ANALIZA SKUPA PODATAKA COFFEE QUALITY DATABASE FROM CQI

ANJA CVETKOVIĆ

ISTRAŽIVANJE PODATAKA I

ANALIZA PODATAKA I PRETPROCESIRANJE

Mere kvaliteta

- Aroma – miris kafe
- Flavor – ukus kafe
- Aftertaste – da li kafa ostavlja prijatnu senzaciju na paleti
- Acidity – prijatna oštrinu u ukusu kafe
- Body – izražava teksturu kafe, za koju je poželjno da ima težinu i da bude kremasta
- Balance – ukus se ne lokaliziju na jednom mestu palete
- Uniformity – konzistentnost ukusa
- Cup Cleanliness – "čista šoljica" označava da se ne javljaju arome koje nisu od kafe a koje su posledica defekata
- Sweetness – slatkoća koja se oseti na vrhu jezika, poželjna osobina
- Moisture – preporučena vlažnost je oko 11.5% potpuno procesuiranog zrna
- Defects – mogu da budu prve i druge kategorije, na osnovu uzorka od 350g, defekti prvi kategorije su skroz crna ili kisela zrna, tragovi rastinja ili kamenčići, dok su defekti druge kategorije insekti u uzorku, oštećenja od vode i slično.
- Total Cup Points – zbir ocean prethodnih mera, vrednost na [1,100] i ciljna promenljiva

ANALIZA PODATAKA I PRETPROCESIRANJE

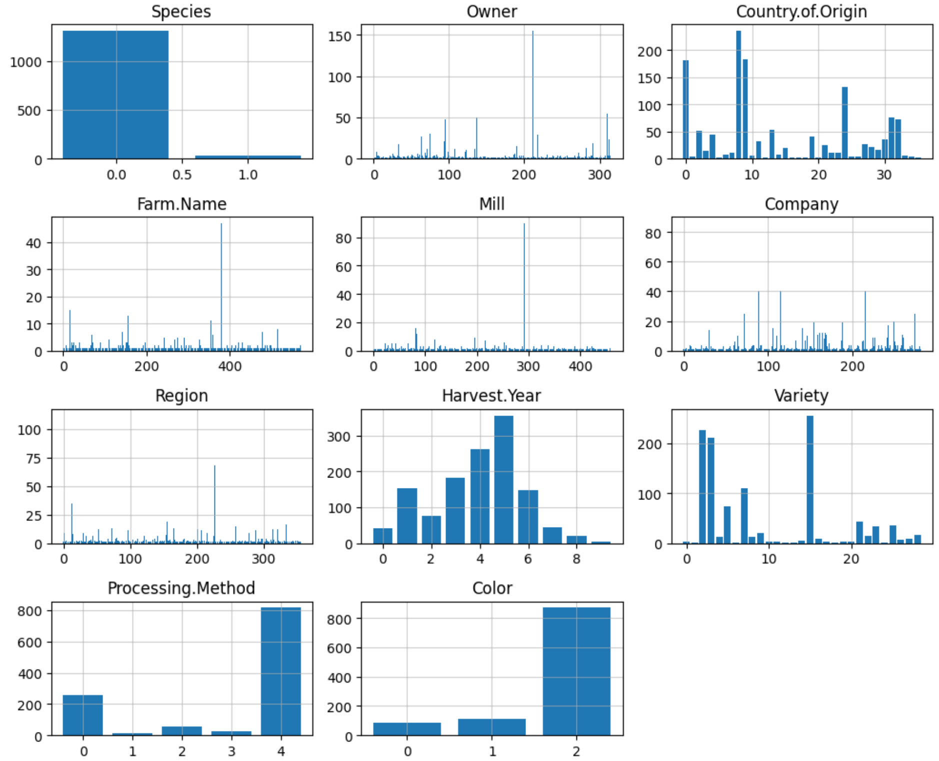
Metapodaci o zrnu

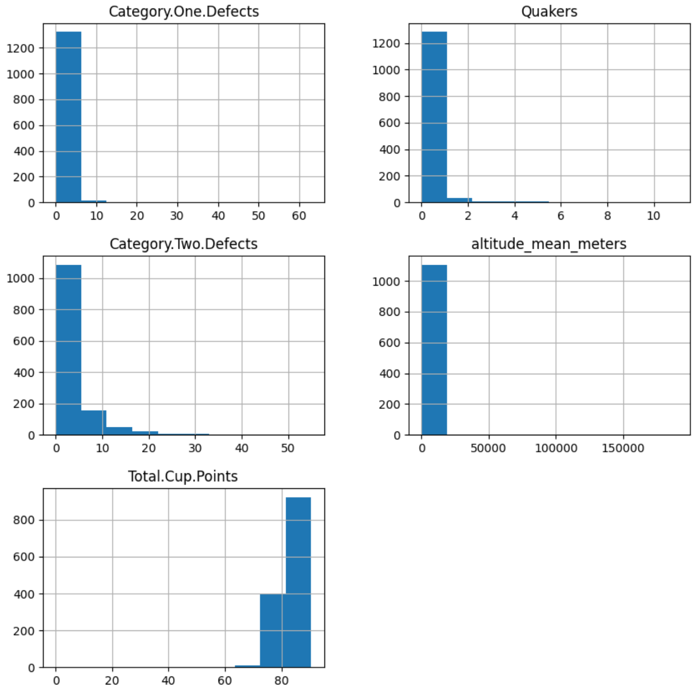
- Processing Method
- Color
- Species (Arabica/Robusta)
- Variety – podvrsta
- Quakers - broj nezrelih zrna

Metapodaci od poreklu

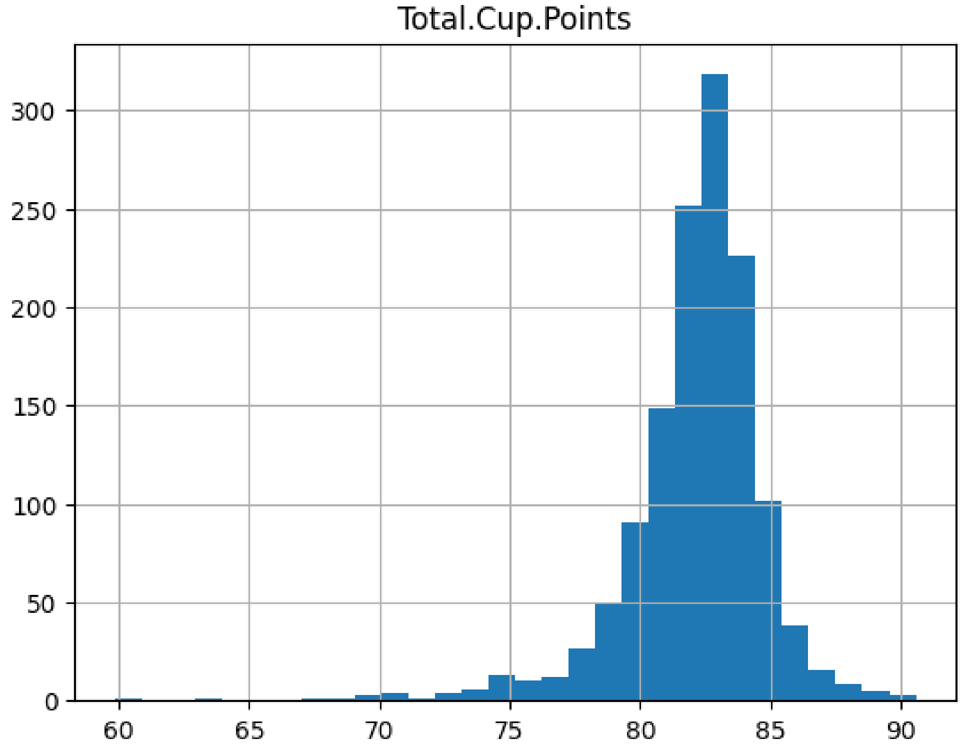
- Owner
- Country of Origin
- Farm Name
- Lot Number
- Mill
- Company
- Altitude
- Region

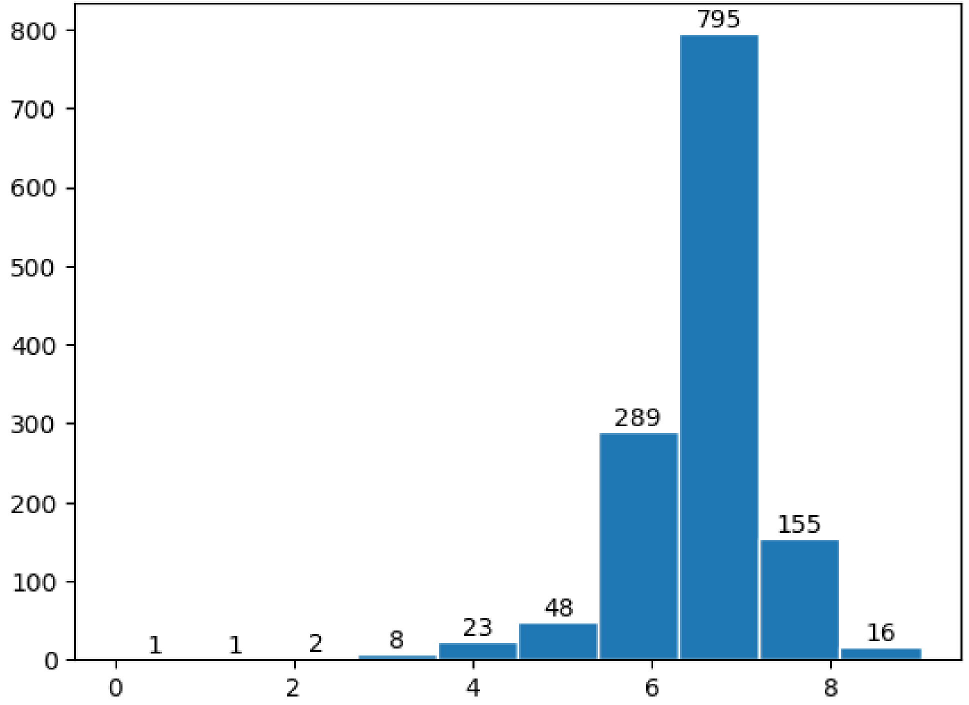
ANALIZA PODATAKA I PRETPROCESIRANJE



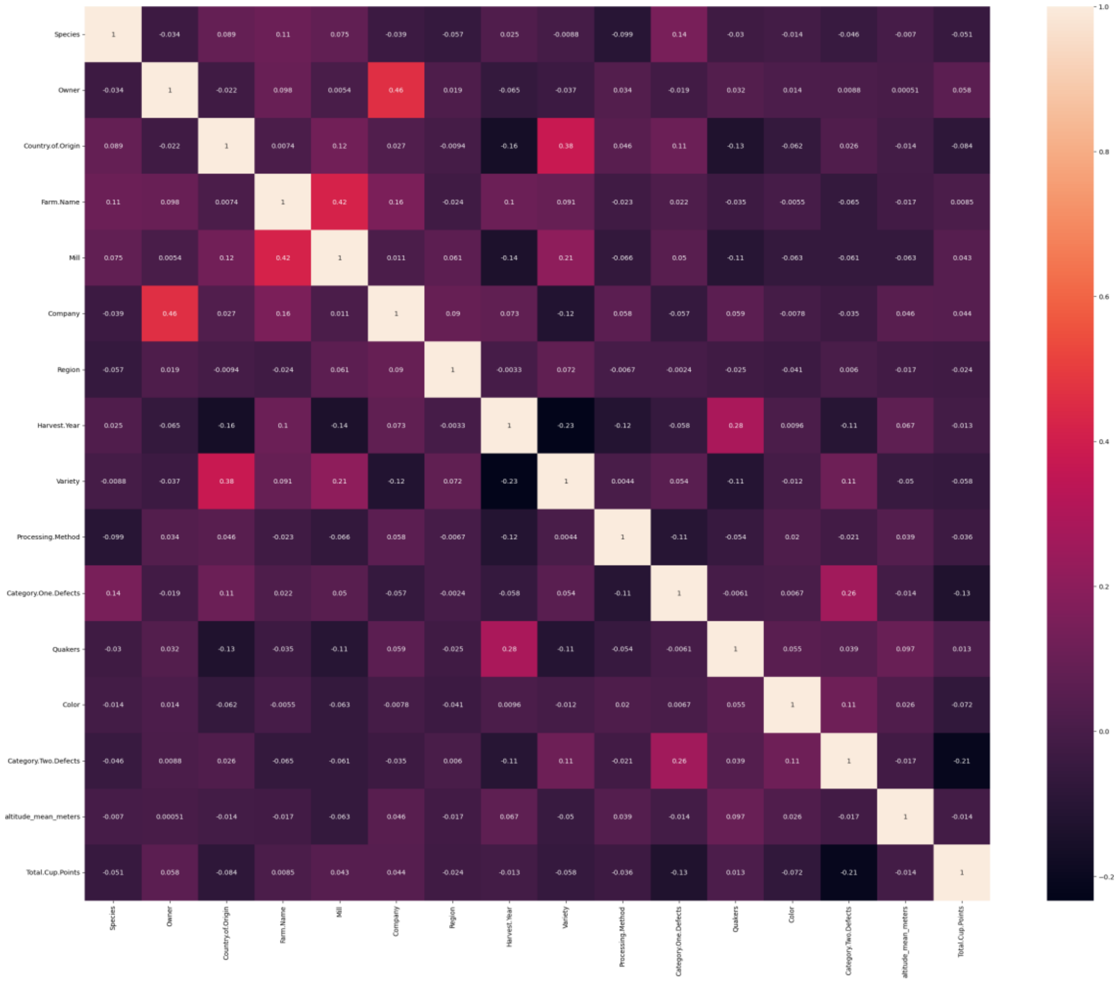


ANALIZA PODATAKA I PRETPROCESIRANJE



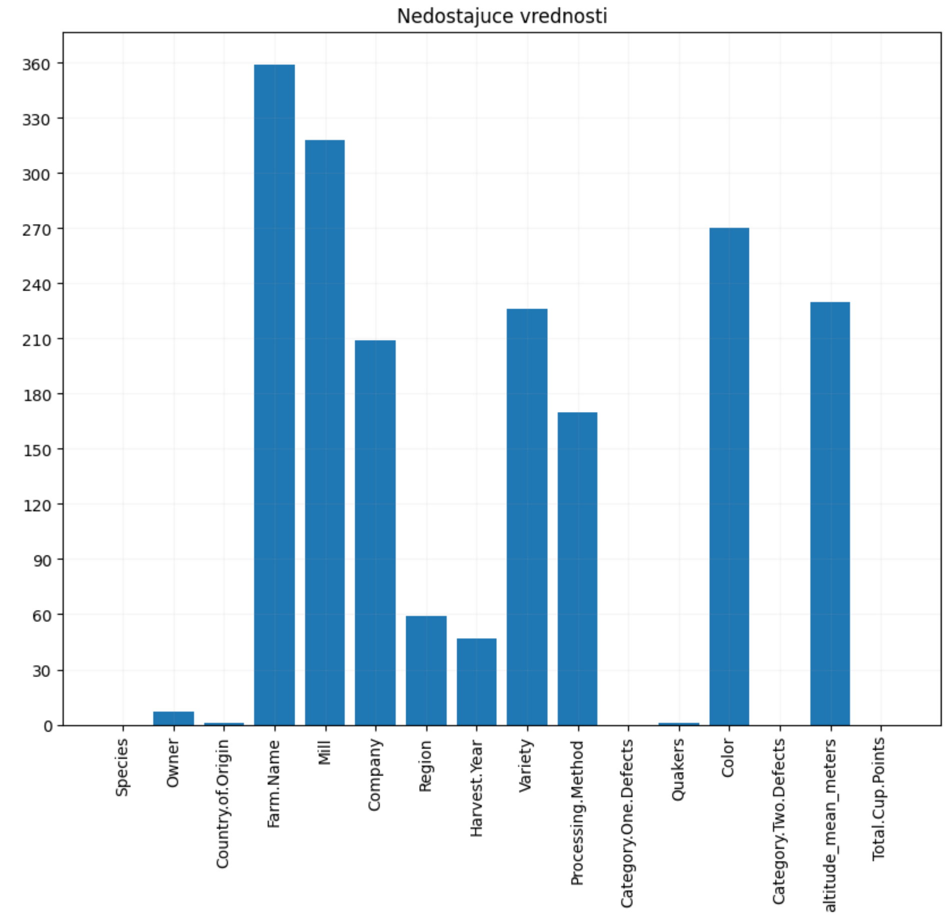


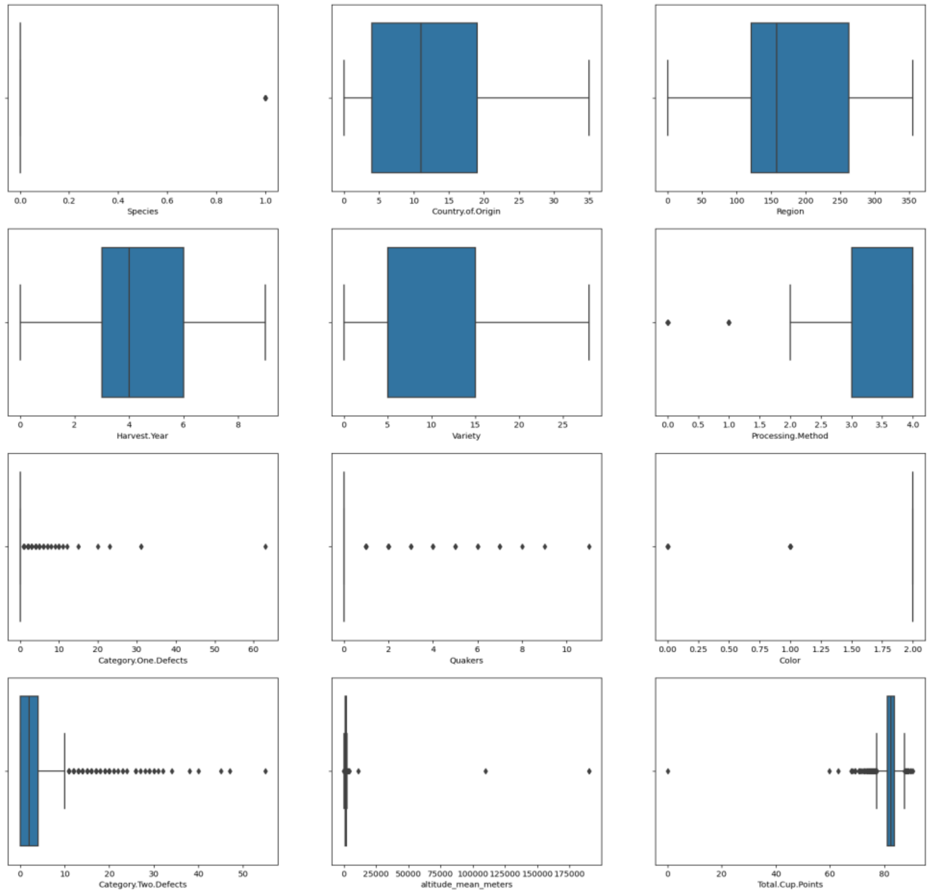
ANALIZA PODATAKA I PRETPROCESIRANJE



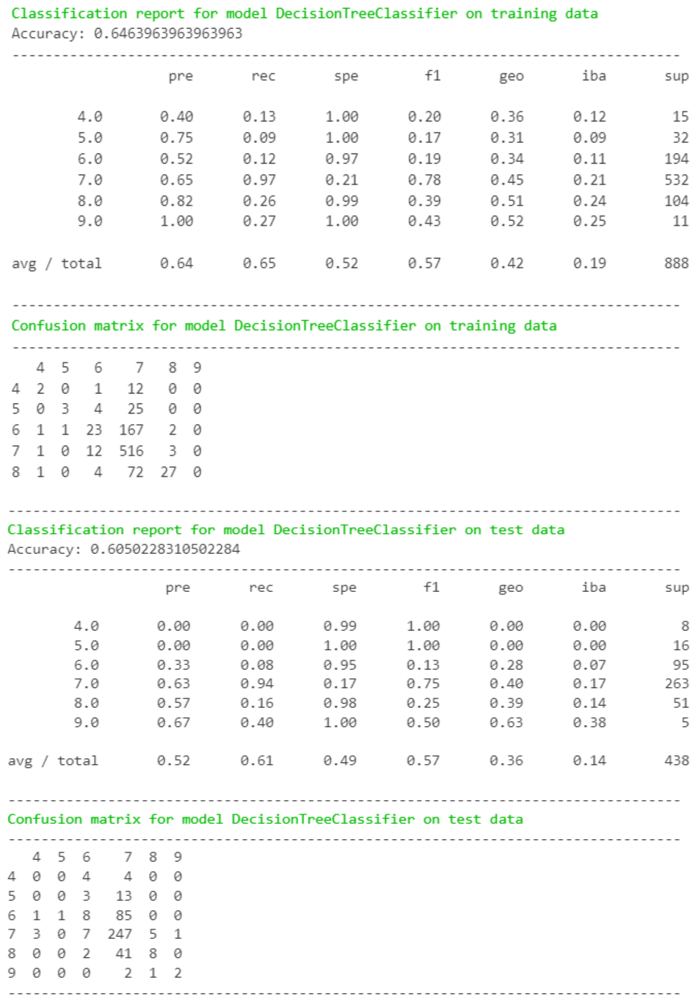
	Species	Owner	Country.of.Origin	Farm.Name	Mill	Company	Region	Harvest.Year	Variety	Processing.Method	Category.One.Defects	Quakers	Color	Category.Two.Defects	altitude_mean_meters	Total.Cup.Points
Species	1	-0.034	0.089	0.11	0.075	-0.039	-0.057	0.025	-0.0088	-0.099	0.14	-0.03	-0.014	-0.046	-0.007	-0.051
Owner	-0.034	1	-0.022	0.098	0.0054	0.46	0.019	-0.065	-0.037	0.034	-0.019	0.032	0.014	0.0088	0.00051	0.058
Country.of.Origin	0.089	-0.022	1	0.0074	0.12	0.027	-0.0094	-0.16	0.38	0.046	0.11	-0.13	-0.062	0.026	-0.014	-0.084
Farm.Name	0.11	0.098	0.0074	1	0.42	0.16	-0.024	0.1	0.091	-0.023	0.022	-0.035	-0.0055	-0.065	-0.017	0.0085
Mill	0.075	0.0054	0.12	0.42	1	0.011	0.061	-0.14	0.21	-0.066	0.05	-0.11	-0.063	-0.061	-0.063	0.043
Company	-0.039	0.46	0.027	0.16	0.011	1	0.09	0.073	-0.12	0.058	-0.057	0.059	-0.0078	-0.035	0.046	0.044
Region	-0.057	0.019	-0.0094	-0.024	0.061	0.09	1	-0.0033	0.072	-0.0067	-0.0024	-0.025	-0.041	0.006	-0.017	-0.024
Harvest.Year	0.025	-0.065	-0.16	0.1	-0.14	0.073	-0.0033	1	-0.23	-0.12	-0.058	0.28	0.0096	-0.11	0.067	-0.013
Variety	-0.0088	-0.037	0.38	0.091	0.21	-0.12	0.072	-0.23	1	0.0044	0.054	-0.11	-0.012	0.11	-0.05	-0.058
Processing.Method	-0.099	0.034	0.046	-0.023	-0.066	0.058	-0.0067	-0.12	0.0044	1	-0.11	-0.054	0.02	-0.021	0.039	-0.036
Category.One.Defects	0.14	-0.019	0.11	0.022	0.05	-0.057	-0.0024	-0.058	0.054	-0.11	1	-0.0061	0.0067	0.26	-0.014	-0.13
Quakers	-0.03	0.032	-0.13	-0.035	-0.11	0.059	-0.025	0.28	-0.11	-0.054	-0.0061	1	0.055	0.039	0.097	0.013
Color	-0.014	0.014	-0.062	-0.0055	-0.063	-0.0078	-0.041	0.0096	-0.012	0.02	0.0067	0.055	1	0.11	0.026	-0.072
Category.Two.Defects	-0.046	0.0088	0.026	-0.065	-0.061	-0.035	0.006	-0.11	0.11	-0.021	0.26	0.039	0.11	1	-0.017	-0.21
altitude_mean_meters	-0.007	0.00051	-0.014	-0.017	-0.063	0.046	-0.017	0.067	-0.05	0.039	-0.014	0.097	0.026	-0.017	1	-0.014
Total.Cup.Points	-0.051	0.058	-0.084	0.0085	0.043	0.044	-0.024	-0.013	-0.058	-0.036	-0.13	0.013	-0.072	-0.21	-0.014	1

ANALIZA PODATAKA I PRETPROCESIRANJE

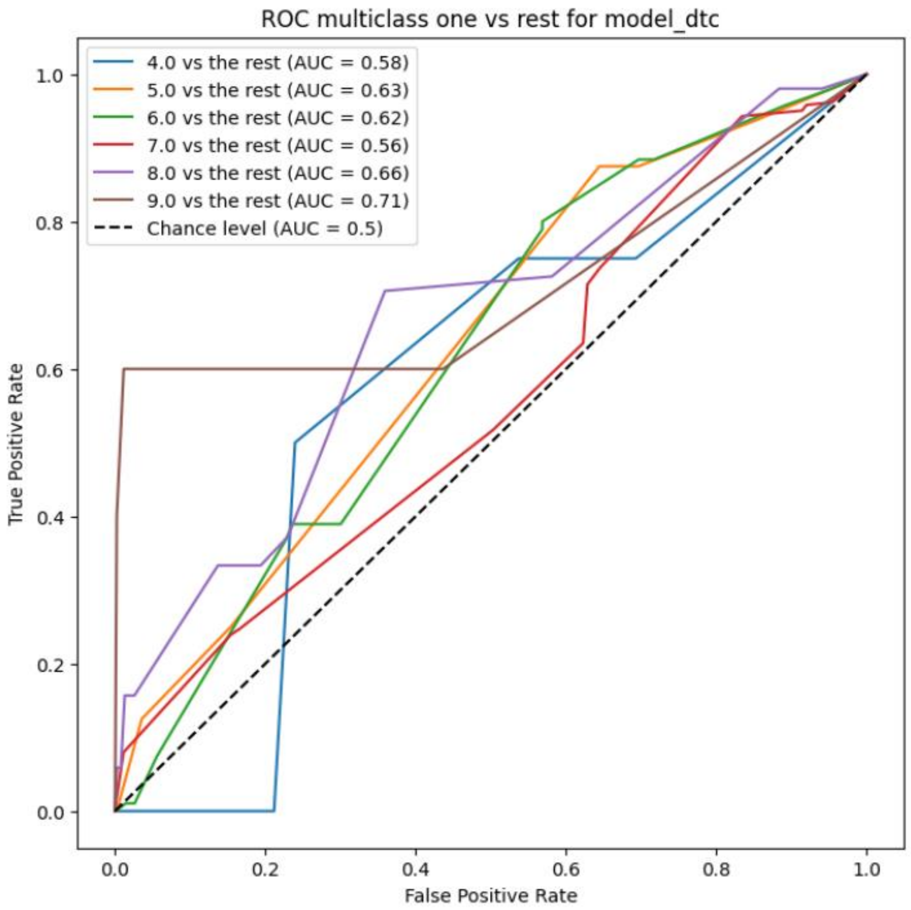




KLASIFIKACIJA-STABLA ODLUČIVANJA

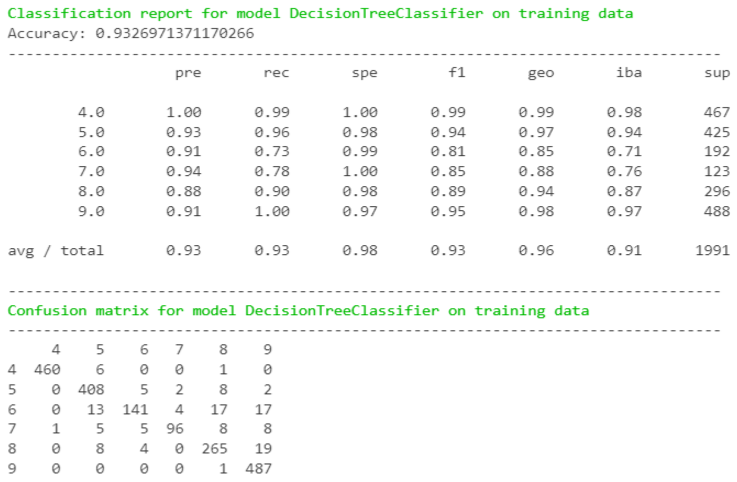


```
Classification report for model DecisionTreeClassifier on training data
Accuracy: 0.6463963963963963
--------------------------------------------------------------------------------
                  pre       rec       spe        f1       geo       iba       sup

        4.0       0.40      0.13      1.00      0.20      0.36      0.12        15
        5.0       0.75      0.09      1.00      0.17      0.31      0.09        32
        6.0       0.52      0.12      0.97      0.19      0.34      0.11       194
        7.0       0.65      0.97      0.21      0.78      0.45      0.21       532
        8.0       0.82      0.26      0.99      0.39      0.51      0.24       104
        9.0       1.00      0.27      1.00      0.43      0.52      0.25        11

avg / total       0.64      0.65      0.52      0.57      0.42      0.19       888

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------------
   4  5   6    7   8  9
4  2  0   1   12   0  0
5  0  3   4   25   0  0
6  1  1  23  167   2  0
7  1  0  12  516   3  0
8  1  0   4   72  27  0

--------------------------------------------------------------------------------
Classification report for model DecisionTreeClassifier on test data
Accuracy: 0.6050228310502284
--------------------------------------------------------------------------------
                  pre       rec       spe        f1       geo       iba       sup

        4.0       0.00      0.00      0.99      1.00      0.00      0.00         8
        5.0       0.00      0.00      1.00      1.00      0.00      0.00        16
        6.0       0.33      0.08      0.95      0.13      0.28      0.07        95
        7.0       0.63      0.94      0.17      0.75      0.40      0.17       263
        8.0       0.57      0.16      0.98      0.25      0.39      0.14        51
        9.0       0.67      0.40      1.00      0.50      0.63      0.38         5

avg / total       0.52      0.61      0.49      0.57      0.36      0.14       438

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------------
   4  5  6    7  8  9
4  0  0  4    4  0  0
5  0  0  3   13  0  0
6  1  1  8   85  0  0
7  3  0  7  247  5  1
8  0  0  2   41  8  0
9  0  0  0    2  1  2
--------------------------------------------------------------------------------
```

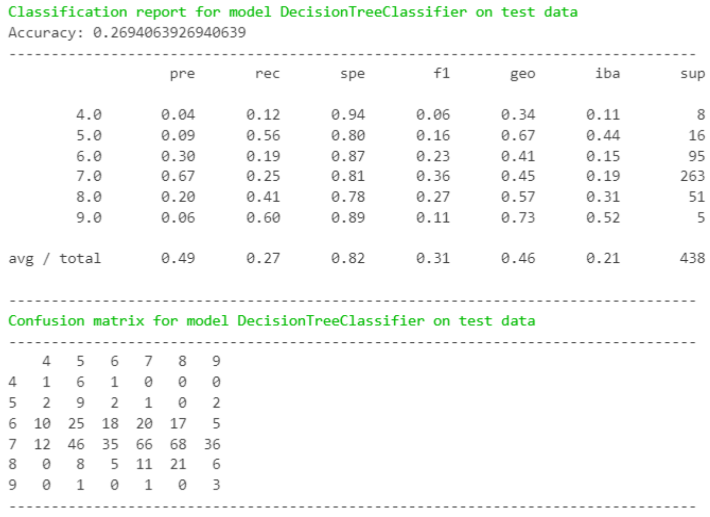


Stablo odlučivanja sa nasumičnim hiper-parametrima i nebalansiranim podacima

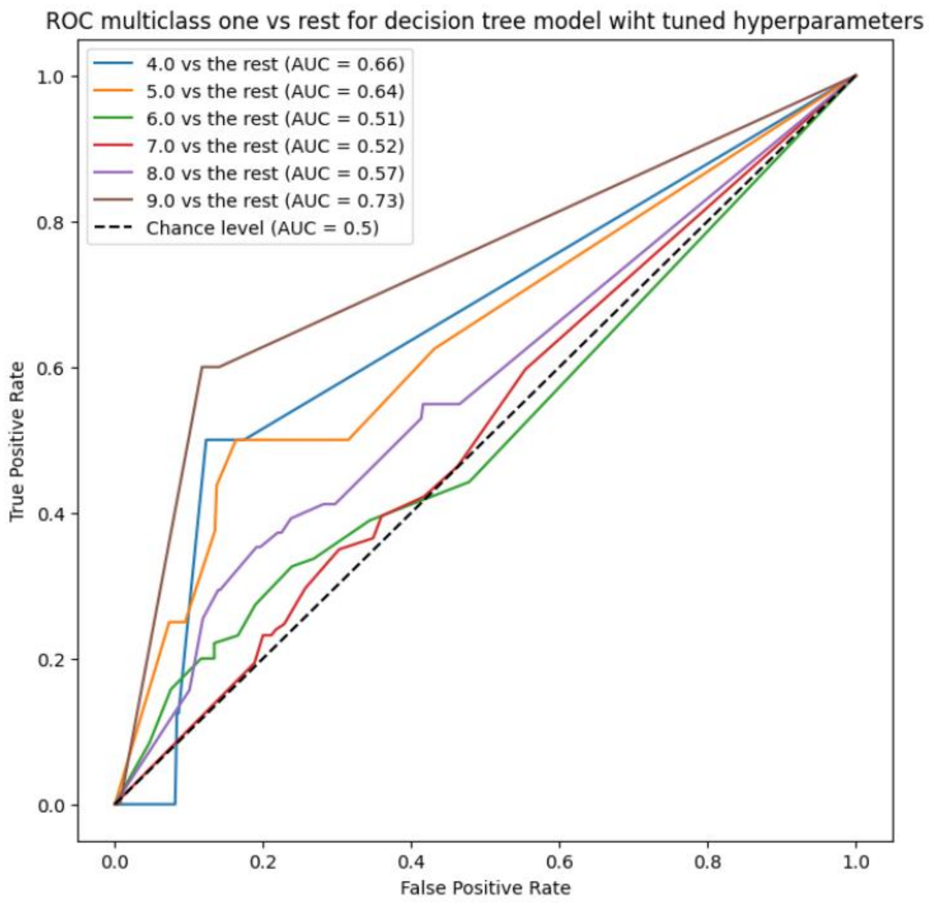
KLASIFIKACIJA-STABLA ODLUČIVANJA



```
Classification report for model DecisionTreeClassifier on training data
Accuracy: 0.9326971371170266
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

        4.0       1.00      0.99      1.00      0.99      0.99      0.98       467
        5.0       0.93      0.96      0.98      0.94      0.97      0.94       425
        6.0       0.91      0.73      0.99      0.81      0.85      0.71       192
        7.0       0.94      0.78      1.00      0.85      0.88      0.76       123
        8.0       0.88      0.90      0.98      0.89      0.94      0.87       296
        9.0       0.91      1.00      0.97      0.95      0.98      0.97       488

avg / total       0.93      0.93      0.98      0.93      0.96      0.91      1991

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------------
     4    5    6   7    8    9
4  460    6    0   0    1    0
5    0  408    5   2    8    2
6    0   13  141   4   17   17
7    1    5    5  96    8    8
8    0    8    4   0  265   19
9    0    0    0   0    1  487
--------------------------------------------------------------------------------
```

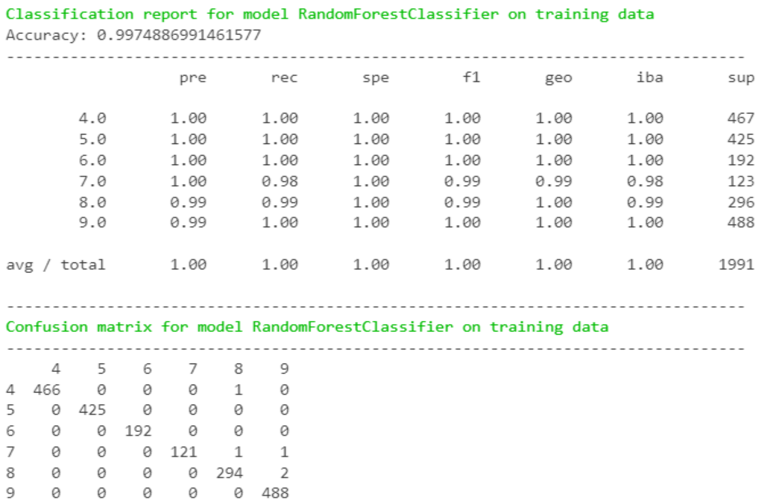


```
Classification report for model DecisionTreeClassifier on test data
Accuracy: 0.2694063926940639
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

        4.0       0.04      0.12      0.94      0.06      0.34      0.11         8
        5.0       0.09      0.56      0.80      0.16      0.67      0.44        16
        6.0       0.30      0.19      0.87      0.23      0.41      0.15        95
        7.0       0.67      0.25      0.81      0.36      0.45      0.19       263
        8.0       0.20      0.41      0.78      0.27      0.57      0.31        51
        9.0       0.06      0.60      0.89      0.11      0.73      0.52         5

avg / total       0.49      0.27      0.82      0.31      0.46      0.21       438

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------------
    4   5   6   7   8   9
4   1   6   1   0   0   0
5   2   9   2   1   0   2
6  10  25  18  20  17   5
7  12  46  35  66  68  36
8   0   8   5  11  21   6
9   0   1   0   1   0   3
--------------------------------------------------------------------------------
```

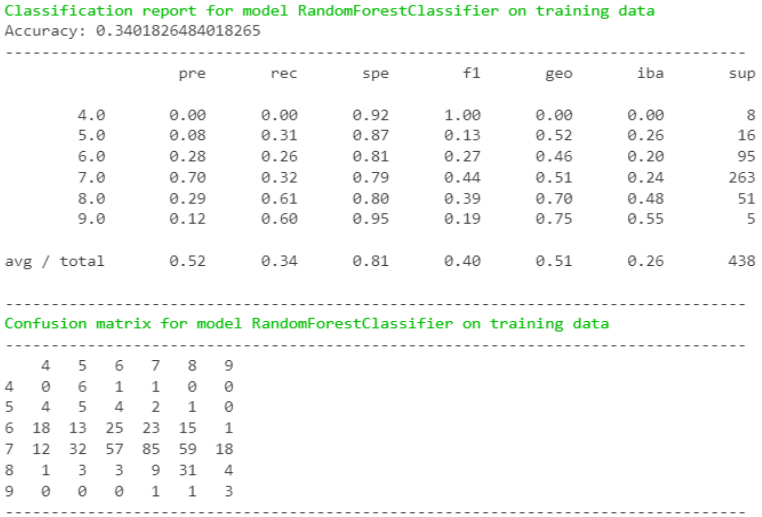


Stablo odlučivanja sa podešenim hiper-parametrima i balansiranim podacima

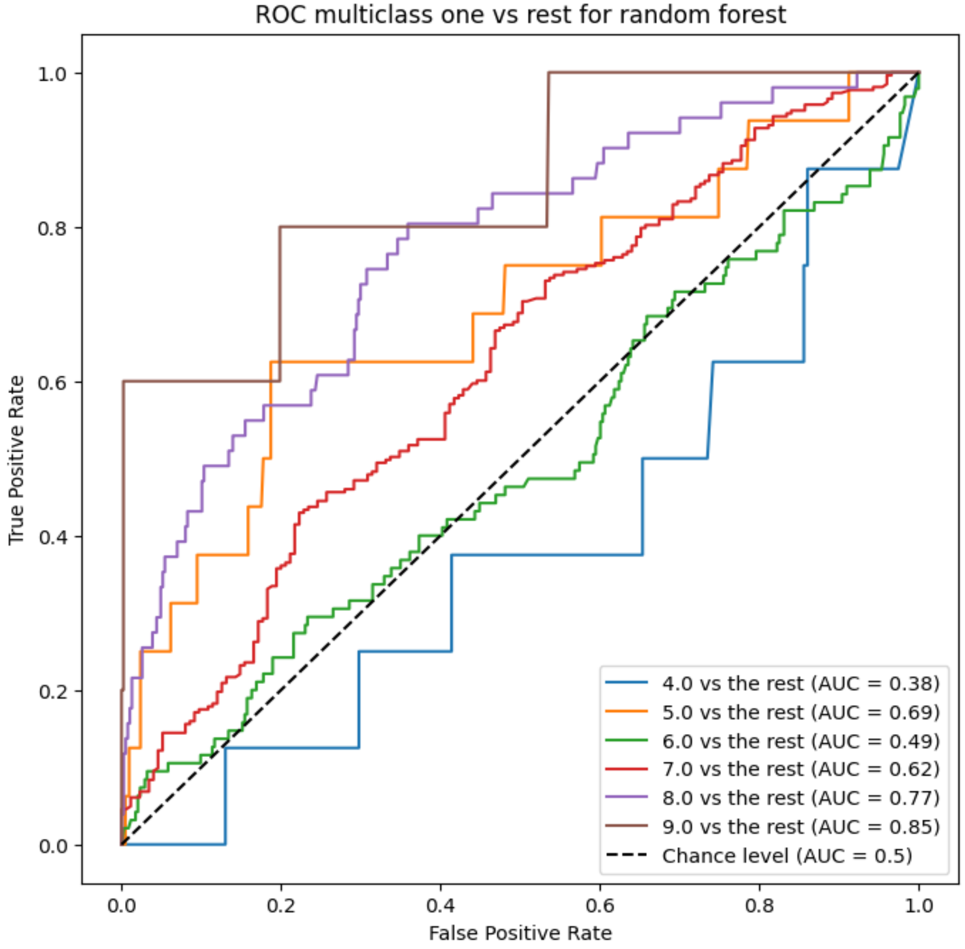
KLASIFIKACIJA-RANDOM FOREST



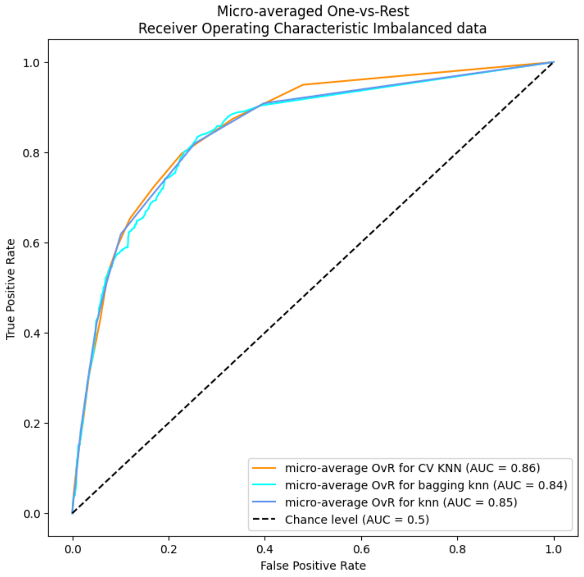
```
Classification report for model RandomForestClassifier on training data
Accuracy: 0.9974886991461577
-----------------------------------------------------------------------------------
                pre       rec       spe        f1       geo       iba       sup

       4.0      1.00      1.00      1.00      1.00      1.00      1.00       467
       5.0      1.00      1.00      1.00      1.00      1.00      1.00       425
       6.0      1.00      1.00      1.00      1.00      1.00      1.00       192
       7.0      1.00      0.98      1.00      0.99      0.99      0.98       123
       8.0      0.99      0.99      1.00      0.99      1.00      0.99       296
       9.0      0.99      1.00      1.00      1.00      1.00      1.00       488

avg / total     1.00      1.00      1.00      1.00      1.00      1.00      1991

-----------------------------------------------------------------------------------
Confusion matrix for model RandomForestClassifier on training data
-----------------------------------------------------------------------------------
     4    5    6    7    8    9
4  466    0    0    0    1    0
5    0  425    0    0    0    0
6    0    0  192    0    0    0
7    0    0    0  121    1    1
8    0    0    0    0  294    2
9    0    0    0    0    0  488
```

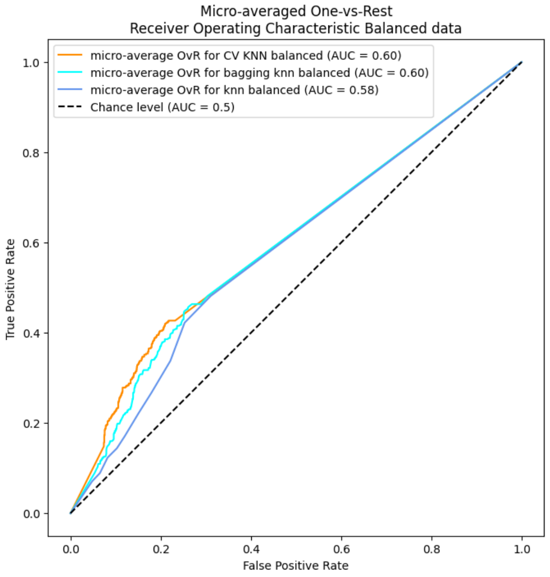


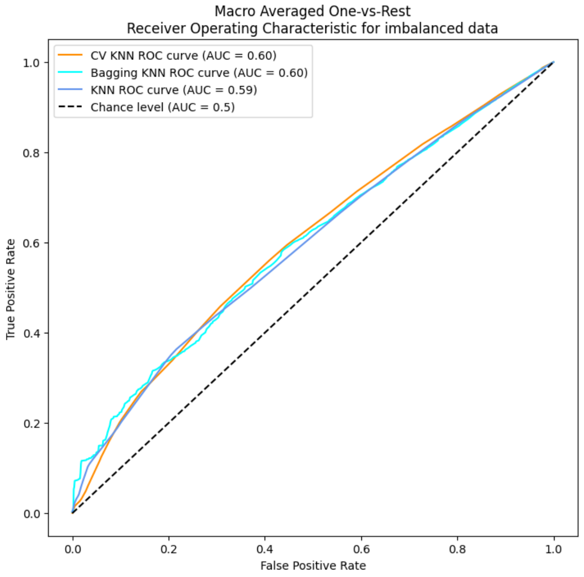
```
Classification report for model RandomForestClassifier on training data
Accuracy: 0.3401826484018265
-----------------------------------------------------------------------------------
                pre       rec       spe        f1       geo       iba       sup

       4.0      0.00      0.00      0.92      1.00      0.00      0.00         8
       5.0      0.08      0.31      0.87      0.13      0.52      0.26        16
       6.0      0.28      0.26      0.81      0.27      0.46      0.20        95
       7.0      0.70      0.32      0.79      0.44      0.51      0.24       263
       8.0      0.29      0.61      0.80      0.39      0.70      0.48        51
       9.0      0.12      0.60      0.95      0.19      0.75      0.55         5

avg / total     0.52      0.34      0.81      0.40      0.51      0.26       438

-----------------------------------------------------------------------------------
Confusion matrix for model RandomForestClassifier on training data
-----------------------------------------------------------------------------------
    4   5   6   7   8   9
4   0   6   1   1   0   0
5   4   5   4   2   1   0
6  18  13  25  23  15   1
7  12  32  57  85  59  18
8   1   3   3   9  31   4
9   0   0   0   1   1   3
-----------------------------------------------------------------------------------
```

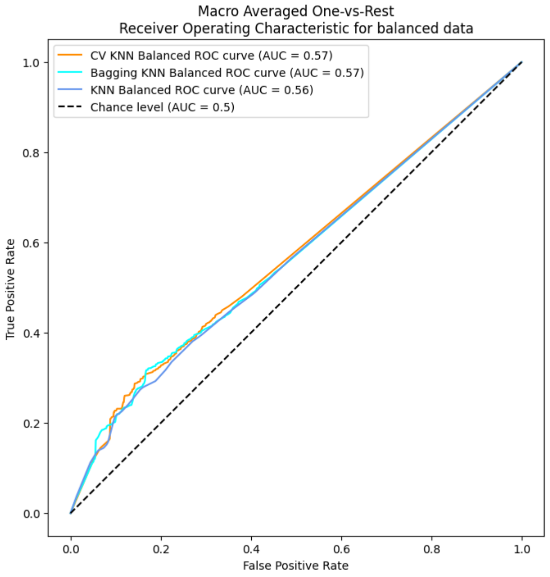


KLASIFIKACIJA-KNN

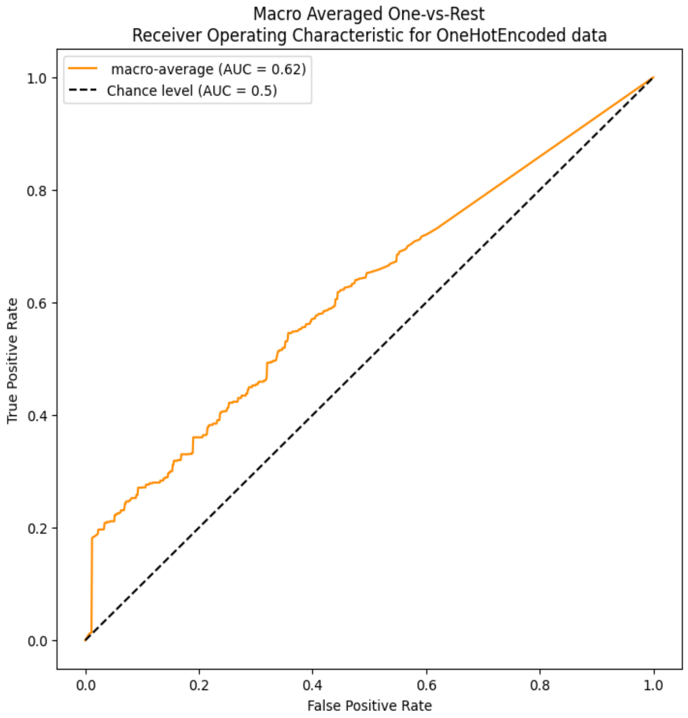


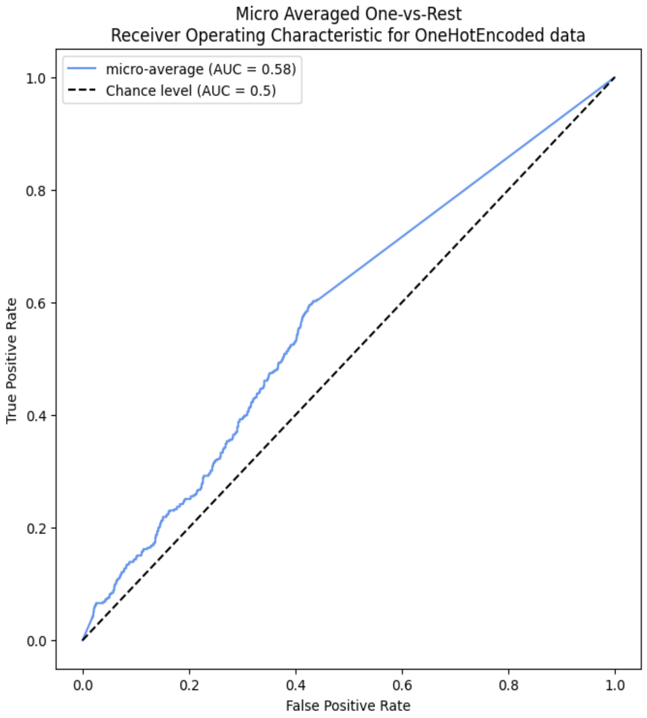


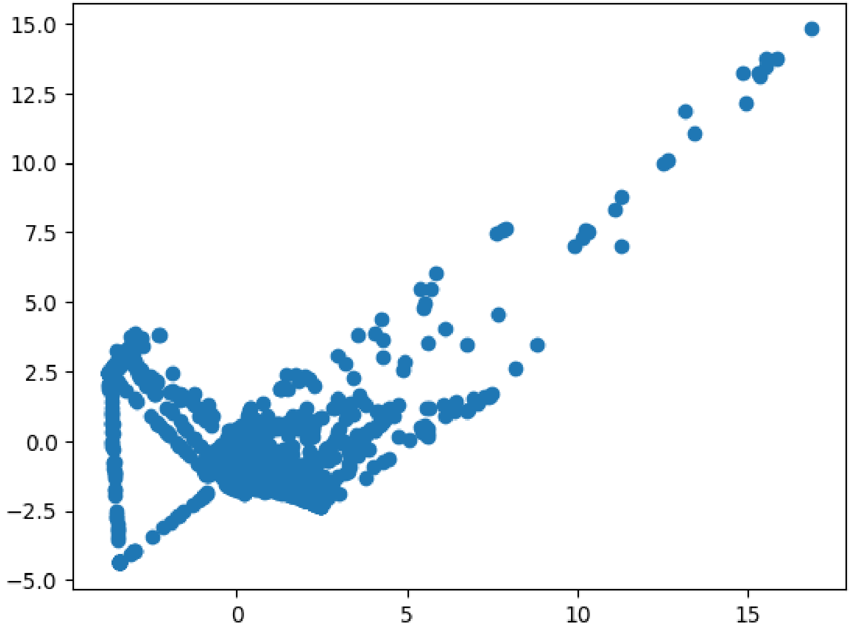




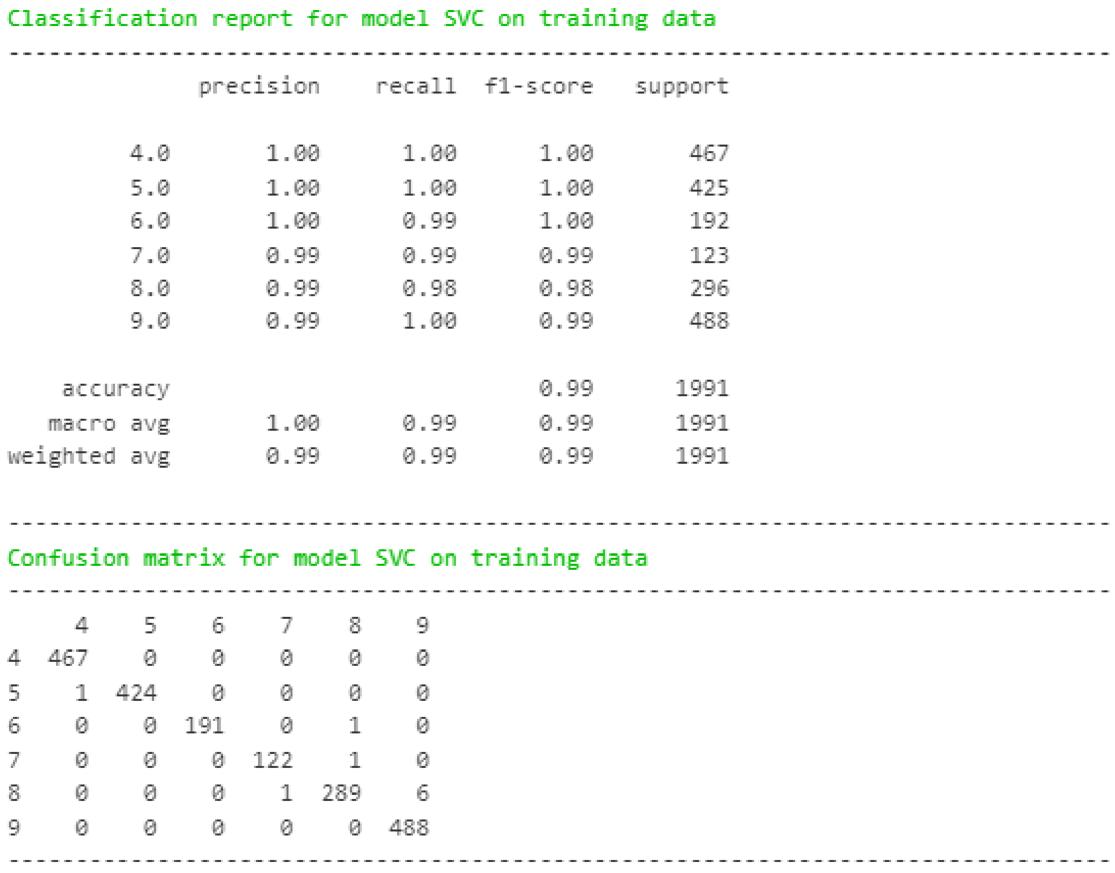
KLASIFIKACIJA-KNN



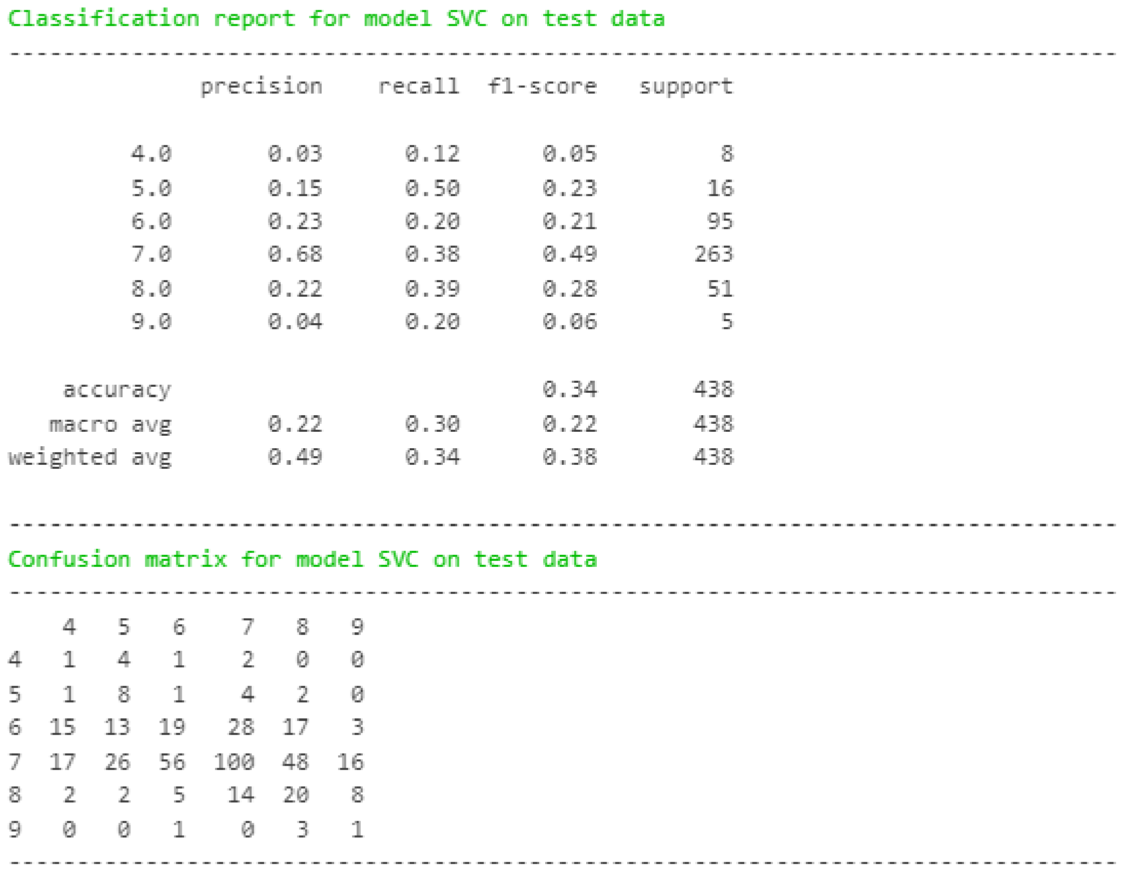




KLASIFIKACIJA-SVM

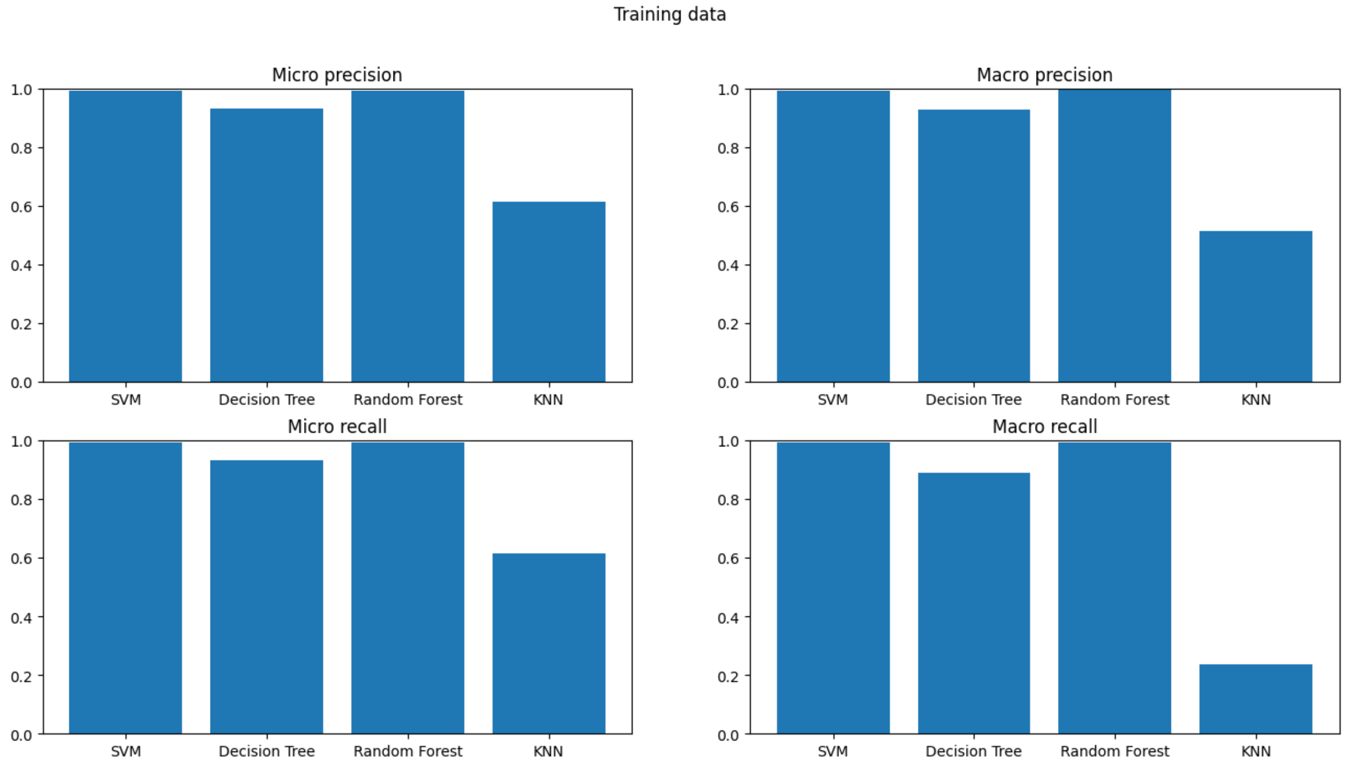


```
Classification report for model SVC on training data
--------------------------------------------------------------------------------
              precision    recall  f1-score   support

         4.0       1.00      1.00      1.00       467
         5.0       1.00      1.00      1.00       425
         6.0       1.00      0.99      1.00       192
         7.0       0.99      0.99      0.99       123
         8.0       0.99      0.98      0.98       296
         9.0       0.99      1.00      0.99       488

    accuracy                           0.99      1991
   macro avg       1.00      0.99      0.99      1991
weighted avg       0.99      0.99      0.99      1991

--------------------------------------------------------------------------------
Confusion matrix for model SVC on training data
--------------------------------------------------------------------------------
     4    5    6    7    8    9
4  467    0    0    0    0    0
5    1  424    0    0    0    0
6    0    0  191    0    1    0
7    0    0    0  122    1    0
8    0    0    0    1  289    6
9    0    0    0    0    0  488
--------------------------------------------------------------------------------
```

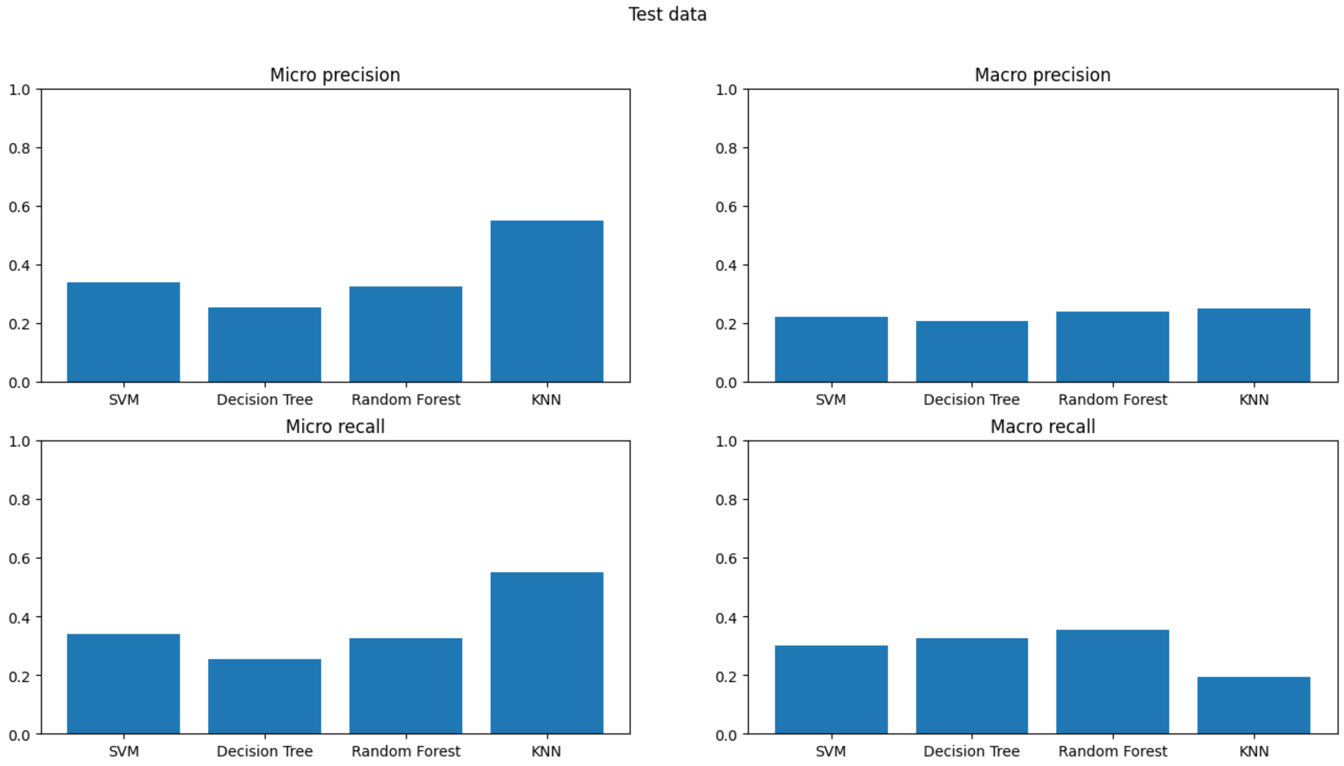


```
Classification report for model SVC on test data
--------------------------------------------------------------------------------
              precision    recall  f1-score   support

         4.0       0.03      0.12      0.05         8
         5.0       0.15      0.50      0.23        16
         6.0       0.23      0.20      0.21        95
         7.0       0.68      0.38      0.49       263
         8.0       0.22      0.39      0.28        51
         9.0       0.04      0.20      0.06         5

    accuracy                           0.34       438
   macro avg       0.22      0.30      0.22       438
weighted avg       0.49      0.34      0.38       438

--------------------------------------------------------------------------------
Confusion matrix for model SVC on test data
--------------------------------------------------------------------------------
    4   5   6    7   8   9
4   1   4   1    2   0   0
5   1   8   1    4   2   0
6  15  13  19   28  17   3
7  17  26  56  100  48  16
8   2   2   5   14  20   8
9   0   0   1    0   3   1
--------------------------------------------------------------------------------
```

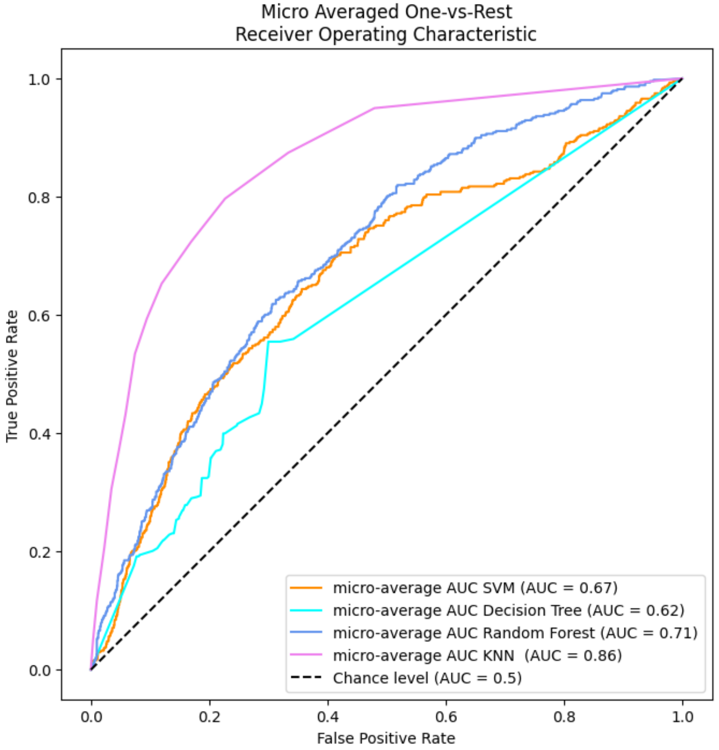
KLASIFIKACIJA

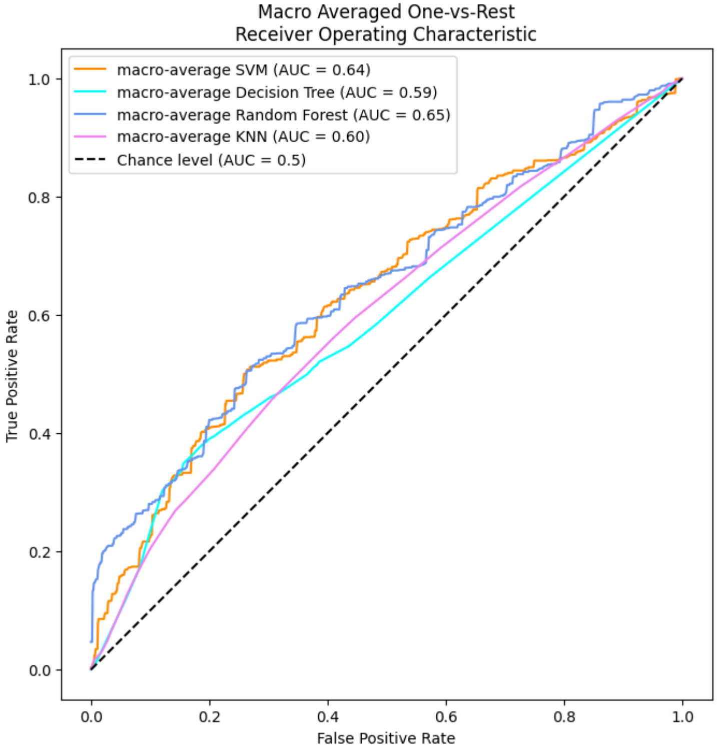


KLASIFIKACIJA

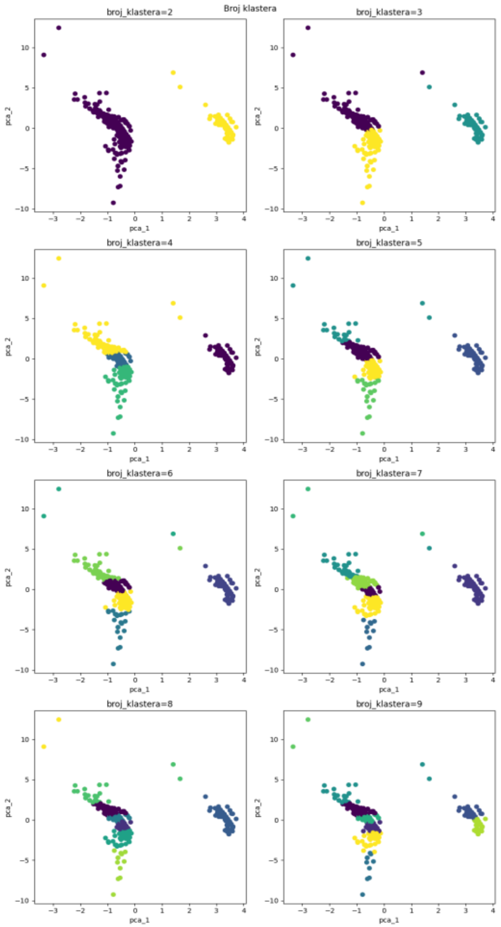


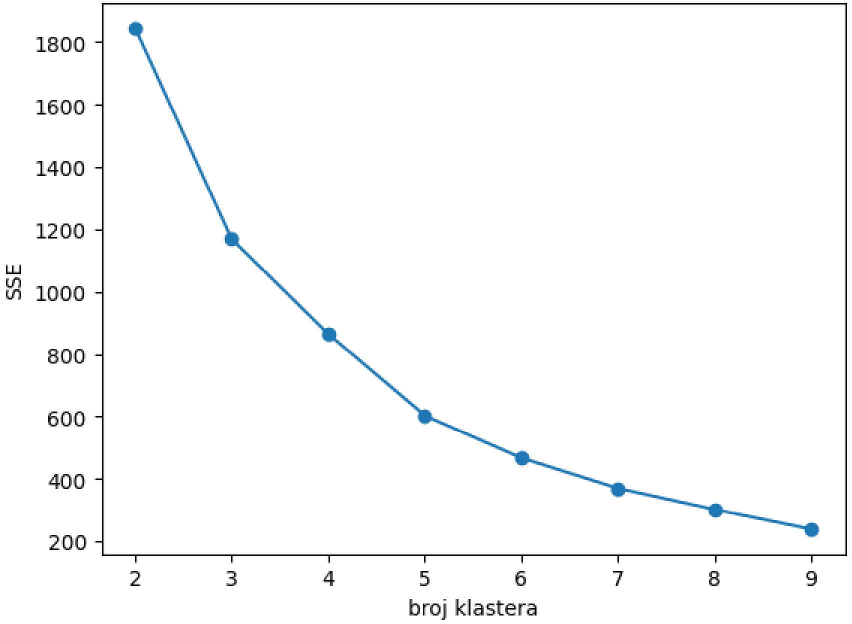
KLASIFIKACIJA

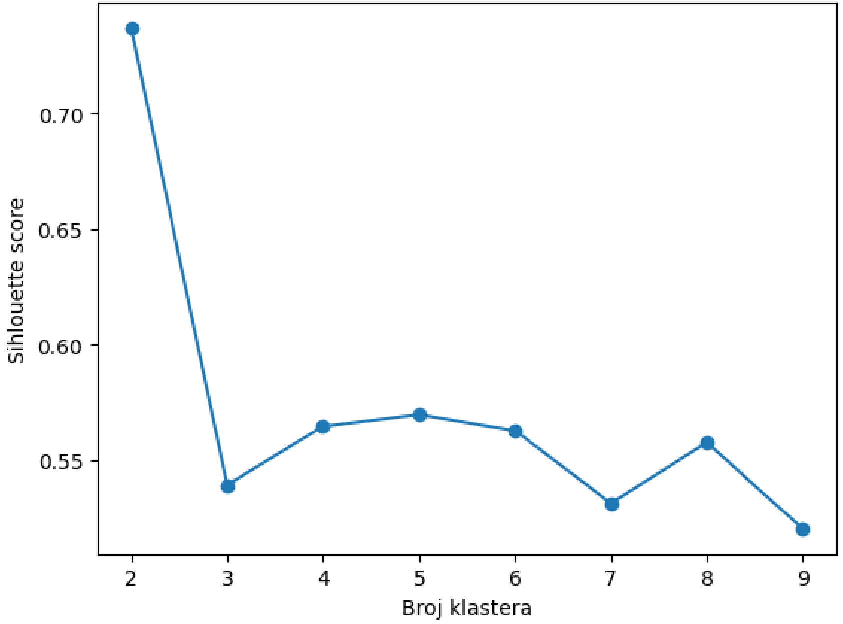




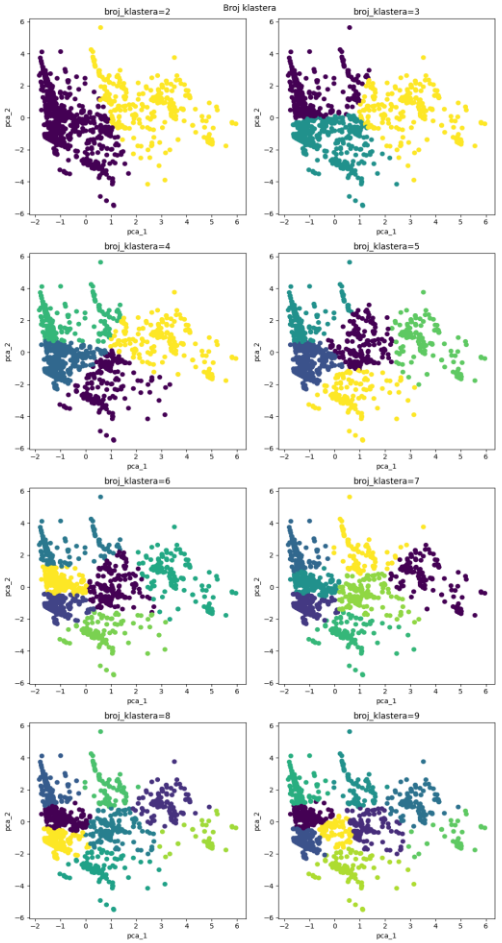
KLASTEROVANJE-KMEANS

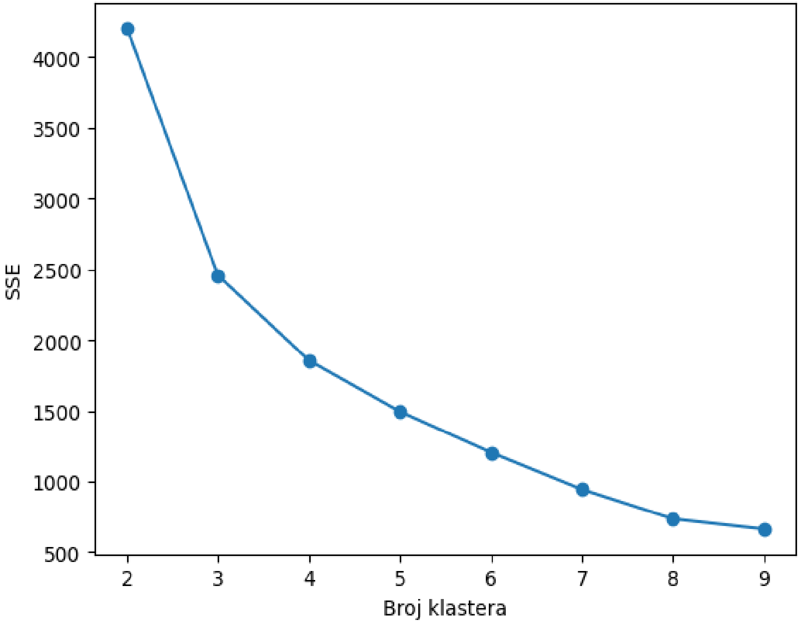


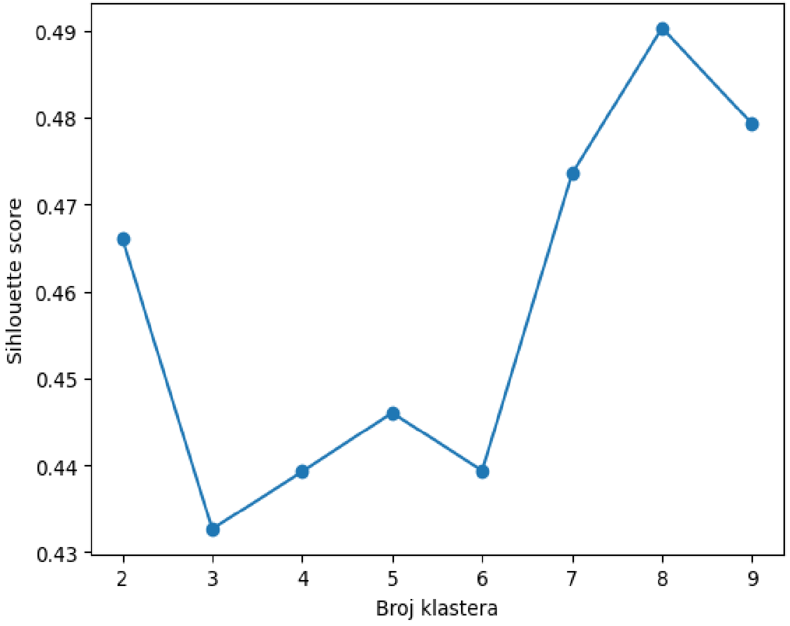




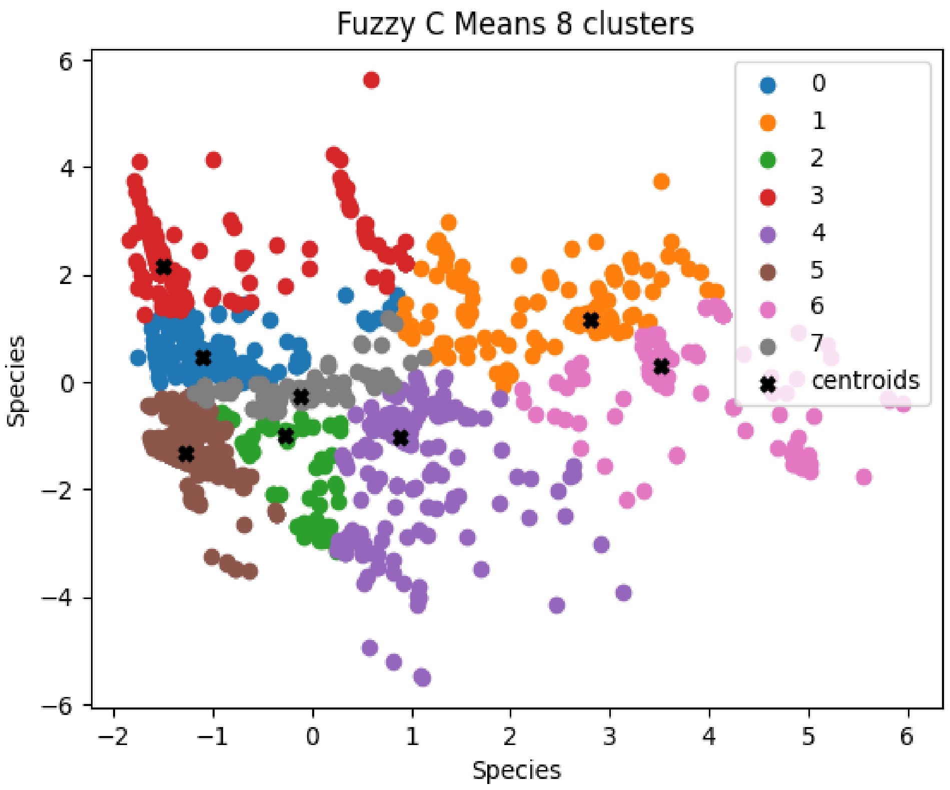
KLASTEROVANJE-KMEANS

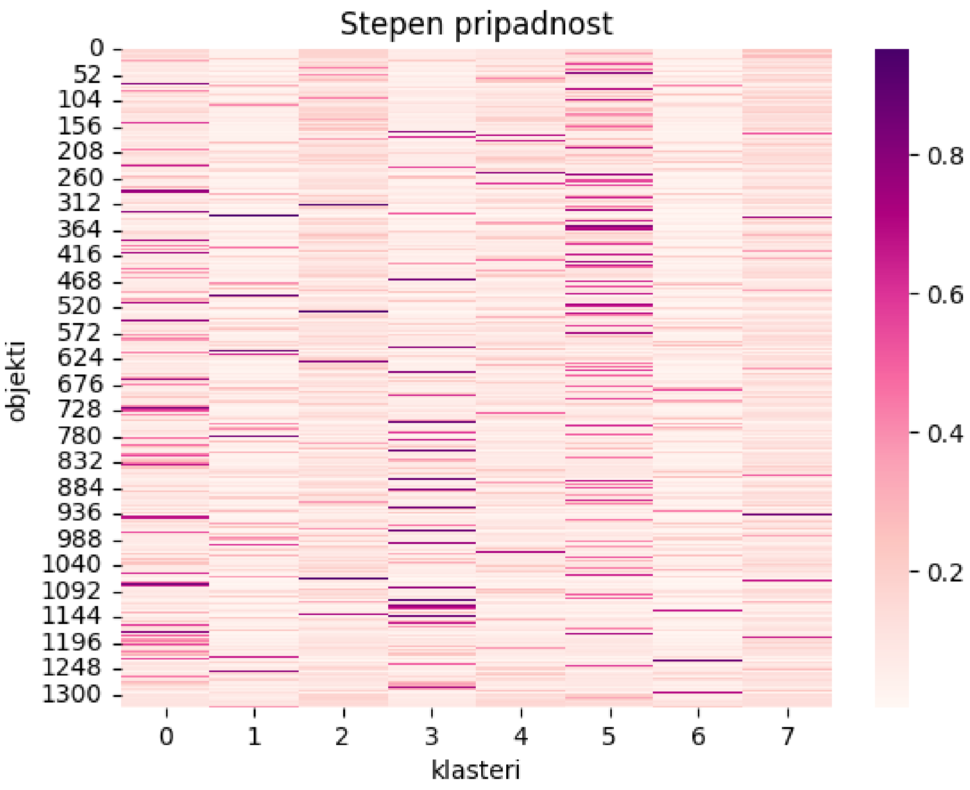




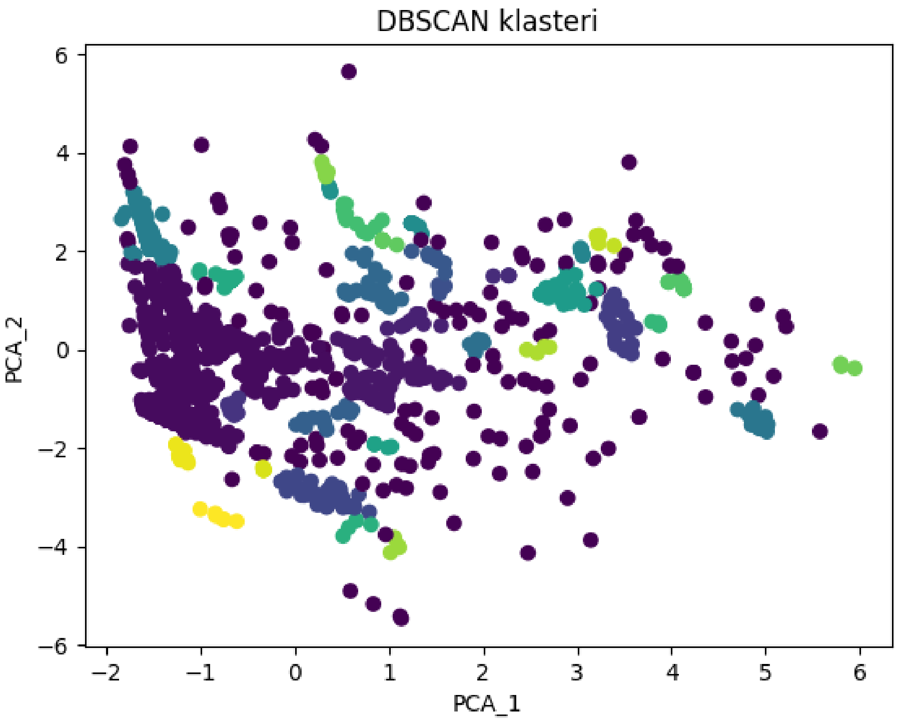


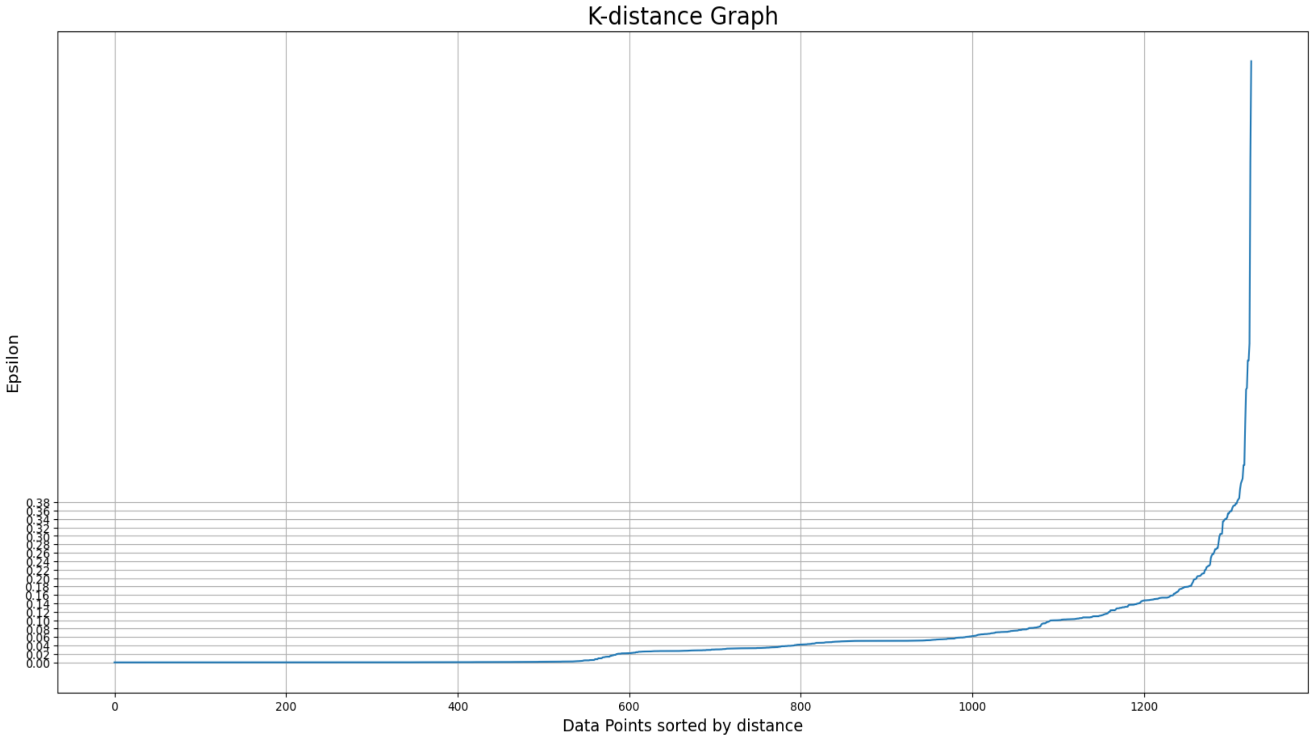
KLASTEROVANJE-FUZZY CMEANS



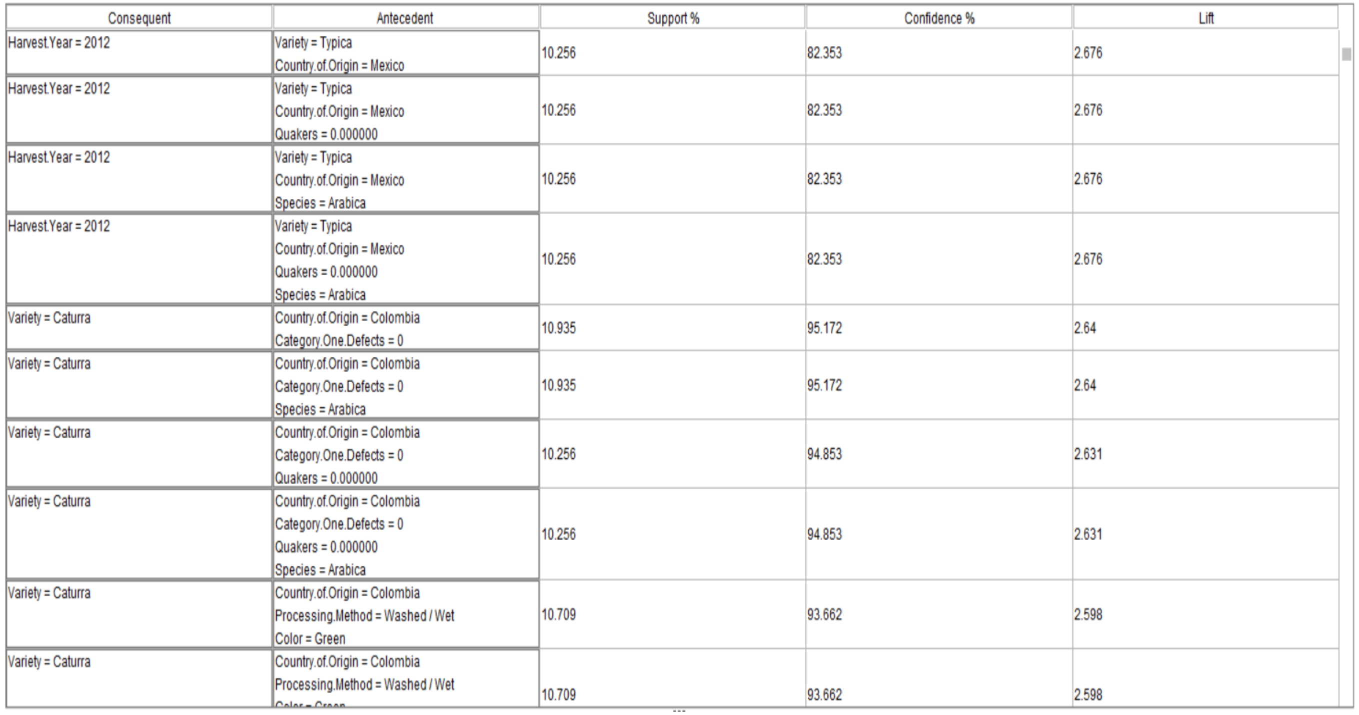


KLASTEROVANJE-DBSCAN





PRAVILA PRIDRUŽIVANJA-APRIORI



Consequent	Antecedent	Support %	Confidence %	Lift

Harvest.Year = 2012	Variety = Typica
	Country.of.Origin = Mexico
		10.256	82.353	2.676

Harvest.Year = 2012	Variety = Typica
	Country.of.Origin = Mexico
	Quakers = 0.000000
		10.256	82.353	2.676

Harvest.Year = 2012	Variety = Typica
	Country.of.Origin = Mexico
	Species = Arabica
		10.256	82.353	2.676

Harvest.Year = 2012	Variety = Typica
	Country.of.Origin = Mexico
	Quakers = 0.000000
	Species = Arabica
		10.256	82.353	2.676

Variety = Caturra	Country.of.Origin = Colombia
	Category.One.Defects = 0
		10.935	95.172	2.64

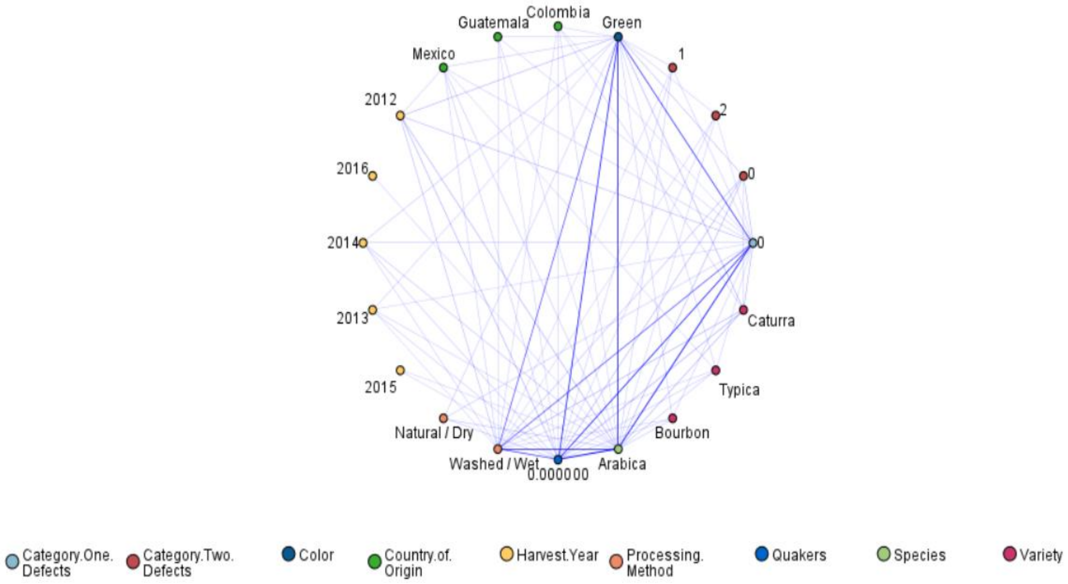
Variety = Caturra	Country.of.Origin = Colombia
	Category.One.Defects = 0
	Species = Arabica
		10.935	95.172	2.64

Variety = Caturra	Country.of.Origin = Colombia
	Category.One.Defects = 0
	Quakers = 0.000000
		10.256	94.853	2.631

Variety = Caturra	Country.of.Origin = Colombia
	Category.One.Defects = 0
	Quakers = 0.000000
	Species = Arabica
		10.256	94.853	2.631

Variety = Caturra	Country.of.Origin = Colombia
	Processing.Method = Washed / Wet
	Color = Green
		10.709	93.662	2.598

Variety = Caturra	Country.of.Origin = Colombia
	Processing.Method = Washed / Wet
	Color = Green
		10.709	93.662	2.598



ZAKLJUČAK

- Poteškoće i izazovi
- Skup podataka ili tehnika?
- Pitanja?