Analiza skupa podataka Coffee Quality database from CQI

ISTRAŽIVANJE PODATAKA 1
MATEMATIČKI FAKULTET

Anja Cvetković | mi20127@alas.matf.bg.ac.rs | Avgust 2023

Uvod

Projekat se bazira na skupu podataka “Coffee Quality database from CQI” koji se može naci na sledećem linku. U izvornom stanju skup sadrzi 44 atributa i 1339 instanci i prikupljeni su od strane “Coffee Quality Institue” u Januru, 2018.

Cilj projekta jeste analiza podataka i primena osnovnih tehnika mašinskog učenja radi procene kvaliteta zrna kafe.

ANALIZA PODATAKA I PRETPROCESIRANJE

Skup sadrži ocene kafe kao i metapodatke o samom zrnu i poreklu istog.

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

Metapodaci o zrnu

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

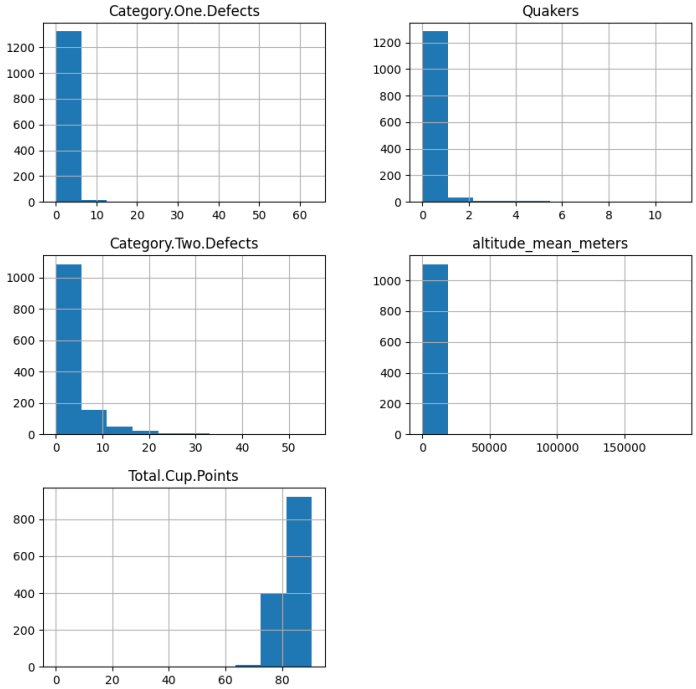
Metapodaci od poreklu

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

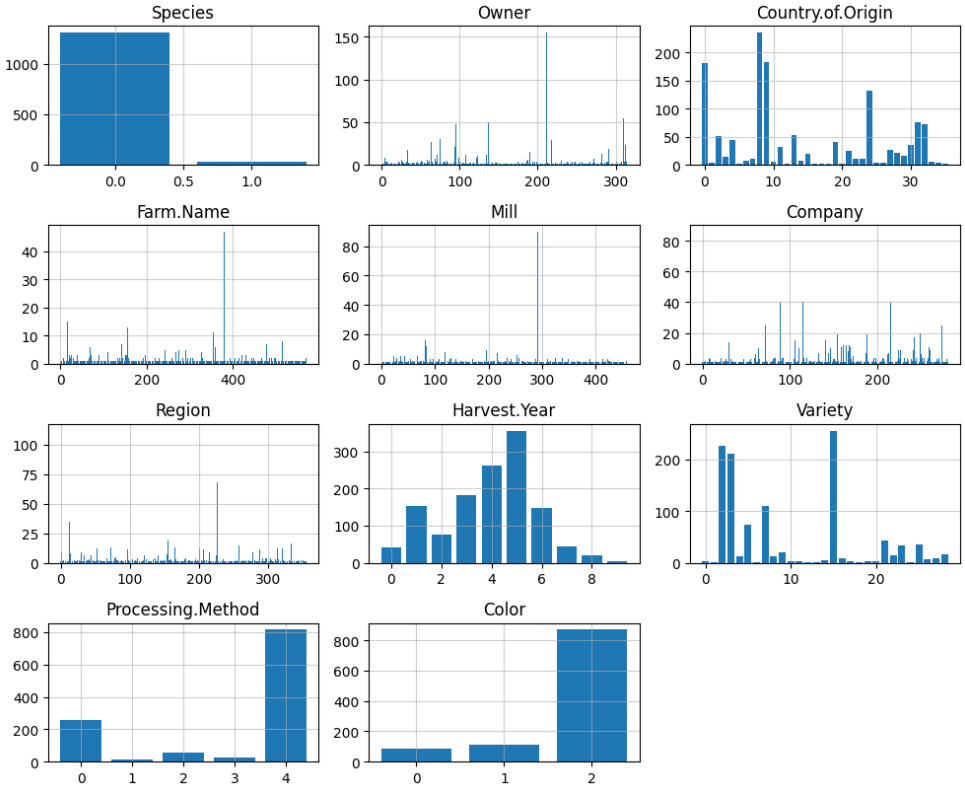
Javlja se jos manje opisnih atributa poput Certification, Certification Address, Expiration i slični, koji nisu od velikog značaja.

Za dalji rad sa podacima izbacujemo sve kolone vezane za mere kvaliteta osim ciljne promenljive Total Cup Points.

Dodatno vršimo izbacivanje atributa Owner, Mill, Company i Farm Name, što su kategorički atributi sa velikim brojem kategorija a malo značajnih informacija. Atribut Region je izbačen na osnovu velikog broja neistinitih podataka i velikog broja kategorija – oslonićemo se na pretpostavku da su podaci iz kolone Country of Origin pretežno tačni.

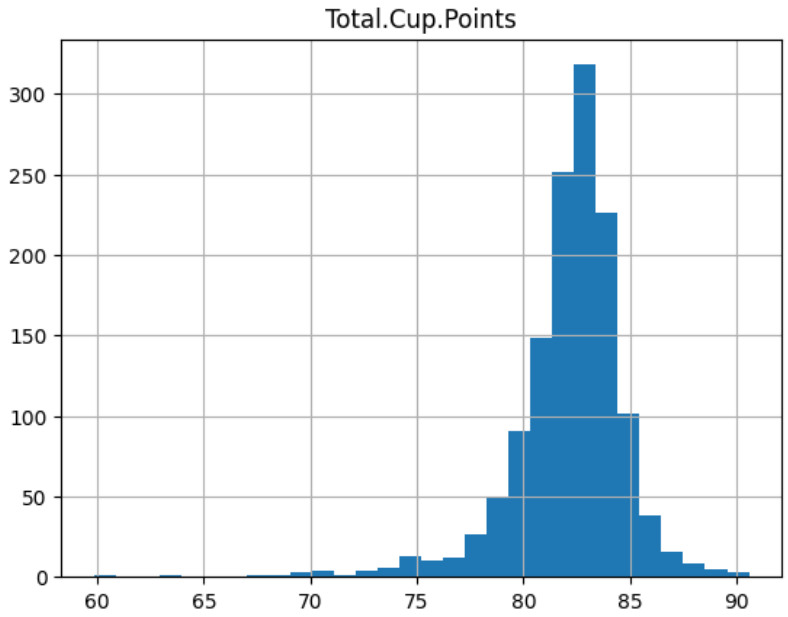


1 Pregled numeričnih atributa



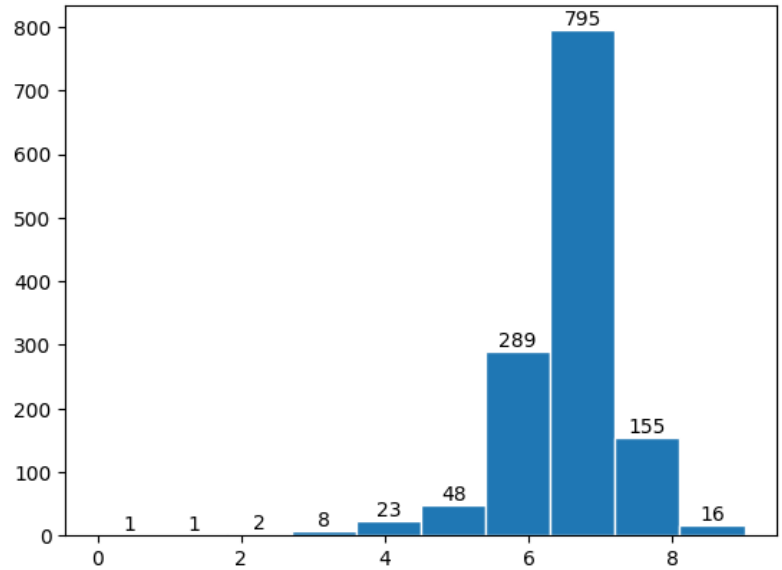
2 Pregled kategoričkih atributa

Ciljna promenljiva Total Cup Points nije celobrojnog tipa, te je za potrebe klasifikacije neophodna diskretizacija. Takođe se javlja izražen disbalans vrednosti, najviše ima ocean iz interval [80,85] što je značajno mali interval u odnosu na ceo domen ocena.



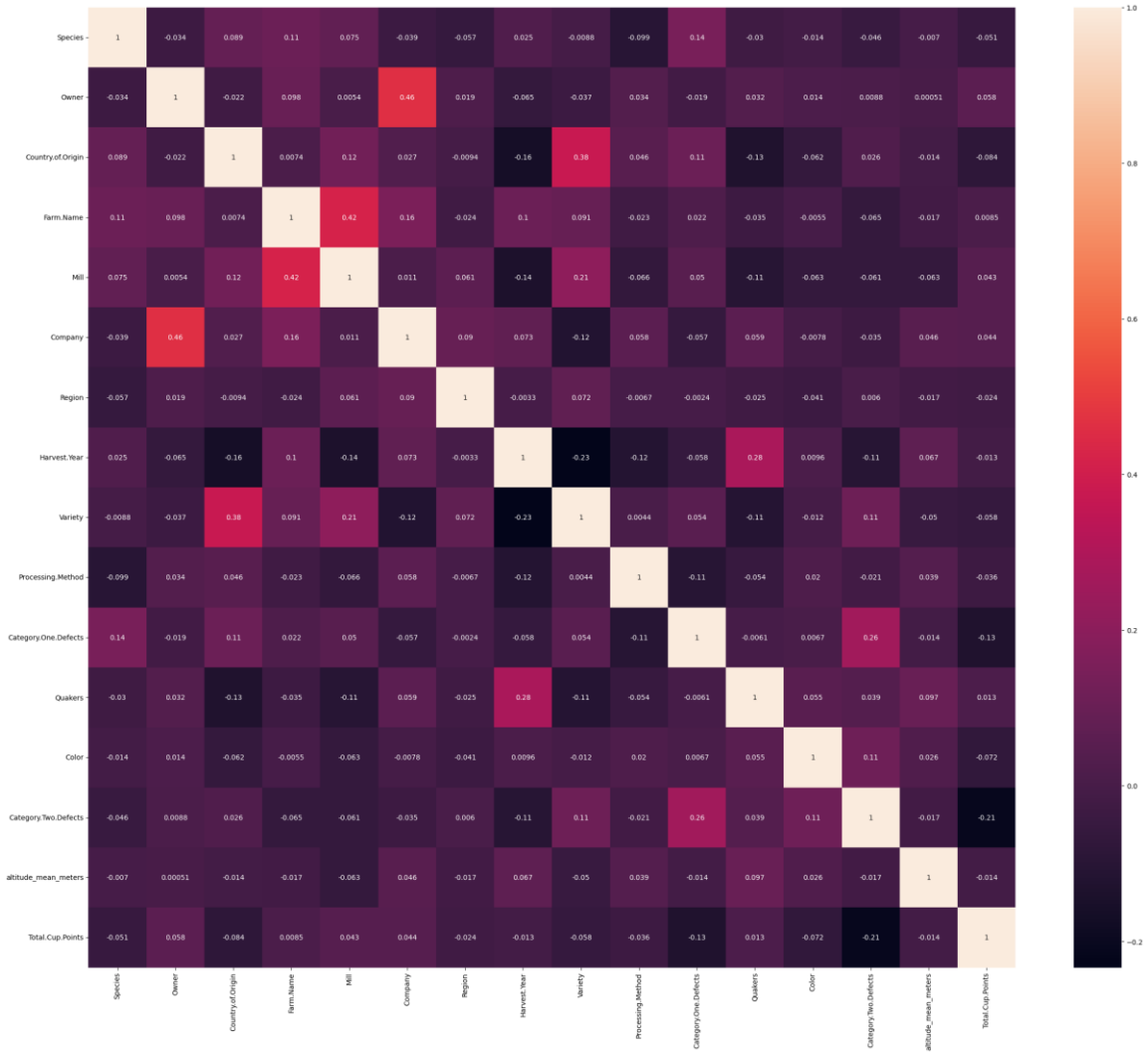
3 Pregled raspodele ciljne promenljive

Diskretizaciju vršimo pomoću *KBinsDiscretizer* – diskretizacija se vrši na osnovu jednake širine interval za koju je uzeta vrednost 10.



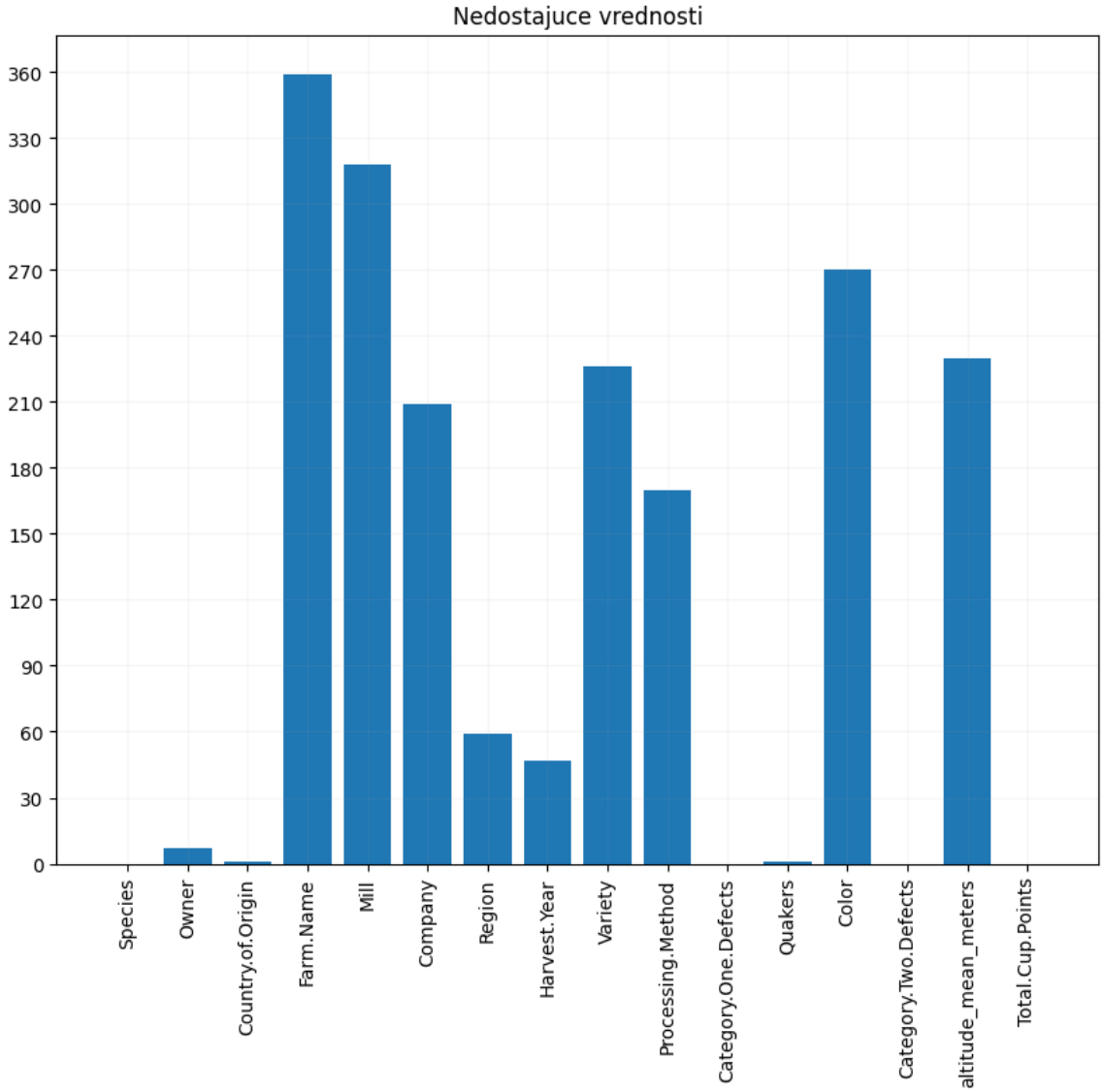
4 Diskretizacija ciljne promenljive

Kako se javlja veliki broj kategorija u kojima se nalazi svega par instanci, izbacujemo one kategorije u kojima se javlja manje od 10 instanci i tada se ciljna promenljiva svela na kategorije [4, 5, 6, 7, 8, 9].



5 Matrica korelacije

Nema izraženih korelacija ni sa ciljnom promenljivom, ni medju atributima tako da ne možemo vršiti dodatna izbacivanja kolona (ali primetimo da su vrednosti u matrici za atribute Mill, Farm Name, Owner i Company sa Total Cup Points vrlo bliske nuli tako da nismo ništa oduzeli od klasifikacije).

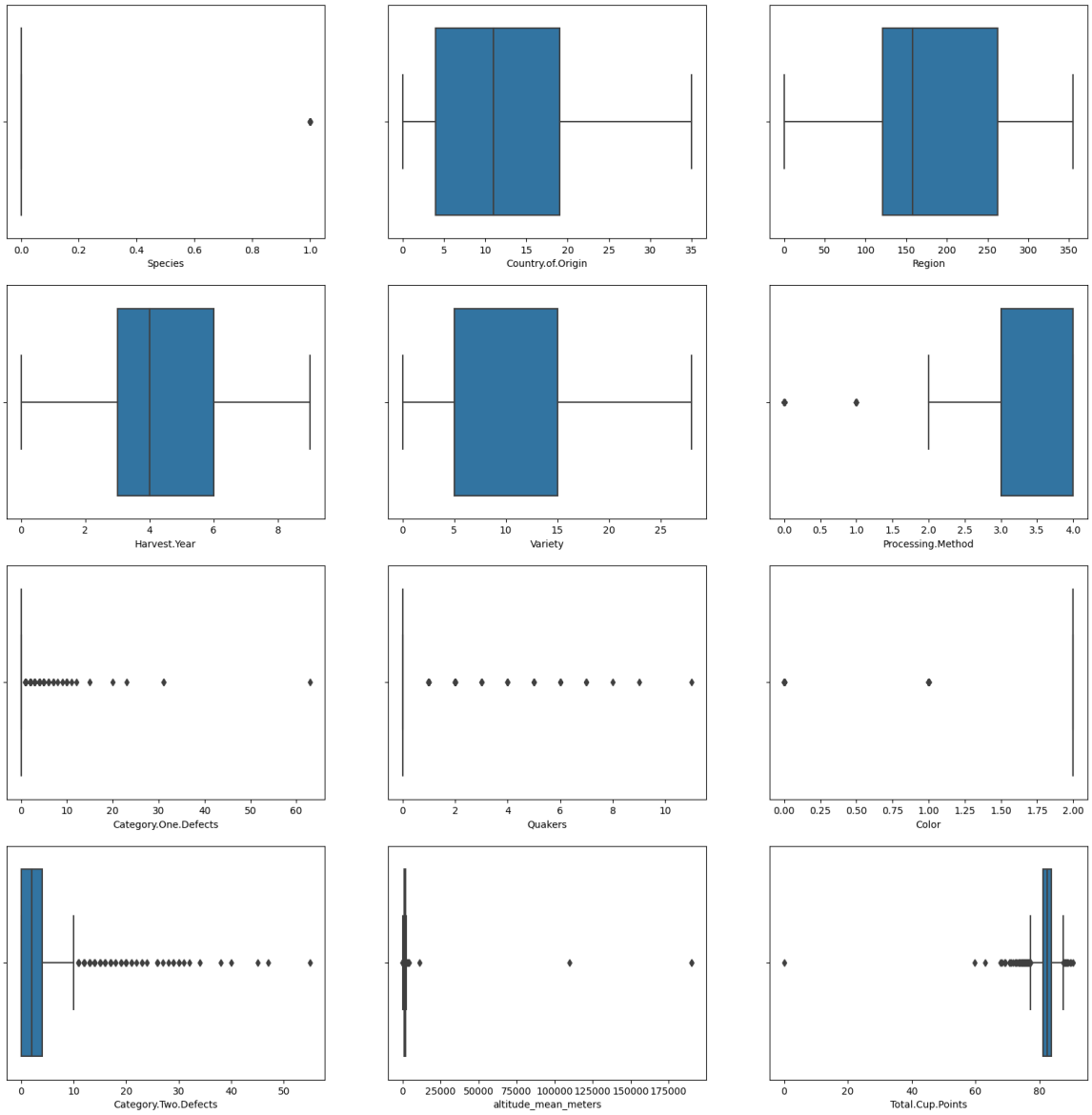


6 Broj nedostajućih vrednosti po kolonama

Nema previse nedostajućih vrednosti u kolonama koje su od interesa – najviše u koloni Color sa 270 nedostajućih vrednosti.

Po pitanju numeričkih atributa, NaN vrednosti popunjene su sa srednjom vrednošću, što ima smisla jer su u pitanju samo kolone Quakers i Altitude Mean Meters, s tim što je ta srednja vrednost zaokružena za Quakers.

Vrednosti kategoričkih atributa su pre svega konvertovane u numeričke primenom *LabelEncoder-a*, a zatim su NaN vrednosti popunjene sa vrednošću koja se najčešće pojavljuje respektivno po koloni.



7 Elemtni van granica

Nema značajan broj elemenata van granica, s obzirom na prirodu podataka – većinom kategoričke promenljive. Doduše, pronađena je jedna instance čija je vrednost ciljne promenljive 0 – ceni se da je anomalija.

Klasifikacija

STABLA ODLUČIVANJA

Stabla odlučivanja (decision trees) su neparametarski metod nadgledanog učenja koji se koriste za klasifikaciju i regresiju. Hijerarhijske su strukture, gde koreni i unutrašnji čvorovi predstavljaju upite koji teže da razdvoje skup podataka na distikntne podskupove. U listovima se javljaju svi mogući rezulati klasifikacije.

Prednosti ovakvog pristupa su u tome sto su drvolike strukture jednostavne za razumevanje i interpretaciju, pa i laku vizuelizaciju. Nije neophodno preveliko pretprocesiranje podataka,a mogu da rade is a numeričkim i kategoričkim atributima (implementacija u scikit learn modulu podržava samo numeričke) i relativno su robusni na outlier-e i nedostajuće vrednosti.

Jedna od mana primene ove metode je što su performanse značajno lose na veoma nebalansiranim skupovima kao što je ovaj.

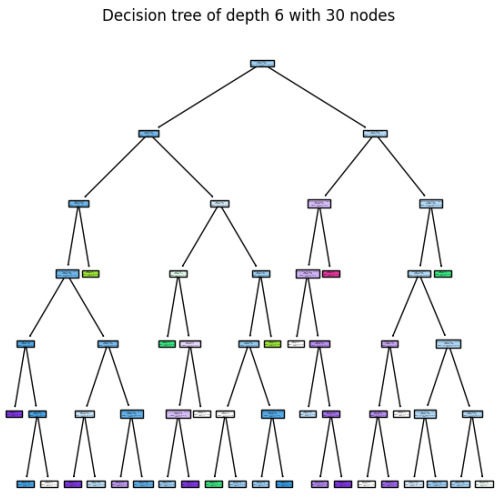
```
Classification report for model DecisionTreeClassifier on training data
Accuracy: 0.6463963963963963
--------------------------------------------------------------------------------
                  pre       rec       spe        f1       geo       iba       sup

        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
```

8 Rezultat primene stabla odlučivanja sa nasumičnim hiperparametrima na nebalansiranom skupu – trening podaci

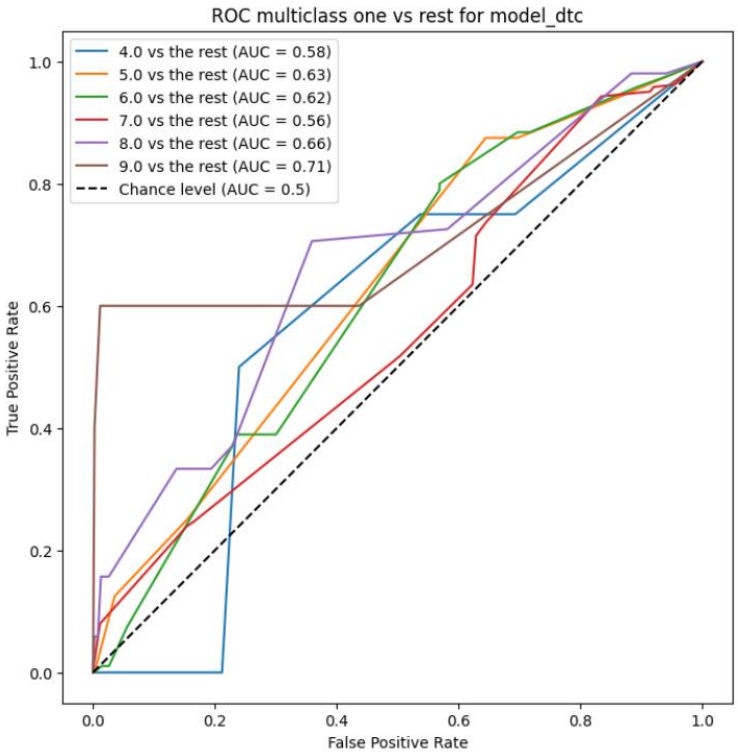
```
--------------------------------------------------------------------------------
Classification report for model DecisionTreeClassifier on test data
Accuracy: 0.6050228310502284
--------------------------------------------------------------------------------
                  pre       rec       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
--------------------------------------------------------------------------------
```

9 Rezultat primene stabla odlučivanja sa nasumičnim hiperparametrima na nebalansiranom skupu - test podaci

Kao što je i predpostavljeno, nedominantne klase se ni ne uzimaju u obzir.



10 Grafički prikaz modela



11 ROC kriva nasumičnog stabla odličivanja na nebalansiranim podacima

Veštački ćemo balansirati skup primenom SMOTEENN tehnike. Bazira se na kombinaciji *over-sampling* i *under-sampling* metoda. Ovu tehniku primenjujemo na prethodno normalizovane podatke.

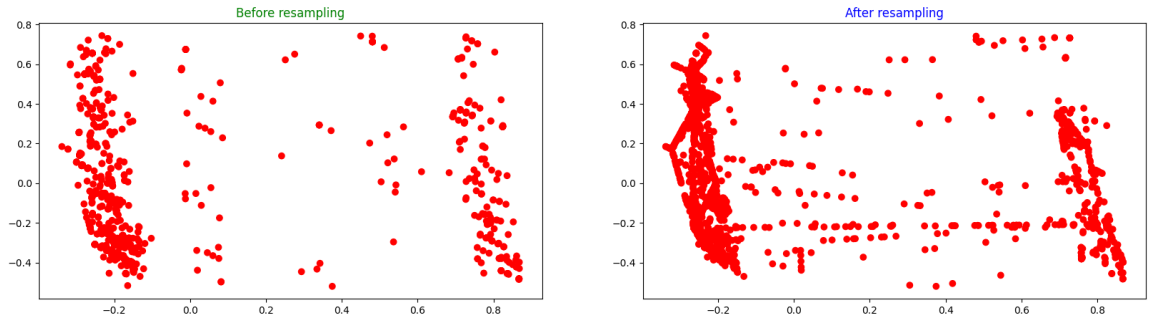
Edited Nearest Neighbor (ENN) radi tako što pronalazi k najbližih suseda svake opservacije, zatim proverava da li je većinska klasa u krugu k najbližih suseda opservacije ista kao i klasa opservacije – ako nije, k najbližih suseda i sama opservacija se brišu iz skupa. Ovaj postupak brisanja se ponavlja dok se ne zadovolji tražena proporcija među klasama. Očigledno, ENN se koristi za under-sampling.

SMOTE je over-sampling metoda i funkcioniše tako što se slučajni uzorak iz manjinske klase izabere, zatim se traži k najbližih suseda tog uzorka,i od tih k suseda se bira jedan i na duži izmedju njega i uzorka se generiše novi sintetički podatak.

Kako SMOTEENN radi? Nakon primene SMOTE algoritma na podatke, primenjuje se ENN. Kako primena SMOTE algoritma unosi šum u podatke interpolacijom novih tačaka sa marginalnim outlajerima i inlajerima, ENN čisti podatke, te odstranjuje šum.

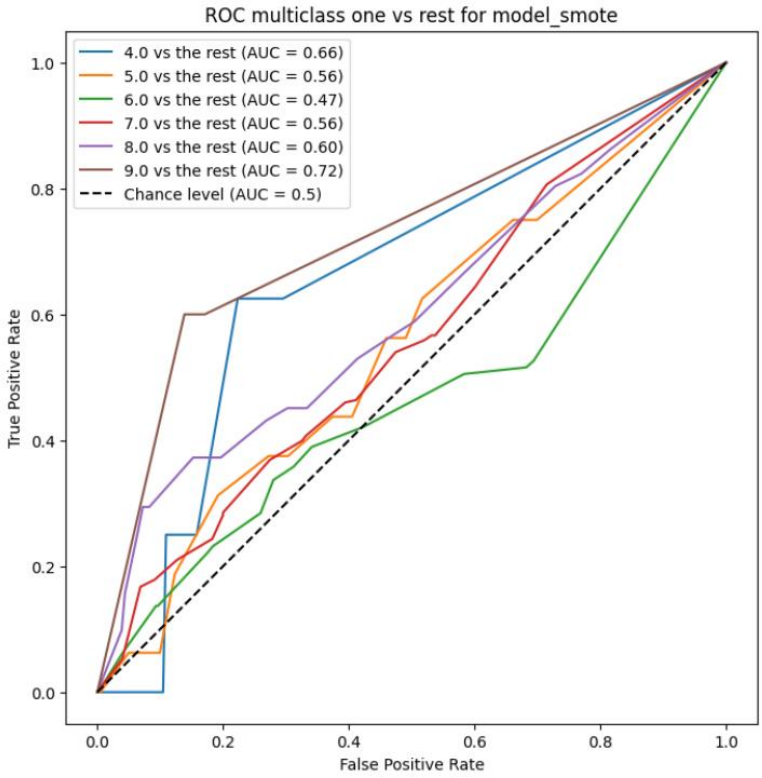
Primenom SMOTEENN-a nismo dobili savršen balans među klasama, ali dizbalans je dovoljno neprimetan.

Razlog zbog kojeg smo primenili ovu kombinovanu tehniku jeste to što je dizbalans prevelik i over-sampling tehnikom će se dobiti previše veštačkih instanci što za posledicu ima preprilagođavanje, a under-sampling stvara preveliki gubitak informacija.




```
Number of instances before resampling with SMOTEENN : dict_items([(8.0, 104), (7.0, 532), (6.0, 194), (4.0, 15), (9.0, 11), (5.0, 32)]).
Number of instances after  resampling with SMOTEENN : dict_items([(4.0, 467), (5.0, 425), (6.0, 192), (7.0, 123), (8.0, 296), (9.0, 488)]).
```

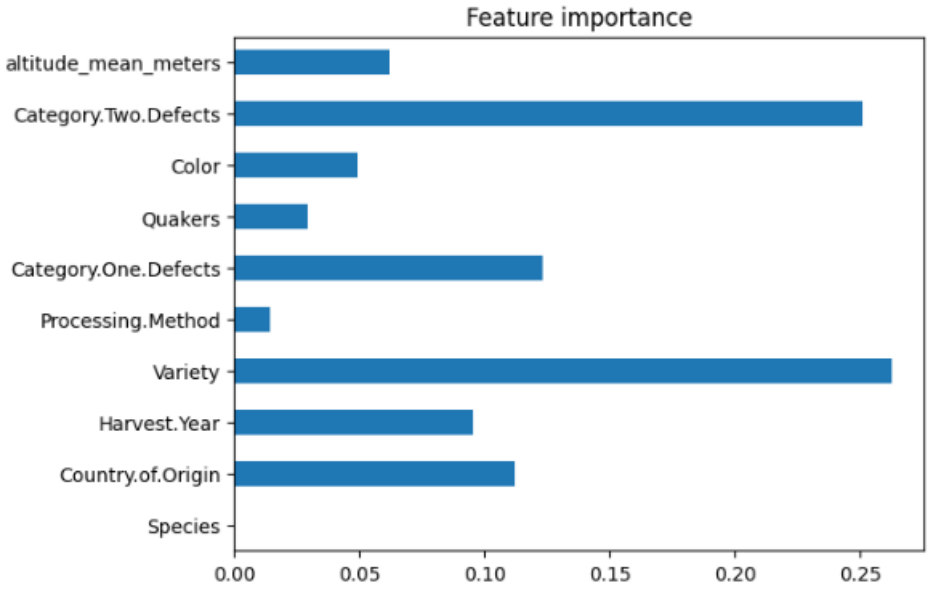
12 Skup nakon primene SMOTEENN algortima



```
Classification report for model DecisionTreeClassifier on training data
Accuracy: 0.8422903063787042
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

        4.0       0.94      0.97      0.98      0.96      0.98      0.95       467
        5.0       0.83      0.84      0.95      0.84      0.90      0.80       425
        6.0       0.62      0.60      0.96      0.61      0.76      0.56       192
        7.0       0.87      0.54      0.99      0.66      0.73      0.51       123
        8.0       0.73      0.67      0.96      0.70      0.80      0.62       296
        9.0       0.89      1.00      0.96      0.94      0.98      0.96       488

avg / total       0.84      0.84      0.97      0.84      0.90      0.81      1991

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on training data
--------------------------------------------------------------------------------
     4    5    6   7    8    9
4  452    9    3   1    2    0
5    2  359   32   0   27    5
6   15    7  115   1   31   23
7    2   12   17  66   14   12
8    8   43   17   7  198   23
9    0    0    0   1    0  487
--------------------------------------------------------------------------------
Classification report for model DecisionTreeClassifier on test data
Accuracy: 0.23515981735159816
--------------------------------------------------------------------------------
                   pre       rec       spe        f1       geo       iba       sup

        4.0       0.00      0.00      0.92      1.00      0.00      0.00         8
        5.0       0.06      0.31      0.81      0.10      0.50      0.24        16
        6.0       0.25      0.22      0.82      0.24      0.43      0.17        95
        7.0       0.71      0.21      0.87      0.32      0.42      0.17       263
        8.0       0.21      0.39      0.80      0.27      0.56      0.30        51
        9.0       0.05      0.60      0.86      0.09      0.72      0.50         5

avg / total       0.51      0.24      0.85      0.30      0.44      0.19       438

--------------------------------------------------------------------------------
Confusion matrix for model DecisionTreeClassifier on test data
--------------------------------------------------------------------------------
    4   5   6   7   8   9
4   0   6   2   0   0   0
5   3   5   3   0   3   2
6  15  22  21  13  16   8
7  17  45  49  54  55  43
8   0   7   8   9  20   7
9   0   0   0   0   2   3
--------------------------------------------------------------------------------
```

13 Rezultati primene stabla odlučivanja na balansirane podatke sa nasumičnim hiperparametrima

Još jedna od prednosti korišćenja ovog modela jeste što možemo dobiti uvid u značajnost svakog od atributa prilikom pravljenja modela.



Ove značajnosti atributa imaju smisla, jer na primer u koloni Species imamo samo dve moguće vrednosti – Arabica i Robusta, od kojih je Arabica dominant, dok Robusta ima tek 28 instanci. Slično, Processing Methods i Color su kategorički atributi sa izraženom dominacijom jedne kategorije.

Na osnovu rezultata nasumičnog modela na balansiranim podacima vidimo da nemamo neko znatno poboljšanje, te pribegavamo podešavanju hiper-parametara korišenjem *cross-validation* tehnike.

```
params = {'criterion': ['gini', 'entropy'],
          'max_depth': [2,4,6,8],
          'min_samples_split' : [2,5,8,10]
         }
```

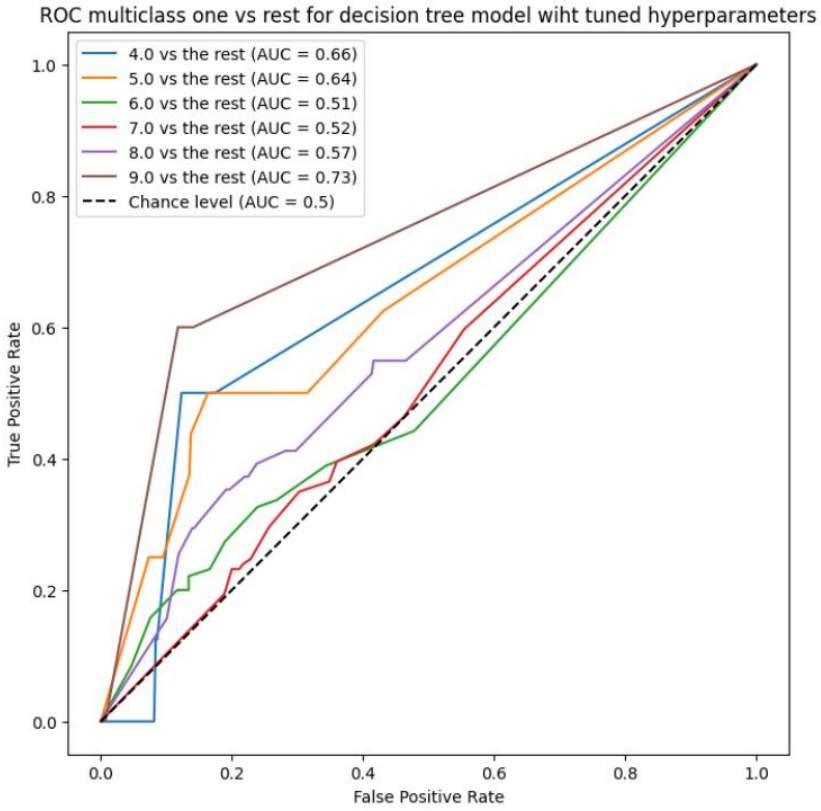
14 Izabrane moguće vrednosti hiper-parametara

GridSearchCV je alat korišćen za iscrpnu pretragu nad mrežnom hiper-parametara. Zasniva se na korišćenju cross-validacije za ocenu performansi modela. Cross-validacija štiti od preprilagođavanja u prediktivnim modelima, naročito kad je skup podataka limitiran – kao ovaj.

```
{'criterion': 'gini', 'max_depth': 8, 'min_samples_split': 2}
0.8734222490900618
```

15 Izabrani hiper-parametri modela sa najboljim performansama i njegova ocena

Šta predstavlja ocena modela? Trening skup se deli na k podskupova (u našem slučaju je to 5) i model uči korišćenjem k-1 podskupova, dok se k-ti koristi kao test i process se ponavlja nad svim mogućim kombinacija trening i test podskupovima. Na kraju se uzima srednja vrednost tačnosti svake iteracije i ona predstavlja ocenu modela sa datim hiper-parametrima (s tim da se mogu koristiti i druge metrike za procenu).



```
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

16 Rezultati modela sa podešenim hiper-parametrima

RANDOM FOREST

Random Forest je pristup nadgledanog učenja zasnovanom na tehnici ansambla. Kombinuje više stabla odlučivanja i odluku donosi na osnovu glasa većine.

Osnovna prednost je smanjena mogućnost od preprilagođavanja, kao i velika tačnost dobijenog modela.

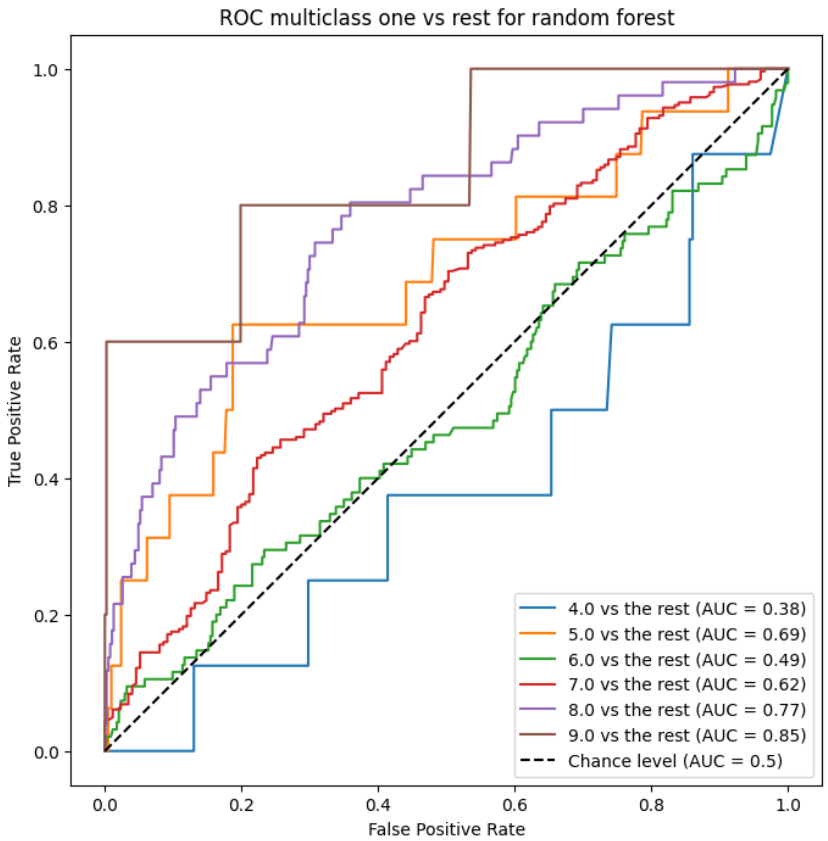
Za podešavanje hiper-parametara korišćen je ponovo GridSearchCV.

```
params_grid={ 'n_estimators': [100, 200, 300],
              'max_depth': [2, 4 , 6, 8, 10]
              }
```

17 Prosleđeni hiper-parametri

```
0.9522814574123751
{'max_depth': 10, 'n_estimators': 200}
```

18 Izabrani hiper-parametri



```
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
--------------------------------------------------------------------------------
```

19 Rezultati Random Forest modela

Iako na osnovu matrica konfuzija izgleda kao da se model preprilagodio, na osnovu ROC krive vidimo da ovaj model radi značajno bolje u odnosu na druge modele zasnovane na stablima odlučivanja.

K-NEAREST NEIGHBOURS

Ovaj model se zasniva na bliskostima između instance i njenih k suseda, a bliskosti su najjčešće izražene kao metrike rastojanja, npr. Euklidsko.

Kako se algoritam zasniva na rastojanjima, za bolje performanse se vrši standardizacija podataka, jer ako su podaci na sličnim skalama manje su šanse da se javi sklonost ka nekim atributima. Ovde je izvršena normalizacija podataka korišćenjem *MinMaxScaler*-a.

Još jedan problem koji se javlja, konkretno kod ovog skupa podataka je veliki broj kategoričkih atributa. Naime, imamo atribut Colors sa mogućim vrednostima koje su numerisane kao 1, 2 ili 3. Ovo nije atribut ordinalne vrste, boje ne možemo poredjati u red, tj reći boja 3 je dalja od boje 1 nego boja 2.

Biće isprobana dva pristupa – bez *OneHotEncoding*-a radi sveobuhvatnog poređenja sa ostalim modelima i sa *OneHotEncoding*-om uz korišćenje *PCA* radi smanjenja dimenzionalnosti.

	lower	min	num_lower	upper	max	num_upper	percantage
Species	0.000000	0	0	0.000000	1	28	2
Country.of.Origin	-21.875000	0	0	45.125000	35	0	0
Harvest.Year	-1.500000	0	0	10.500000	9	0	0
Variety	-10.000000	0	0	30.000000	28	0	0
Processing.Method	1.500000	0	281	5.500000	4	0	21
Category.One.Defects	0.000000	0	0	0.000000	63	199	15
Quakers	0.000000	0	0	0.000000	11	93	7
Color	2.000000	0	198	2.000000	2	0	15
Category.Two.Defects	-6.000000	0	0	10.000000	47	90	7
altitude_mean_meters	337.454183	1	44	2637.576362	190164	18	5

20 Outlajeri u skupu

Pre svega analizirajmo outlajere na osnovu interkvartilnog rastojanja. Maksimalni procenat outlajera je u koloni Processing Method, koja je kategoričkog tipa i ima jednu dominantu kategoriju. Izbacivanje outlajera ne rešava problem jer onda cela kolona postaje beskorisna, ali kako ih nema previše – samo 21%, outlajere nećemo sanirati.

Za pristup bez OneHotEncoding-a koristićemo MinMaxScaler za skaliranje podataka.

Napravićemo par modela – nasumičan KNN model, model sa podešenim hiper-parametrima i model zasnovan na ansamblu i to sa verzijom sa balansiranim i ne balansiranim podacima.

Rezultati ovih modela prikazani su u nastavku:

Classification report for model KNeighborsClassifier on training data
Accuracy: 0.6328828828828829

	pre	rec	spe	f1	geo	iba	sup
4.0	0.50	0.20	1.00	0.29	0.45	0.18	15
5.0	0.31	0.12	0.99	0.18	0.35	0.11	32
6.0	0.50	0.31	0.91	0.39	0.54	0.27	194
7.0	0.67	0.90	0.33	0.77	0.55	0.31	532
8.0	0.46	0.12	0.98	0.20	0.35	0.11	104
9.0	1.00	0.00	1.00	0.00	0.00	0.00	11
avg / total	0.60	0.63	0.58	0.58	0.50	0.27	888

Confusion matrix for model KNeighborsClassifier on training data

	4	5	6	7	8	9
4	3	0	2	9	1	0
5	0	4	8	19	1	0
6	3	4	61	121	5	0
7	0	3	42	481	6	0
8	0	2	7	82	13	0
9	0	0	1	8	2	0

Classification report for model KNeighborsClassifier on test data
Accuracy: 0.5525114155251142

	pre	rec	spe	f1	geo	iba	sup
4.0	0.00	0.00	1.00	1.00	0.00	0.00	8
5.0	0.20	0.19	0.97	0.19	0.43	0.17	16
6.0	0.30	0.20	0.87	0.24	0.42	0.16	95
7.0	0.62	0.82	0.25	0.71	0.45	0.22	263
8.0	0.40	0.08	0.98	0.13	0.28	0.07	51
9.0	1.00	0.00	1.00	0.00	0.00	0.00	5
avg / total	0.50	0.55	0.52	0.52	0.41	0.18	438

Confusion matrix for model KNeighborsClassifier on test data

	4	5	6	7	8	9
4	0	3	2	3	0	0
5	0	3	2	11	0	0
6	2	3	19	71	0	0
7	0	5	38	216	4	0
8	0	1	2	44	4	0
9	0	0	1	2	2	0

21 KNN na nebalansiranim podacima, k=10

Classification report for model KNeighborsClassifier on training data
Accuracy: 0.6148648648648649

	pre	rec	spe	f1	geo	iba	sup
4.0	1.00	0.07	1.00	0.12	0.26	0.06	15
5.0	0.50	0.09	1.00	0.16	0.31	0.08	32
6.0	0.42	0.26	0.90	0.32	0.49	0.22	194
7.0	0.65	0.90	0.27	0.76	0.49	0.26	532
8.0	0.53	0.10	0.99	0.16	0.31	0.09	104
9.0	1.00	0.00	1.00	0.00	0.00	0.00	11
avg / total	0.59	0.61	0.54	0.55	0.45	0.22	888

Confusion matrix for model KNeighborsClassifier on training data

	4	5	6	7	8	9
4	1	0	2	11	1	0
5	0	3	6	23	0	0
6	0	2	51	139	2	0
7	0	0	47	481	4	0
8	0	1	11	82	10	0
9	0	0	4	5	2	0

Classification report for model KNeighborsClassifier on test data
Accuracy: 0.5525114155251142

	pre	rec	spe	f1	geo	iba	sup
4.0	1.00	0.00	1.00	0.00	0.00	0.00	8
5.0	0.20	0.06	0.99	0.10	0.25	0.06	16
6.0	0.26	0.21	0.84	0.23	0.42	0.17	95
7.0	0.62	0.83	0.25	0.71	0.45	0.22	263
8.0	0.43	0.06	0.99	0.10	0.24	0.05	51
9.0	1.00	0.00	1.00	0.00	0.00	0.00	5
avg / total	0.52	0.55	0.51	0.49	0.40	0.17	438

Confusion matrix for model KNeighborsClassifier on test data

	4	5	6	7	8	9
4	0	3	2	3	0	0
5	0	1	4	11	0	0
6	0	0	20	74	1	0
7	0	1	42	218	2	0
8	0	0	8	40	3	0
9	0	0	0	4	1	0

22 KNN sa podešenim hiper-parametrima na nebalansiranim podacima

```
{'algorithm': 'auto', 'n_neighbors': 15, 'p': 2, 'weights': 'uniform'}
0.5867117117117117
```

23 Izabrani hiper-parametri i CV ocena prethodnog modela

Classification report for model BaggingClassifier on training data
Accuracy: 0.7105855855855856

	pre	rec	spe	f1	geo	iba	sup
4.0	0.60	0.20	1.00	0.30	0.45	0.18	15
5.0	0.47	0.22	0.99	0.30	0.47	0.20	32
6.0	0.67	0.42	0.94	0.52	0.63	0.38	194
7.0	0.73	0.92	0.49	0.82	0.67	0.47	532
8.0	0.68	0.42	0.97	0.52	0.64	0.39	104
9.0	0.43	0.27	1.00	0.33	0.52	0.25	11
avg / total	0.70	0.71	0.68	0.68	0.65	0.43	888

Confusion matrix for model BaggingClassifier on training data

	4	5	6	7	8	9
4	3	0	1	10	1	0
5	0	7	8	17	0	0
6	2	5	82	100	5	0
7	0	2	25	492	11	2
8	0	1	6	51	44	2
9	0	0	1	3	4	3

Classification report for model BaggingClassifier on test data
Accuracy: 0.541095890410959

	pre	rec	spe	f1	geo	iba	sup
4.0	0.00	0.00	0.99	1.00	0.00	0.00	8
5.0	0.20	0.12	0.98	0.15	0.35	0.11	16
6.0	0.31	0.24	0.85	0.27	0.45	0.19	95
7.0	0.64	0.78	0.35	0.70	0.52	0.28	263
8.0	0.30	0.16	0.95	0.21	0.39	0.14	51
9.0	0.00	0.00	0.99	1.00	0.00	0.00	5
avg / total	0.49	0.54	0.57	0.54	0.47	0.23	438

Confusion matrix for model BaggingClassifier on test data

	4	5	6	7	8	9
4	0	2	2	3	1	0
5	0	2	4	9	1	0
6	4	3	23	64	1	0
7	0	2	39	204	15	3
8	0	0	6	36	8	1
9	0	1	1	2	1	0

24 KNN bagging classifier na nebalansiranim podacima

Balansiranje podataka kao i kod stabla odlučivanja radimo pomoću SMOTEENN tehnike.

Classification report for model KNeighborsClassifier on training data

	precision	recall	f1-score	support
4.0	0.95	0.99	0.97	467
5.0	0.95	0.98	0.96	425
6.0	0.88	0.88	0.88	192
7.0	0.84	0.60	0.70	123
8.0	0.94	0.88	0.91	296
9.0	0.95	0.99	0.97	488
accuracy			0.94	1991
macro avg	0.92	0.89	0.90	1991
weighted avg	0.93	0.94	0.93	1991

Confusion matrix for model KNeighborsClassifier on training data

	4	5	6	7	8	9
4	460	1	0	0	0	6
5	9	416	0	0	0	0
6	0	5	169	7	7	4
7	11	10	9	74	11	8
8	2	6	10	7	261	10
9	0	0	3	0	0	485

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
4.0	0.00	0.00	1.00	8
5.0	0.10	0.50	0.16	16
6.0	0.26	0.22	0.24	95
7.0	0.67	0.18	0.29	263
8.0	0.20	0.39	0.26	51
9.0	0.04	0.40	0.08	5
accuracy			0.23	438
macro avg	0.21	0.28	0.34	438
weighted avg	0.48	0.23	0.28	438

Confusion matrix for model KNeighborsClassifier on test data

	4	5	6	7	8	9
4	0	7	1	0	0	0
5	2	8	2	3	1	0
6	22	20	21	10	18	4
7	28	43	52	48	61	31
8	2	4	4	11	20	10
9	1	1	0	0	1	2

25 KNN na balansiranim podacima, k=10

Classification report for model KNeighborsClassifier 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	1.00	1.00	192
7.0	1.00	1.00	1.00	123
8.0	1.00	1.00	1.00	296
9.0	1.00	1.00	1.00	488
accuracy			1.00	1991
macro avg	1.00	1.00	1.00	1991
weighted avg	1.00	1.00	1.00	1991

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
4.0	0.00	0.00	1.00	8
5.0	0.12	0.38	0.18	16
6.0	0.25	0.22	0.23	95
7.0	0.71	0.33	0.45	263
8.0	0.20	0.41	0.27	51
9.0	0.03	0.20	0.05	5
accuracy			0.31	438
macro avg	0.22	0.26	0.36	438
weighted avg	0.51	0.31	0.38	438

Confusion matrix for model KNeighborsClassifier on training data

	4	5	6	7	8	9
4	467	0	0	0	0	0
5	0	425	0	0	0	0
6	0	0	192	0	0	0
7	0	0	0	123	0	0
8	0	0	0	0	296	0
9	0	0	0	0	0	488

Confusion matrix for model KNeighborsClassifier on test data

	4	5	6	7	8	9
4	0	7	0	0	1	0
5	2	6	1	5	2	0
6	18	11	21	22	20	3
7	16	21	55	86	59	26
8	2	5	6	8	21	9
9	0	0	1	0	3	1

26 KNN sa podešenim hiper-parametrima na balansiranim podacima

```
{'algorithm': 'auto', 'n_neighbors': 5, 'p': 1, 'weights': 'distance'}
0.9598181564871856
```

27 Izbor hiper-parametara i CV ocena prethodnog modela

Classification report for model BaggingClassifier on training data

	precision	recall	f1-score	support
4.0	0.99	1.00	0.99	467
5.0	0.99	0.99	0.99	425
6.0	0.94	0.95	0.94	192
7.0	0.95	0.85	0.90	123
8.0	0.96	0.96	0.96	296
9.0	0.99	1.00	0.99	488
accuracy			0.98	1991
macro avg	0.97	0.96	0.96	1991
weighted avg	0.98	0.98	0.98	1991

Classification report for model BaggingClassifier on test data

	precision	recall	f1-score	support
4.0	0.00	0.00	1.00	8
5.0	0.12	0.50	0.20	16
6.0	0.25	0.21	0.23	95
7.0	0.68	0.25	0.37	263
8.0	0.19	0.41	0.26	51
9.0	0.05	0.40	0.09	5
accuracy			0.27	438
macro avg	0.22	0.30	0.36	438
weighted avg	0.49	0.27	0.33	438

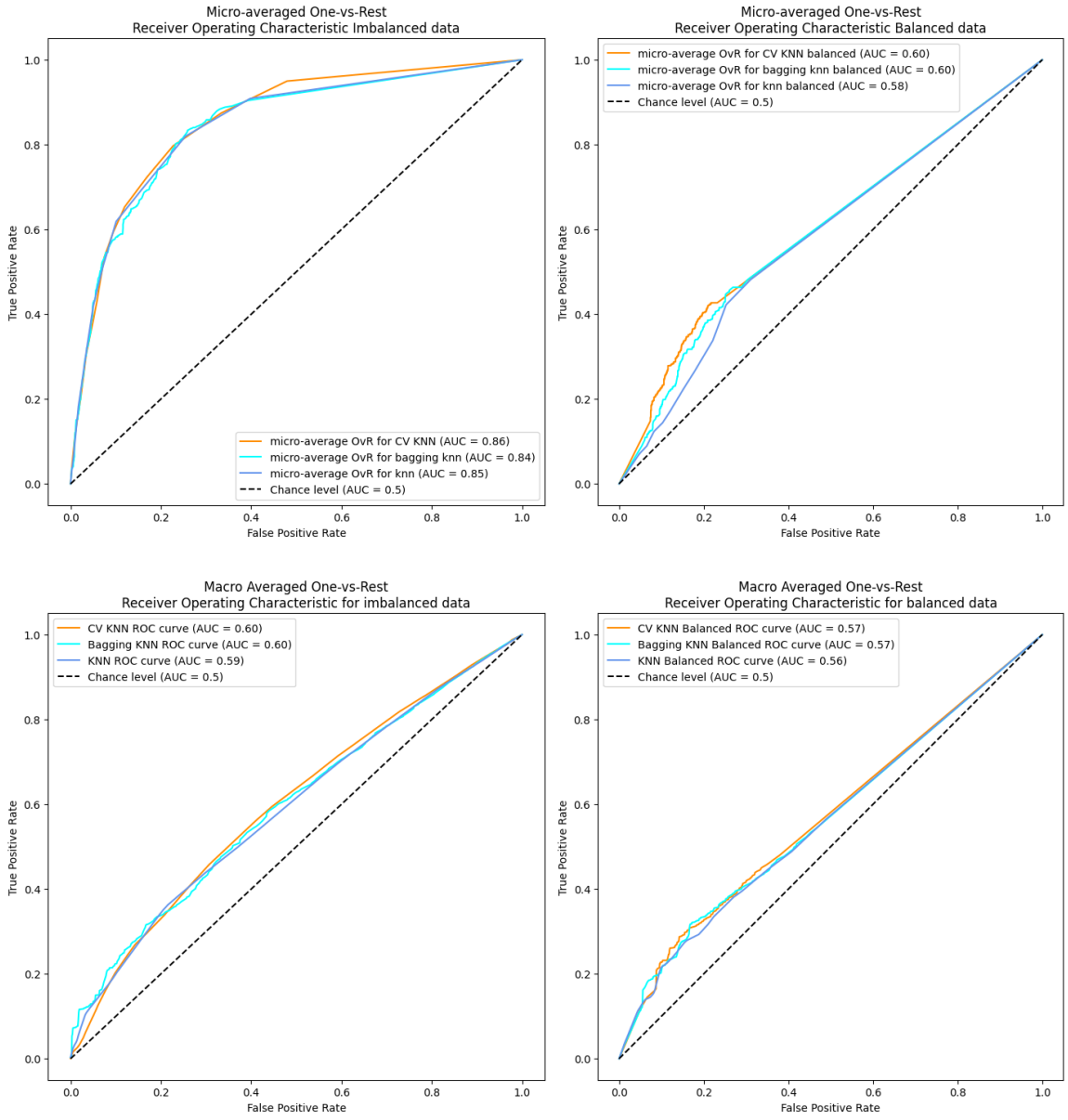
Confusion matrix for model BaggingClassifier on training data

	4	5	6	7	8	9
4	466	0	0	0	0	1
5	2	420	0	3	0	0
6	0	0	182	3	7	0
7	3	2	4	105	6	3
8	1	1	7	0	285	2
9	0	0	1	0	0	487

Confusion matrix for model BaggingClassifier on test data

	4	5	6	7	8	9
4	0	7	0	0	1	0
5	2	8	0	3	3	0
6	20	14	20	19	19	3
7	23	30	55	67	63	25
8	2	6	4	9	21	9
9	0	1	0	0	2	2

28 KNN bagging classifier na balansiranim klasama



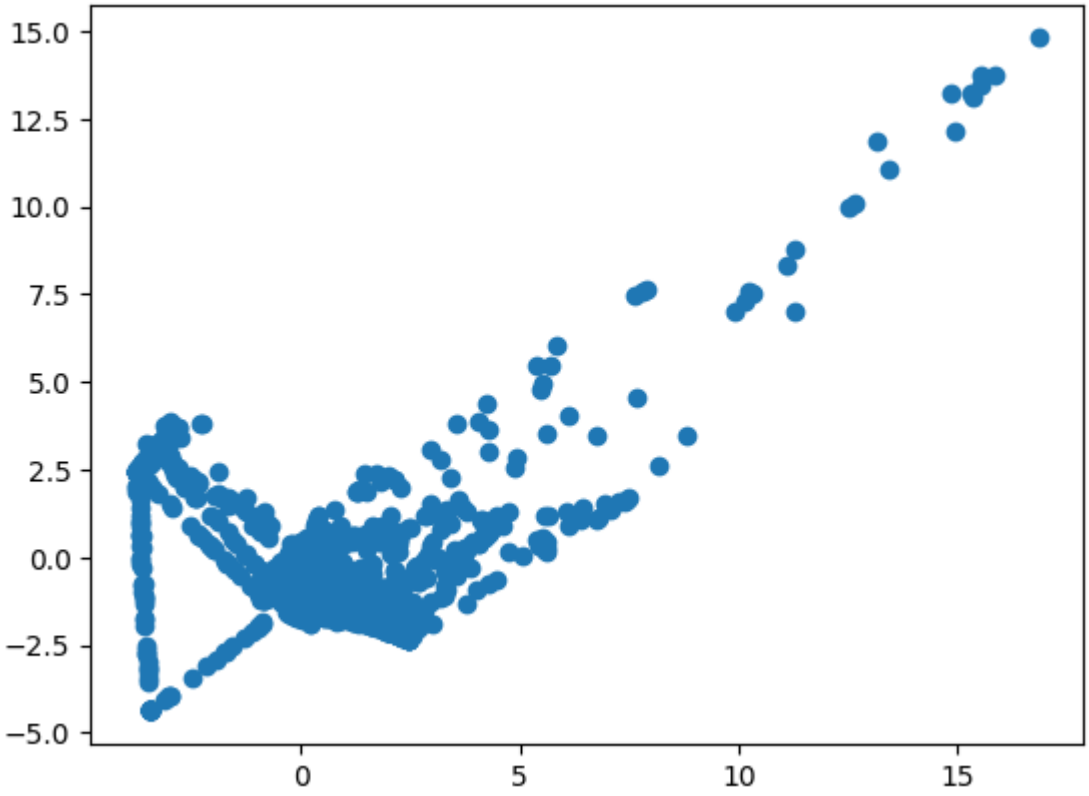
29 ROC krive zasnovane na micro i macro uprosečavanju

Vidimo da su generalno performanse loše i da se uvek javlja izvestan *overfitting*, međutim modeli koji rade sa nebalansiranim podacima iskazuju neku vrstu konzistentnosti i kao najbolji se pokazao model sa podešenim hiper-parametrima.

Kako je ovo višeklasna klasifikacija, predstavljene su ROC krive zasnovane na One vs Rest tehnici i micro i macro uprosečavanju – micro sve instance posmatra ravnopravno, a macro average uprosečavanje radi na nivou klasa – kalkuliše ocene za svaku klasu i onda se računa prosek tih ocena. Ovo daje jedan valjan uvid u to koliko micro average ocenjivanje može da zavara, naime na osnovu prvog grafa rekli bi da su modeli

zadovoljavajući, a kada se pogleda njegov ekvivalent zasnovan na macro average ocenjivanju vidimo da stanje nije uopšte tako.

Pokušajmo sada pristum sa OneHotEncoding-om.



30 Izgled balansiranog trening skupa nakon primene OneHotEncoding i PCA tehnika

```
{'algorithm': 'auto', 'n_neighbors': 8, 'p': 1, 'weights': 'distance'}
0.768051434223541
```

31 Izbor hiper-parametara i CV ocena modela

Classification report for model KNeighborsClassifier 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	1.00	1.00	206
7.0	1.00	1.00	1.00	133
8.0	1.00	1.00	1.00	300
9.0	1.00	1.00	1.00	491
accuracy			1.00	2022
macro avg	1.00	1.00	1.00	2022
weighted avg	1.00	1.00	1.00	2022

Classification report for model KNeighborsClassifier on test data

	precision	recall	f1-score	support
4.0	0.00	0.00	1.00	6
5.0	0.04	0.27	0.07	15
6.0	0.28	0.22	0.24	92
7.0	0.65	0.16	0.26	278
8.0	0.15	0.41	0.22	46
9.0	0.04	1.00	0.08	1
accuracy			0.20	438
macro avg	0.19	0.34	0.31	438
weighted avg	0.49	0.20	0.25	438

Confusion matrix for model KNeighborsClassifier on training data

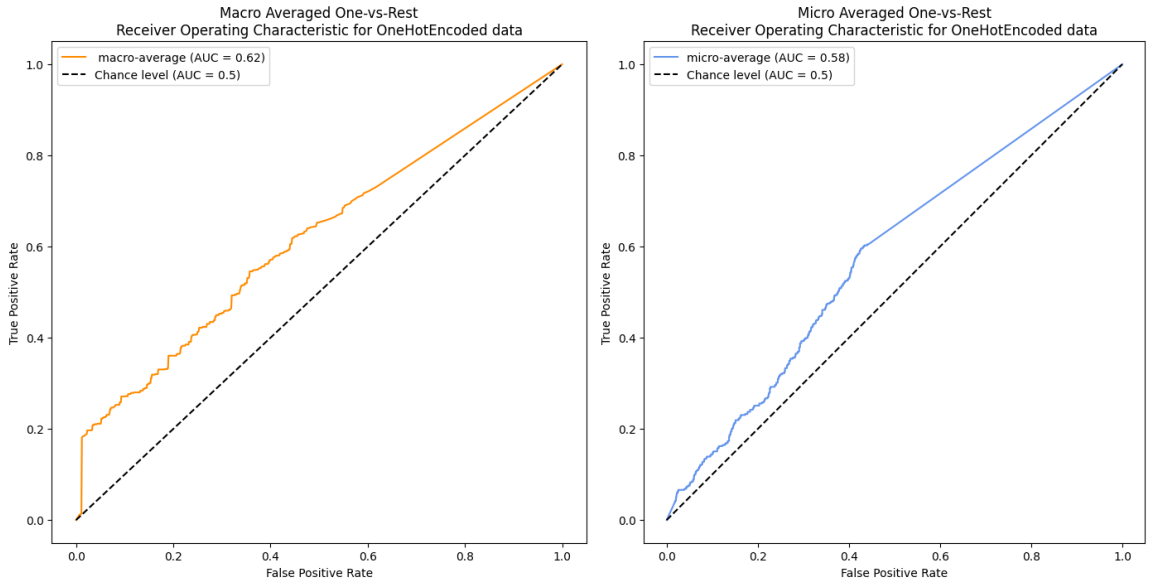
	4	5	6	7	8	9
4	467	0	0	0	0	0
5	0	425	0	0	0	0
6	0	0	206	0	0	0
7	0	0	0	133	0	0
8	0	0	0	0	300	0
9	0	0	0	0	0	491

Confusion matrix for model KNeighborsClassifier on test data

	4	5	6	7	8	9
4	0	3	2	0	1	0
5	2	4	4	0	4	1
6	12	23	20	17	17	3
7	25	70	38	45	88	12
8	0	4	8	7	19	8
9	0	0	0	0	0	1

32 Rezultati primene prethodnog modela

Model se ponaša slično kao i prethodni.



33 ROC krive prethodnog modela

SUPPORT VECTOR MACHINES

SVM mapira trening instance u tačke u prostoru tako da se maksimizuje širina prozanog prostora između kategorija – prostor se deli pomoću pravih, tj. hiperravni u više dimenzija, a zatim nove instance se mapiraju u taj isti prostor i dodeljuje im se klase u odnosu na potprostor kojem pripadaju.

Mogućnost rada sa hiperravnima čini ovaj metod jako pogodnim za višeklasnu klasifikaciju.

Podatke ćemo kao i ranije normalizovati upotrebom MinMaxScalera, a zatim vršimo balansiranje SMOTEEN tehnikom.

```
params = [
        {
            'kernel': ['linear'],
            'C': [0.01, 0.1, 1, 10,100],
        },
        {
            'kernel': ['rbf'],
            'C': [0.01, 0.1, 1, 10,100],
            'gamma': [0.01, 0.1, 1, 10],
        },
    ]
```

34 Mogući parametri prosleđeni GridSearchCV

```
{'C': 100, 'gamma': 10, 'kernel': 'rbf'}
```

```
0.9668523263491372
```

35 Izabrani parametri i ocena

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

36 Rezultati datog modela

Čisto poređenja radi, napravljen je još jedan model gde je urađen OneHotEncoding kategoričkih atributa i primenjen PCA na takav model.

Classification report for model SVC on training data

	precision	recall	f1-score	support
4.0	0.89	0.88	0.89	467
5.0	0.83	0.77	0.80	425
6.0	0.75	0.59	0.66	206
7.0	0.78	0.47	0.58	133
8.0	0.64	0.93	0.76	300
9.0	0.97	0.97	0.97	491
accuracy			0.83	2022
macro avg	0.81	0.77	0.78	2022
weighted avg	0.84	0.83	0.83	2022

Confusion matrix for model SVC on training data

	4	5	6	7	8	9
4	413	20	7	2	25	0
5	29	327	14	3	47	5
6	15	27	121	7	31	5
7	4	16	14	62	36	1
8	3	2	5	6	279	5
9	0	2	0	0	15	474

Classification report for model SVC on test data

	precision	recall	f1-score	support
4.0	0.00	0.00	1.00	6
5.0	0.05	0.33	0.09	15
6.0	0.22	0.14	0.17	92
7.0	0.60	0.12	0.20	278
8.0	0.15	0.54	0.23	46
9.0	0.05	1.00	0.09	1
accuracy			0.18	438
macro avg	0.18	0.36	0.30	438
weighted avg	0.44	0.18	0.20	438

Confusion matrix for model SVC on test data

	4	5	6	7	8	9
4	0	5	0	0	1	0
5	4	5	2	0	3	1
6	16	21	13	18	21	3
7	21	59	39	33	118	8
8	0	4	5	4	25	8
9	0	0	0	0	0	1

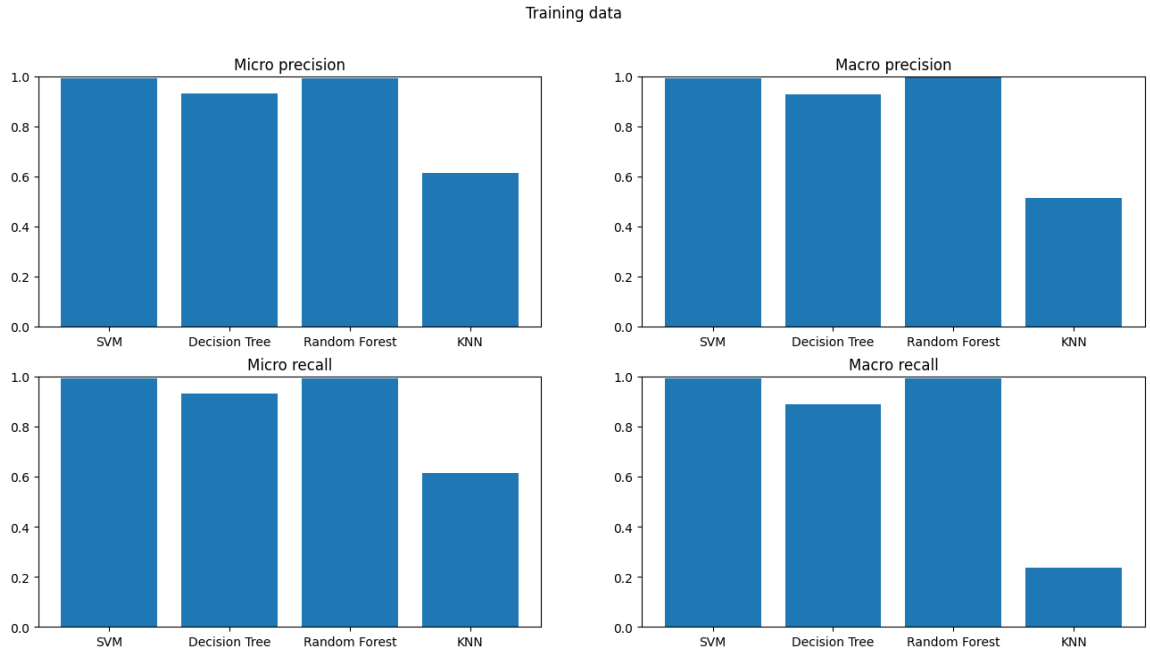
37 Parametri dobijem CV tehnikom: {'C': 100, 'gamma': 10, 'kernel': 'rbf'}

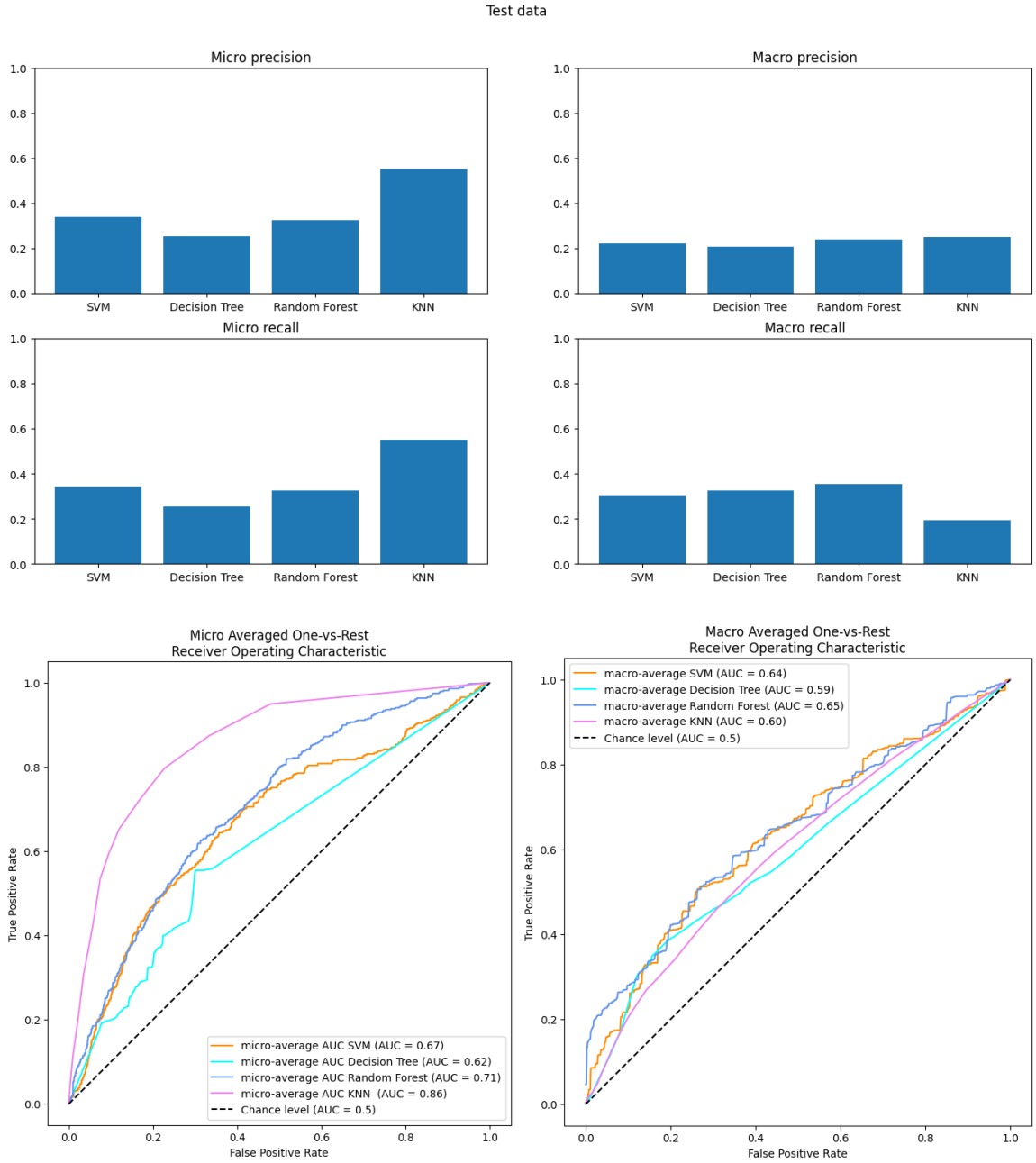
Cilj je bio da pokažemo da se nije napravila neka značajna razlika, tj. da manipulisanjem kategoričkim atributima ipak ne možemo u potpunosti da simuliramo svojstva koja ispoljavaju pravi numerički atributi.

POREĐENJE MODELA

Za kraj, uporedimo par prethodno prikazanih modela.

Kao reprezentativni modeli korišćeni su Decision Tree Classifier sa podešenim hiper-parametrima, Random Forest, SVM bez OneHotEncoding-a i KNN sa podešenim hiper-parametrima na nebalansiranim podacima.





38 Grafički prikaz raznih ocena

Iako se ni jedan od modela nije istakao kao značajno precizan ili dobar, možemo dati prednost Random Forest-u ili uopšte samom pristupu stabala odlučivanja za ovakav skup podataka.

Jedna od verovatnih mana jeste ogroman broj kategorija, što otežava posao modelima zasnovanim na odlulivanju, ali primena takvih modela je daleko jednostavnija nego forsiranje numeričkih osobina na podatke koji su čisto kategorički.

Klasterovanje

Klasterovanje je oblik nenadgledanog učenja i predstavlja grupisanje medjusobno sličnih instance. U nastavku će biti prikazani algortimi KMeans i DBSCAN.

KMEANS

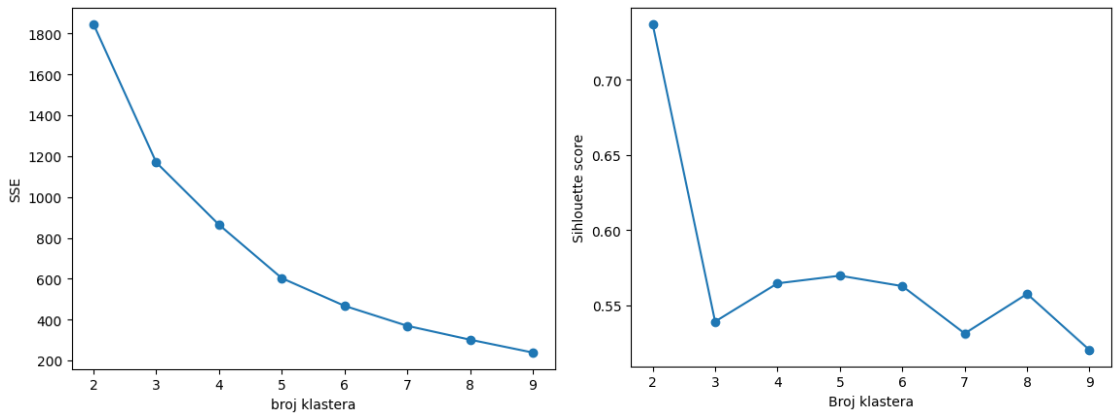
Za KMeans algoritam broj klastera mora biti unapred poznat, što može ujedno da se interpretira i kao mana i kao prednost. Ovo je jednostavna tehnika koja se bazira na pronalaženju reprezentativnih predstavnika, konkretno u ovoj implementaciji to je centroid.

	Species	Country.of.Origin	Harvest.Year	Variety	Processing.Method	Category.One.Defects	Quakers	Color	Category.Two.Defects	altitude_mean_meters
0	0	8	5	5	4	0	0.0	2	0	2075.0
1	0	8	5	15	4	0	0.0	2	1	2075.0
2	0	9	3	2	4	0	0.0	2	0	1700.0
3	0	8	5	5	0	0	0.0	2	2	2000.0
4	0	8	5	15	4	0	0.0	2	2	2075.0

39 Podaci za klasterovanje

Kako atributi poput Country of Origin, Variety i Processing Method imaju dosta kategorija, isprobaćemo dva pristupa: izbacujemo te attribute, primenjujemo OneHotEncoding na preostale kategoričke atributa i standardizujemo podatke i vršimo PCA; drugi pristup je da izvršimo OneHotEncoding nad svim kategoričkim atributima, a zatim smanjimo dimenzionalnost prethodno standardizovanih podataka.

U prvom pristupu, primenom PCA ukupno objašnjene varijanse je oko 0.37.

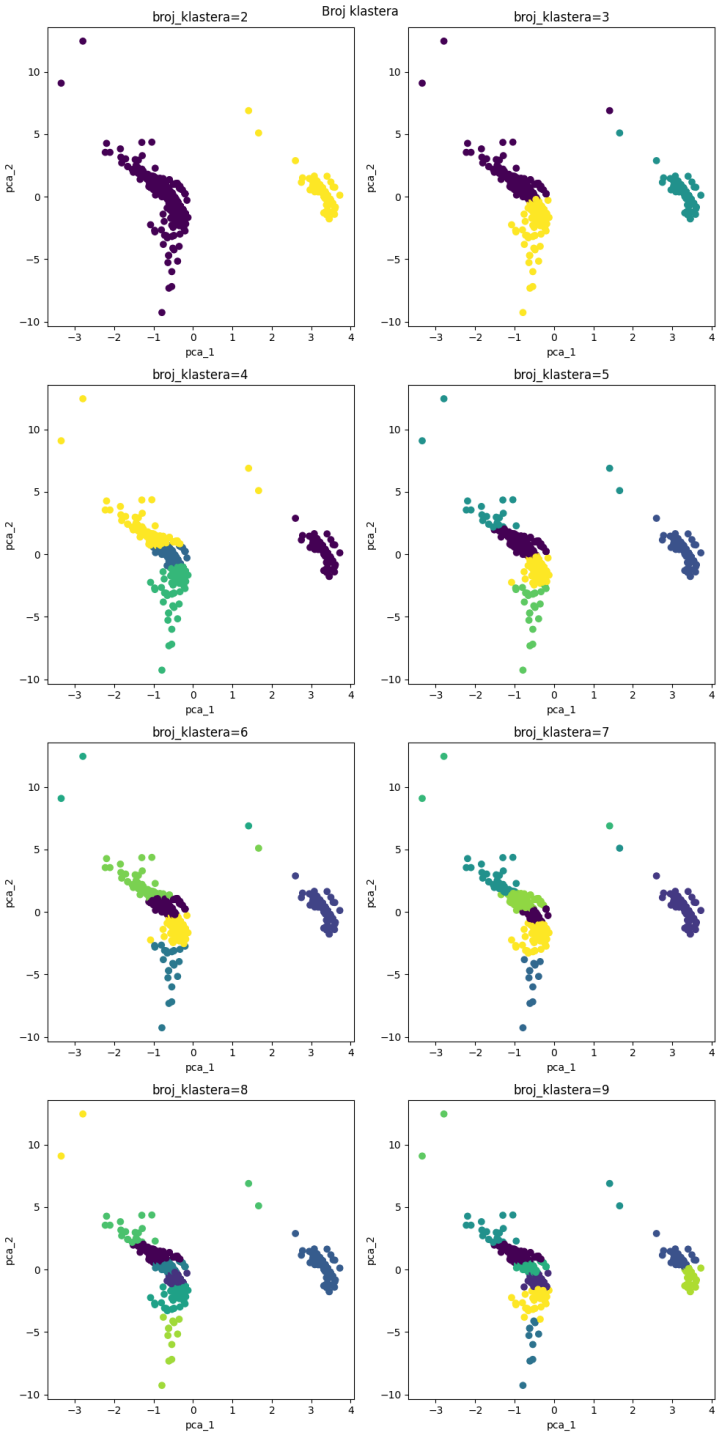


40 SSE i koeficijent senke sa porastom broja klastera

SSE je suma kvadrata grešaka – gde je greška zapravo Euklidsko rastojanje od najbližeg centroida. Teži se da se ova vrednost minimizuje i očekivano porastom broja klastera SSE opada.

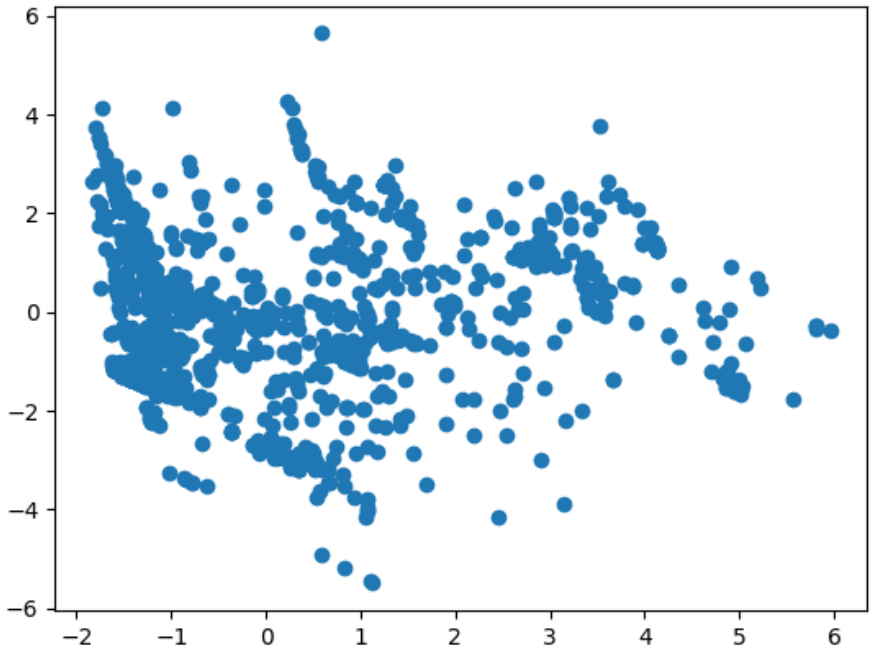
Pravilo lakta je heuristika kojom se traži tačka “na laktu” gde SSE najbrže opadne, da bi se dobila optimalna vrednost parametra K.

Koeficijent senke objasnjava koliko je objekat sličan svom klasteru (kohezija) u odnosu na druge klastere (separacija). Težimo da maksimizujemo ovu vrednost.

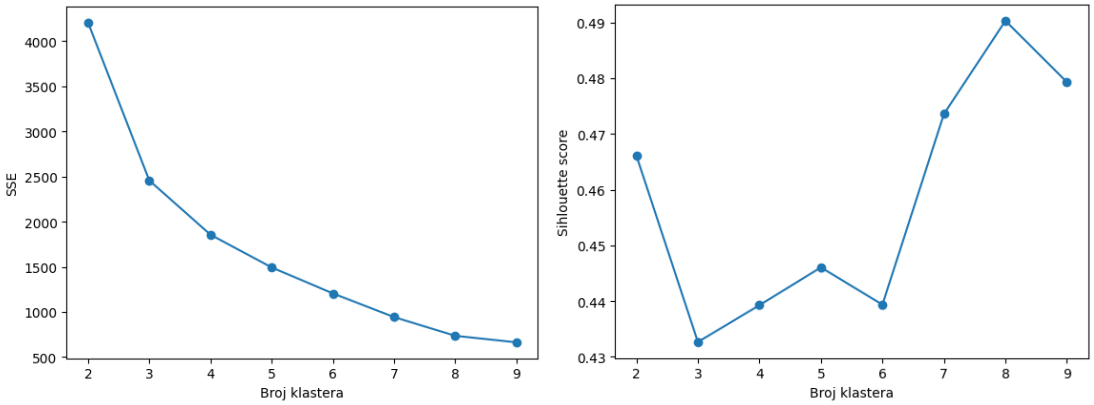


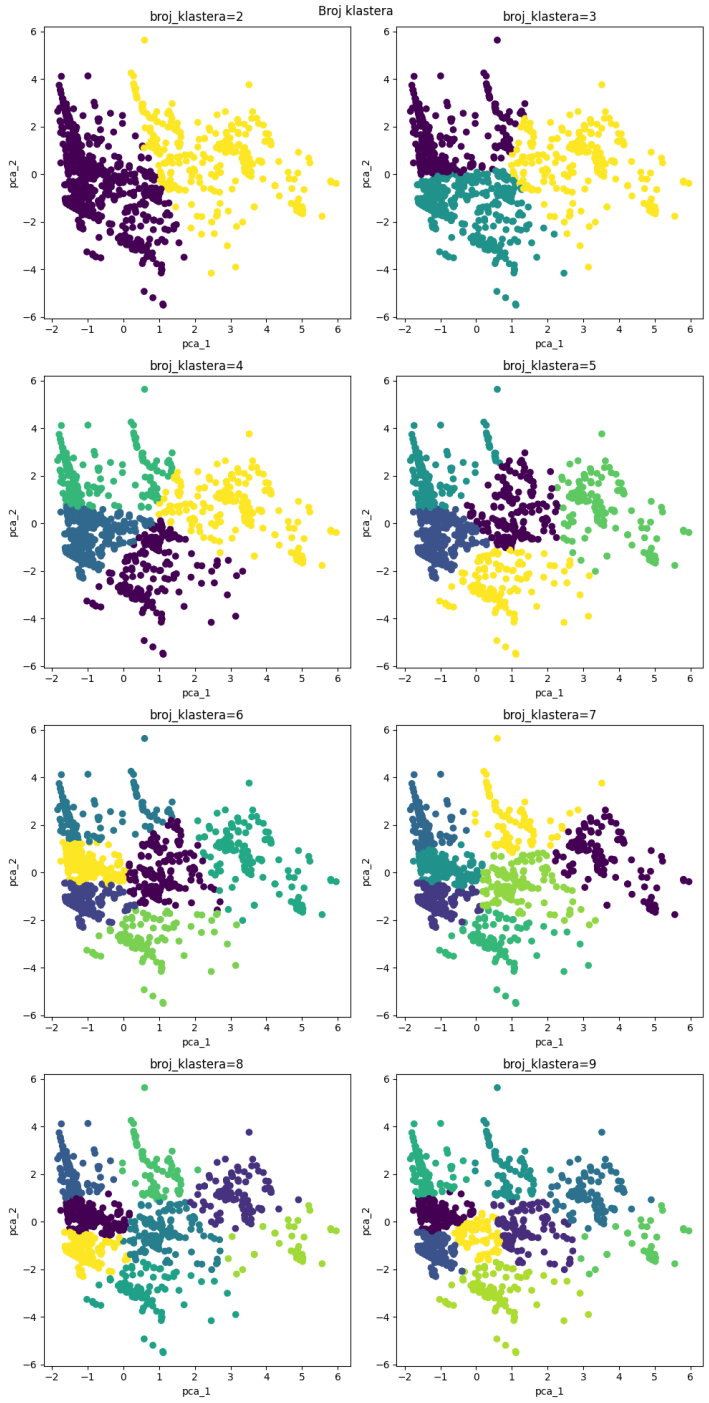
Na osnovu pravila lakta, mogli bismo da kažemo da je optimalna vrednost 3, a može i 4 – promena nije velika i ovaj pristup je jako subjektivan. Nasuprot njemu, na osnovu koeficijenata senke, najveća vrednost se dobija za dva klastera - što–se podudara sa vizuelizovanim podacima.

U drugom pristupu primenom PCA dobijemo udeo objasnjene varijanse od neverovatnih 0.068 – što je očekivano jer smo imali ogroman broj binarnih atributa nakon primene OneHotEncoder-a.



41 Vizuelizacija podataka



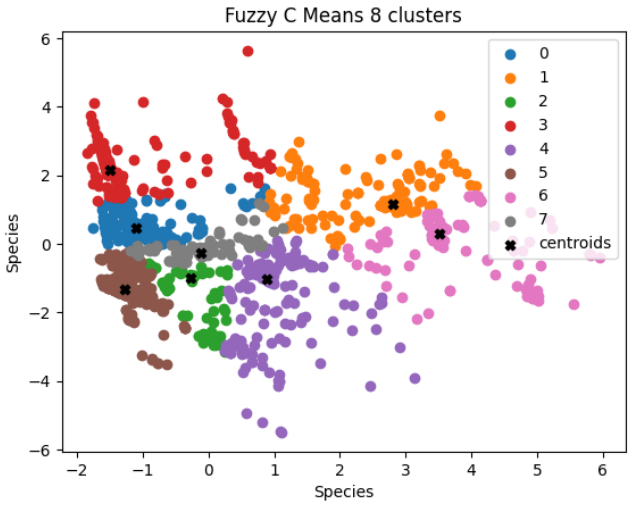


Koeficijent senke se maksimizuje za 8 klastera što izgleda kao razuman izbor.

FUZZY CMEANS

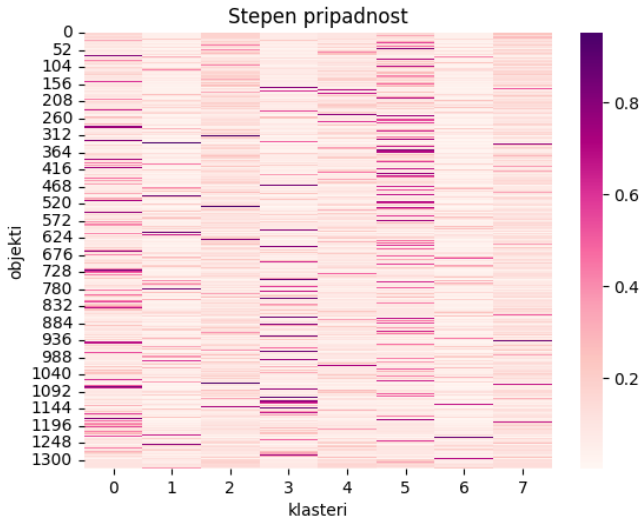
Ovo je predstavnik *soft clustering* algortiama koja omogućava da instance pripada većem broju klastera. Na osnovu težinskih sumi rastojanja se određuje pripadnost svakom od C klastera.

Ako za dodeljeni klaster uzmemo onaj za koji je stepen pripadnost najveći dobijamo efekat hard clustering algortima.



Dobili smo prilično slične rezultate kao kad smo na isti skup primenili KMeans gde je K=8.

Pošto soft clustering ne možemo vizuelizovati, možemo prikazati matricu pripadnosti.

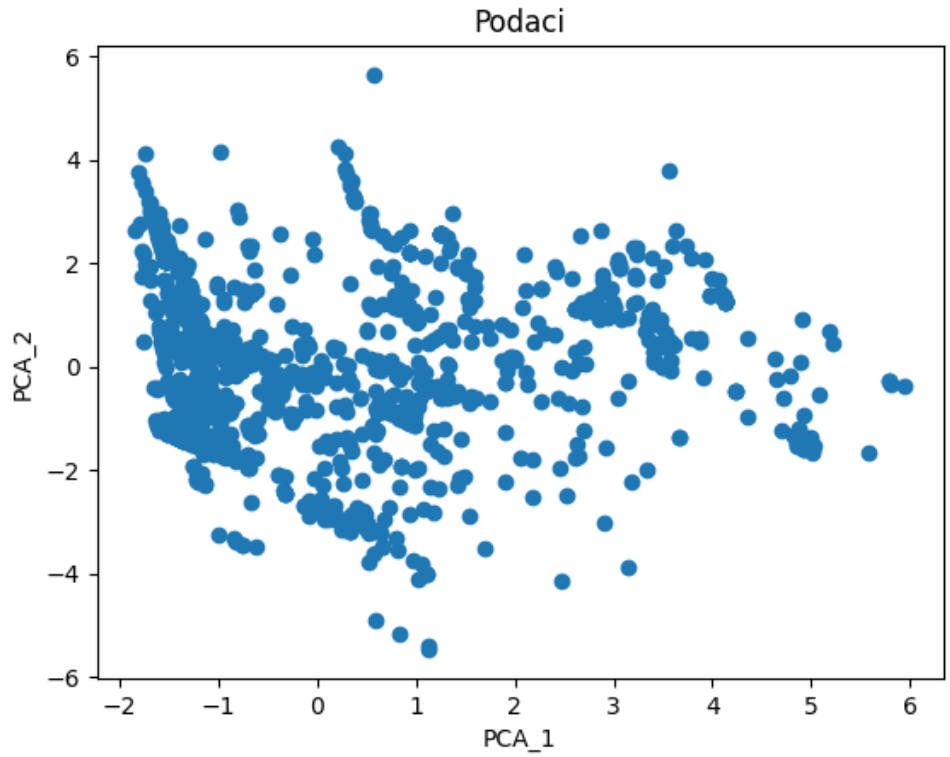


U ovom modelu dobijen je koeficijent senke od 0.65 što je prihvatljivo.

DBSCAN

DBSCAN je algoritam zasnovan na gustini koji se bazira na principu da su glasteri grusto populisane regije razdvojene regijama sa manjom gustinom populacije. Ono što ga naročito razlikuje od prethodna dva algortima jeste to što se unapred ne zadaje broj klastera i rakodje je robusan na outlajere.

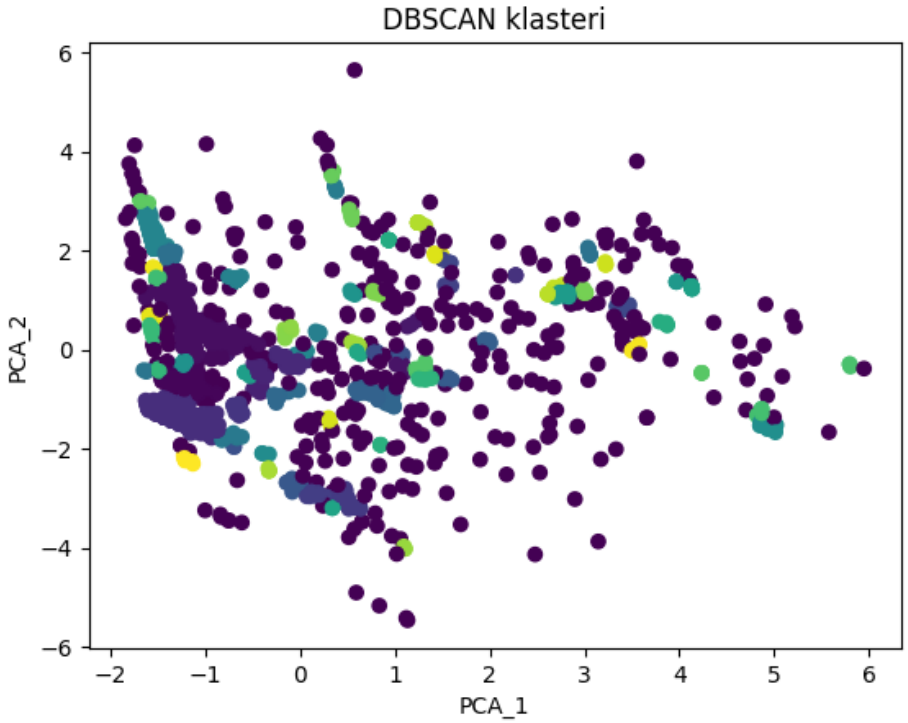
U ovom primeru ćemo klasterovanje primeniti na OneHotEncoded kategoričke atribute, sa smanjenom dimenzionalnošću na 2.



42 Kako izgledaju podaci

DBSCAN uzima dva parametra – epsilon i MinPoints. Epsilon predstavlja poluprečnik kruga koji se "iscrtava" oko svake instance i proverava se gustina u epsilon okolini, dok je MinPoints minimalni broj instanci koji treba da se nadje u epsilon okolini neke instance da bi ona bila kategorisana kao *jezgro*. (U slučaju više dimenzija epsilon jer radijuh hiper-sfere)

Insance koje se kategorišu kao *granice* su one koj kojih se u epsilon okolini nadje manje od MinPoints instance, dok je *šum* svaka insatnca kod koje se u epsilon okolini ne nadje ni jedna druga instanca.



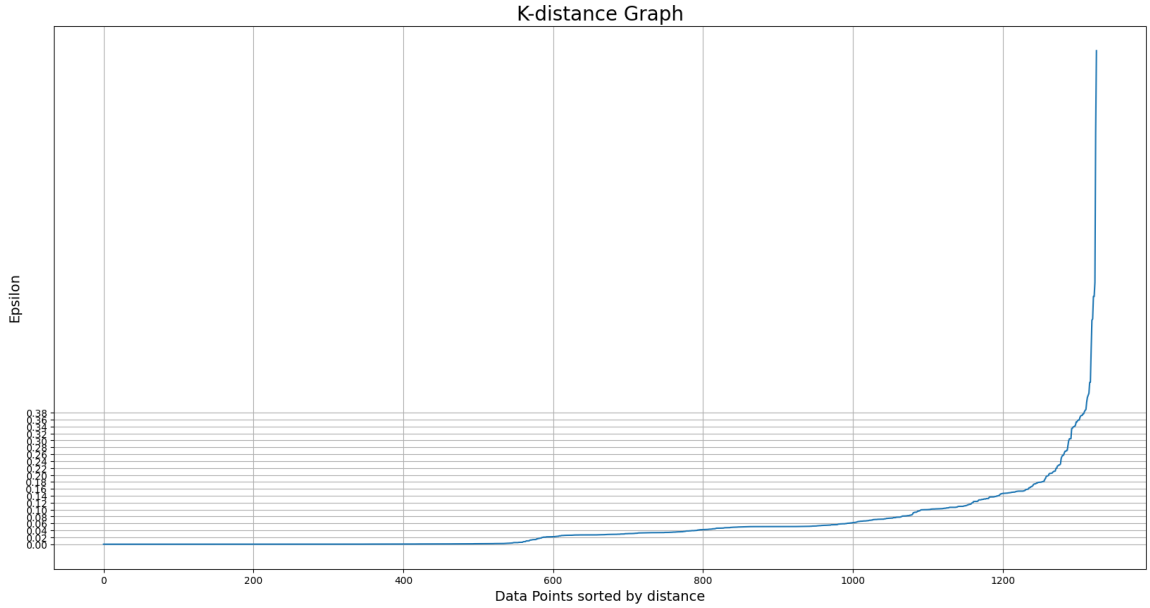
43 Klasteri dobijenim DBSCAN algortimom sa parametrima eps=0.1 i minPoints=3

Imamo dosta klastera koji se preklapaju.

Kako izabrati optimalne parameter?

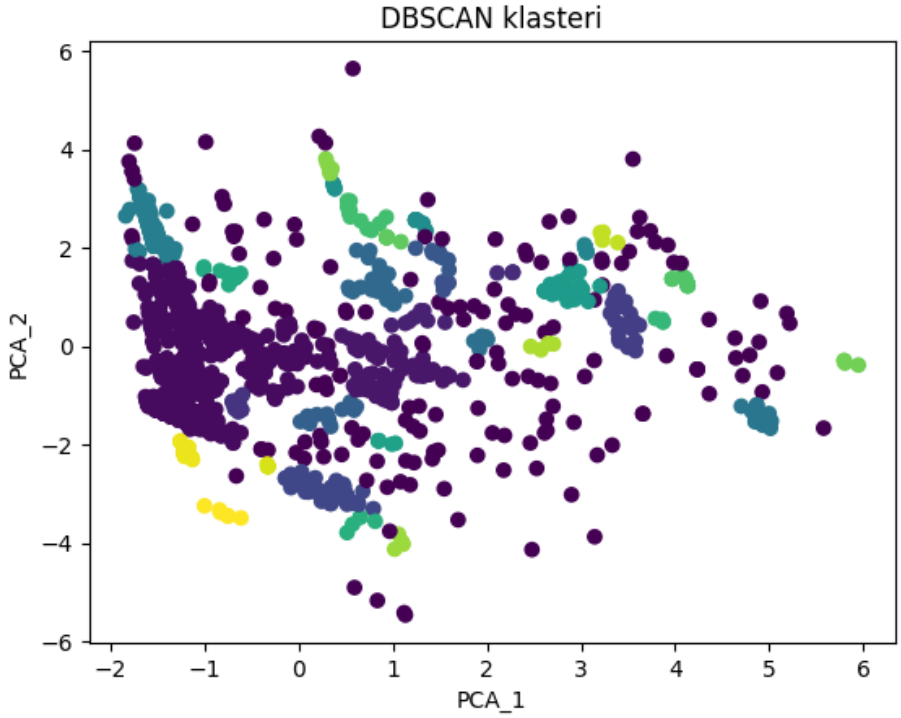
MinPoints bi trebalo da bude bar za jedan veći od broja dimenzija podataka. Često se uzima broj dimenzija puta dva.

Da bi odredili epsilon potrebno nam je da znamo koliko su gusto,tj retko rasporedjeni podaci. Ovo možemo dobiti primenom pravila lakta na graf najbližih suseda.



Tražimo tačku maksimalnog zakrivljenja što je ovde oko 0.2.

Dakle za MinPoints je uzeto 4 jer imamo dve dimenzije, a za epsilon 0.2



44 Klasteri

Koeficijent senke je oko 0.16 što je prilično malo i indikuje da imamo dosta preklapajućih klastera, a sam broj klastera je 34.

Ovaj pristup ima znatno lošije performanse na ovom skupu u odnosu na KMeans i Fuzzy CMeans. To smo donekle i mogli da predividmo na osnovu samog grafičkog prikaza podataka i principa na kojem je zasnovan DBSCAN, naime imamo jako sitne prazne regije i dosta preklapajućih instanci.

Pravila pridruživanja

Za određivanje pravila pridruživanja korišćen je Apriori algoritam pomoću alata *SPSS Modeler*.

Kakos u potrebni kategorički atributi, odlučeno je da se atribut Altitude Mean Meters izostavi.

Za minimalnu podršku uzeto je 10%, a za minimalnu pozdanost 80%.

Prvi pokušaj jeste da otkrijemo veze sa ciljnim atributem Total Cup Points.

Consequent	Antecedent	Support %	Confidence %	Lift
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Variety = Caturra Processing.Method = Washed / Wet	11.011	84.247	1.405
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Variety = Caturra Processing.Method = Washed / Wet Species = Arabica	11.011	84.247	1.405
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Variety = Caturra Quakers = 0.000000	11.84	84.076	1.402
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Variety = Caturra Quakers = 0.000000 Species = Arabica	11.84	84.076	1.402
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Processing.Method = Washed / Wet Quakers = 0.000000	10.709	83.803	1.398
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Variety = Caturra Color = Green Quakers = 0.000000	10.709	83.803	1.398
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Processing.Method = Washed / Wet Quakers = 0.000000 Species = Arabica	10.709	83.803	1.398
Total.Cup.Points = 7.000000	Country.of.Origin = Colombia Variety = Caturra Color = Green	10.709	83.803	1.398

45 Pravila pridruživanja tipa telo -> Total Cup Points

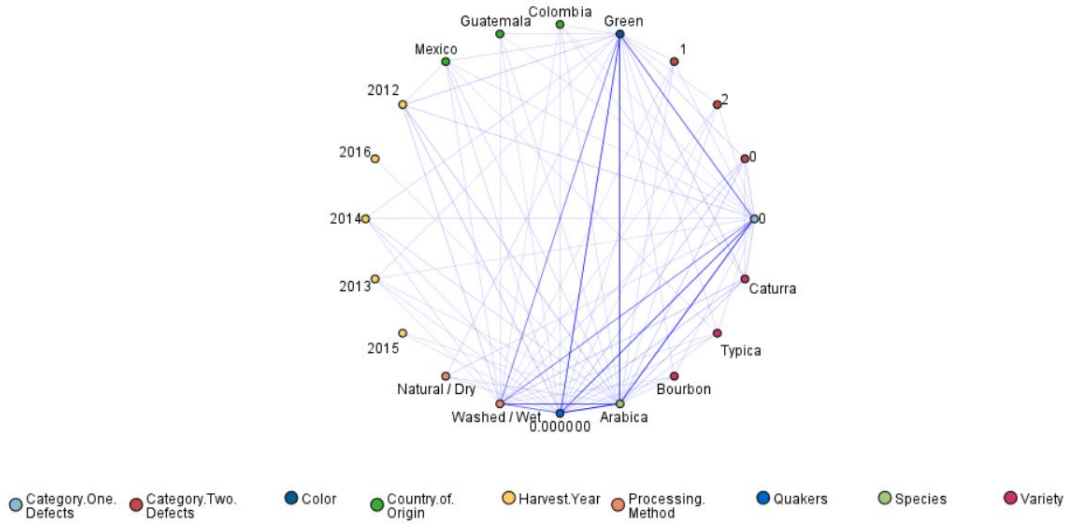
Prona]eno je 32 pravila i sva su vezana za ocenu 7 što ih ne čini naročito interesantnim, mada je lift svuda veći od 1.

Drugi pokušaj je da se uoče veze među atributima. Dakle izbačena ciljna promenljiva.

Pronađeno je 720 pravila, od kojih svega 20-ak ima lift veći od 2, a preostala pravila

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	10.709	93.662	2.598

46 Pravila pridruživanja sortirana po Lift-u



47 Uočene najjače veze

Zaključak

Nakon pregleda različitih modela i primena raznih tehniki, klasifikacija se pokazala kao izuzetno izazovna. Kao što je i pre napomenuto teško je striktnim kategorijama dodeliti osobine brojeva, te algoritmi zasnovani na rastojanjima nisu zablistali na ovom skupu. Međutim, nisu se nešto bolje pokazali ni algoritmi zasnovani na odlučivanju, što bi se moglo pripisati velikom broju kategorija, i maloj veličini skupa.

Naravno ulogu igra i način diskretizacije ciljne promenljive, i možda bi stvari bile drugačije da intuitivno najizraženija ocena nije viša-srednja.

Ovo dovodi do jednog jako bitnog zaključka, a to je da ćemo, za sada, precizno ocenjivanje kvaliteta zrna kafe ostaviti ekspretima.