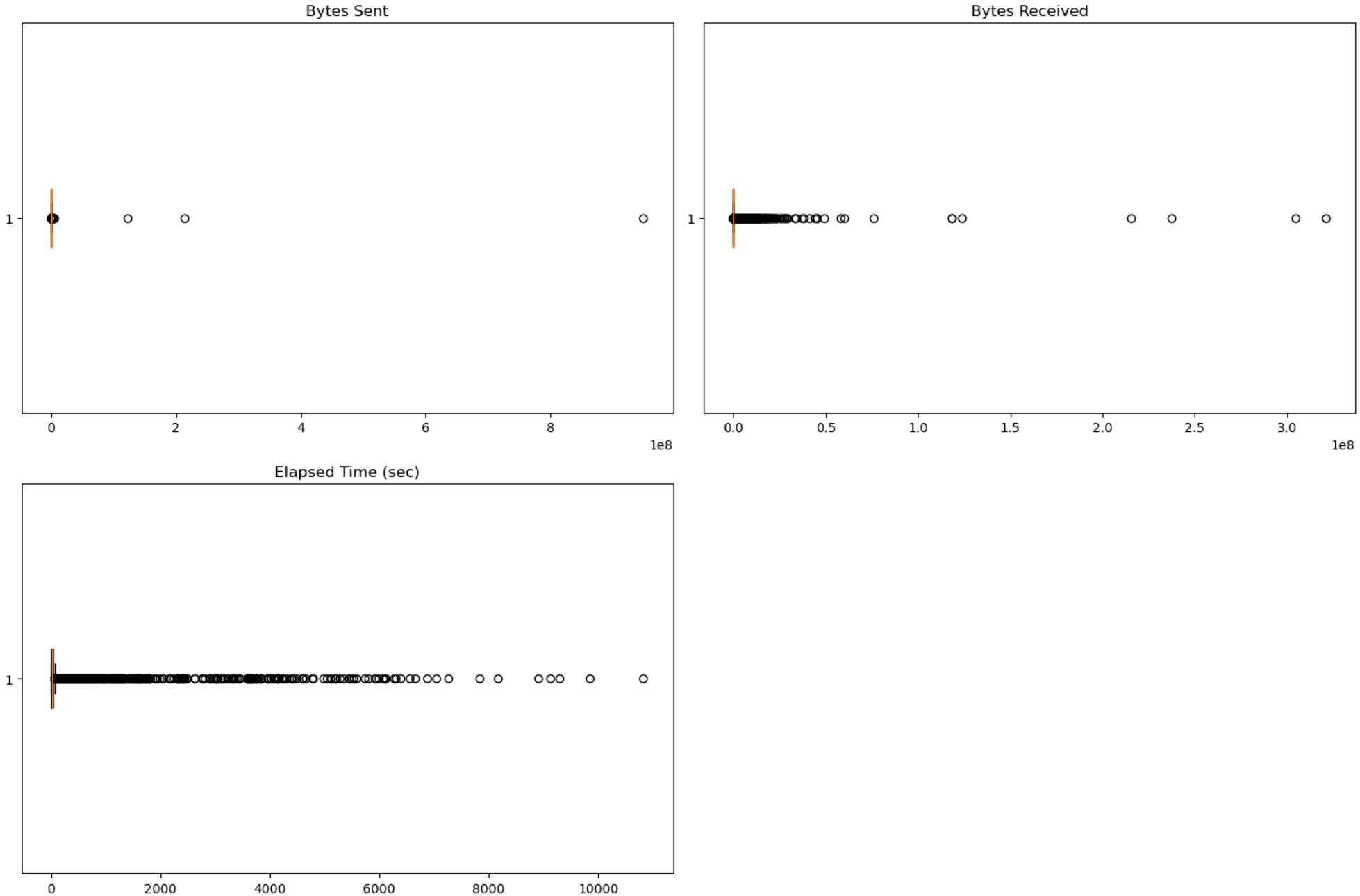

Actions and Destination Ports
reset-both
deny
drop
allow
Action
0
10000
20000
30000
40000
50000
60000
Destination Port



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
```

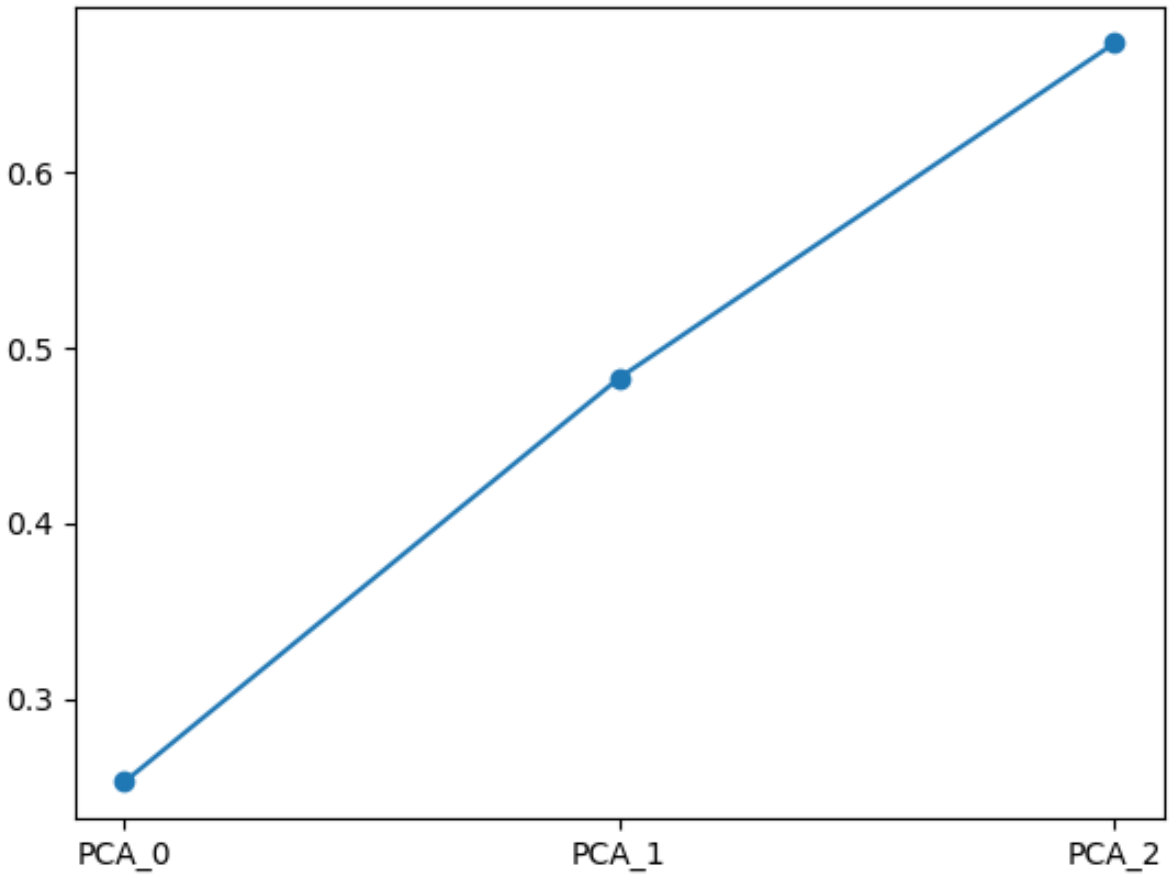

Bytes Sent
1
0
2
4
6
8
1e8
Bytes Received
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e8
Elapsed Time (sec)
1
0
2000
4000
6000
8000
10000

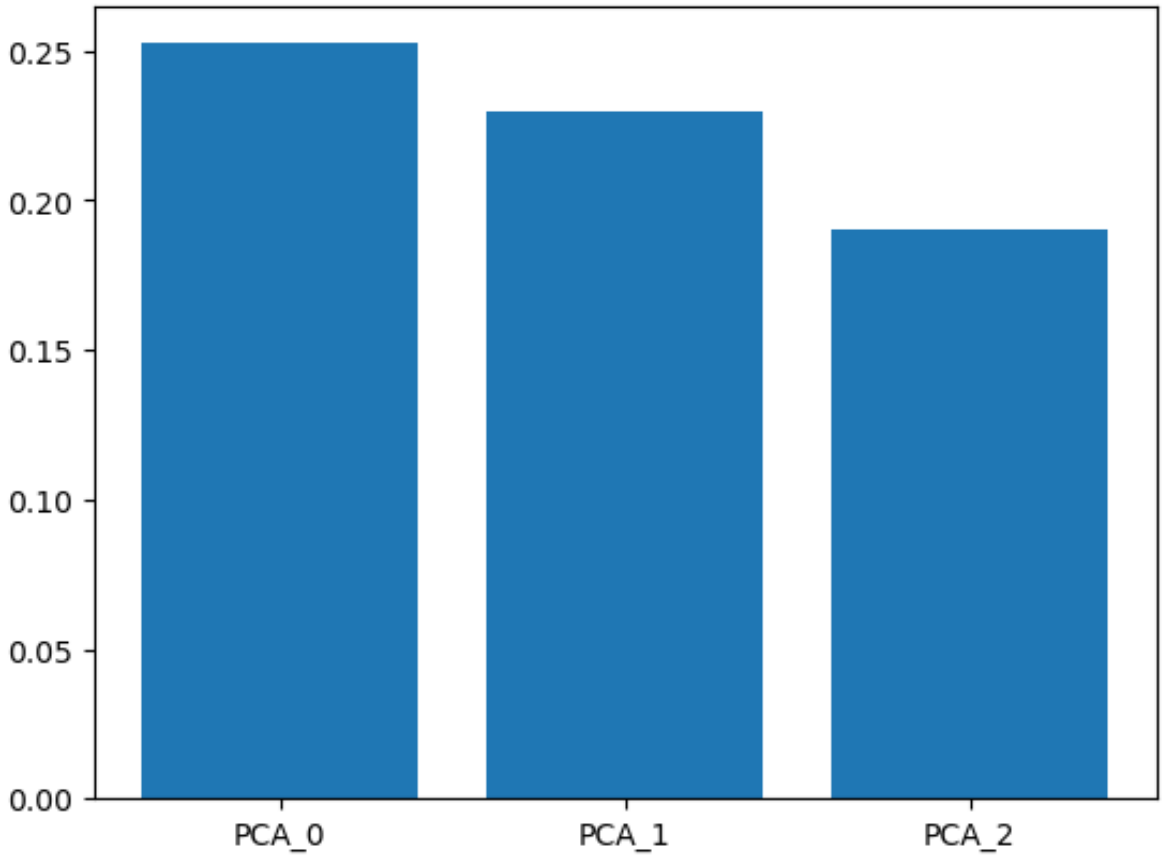
```
def classify_port(port):
    if port >= 0 and port <= 1023:
        return 'Well known'
    elif port >= 1024 and port <= 49151:
        return 'Registered'
    elif port >= 49152 and port <= 65535:
        return 'Private'
    else:
        return 'Unknown'
```

Out[24]:

	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

PCA


0.6
0.5
0.4
0.3
PCA_0
PCA_1
PCA_2


0.25
0.20
0.15
0.10
0.05
0.00
PCA_0
PCA_1
PCA_2

Stabla odlučivanja

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \qquad \text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

- Pre nameštanja hiperparametara

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

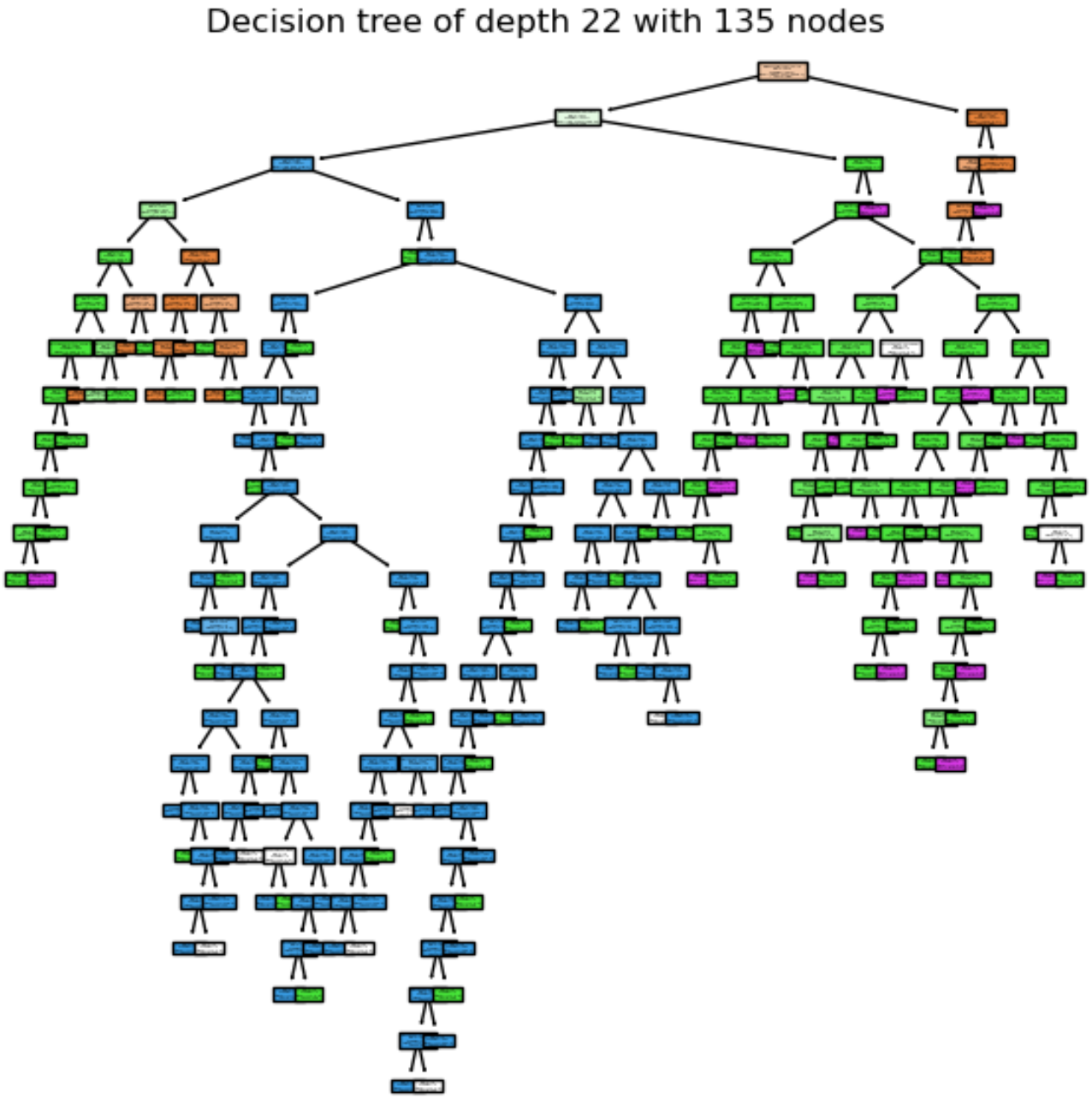
	allow	drop	deny	reset-both
allow	28220	7	0	0
drop	0	11240	0	0
deny	0	6	9632	0
reset-both	0	0	0	41

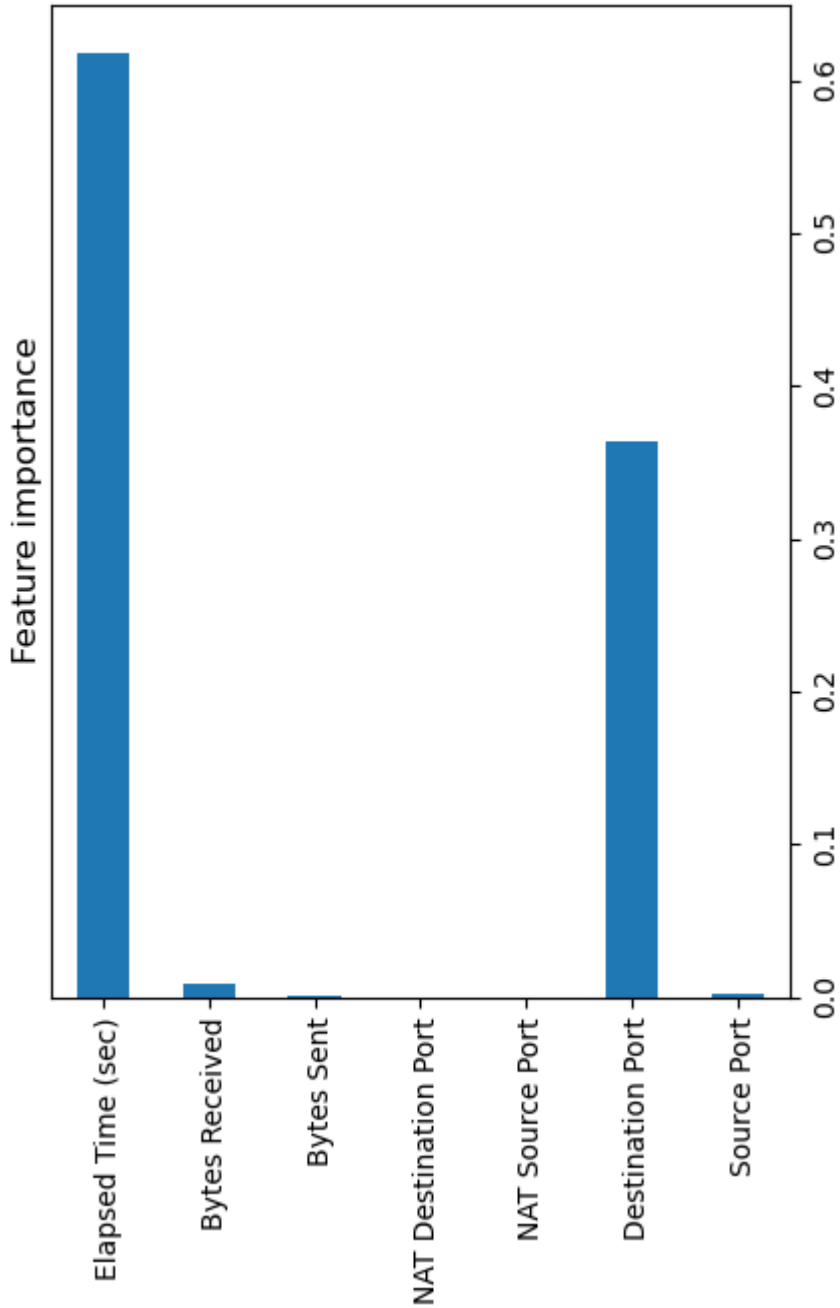
Classification report for model DecisionTreeClassifier on test data

	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


Decision tree of depth 22 with 135 nodes


Feature importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Elapsed Time (sec)
Bytes Received
Bytes Sent
NAT Destination Port
NAT Source Port
Destination Port
Source Port

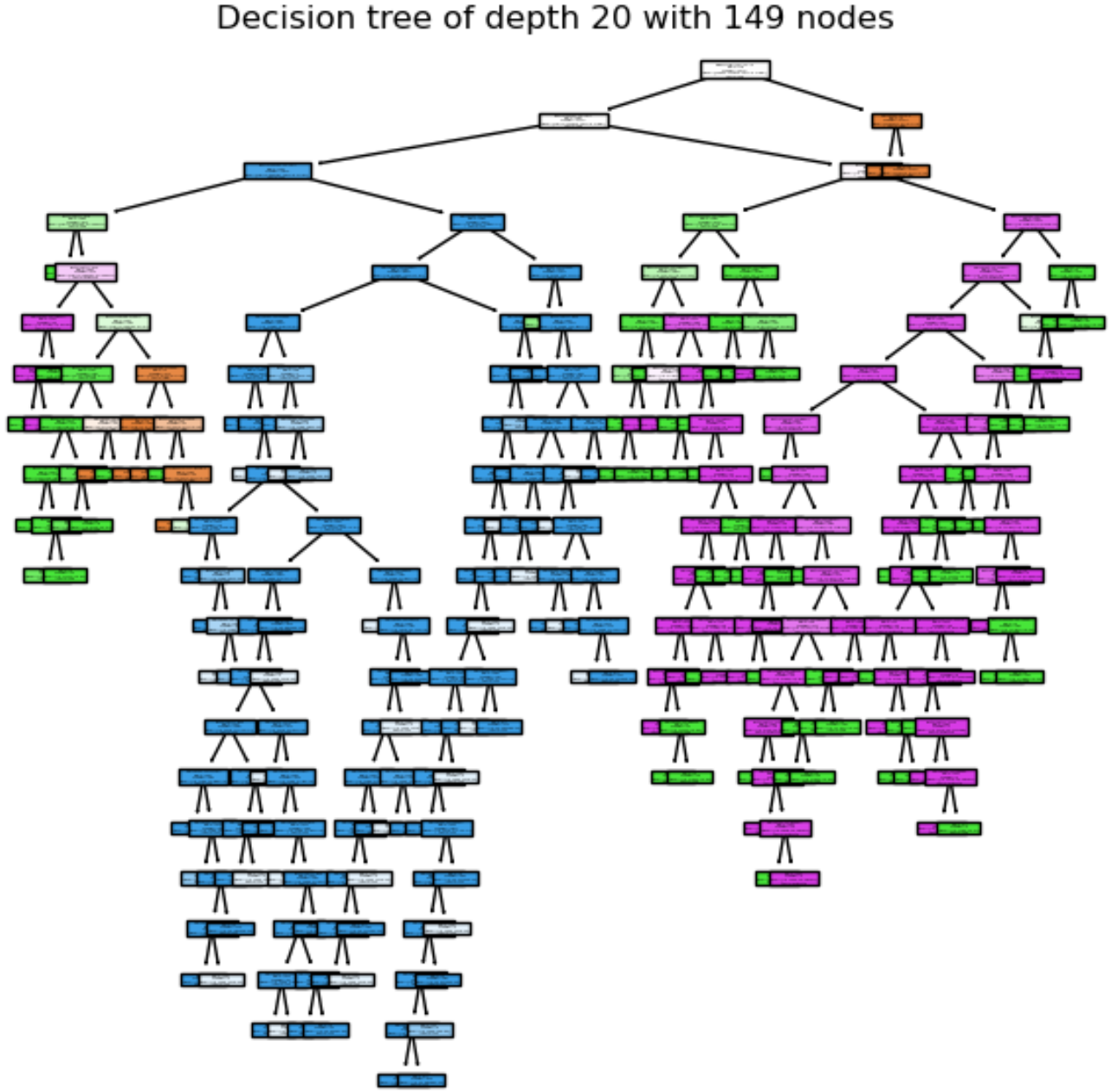
```
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [12,15,18,20,None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'splitter': ['best', 'random']
}
```

Parametri za GridSearch

```
print(estimator.best_params_, '\n')
print(estimator.best_score_)
```

```
{'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'splitter': 'best'}

0.9978024661213527
```


Decision tree of depth 20 with 149 nodes

Parametri, metrike i matrice konfuzije 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
-----------------------------------------
```

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

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

Random forest

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

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

K-najbližih suseda bez nameštanja hiperparametara

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

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

KNN + GridSearch

In [14]:
```
print(estimator.best_params_, '\n')
print(estimator.best_score_)
```

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

0.9956252797786188
```

```
params_grid = {'n_neighbors': range(5, 50, 5),
               'weights': ['uniform', 'distance'],
               'p': [1, 2],
               'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute']
              }
```

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

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

Bagging KNN

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

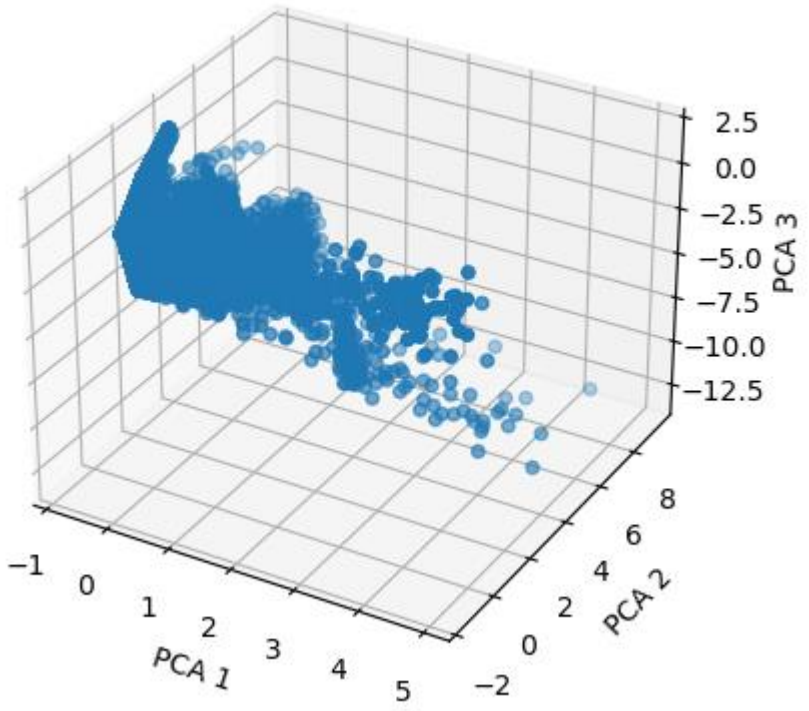
Classification report for model BaggingClassifier on test data

	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

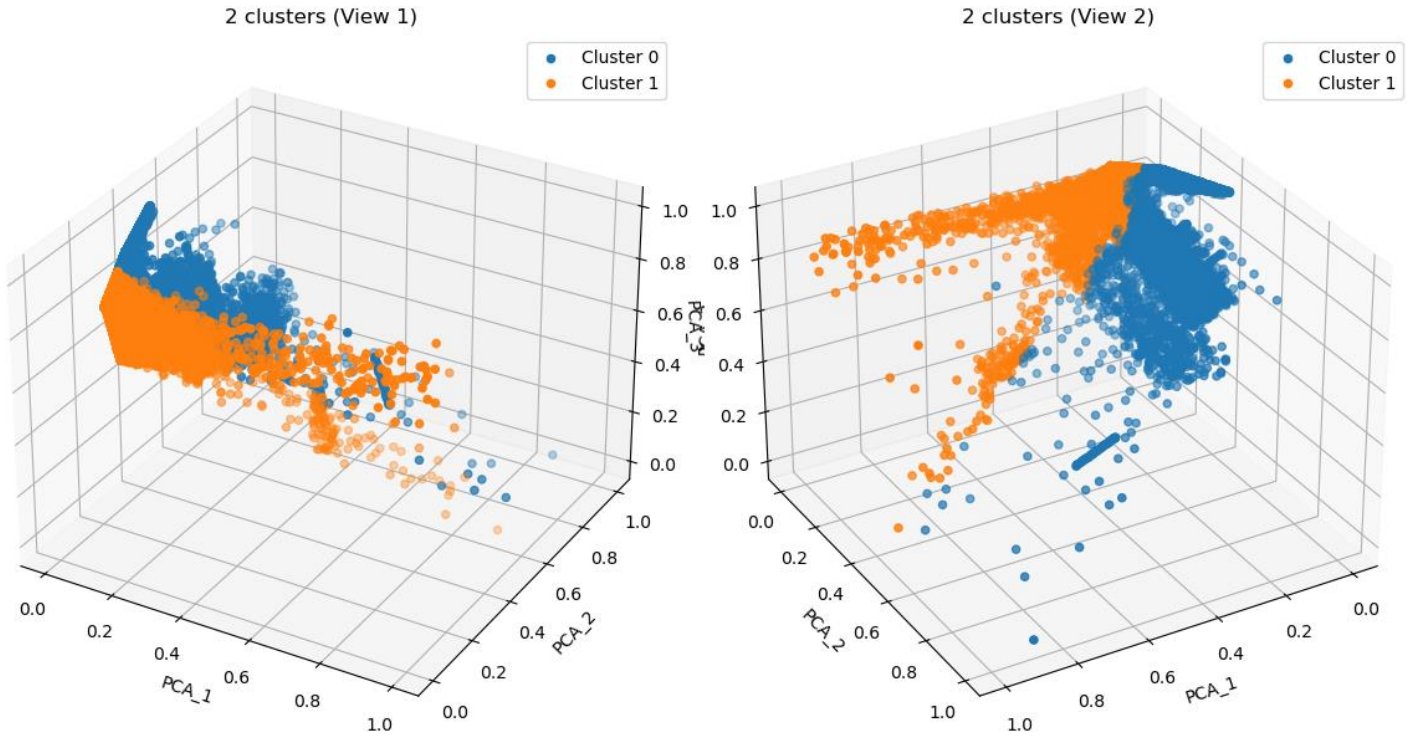
Confusion matrix for model BaggingClassifier on test data

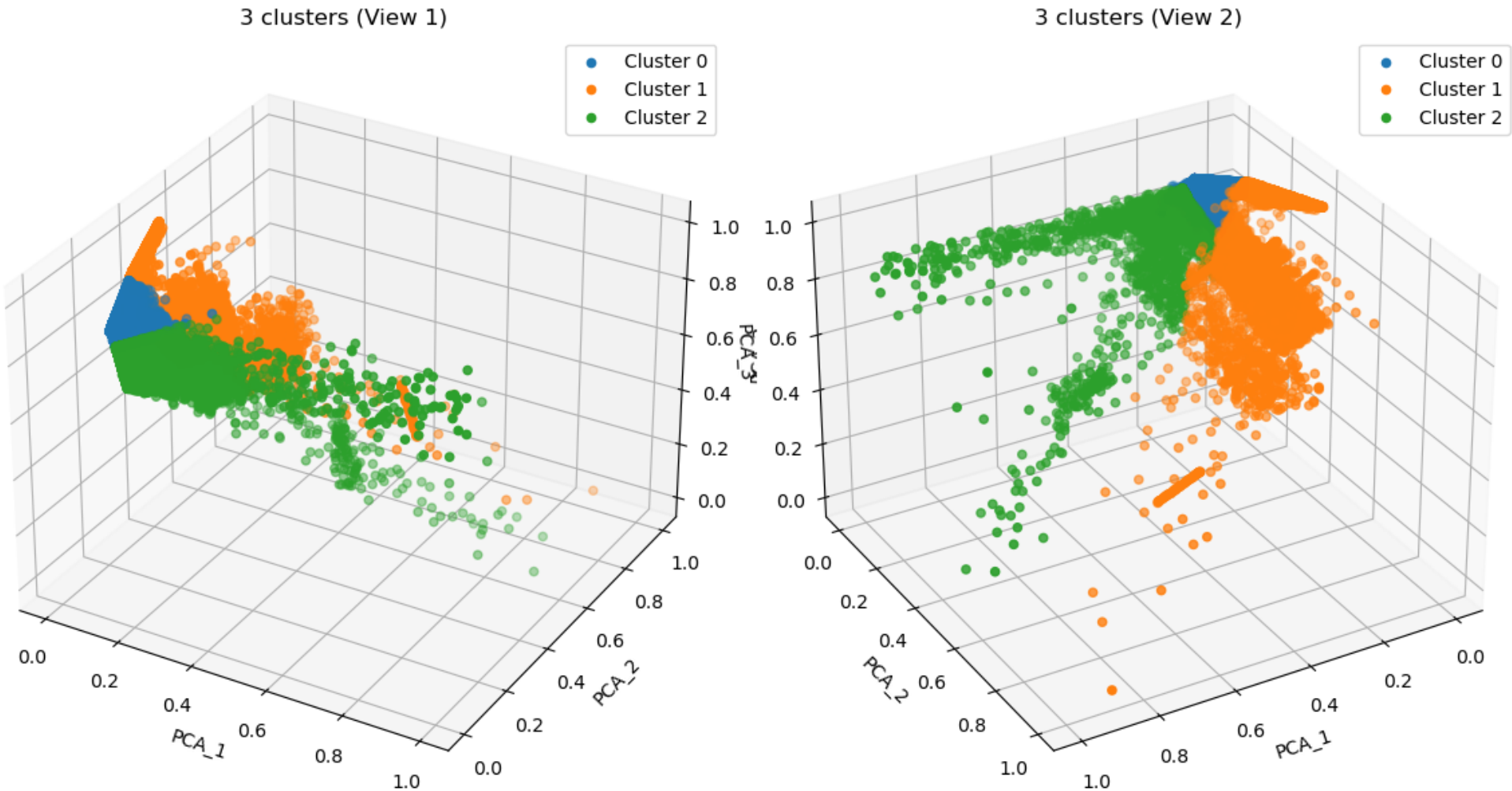
	allow	drop	deny	reset-both
allow	9378	31	1	0
drop	12	3725	10	0
deny	0	0	3213	0
reset-both	3	10	0	0

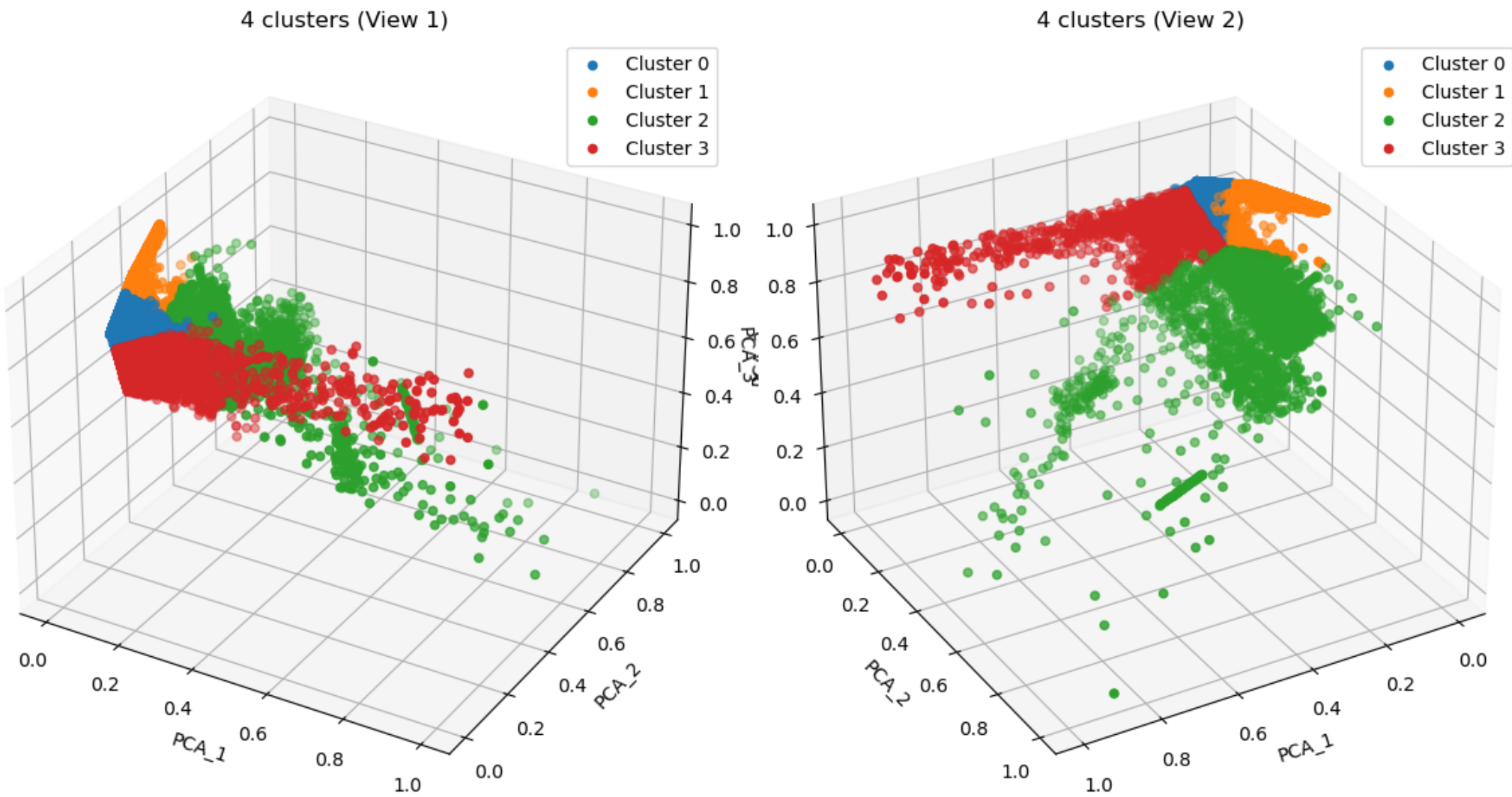
K-means

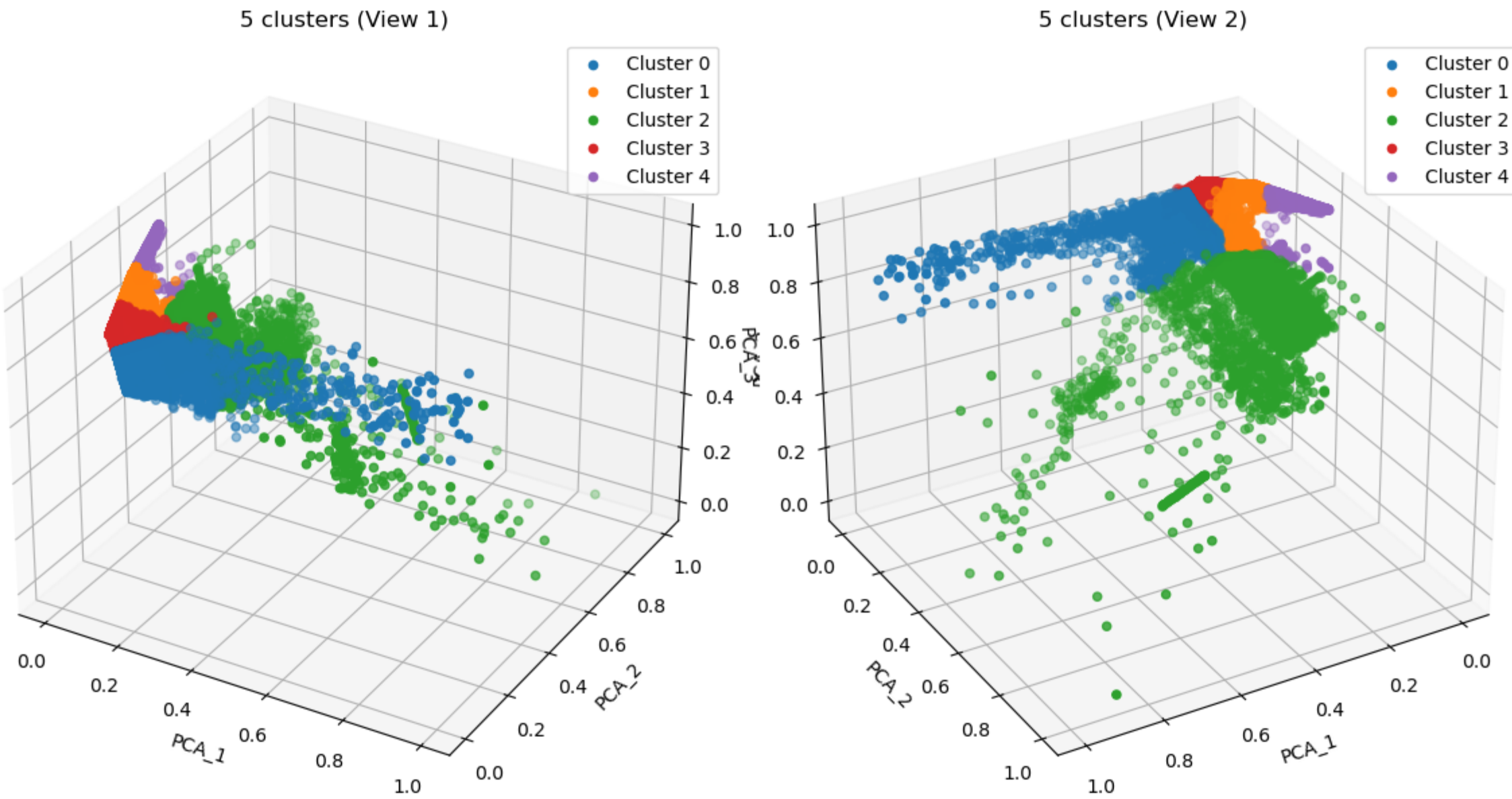


Pre klasterovanja




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

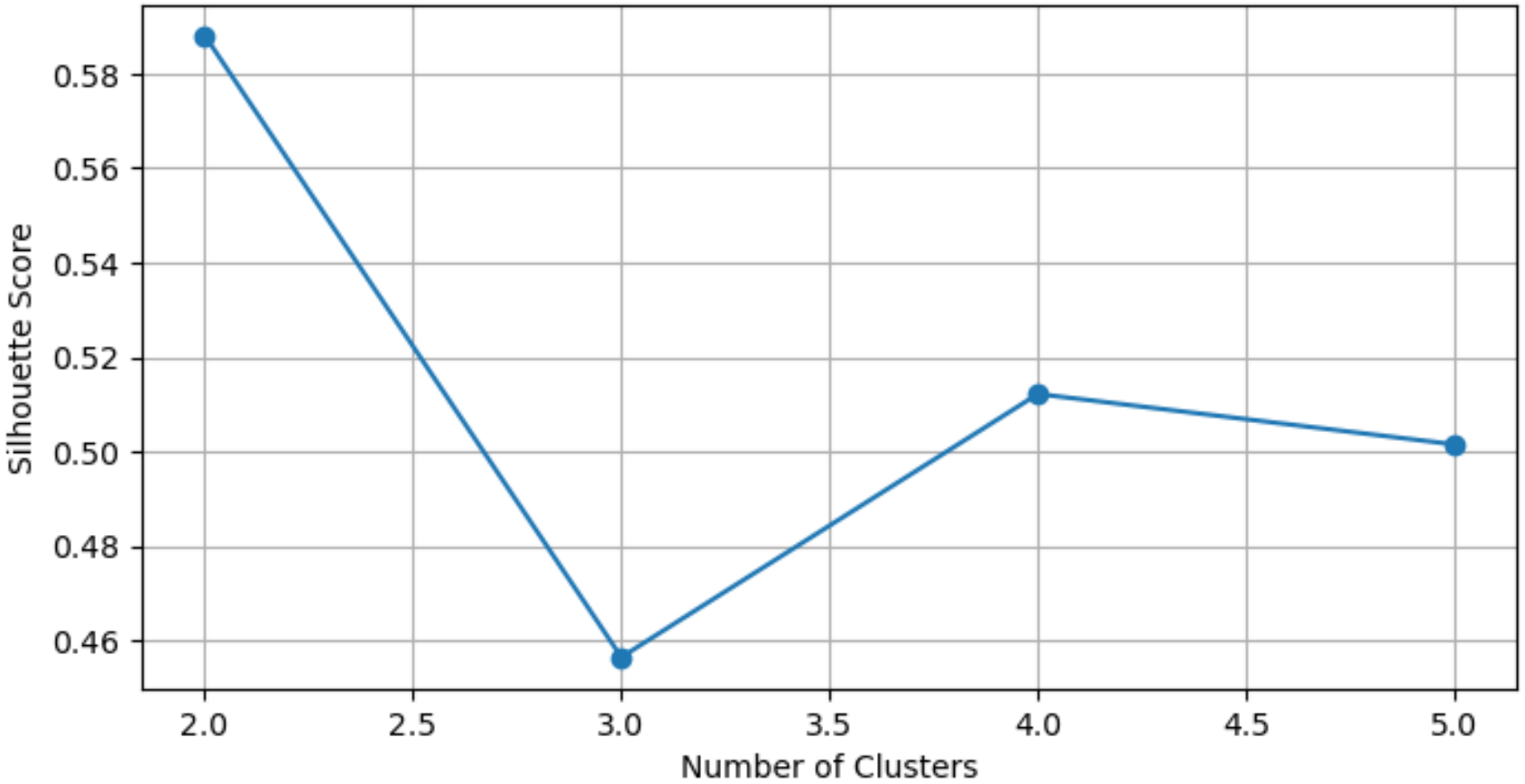

4 clusters (View 1)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
PCA_1
PCA_2
PCA_3
4 clusters (View 2)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
PCA_1
PCA_2


5 clusters (View 1)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
PCA_1
PCA_2
PCA_3
0.0
0.2
0.4
0.6
0.8
1.0
5 clusters (View 2)
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
PCA_1
PCA_2
0.0
0.2
0.4
0.6
0.8
1.0

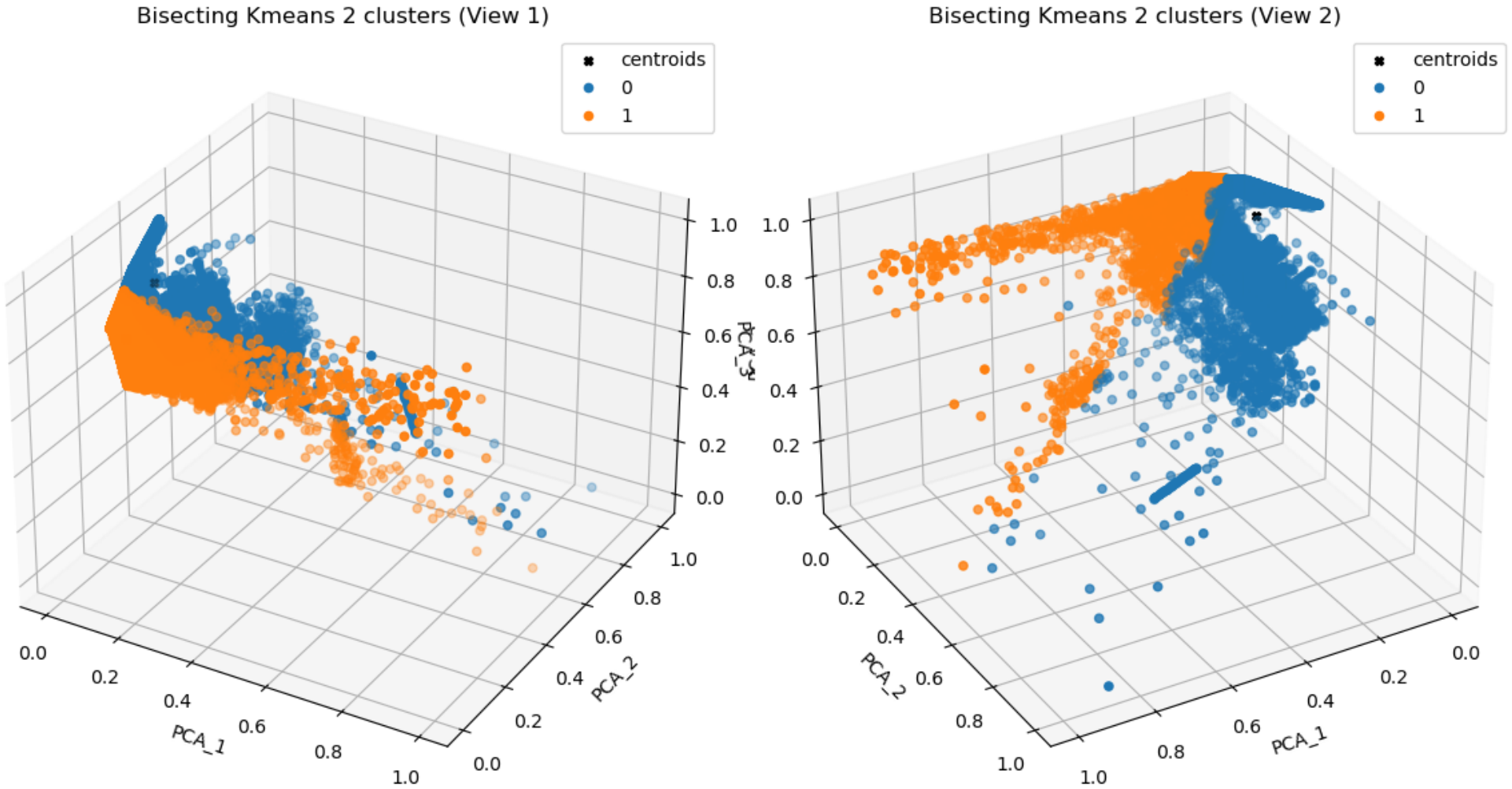
Silhouette score

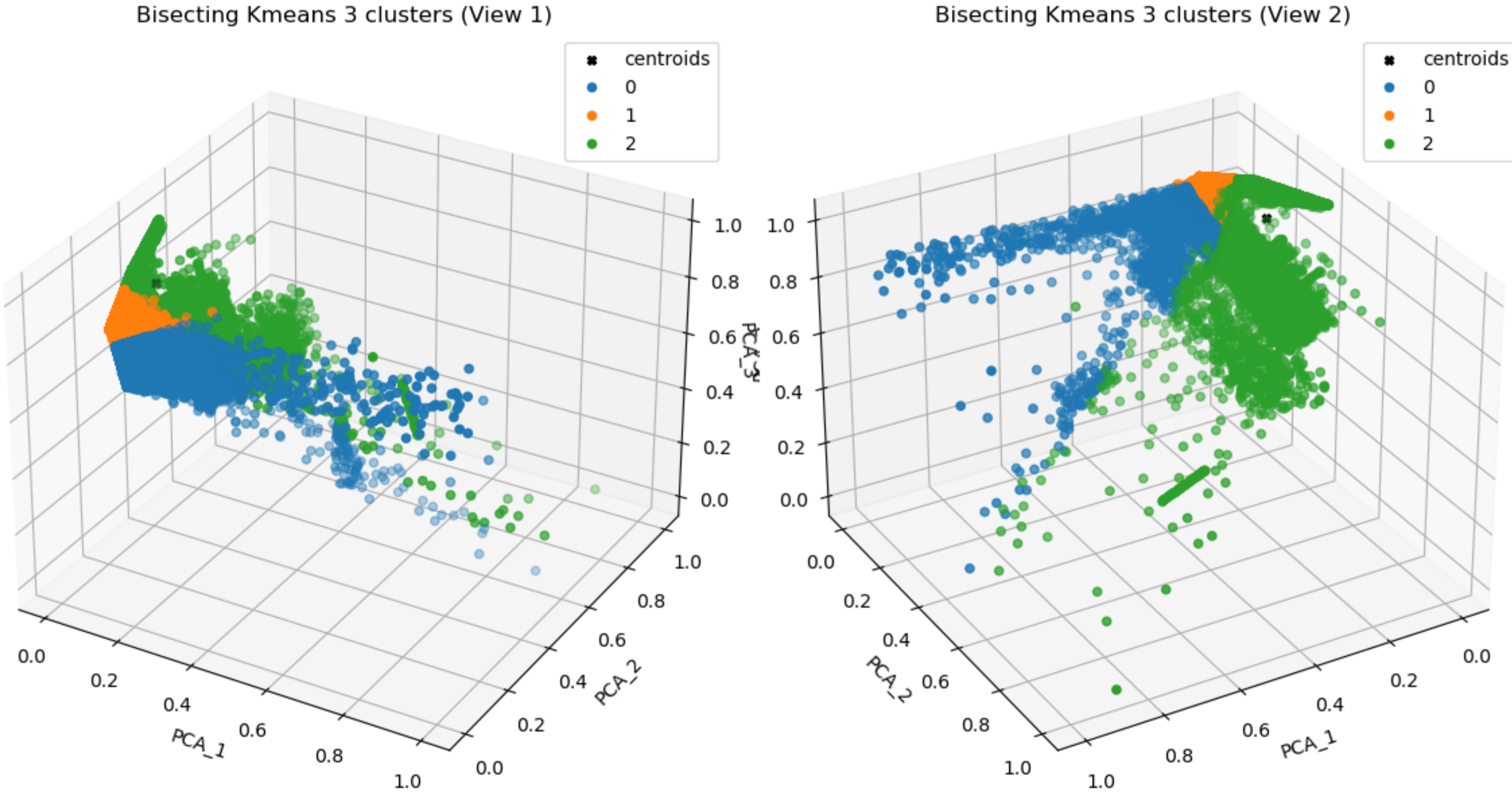
$$s = \frac{b - a}{\max(a, b)}$$

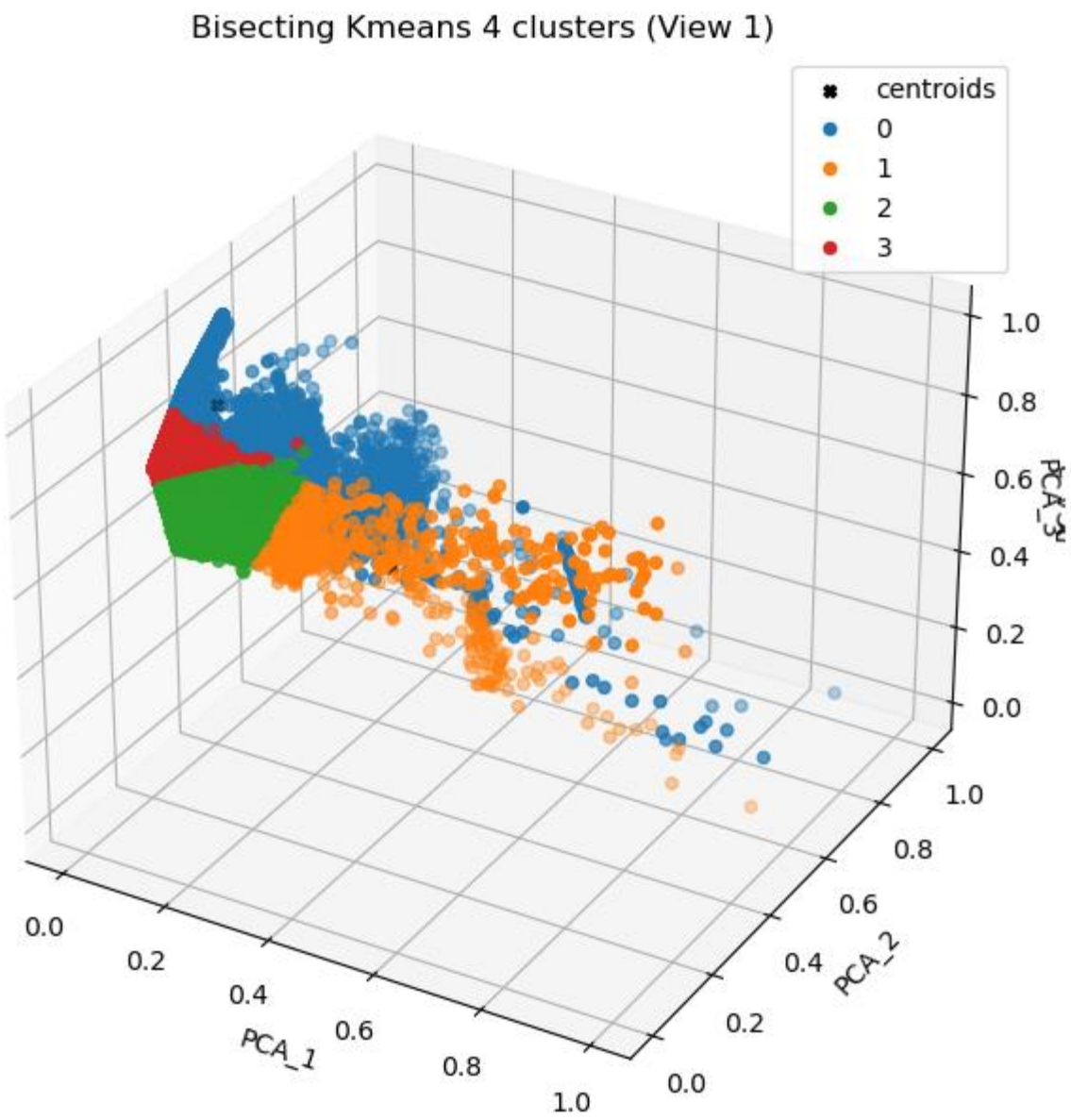
- a - prosečno rastojanje između instance i ostalih instanci u istom klasteru
- b - prosečno rastojanje između instance i svih instanci iz najbližeg susednog klastera

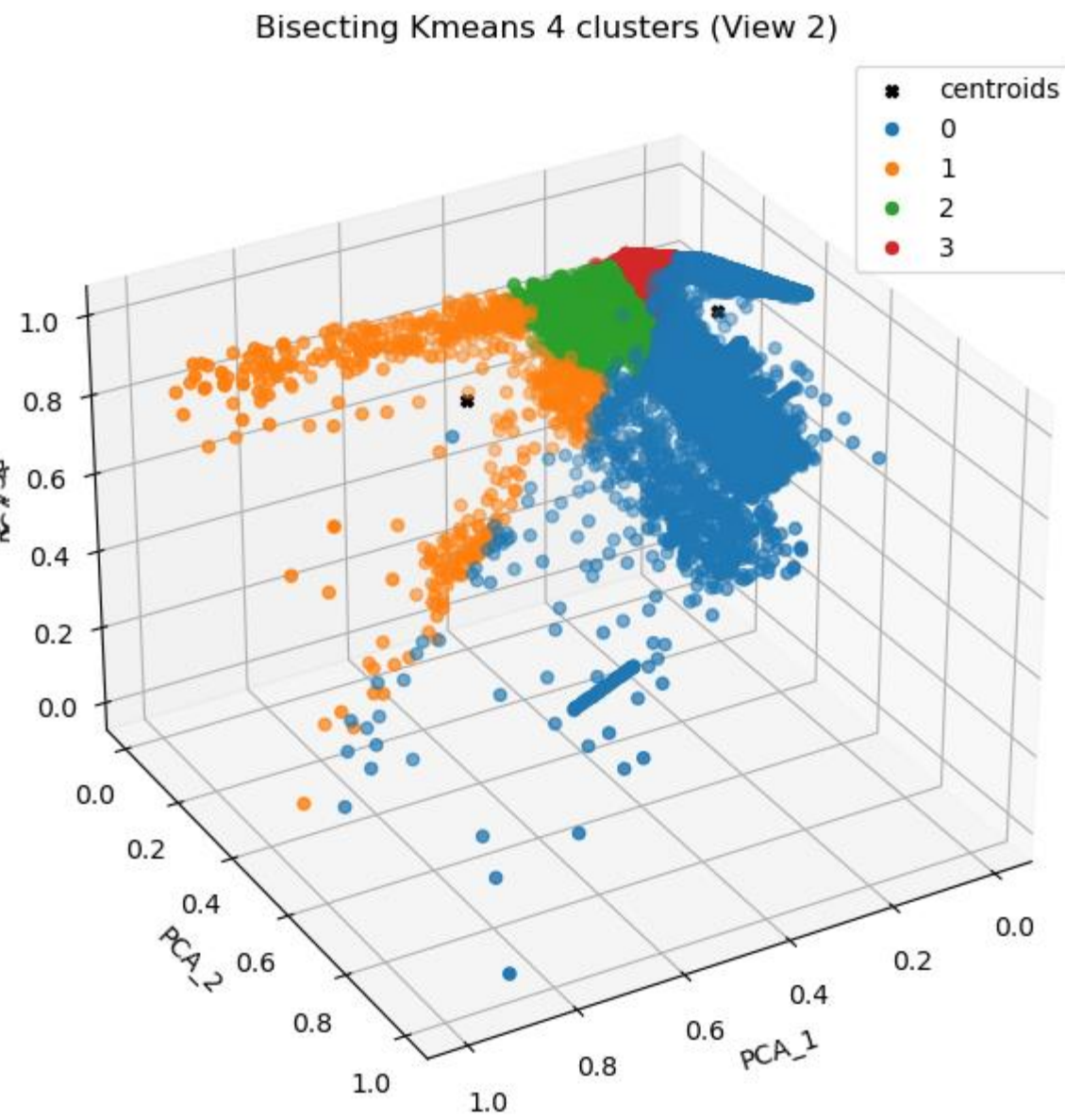


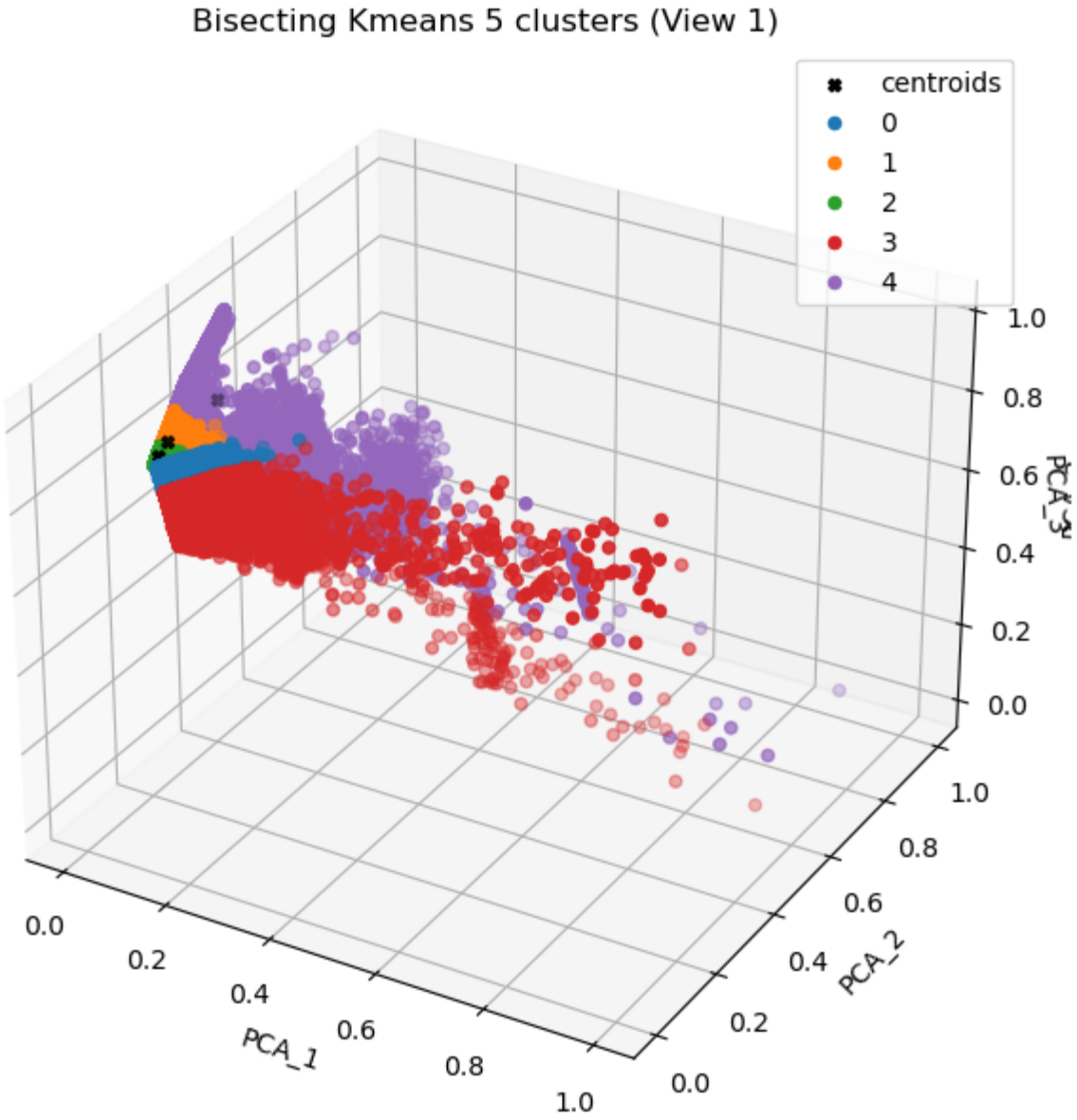
Bisecting K-means

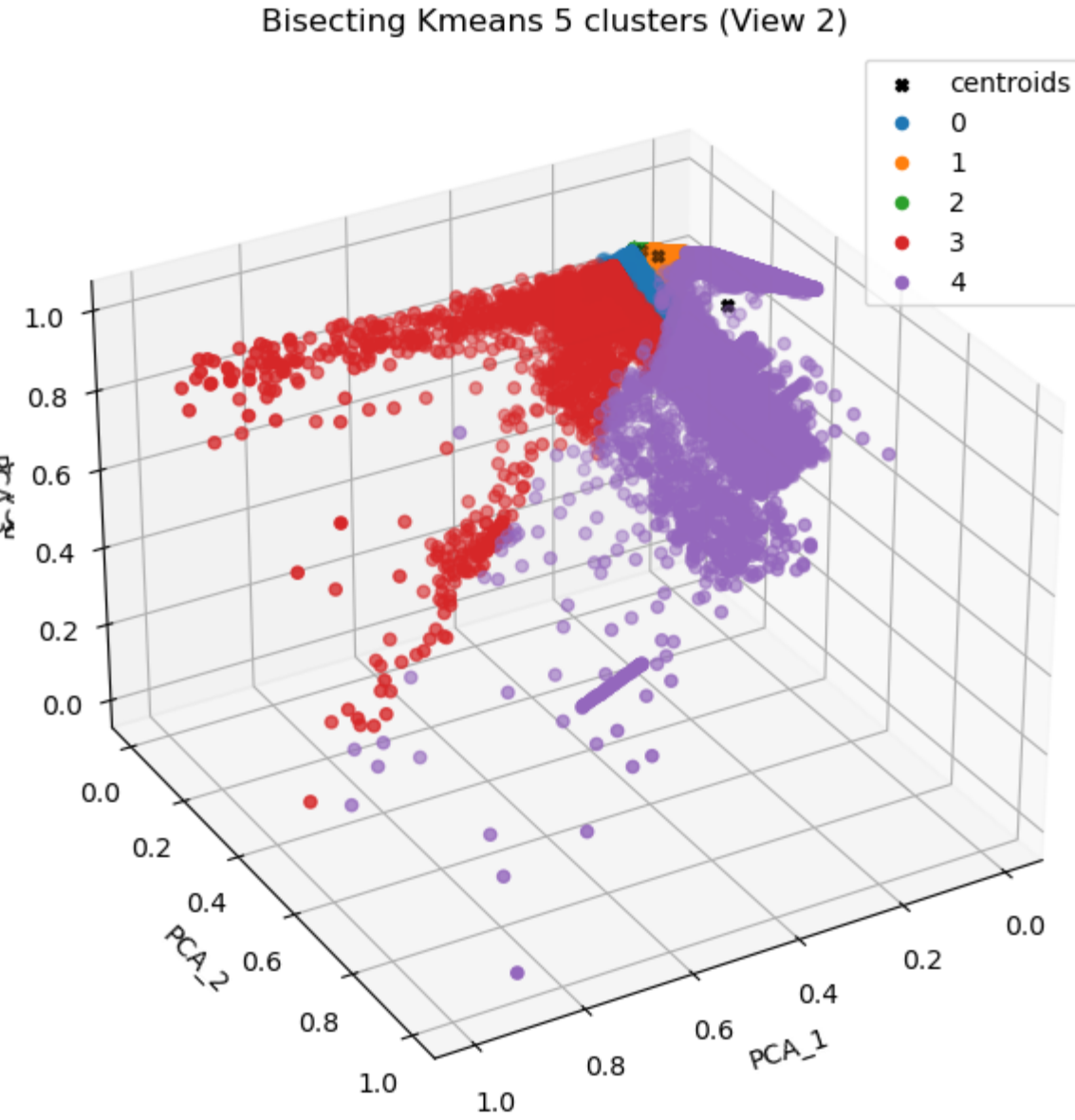



Bisecting Kmeans 3 clusters (View 1)
centroids
0
1
2
PCA_1
PCA_2
PCA_3
Bisecting Kmeans 3 clusters (View 2)
centroids
0
1
2
PCA_1
PCA_2


Bisecting Kmeans 4 clusters (View 1)
centroids
0
1
2
3
PCA_1
PCA_2
PCA_3
0.0
0.2
0.4
0.6
0.8
1.0

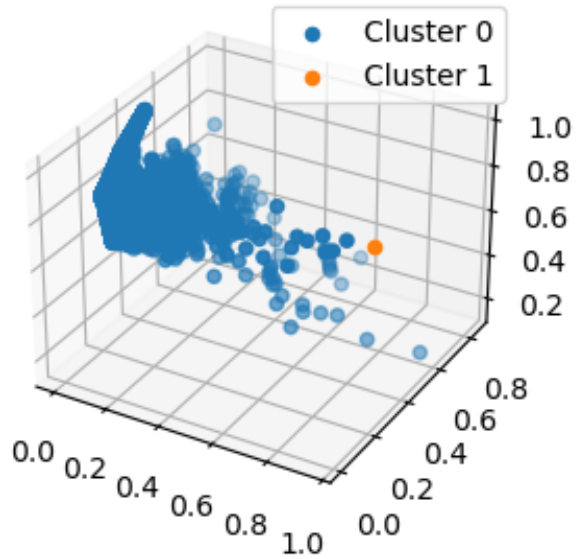

Bisecting Kmeans 4 clusters (View 2)
centroids
0
1
2
3
PCA_1
PCA_2
0.0
0.2
0.4
0.6
0.8
1.0


Bisecting Kmeans 5 clusters (View 1)
centroids
0
1
2
3
4
PCA_1
PCA_2
PCA_3
0.0
0.2
0.4
0.6
0.8
1.0

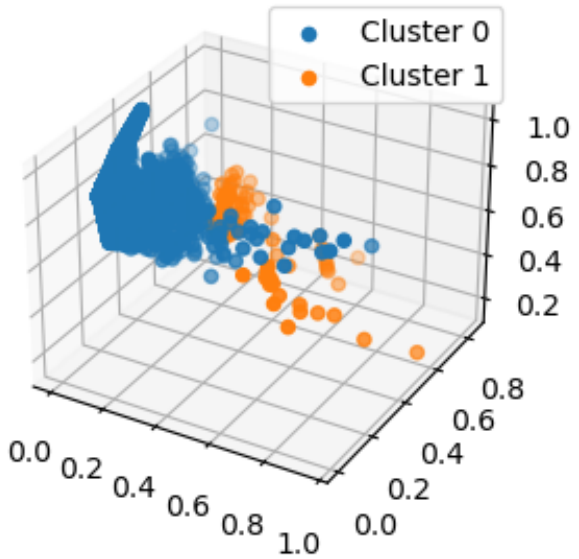

Bisecting Kmeans 5 clusters (View 2)
centroids
0
1
2
3
4
PCA_1
PCA_2
0.0
0.2
0.4
0.6
0.8
1.0

Algoritam sakupljajućeg hijerarhijskog klasterovanja

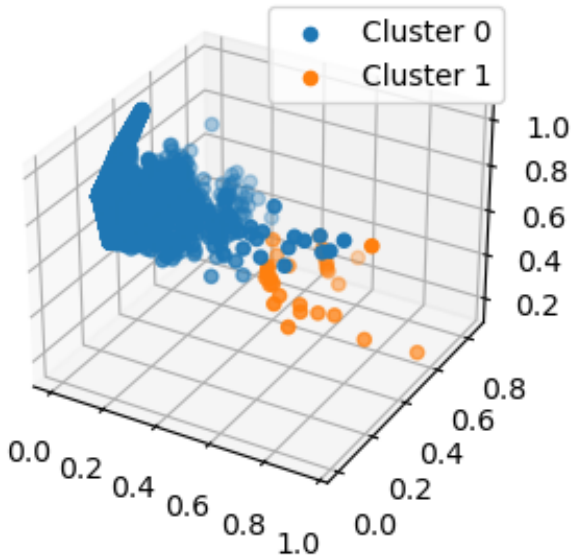
n_clusters: 2, Linkage: single



n_clusters: 2, Linkage: complete



n_clusters: 2, Linkage: average

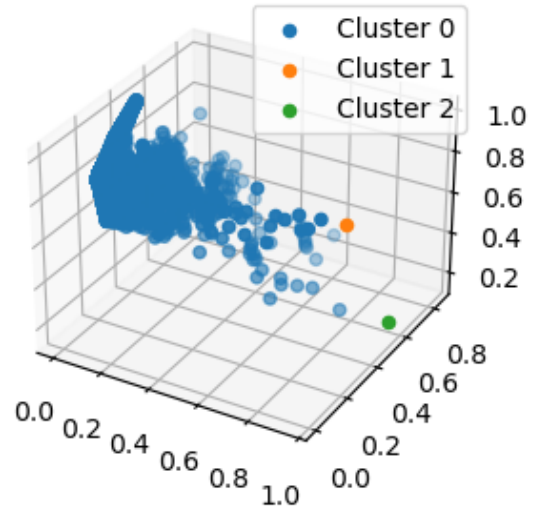


n_clusters: 3, Linkage: single

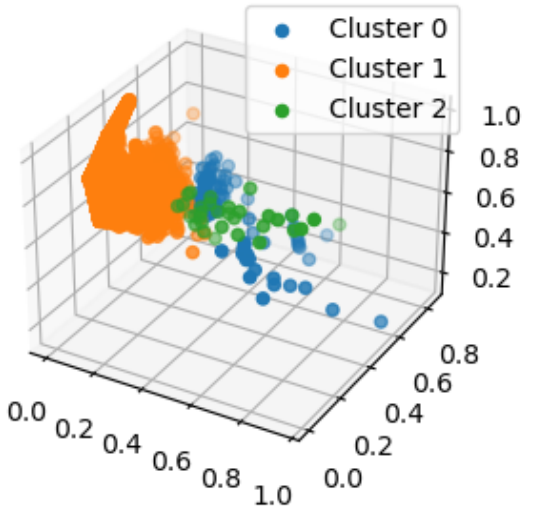
n_clusters: 3, Linkage: complete

n_clusters: 3, Linkage: average

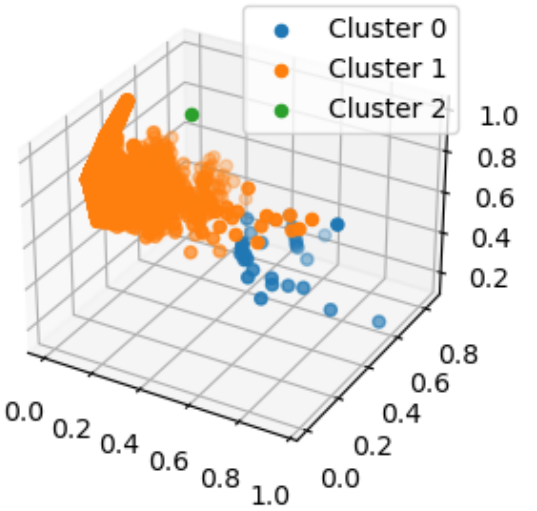
n_clusters: 3, Linkage: single



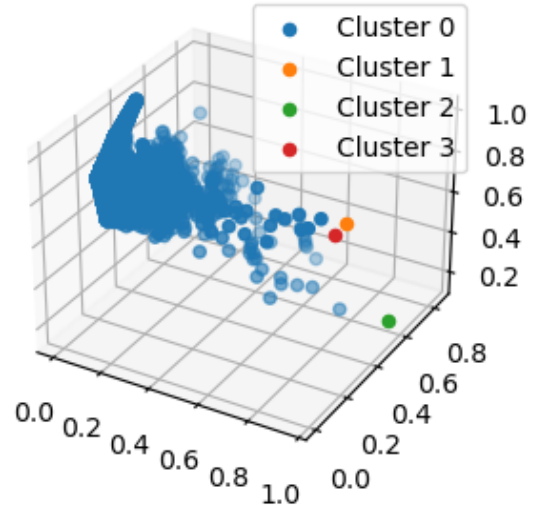
n_clusters: 3, Linkage: complete



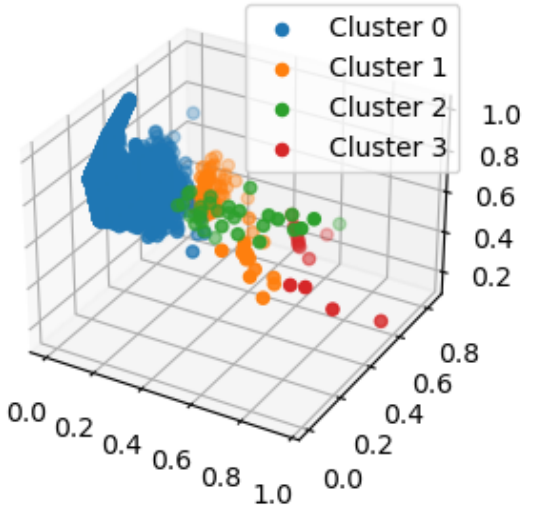
n_clusters: 3, Linkage: average



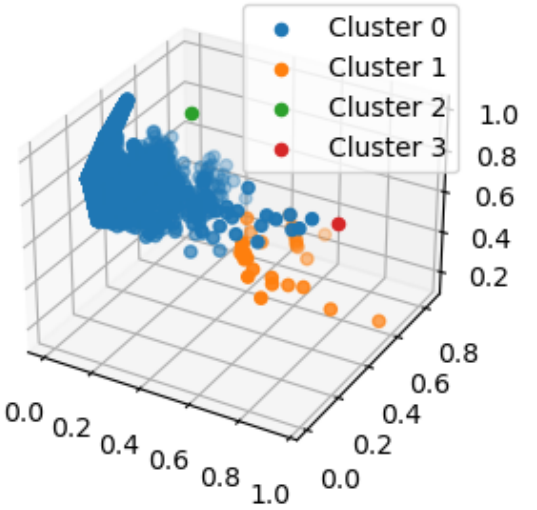
n_clusters: 4, Linkage: single



n_clusters: 4, Linkage: complete



n_clusters: 4, Linkage: average

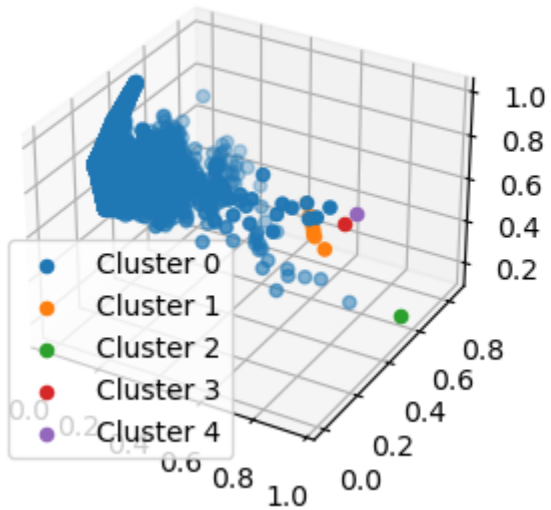


n_clusters: 5, Linkage: single

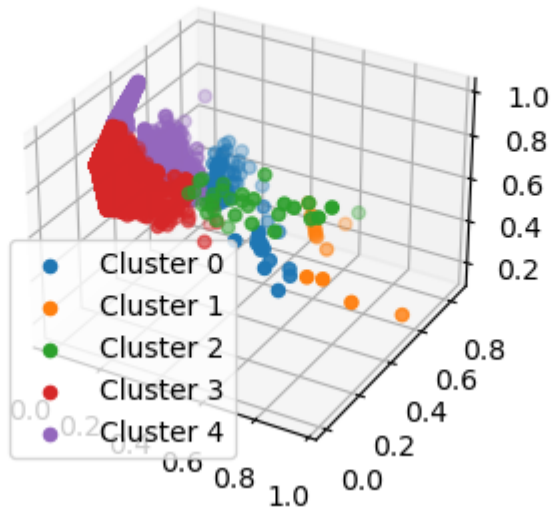
n_clusters: 5, Linkage: complete

n_clusters: 5, Linkage: average

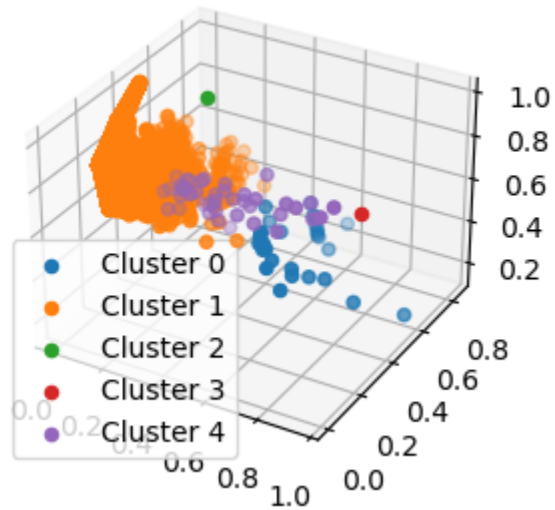
n_clusters: 5, Linkage: single

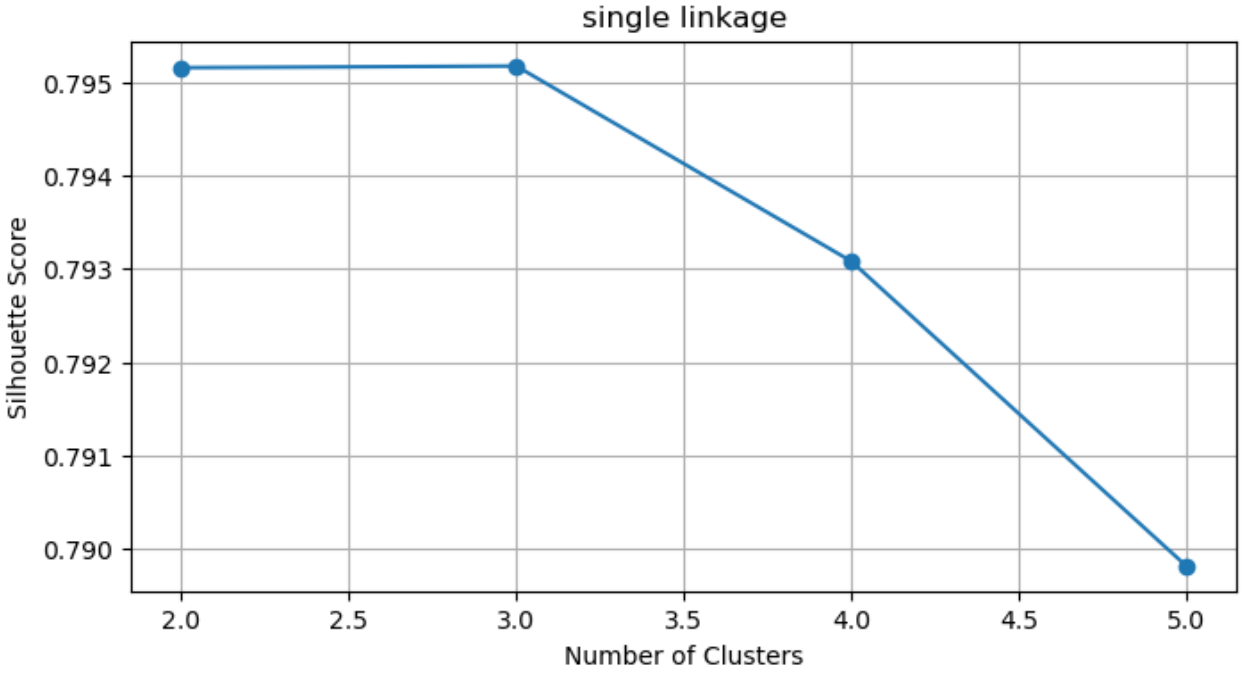


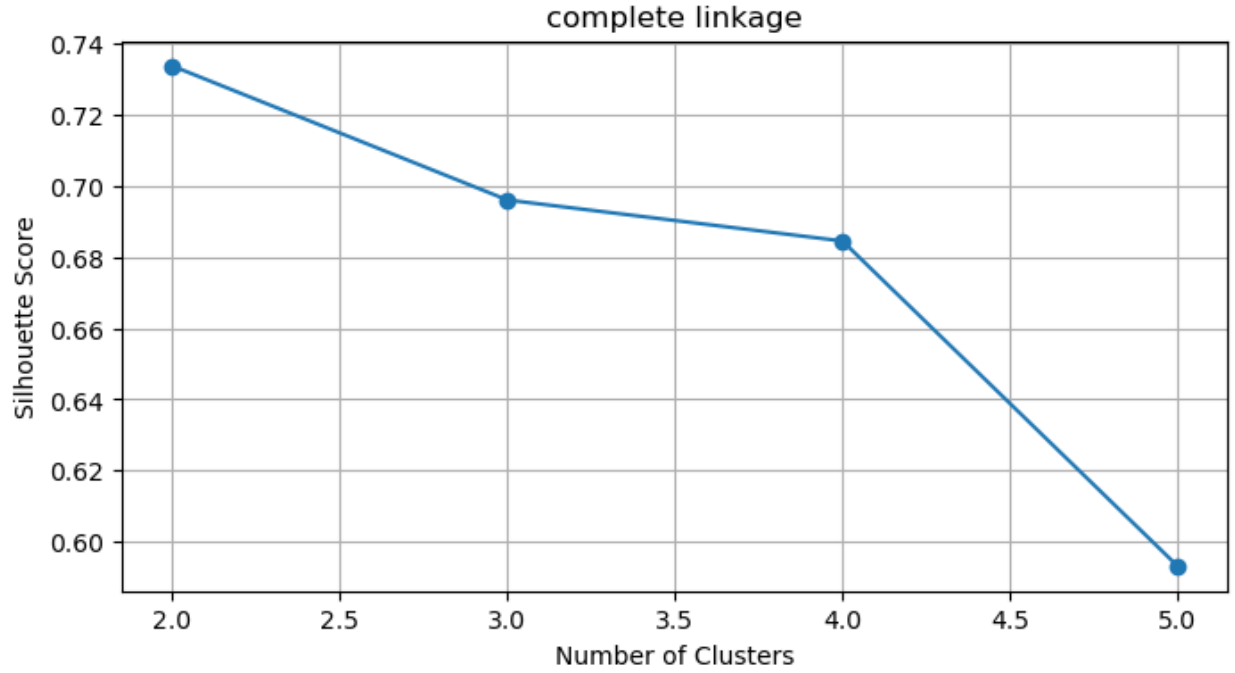
n_clusters: 5, Linkage: complete

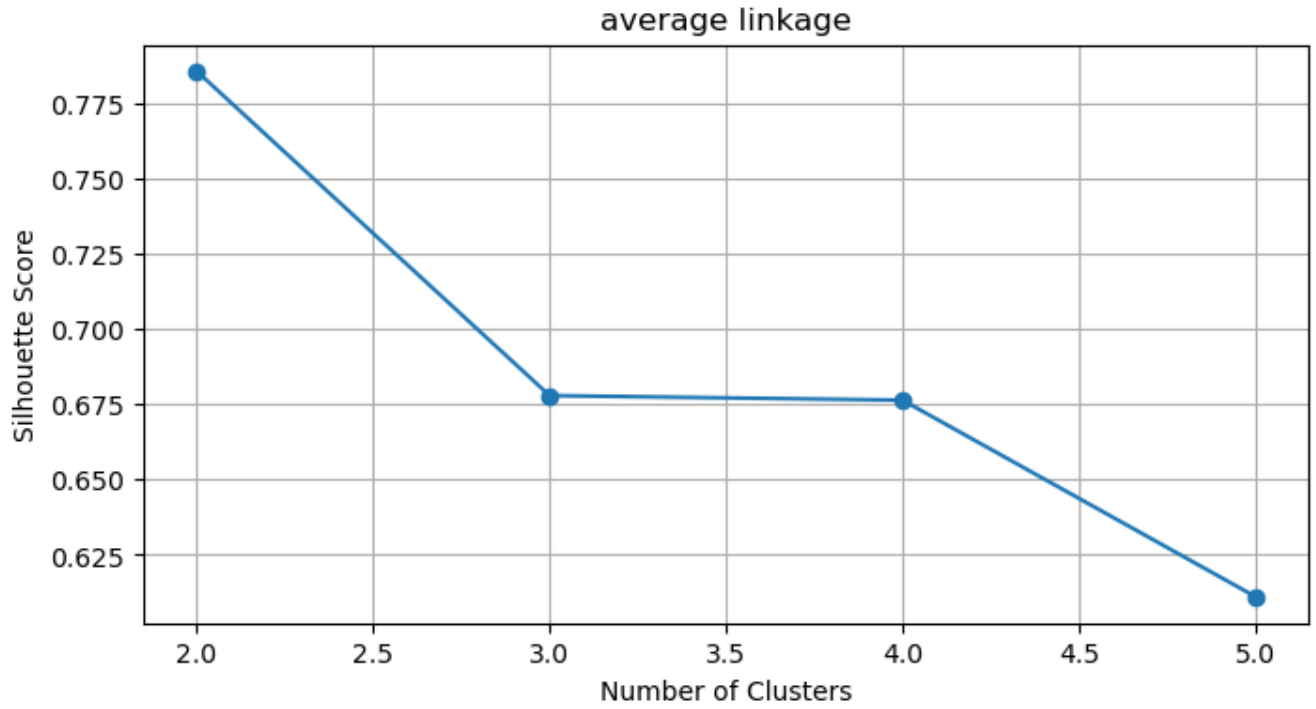


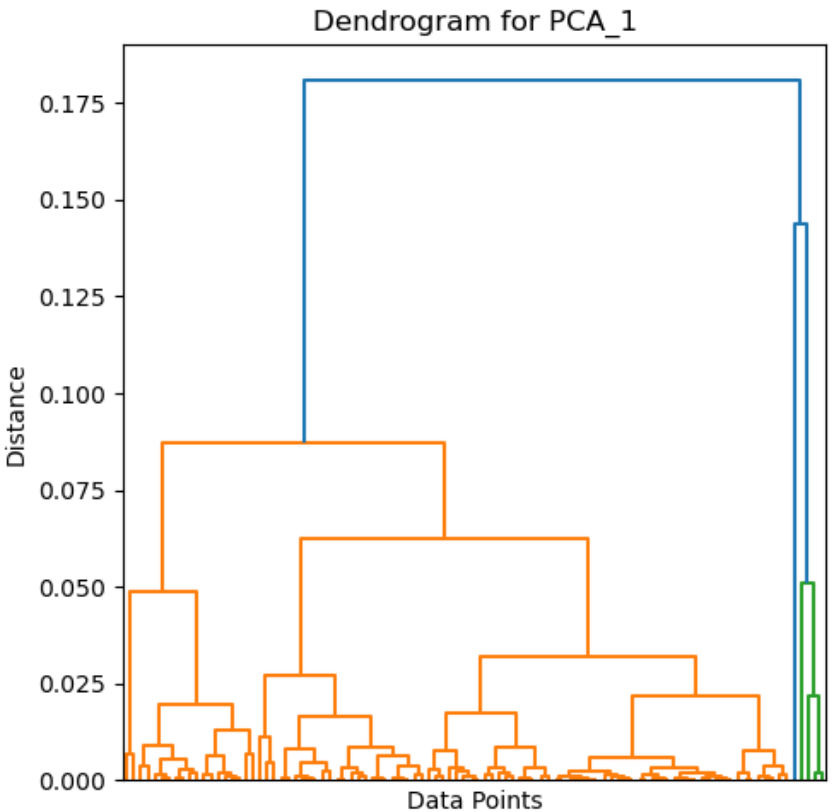
n_clusters: 5, Linkage: average

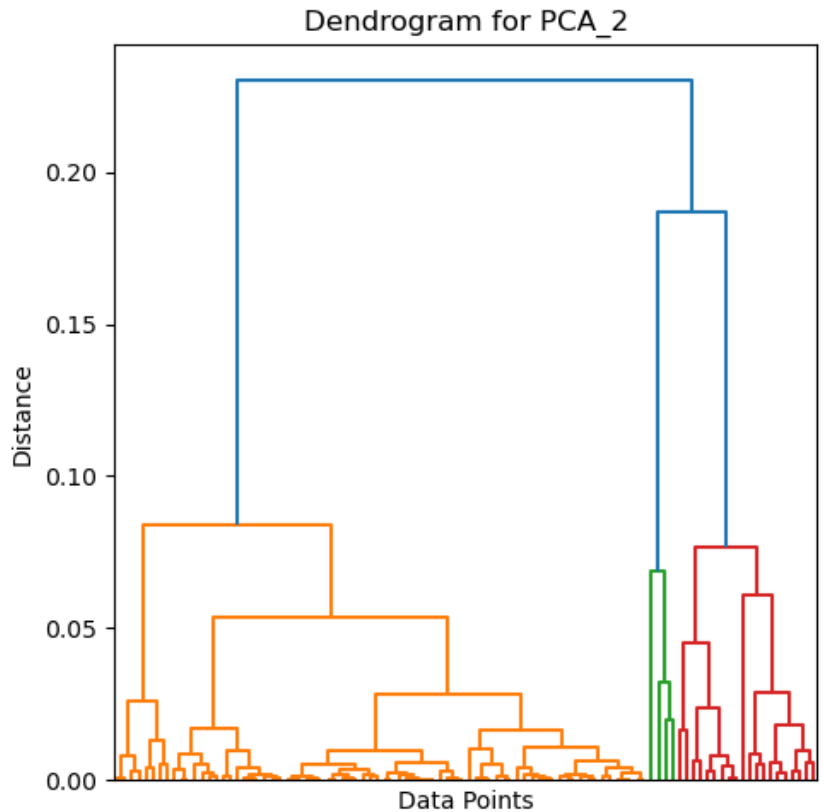


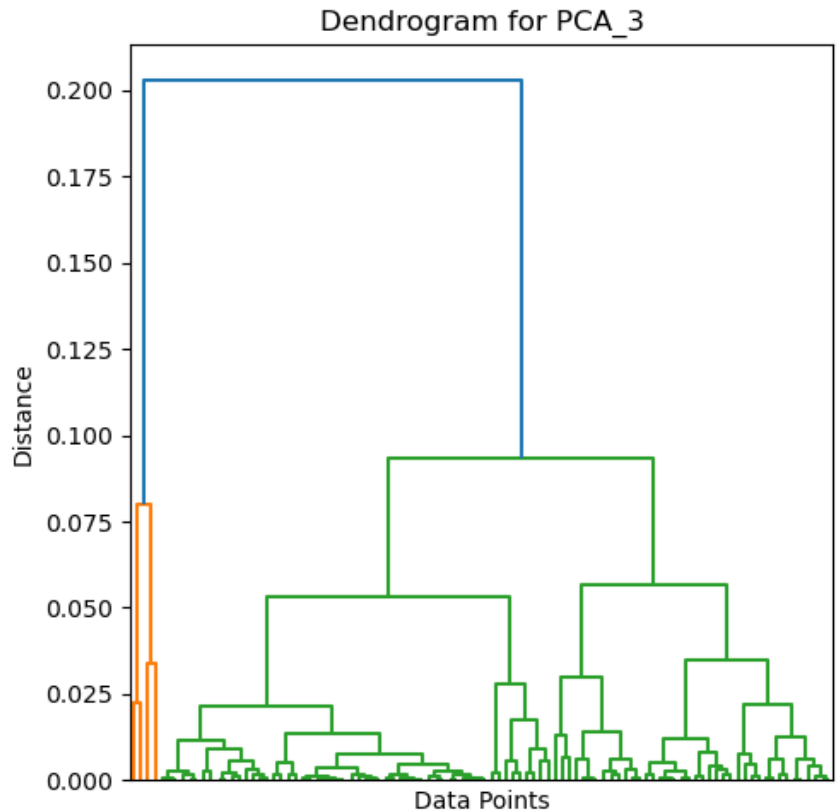

single linkage
Silhouette Score
0.795
0.794
0.793
0.792
0.791
0.790
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Number of Clusters


complete linkage
Silhouette Score
0.74
0.72
0.70
0.68
0.66
0.64
0.62
0.60
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Number of Clusters

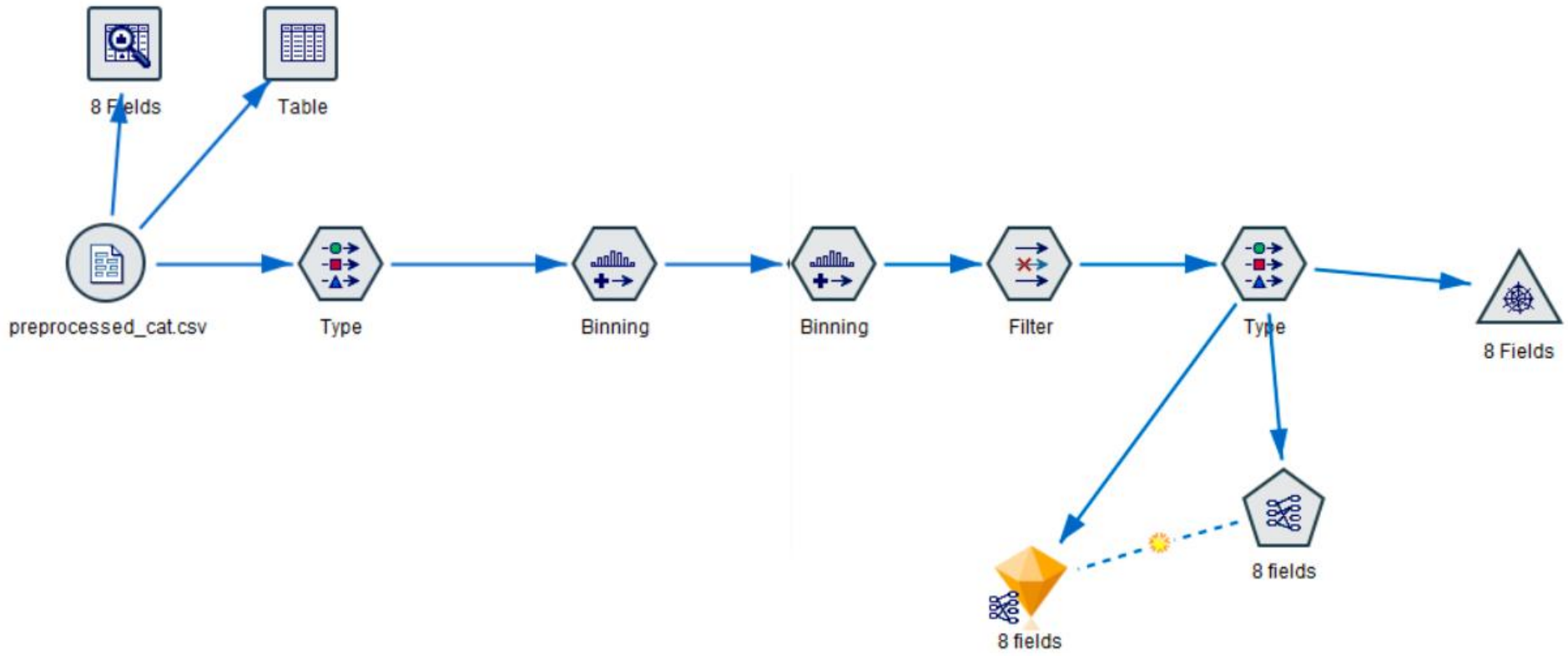

average linkage
Silhouette Score
0.775
0.750
0.725
0.700
0.675
0.650
0.625
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Number of Clusters


Dendrogram for PCA_1
Distance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Data Points


Dendrogram for PCA_2
Distance
0.20
0.15
0.10
0.05
0.00
Data Points


Dendrogram for PCA_3
Distance
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Data Points

Apriori algoritam



Type

Preview

Types | Format | Annotations

Read Values | Clear Values | Clear All Values

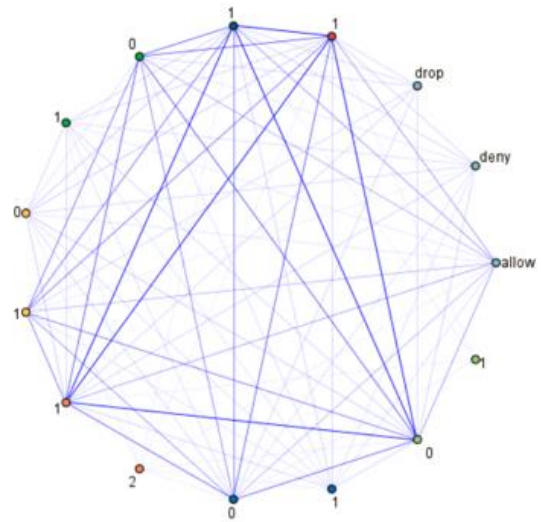
Field	Measurement	Values	Missing	Check	Role
Action	Nominal	allow,deny...		None	Both
Bytes Sent	Continuous	[60,53069...		None	Both
Bytes Receiv...	Continuous	[0,304353...		None	Both
Elapsed Tim...	Continuous	[0,10824]		None	Both
Source Port_...	Flag	1/0		None	Both
Source Port_...	Flag	1/0		None	Both
Destination ...	Flag	1/0		None	Both
Destination ...	Flag	1/0		None	Both

View current fields | View unused field settings

OK | Cancel | Apply | Reset

Sort by: Confidence % 7 of 7

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



Action | Bytes Received_BIN | Bytes Sent_BIN | Destination Port_Registered | Destination Port_Well known | Elapsed Time (sec)_BIN | Source Port_Registered | Source Port_Well known