

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](https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi). 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

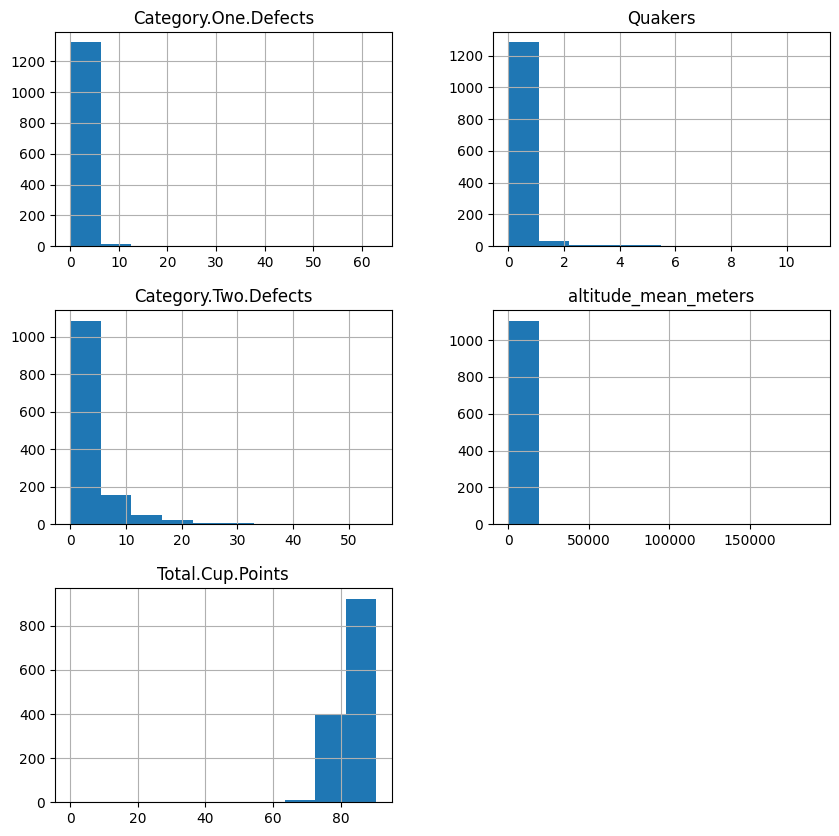
### Metapodaci od poreklu

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

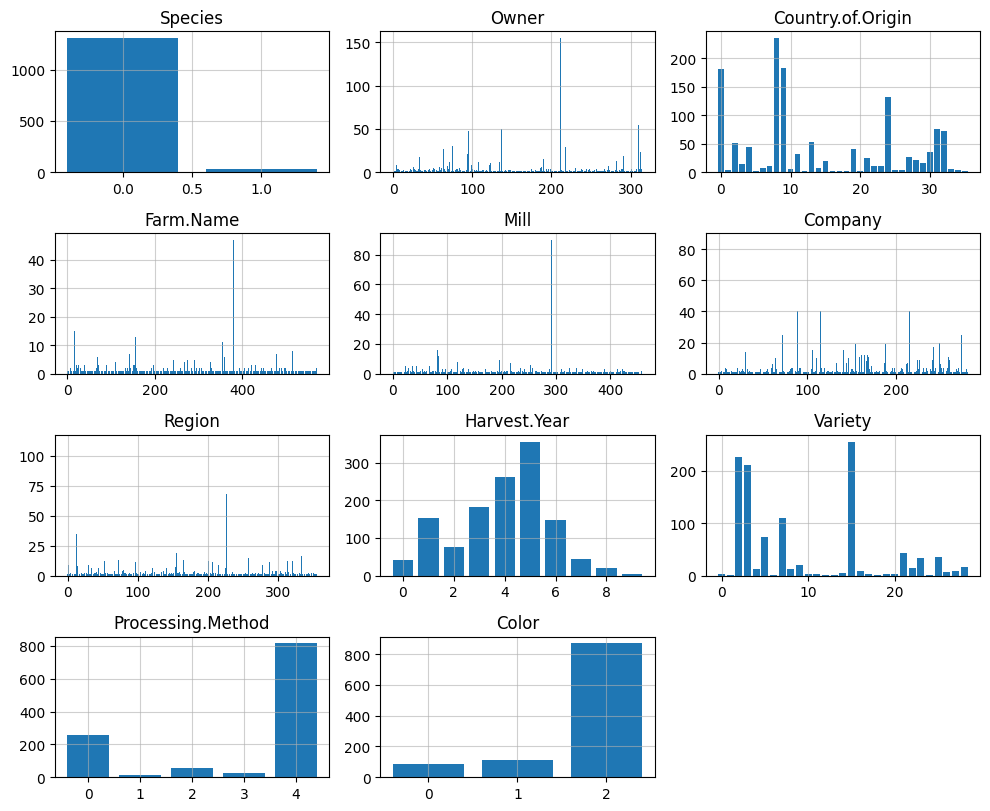
Javlja se jos manje opisnih atributa poput Certification, Certification Address, Expiration i slični, koji nisu od velikog značaja.

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.

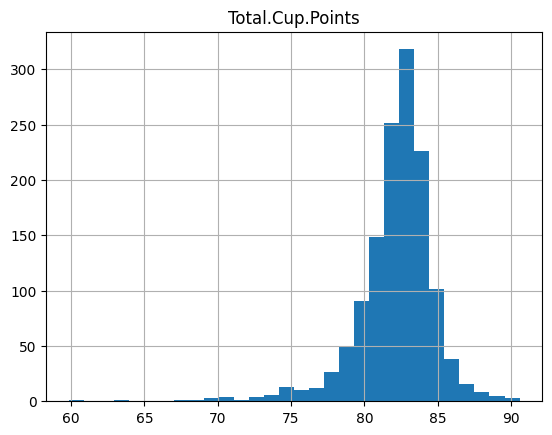


Pregled numeričnih atributa



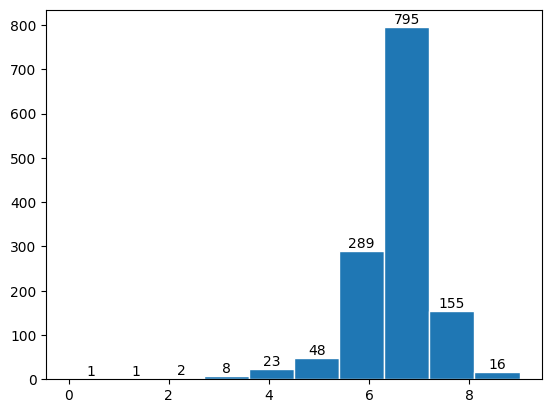
Pregled kategoričkih atributa

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.



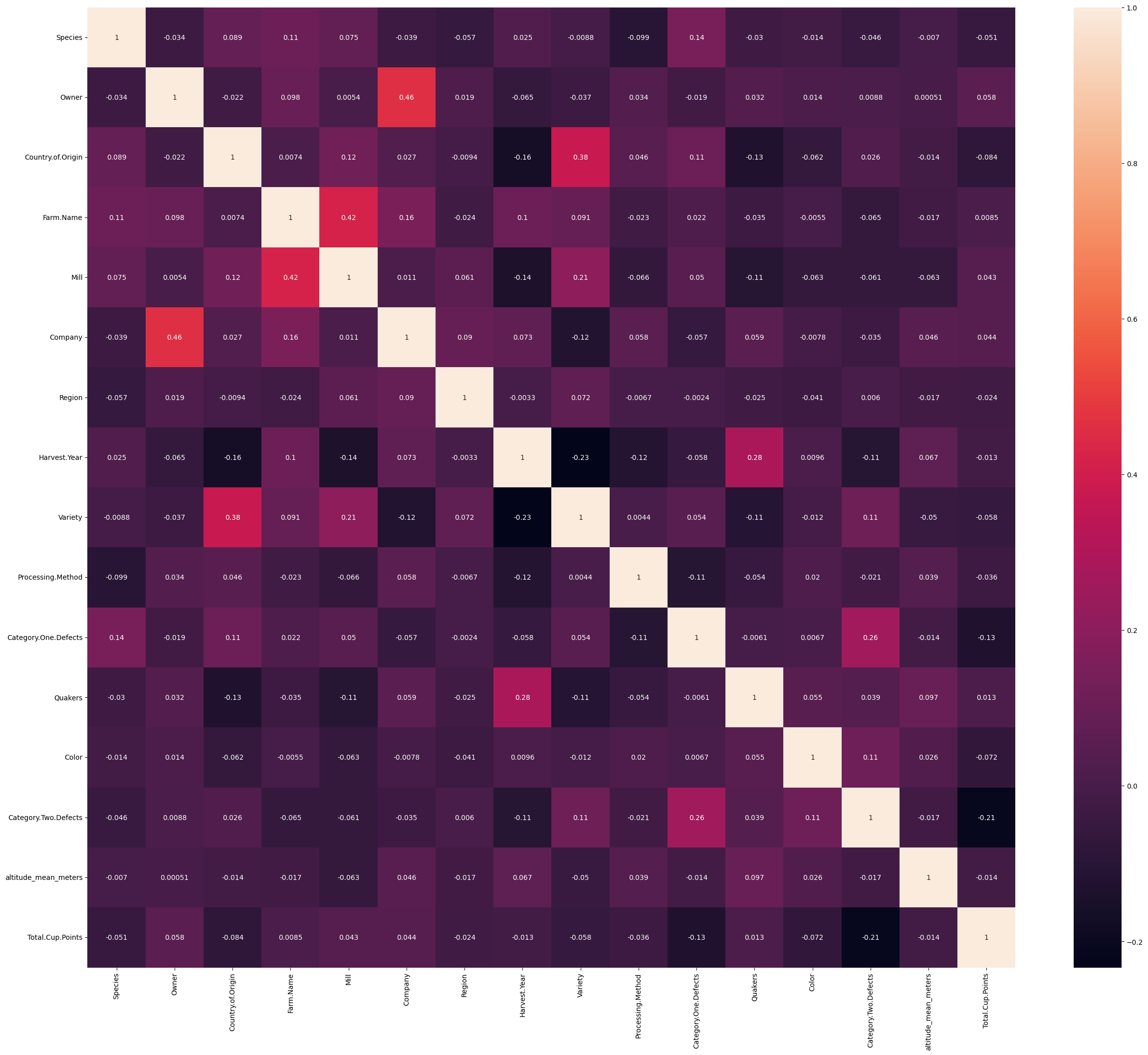
Pregled raspodele ciljne promenljive

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



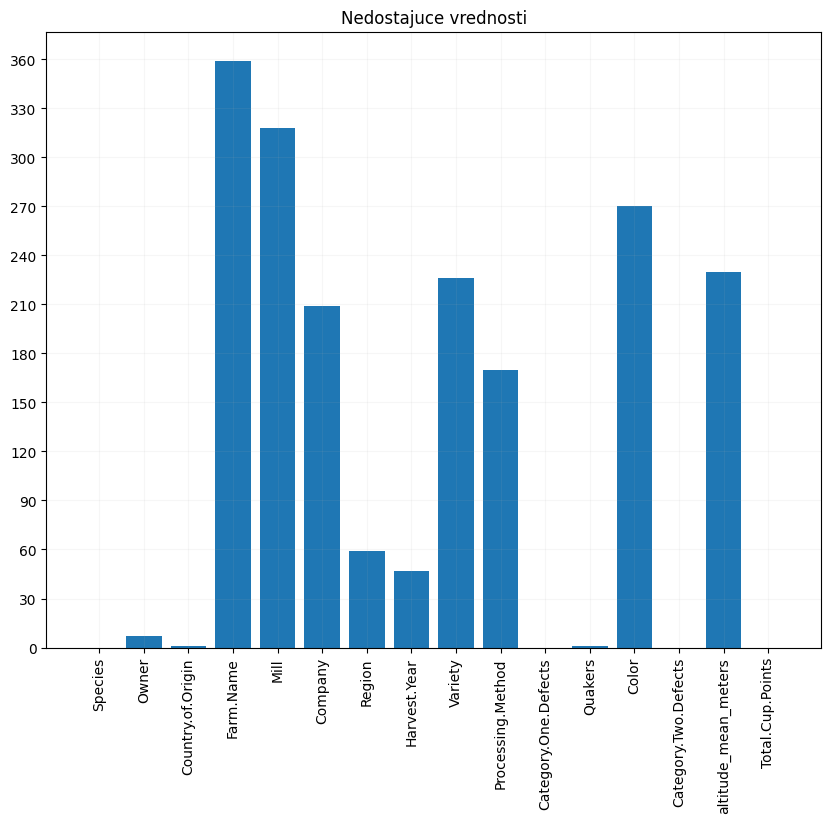
Diskretizacija ciljne promenljive

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].



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 ).

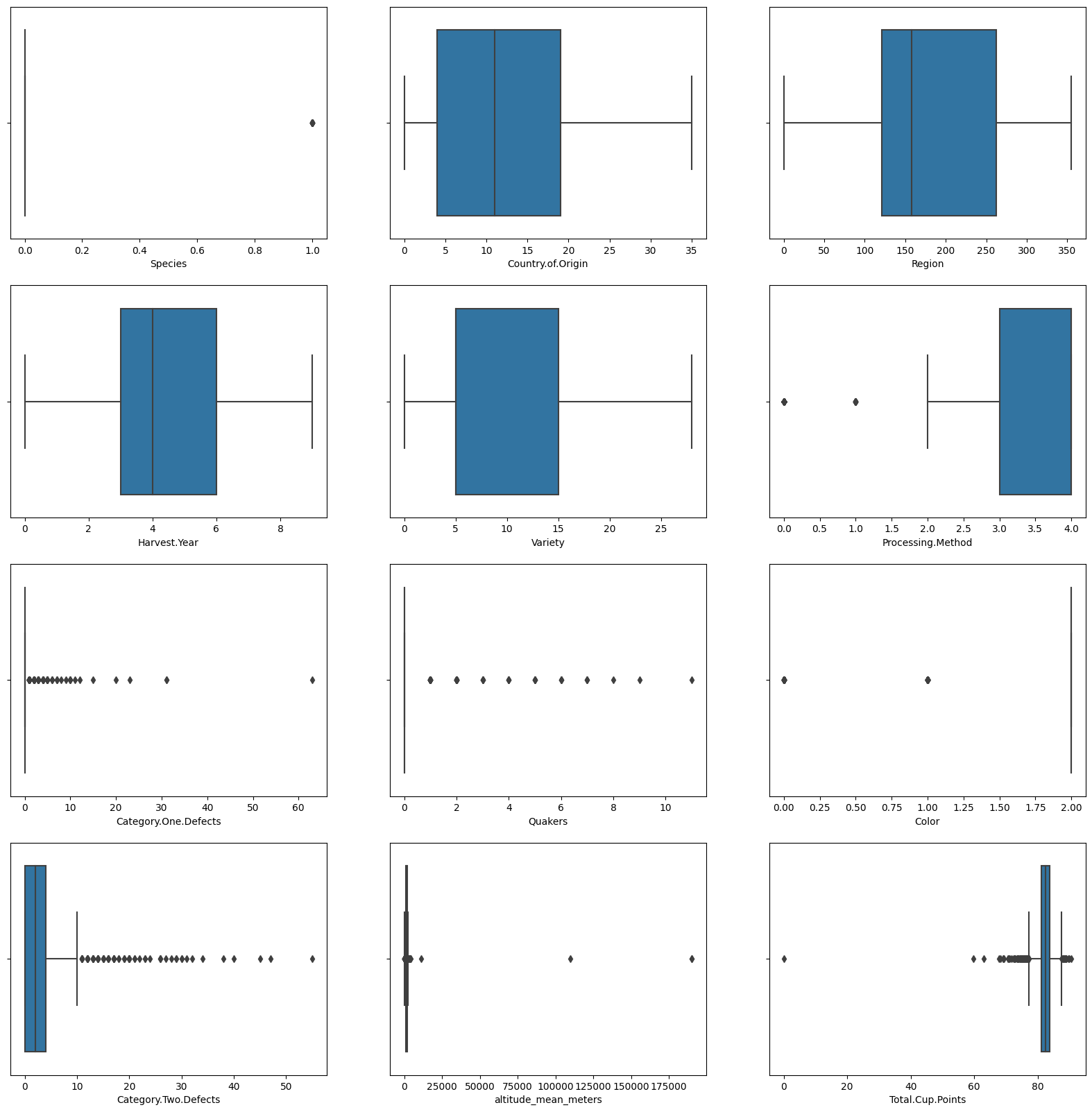


Broj nedostajućih vrednosti po kolonama

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.



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.

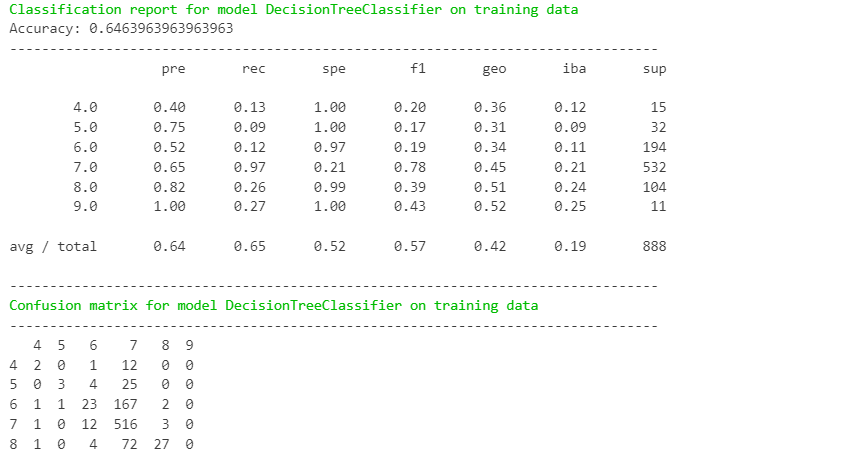
# Klasifikacija

## Stabla odlučivanja

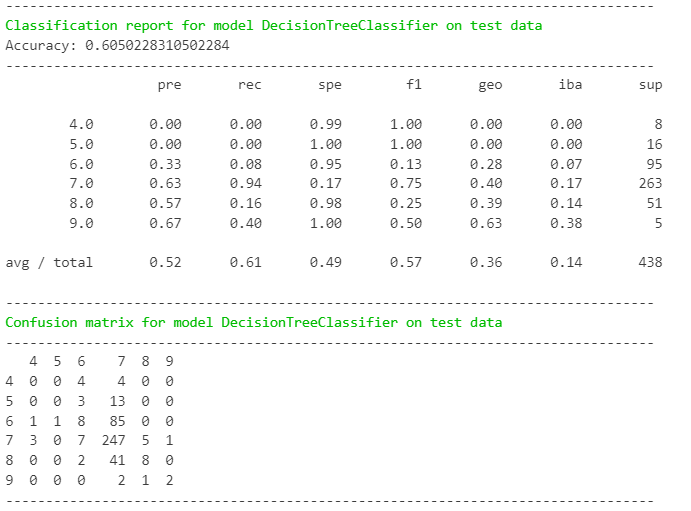
Stabla odlučivanja (decision trees) su neparametarski metod nadgledanog učenja koji se koriste za klasifikaciju i regresiju. Hijerarhijske su strukture, gde koreni i unutrašnji čvorovi predstavljaju upite koji teže da razdvoje skup podataka na distikntne podskupove. U listovima se javljaju svi mogući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

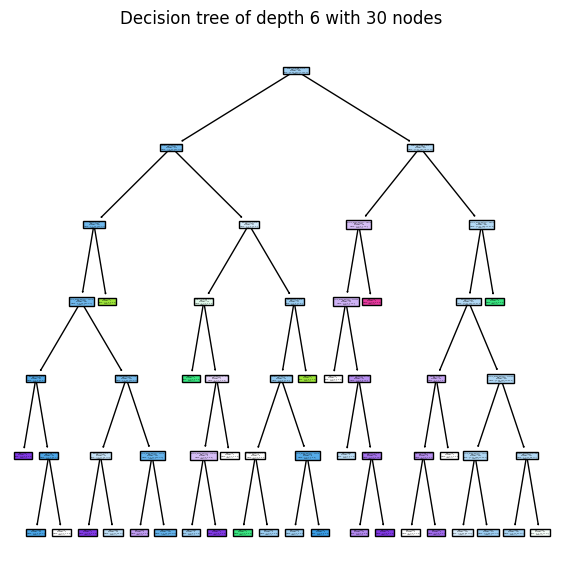


Rezultat primene stabla odlučivanja sa nasumičnim hiperparametrima na nebalansiranom skupu – trening podaci

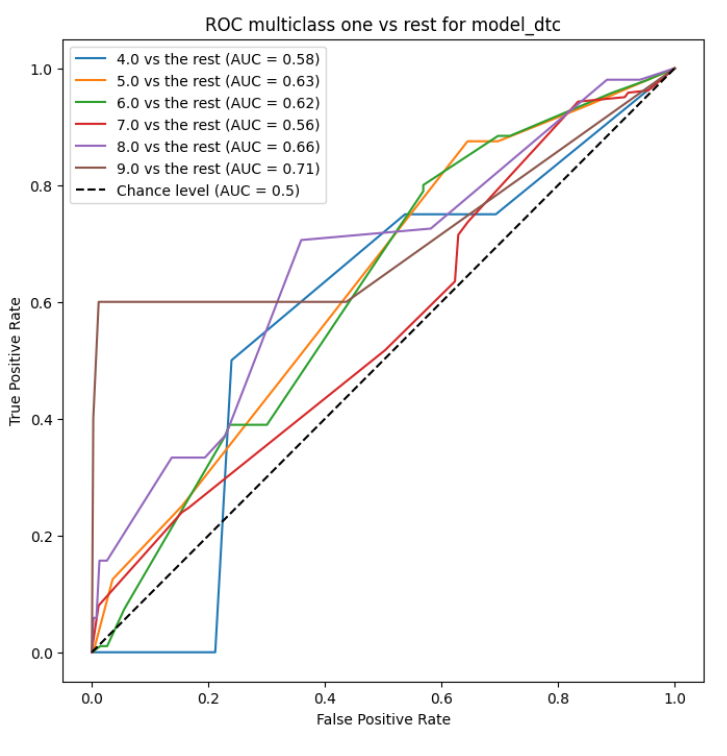


Rezultat primene stabla odlučivanja sa nasumičnim hiperparametrima na nebalansiranom skupu - test podaci

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



Grafički prikaz modela



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.

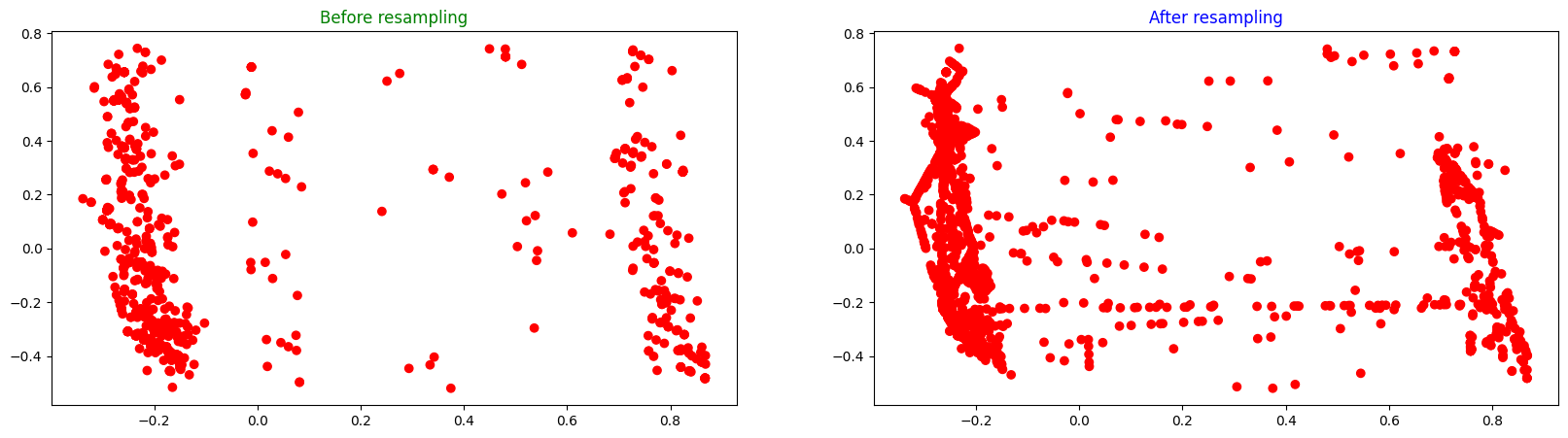
Edited Nearest Neighbor (ENN) radi tako što pronalazi k najbližih suseda svake opservacije, zatim proverava da li je većinska klasa u krugu k najbližih suseda opservacije ista kao i klasa opservacije – ako nije, k najbližih suseda i sama opservacija se brišu iz skupa. Ovaj postupak brisanja se ponavlja 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.

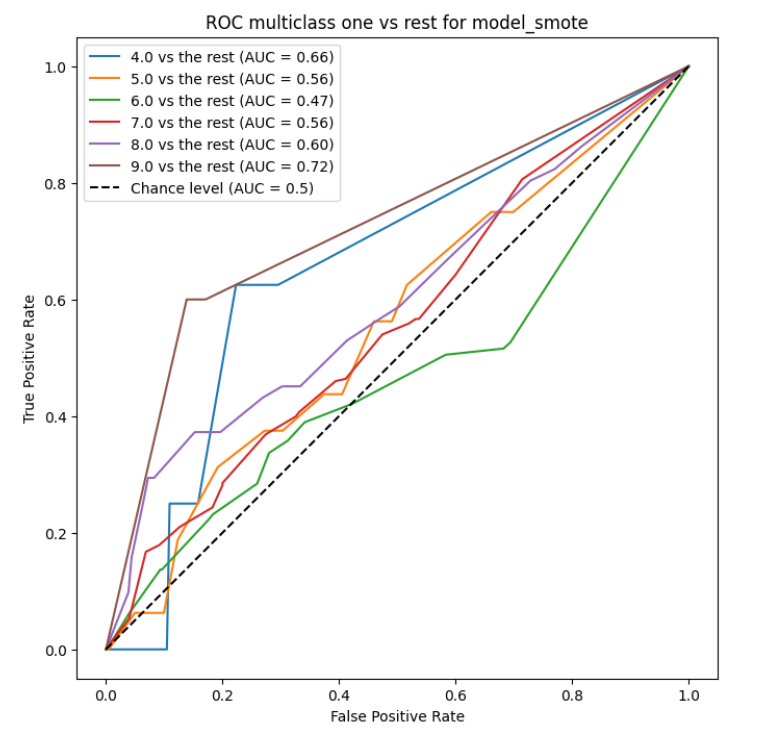
Kako SMOTEENN radi? Nakon primene SMOTE algoritma na podatke, primenjuje se ENN. Kako primena SMOTE algoritma unosi šum u podatke interpolacijom novih tačaka sa marginalnim outlajerima i inlajerima, ENN čisti podatke, te odstranjuje šum.

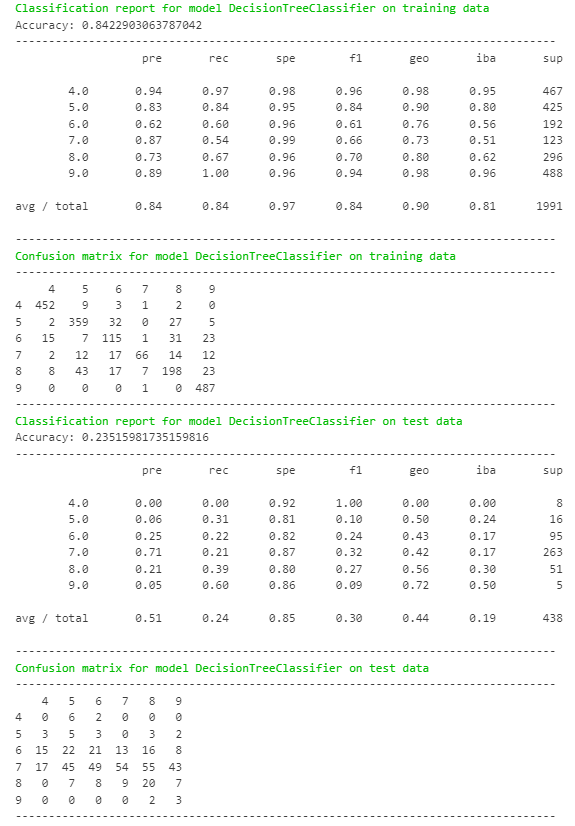
Primenom SMOTEENN-a nismo dobili savršen balans među klasama, ali dizbalans je dovoljno neprimetan.

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.



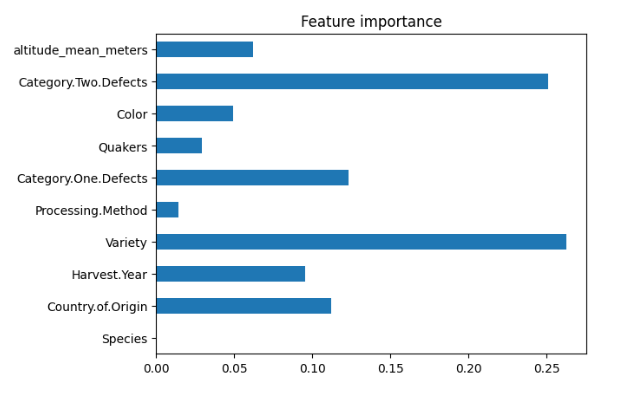
Skup nakon primene SMOTEENN algortima





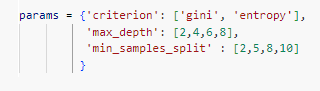
Rezultati primene stabla odlučivanja na balansirane podatke sa nasumičnim hiperparametrima

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.



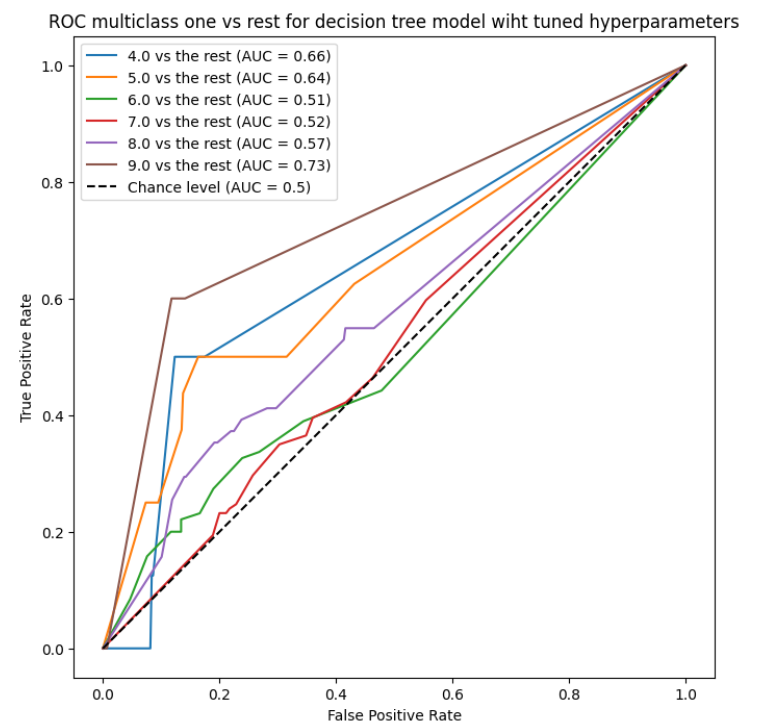
Izabrane moguće vrednosti hiper-parametara

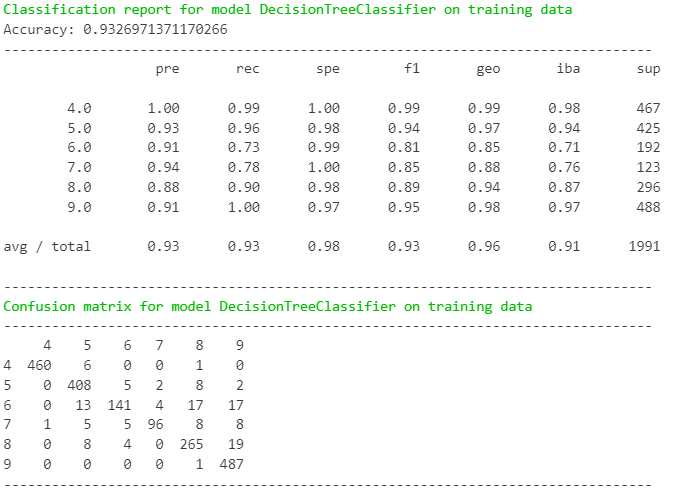
*GridSearchCV*  je alat korišćen za iscrpnu pretragu nad mrežnom hiper-parametara. Zasniva se na korišćenju cross-validacije za ocenu performansi modela. Cross-validacija štiti od preprilagođavanja u prediktivnim modelima, naročito kad je skup podataka limitiran – kao ovaj.

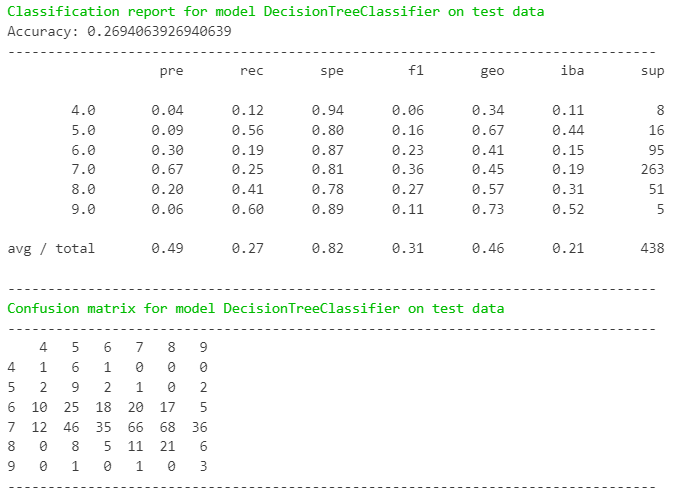


Izabrani hiper-parametri modela sa najboljim performansama i njegova ocena

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).







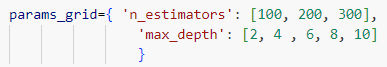
Rezultati modela sa podešenim hiper-parametrima

## 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.

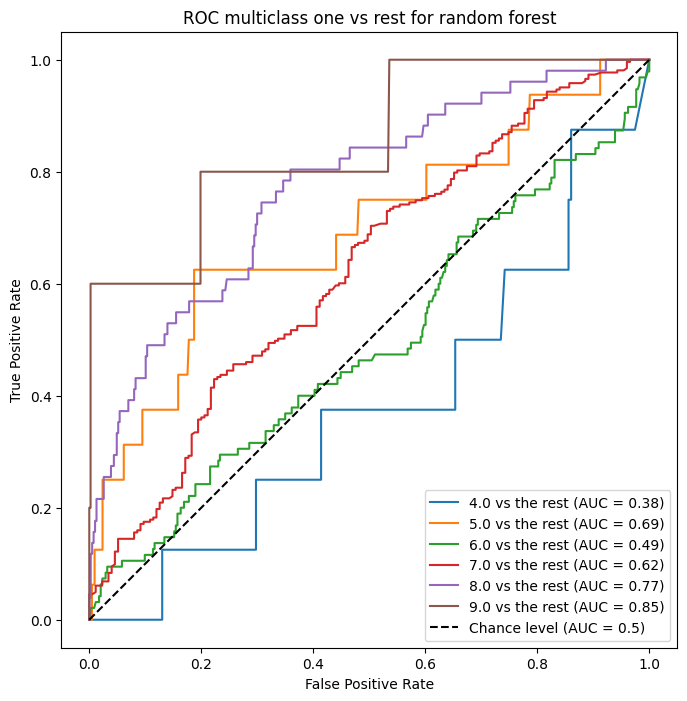
Za podešavanje hiper-parametara korišćen je ponovo GridSearchCV.

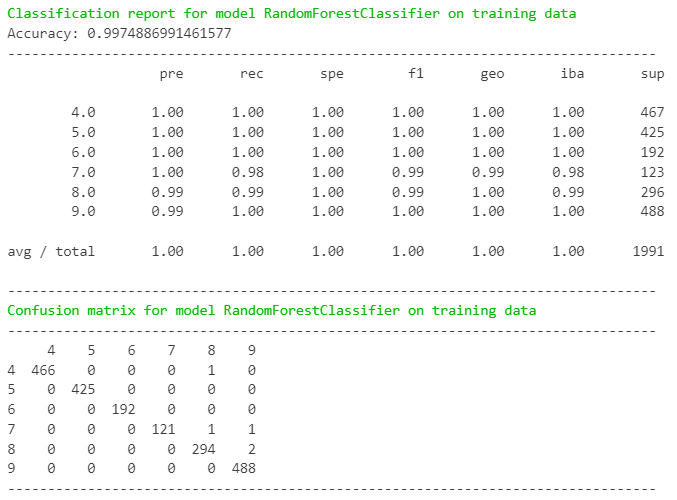


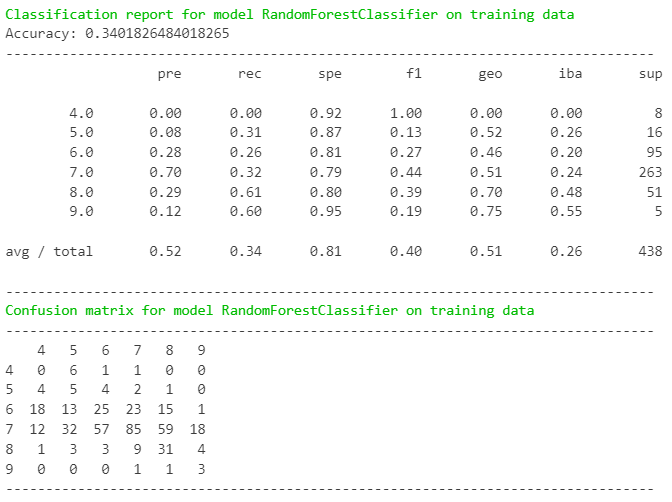
Prosleđeni hiper-parametri



Izabrani hiper-parametri







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.

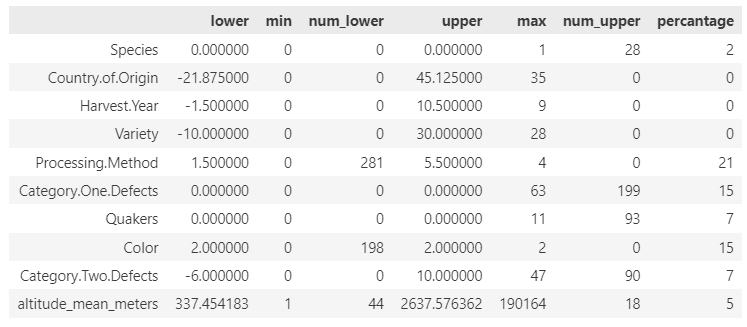
## K-Nearest neighbours

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.



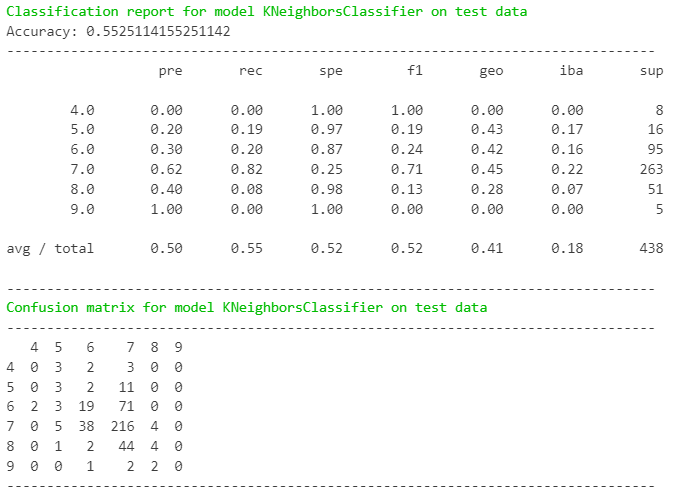
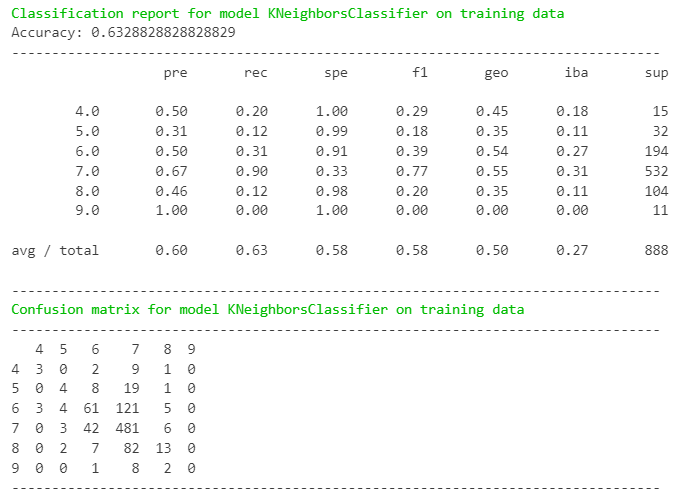
20 Outlajeri u skupu

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.

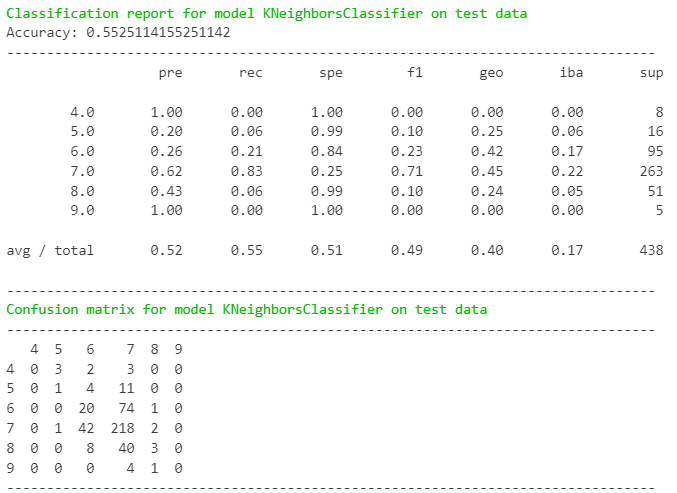
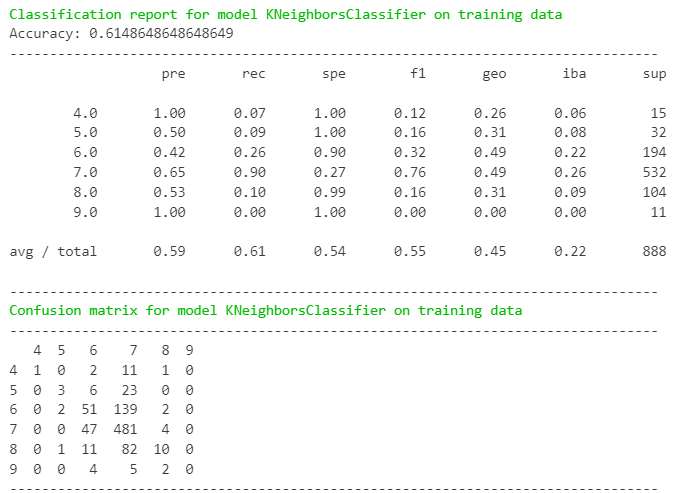
Za pristup bez OneHotEncoding-a koristićemo MinMaxScaler za skaliranje podataka.

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:



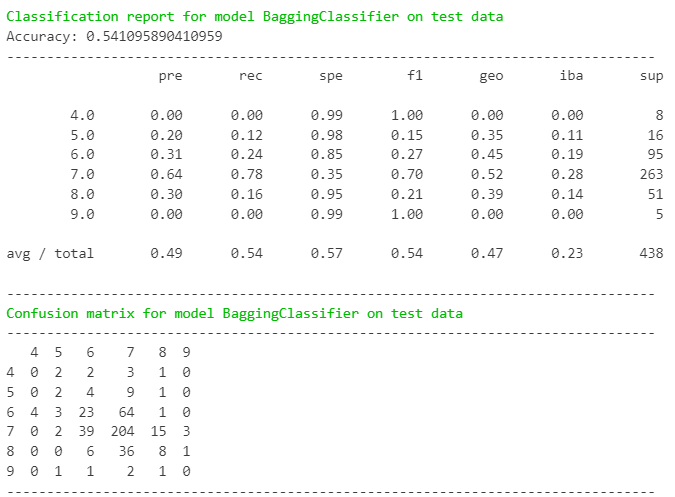
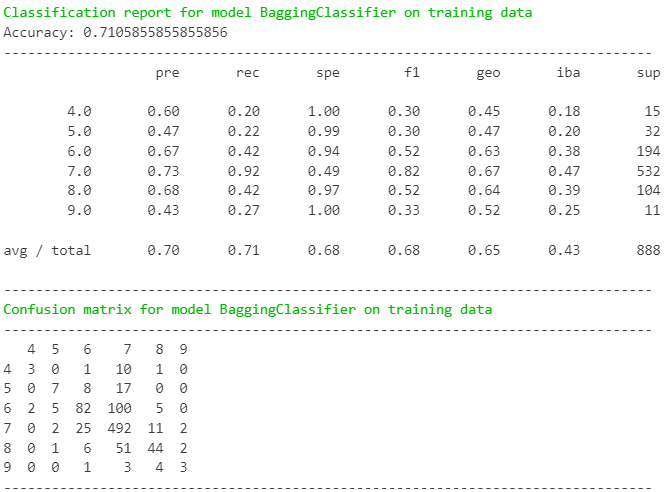
21 KNN na nebalansiranim podacima, k=10



22 KNN sa podešenim hiper-parametrima na nebalansiranim podacima

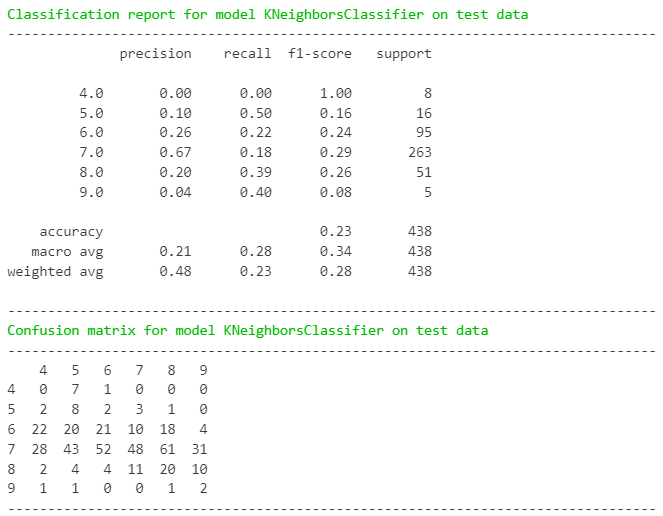
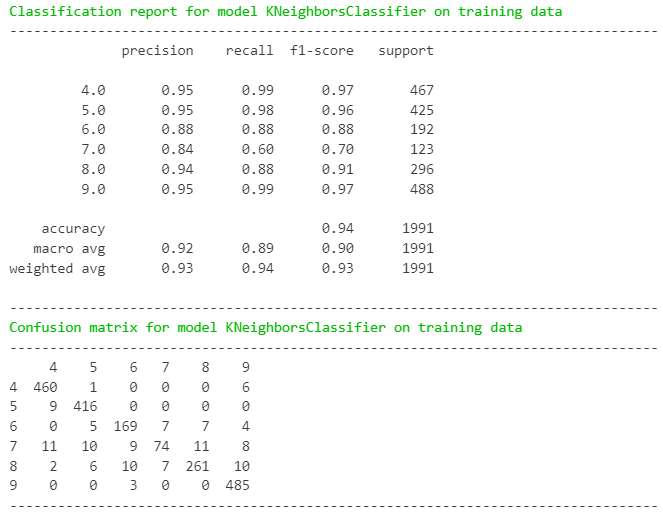


23 Izabrani hiper-parametri i CV ocena prethodnog modela

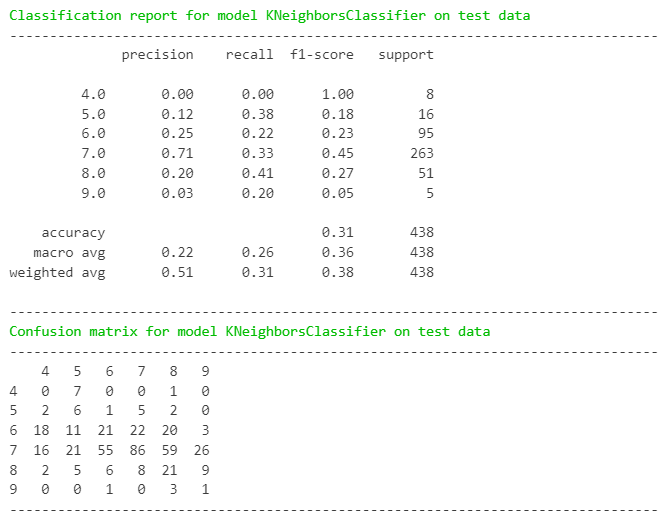
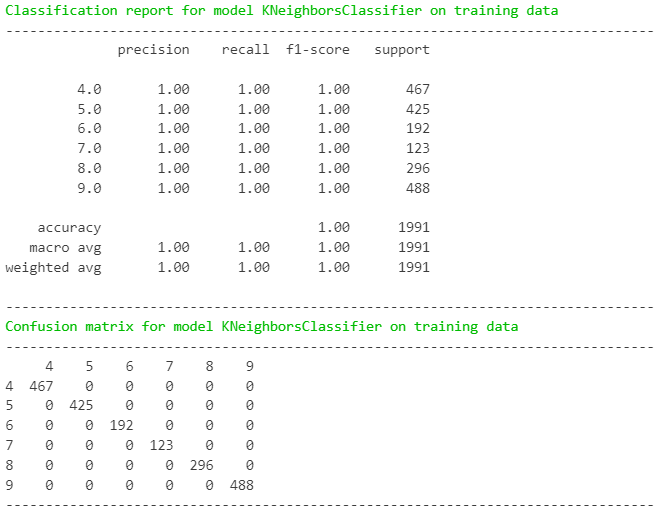


24 KNN bagging classifier na nebalansiranim podacima

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



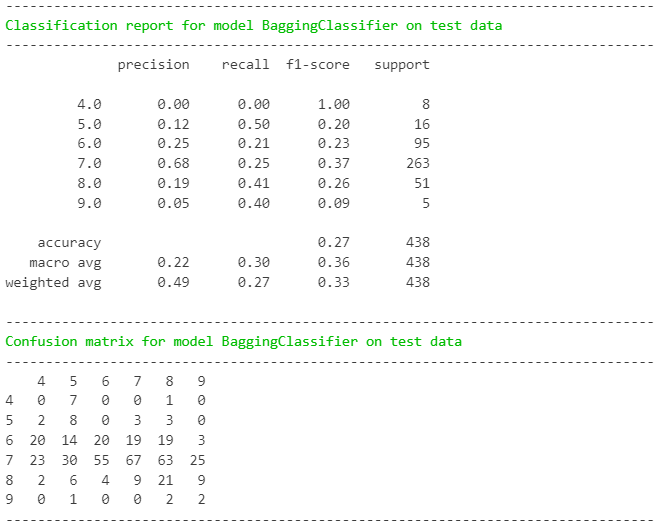
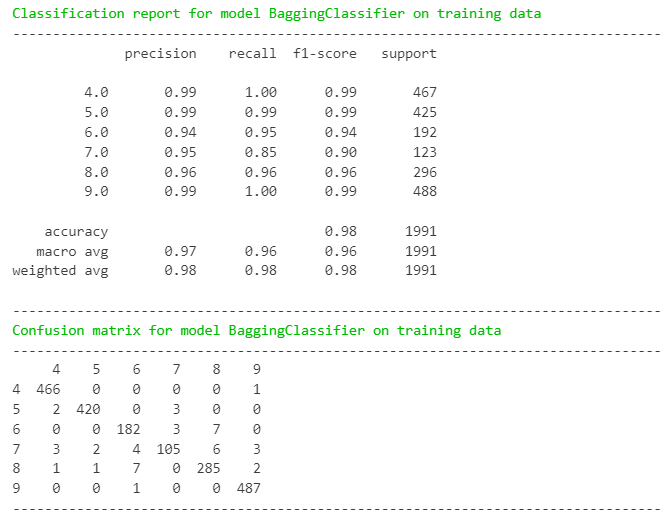
25 KNN na balansiranim podacima, k=10



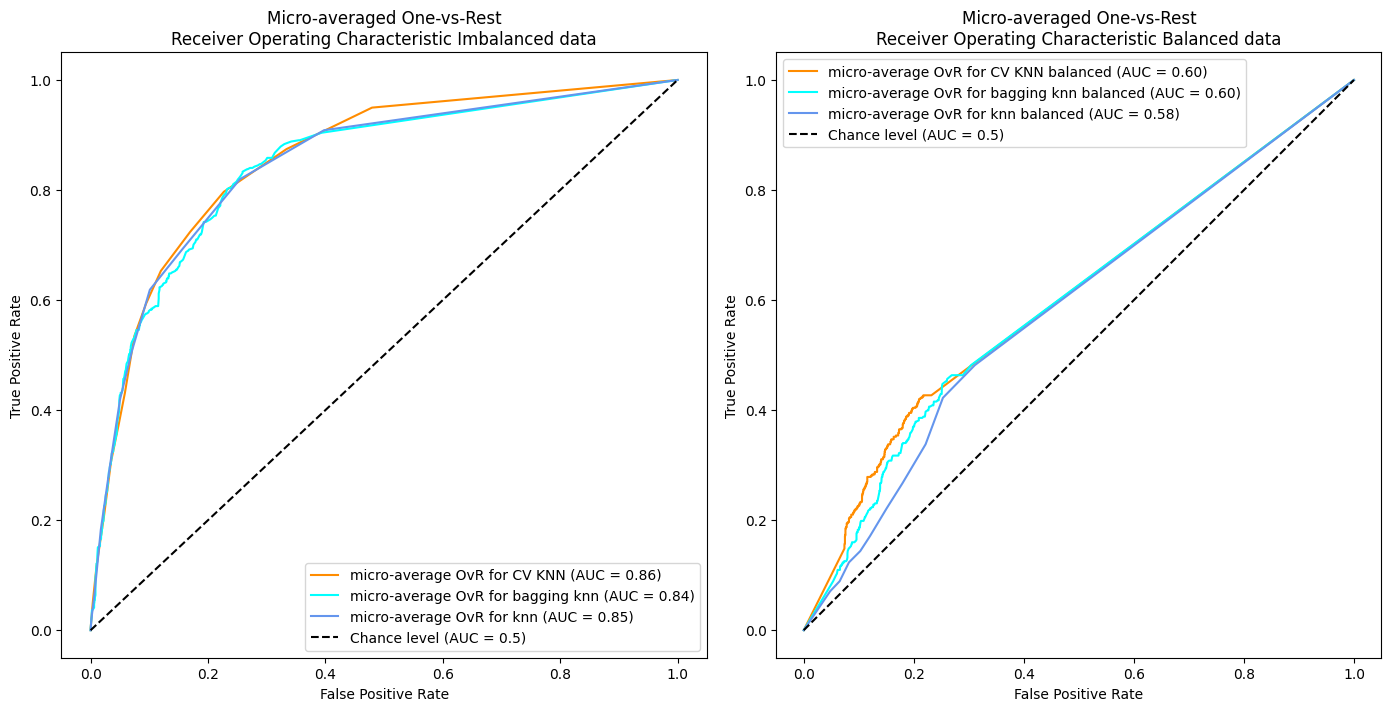
26 KNN sa podešenim hiper-parametrima na balansiranim podacima

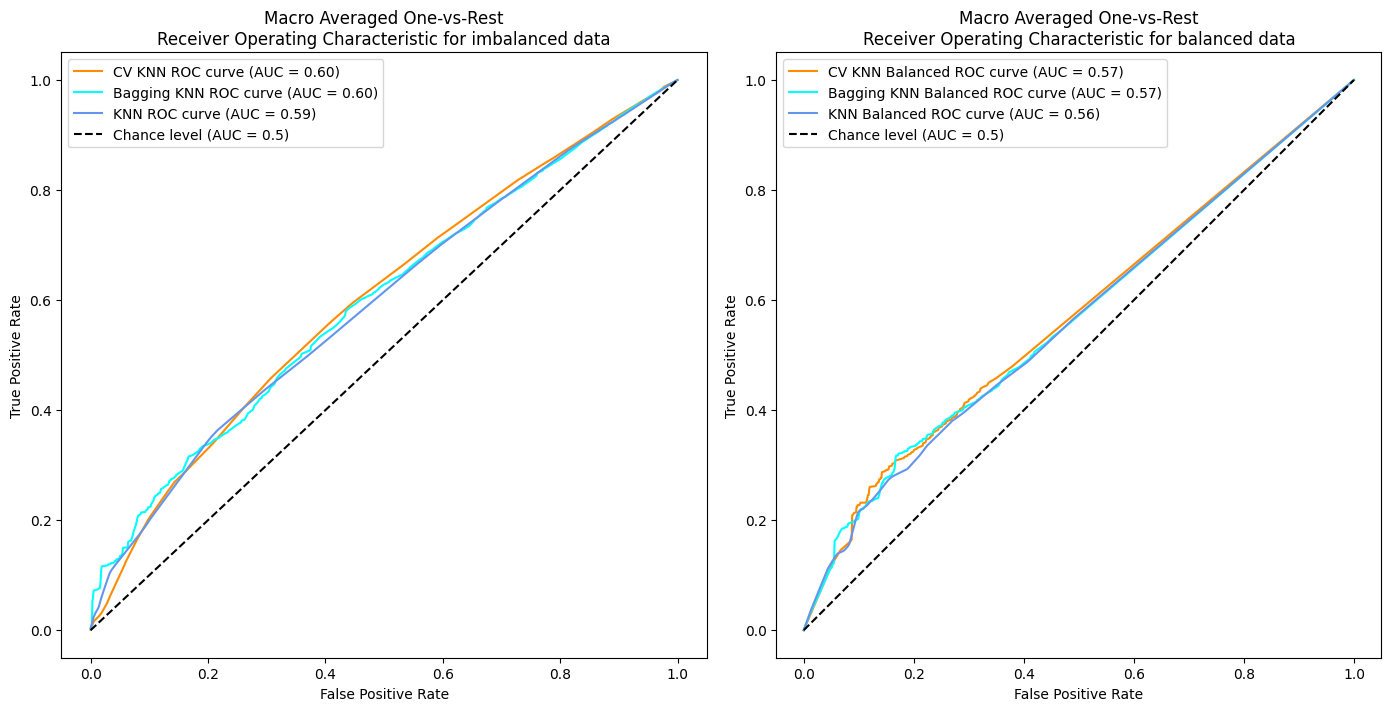


27 Izbor hiper-parametara i CV ocena prethodnog modela



28 KNN bagging classifier na balansiranim klasama



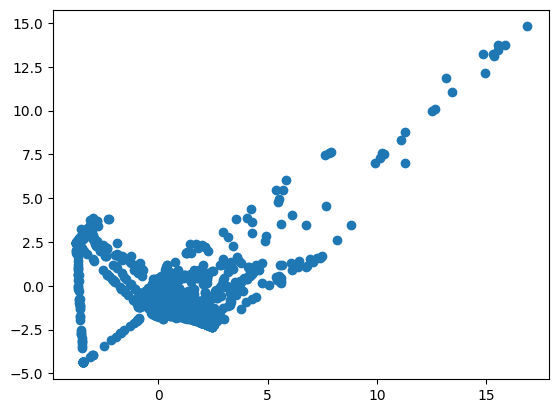


29 ROC krive zasnovane na micro i macro uprosečavanju

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.

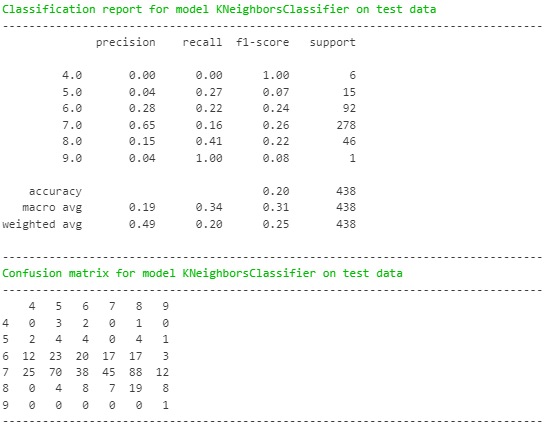
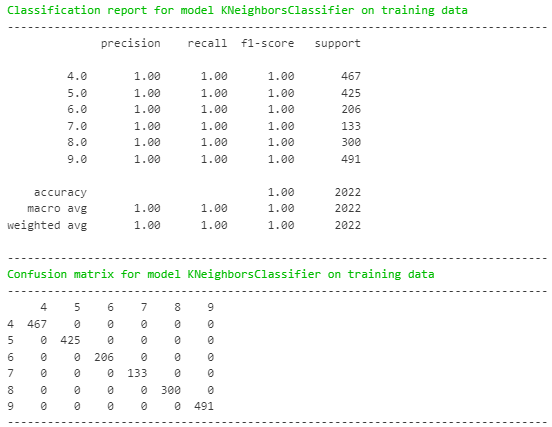
Pokušajmo sada pristum sa OneHotEncoding-om.



30 Izgled balansiranog trening skupa nakon primene OneHotEncoding i PCA tehnika

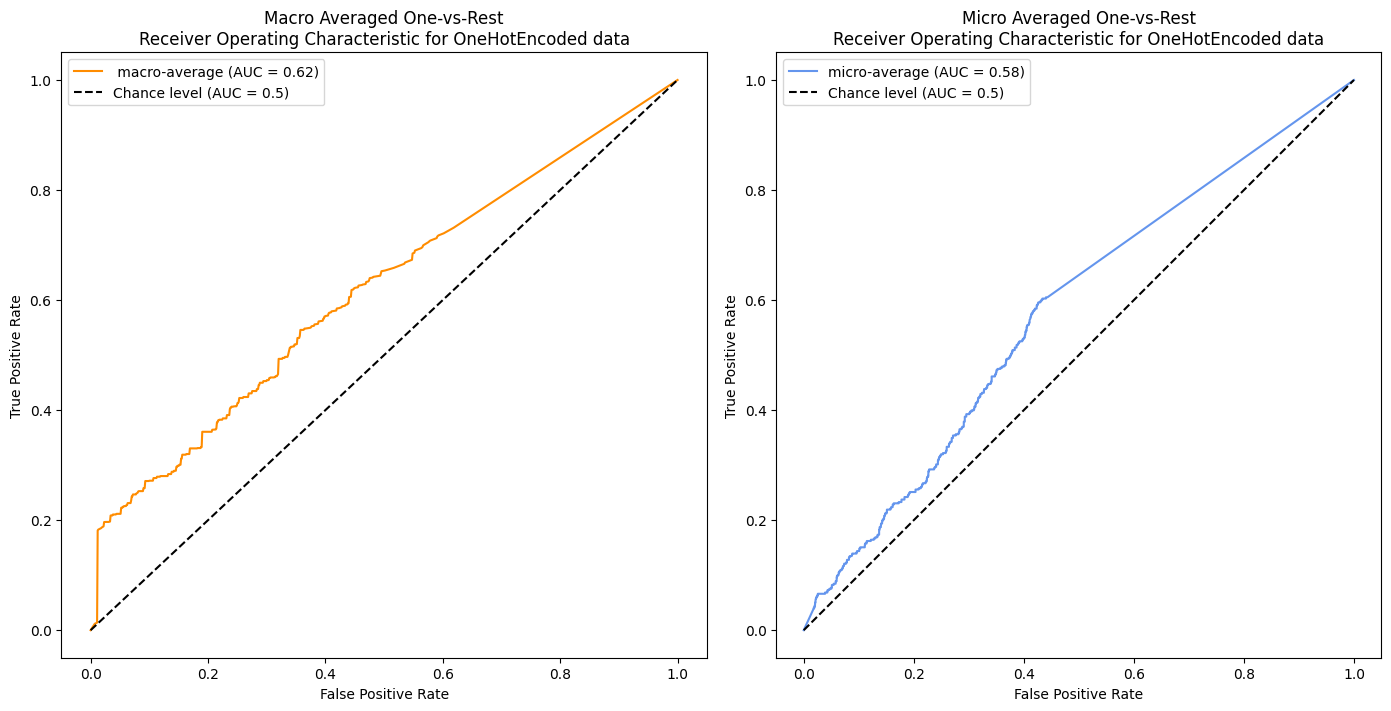


31 Izbor hiper-parametara i CV ocena modela



32 Rezultati primene prethodnog modela

Model se ponaša slično kao i prethodni.



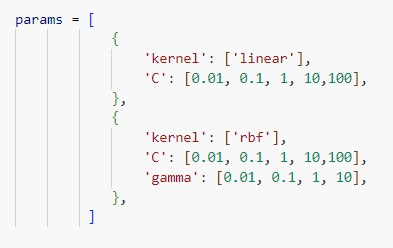
33 ROC krive prethodnog modela

## 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.

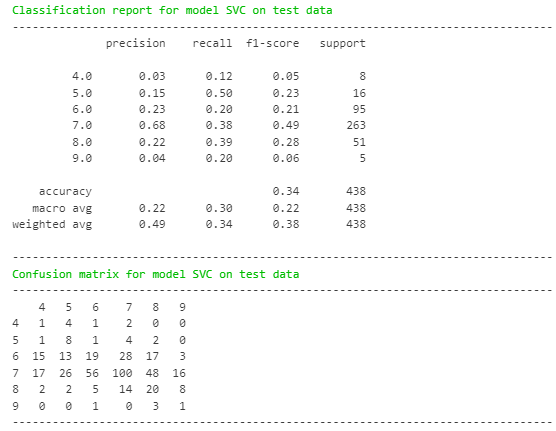
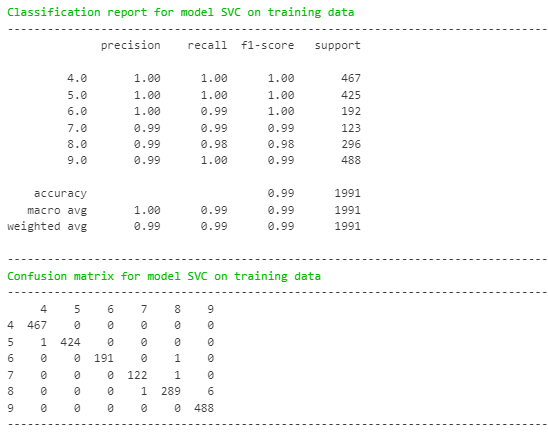


34 Mogući parametri prosleđeni GridSearchCV



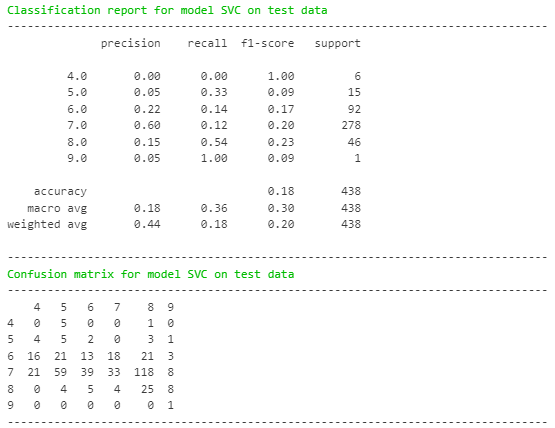
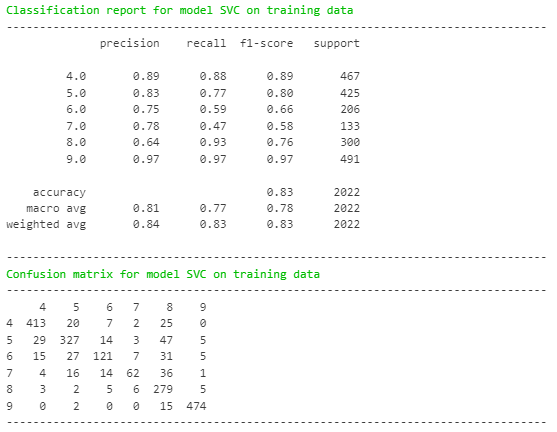


35 Izabrani parametri i ocena



Rezultati datog modela

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



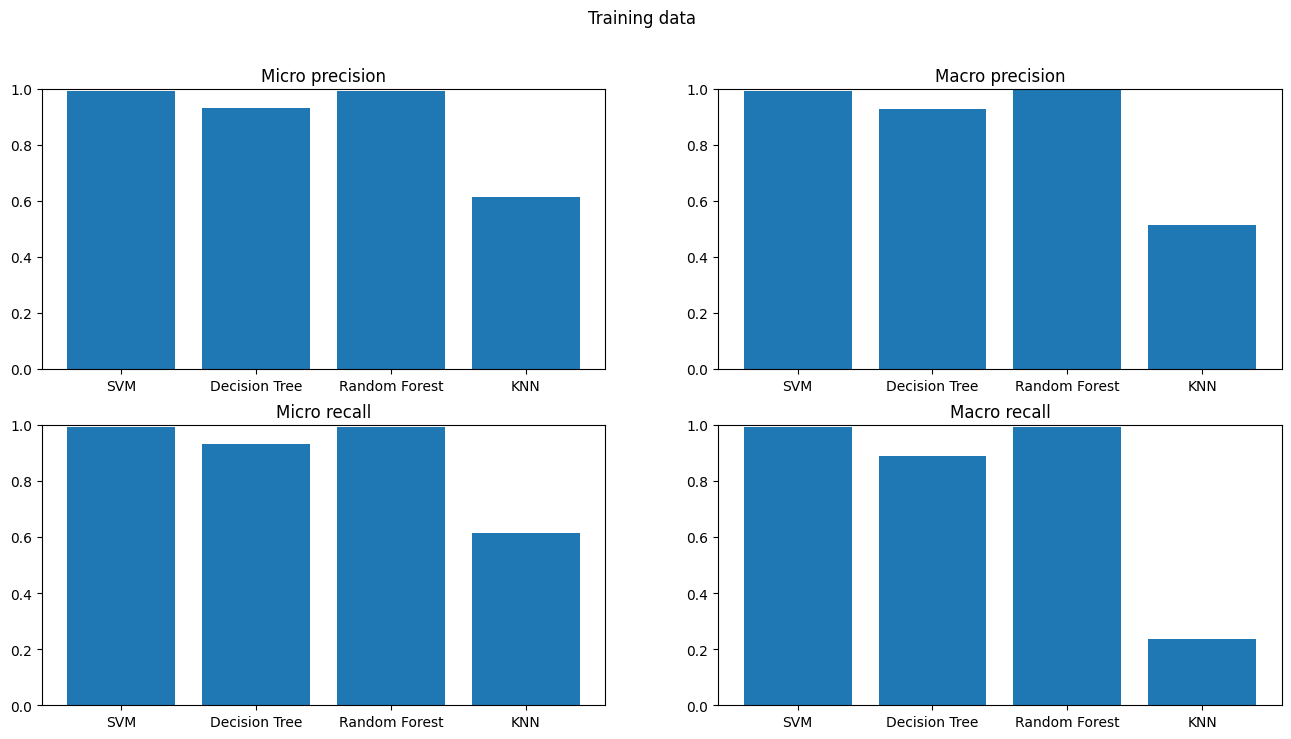
Parametri dobijem CV tehnikom: {'C': 100, 'gamma': 10, 'kernel': 'rbf'}

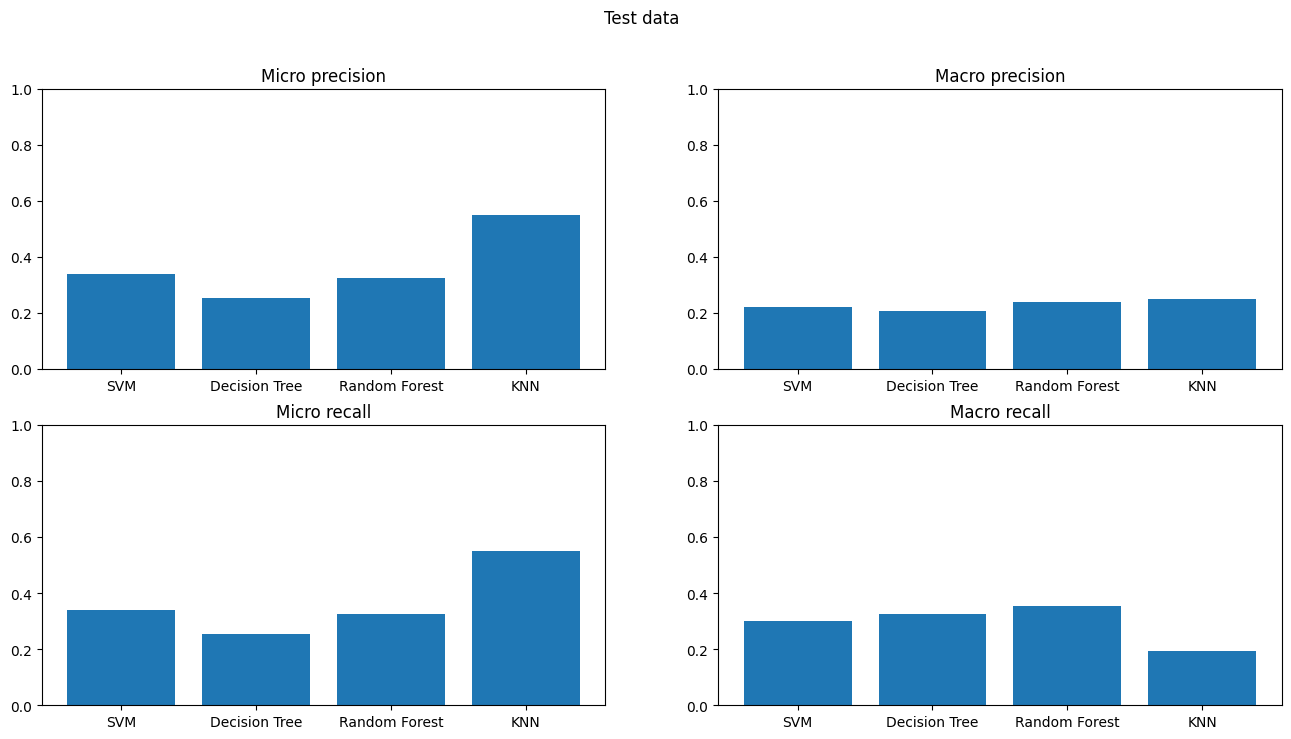
Cilj je bio da pokažemo da se nije napravila neka značajna razlika, tj. da manipulisanjem kategoričkim atributima ipak ne možemo u potpunosti da simuliramo svojstva koja ispoljavaju pravi numerički atributi.

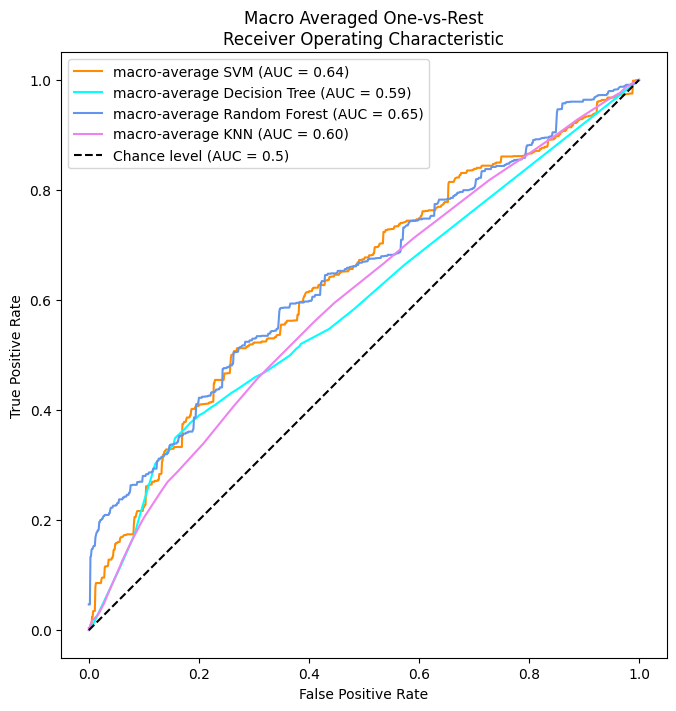
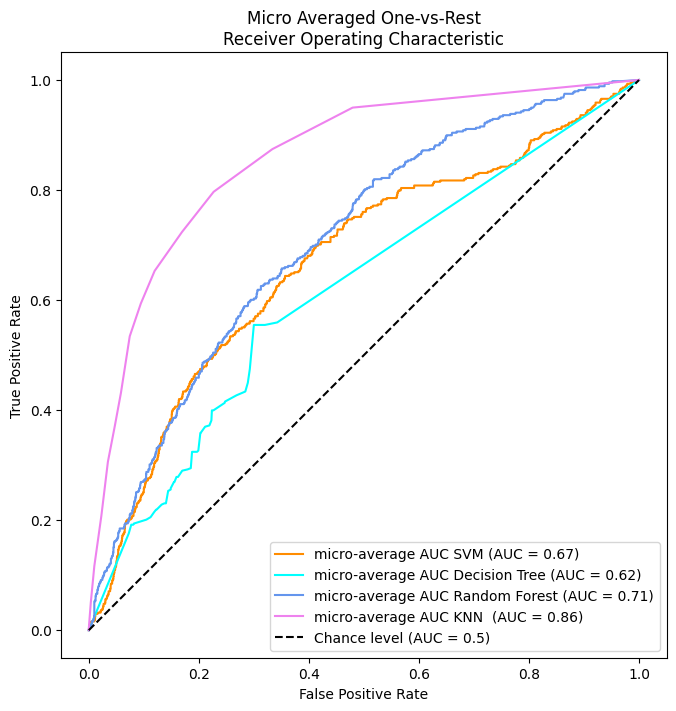
## poređenje modela

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.

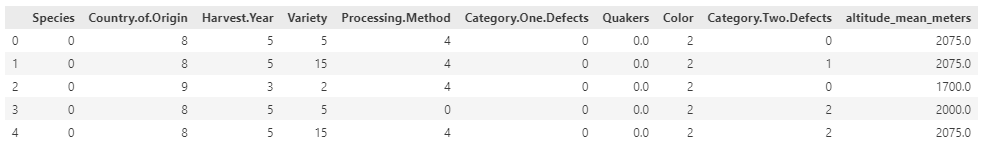
Jedna od verovatnih mana jeste ogroman broj kategorija, što otežava posao modelima zasnovanim na odlulivanju, ali primena takvih modela je daleko jednostavnija nego forsiranje numeričkih osobina na podatke koji su čisto kategorički.

# Klasterovanje

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

## Kmeans

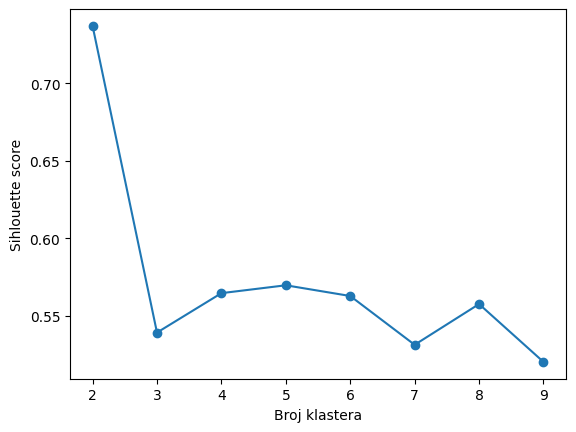
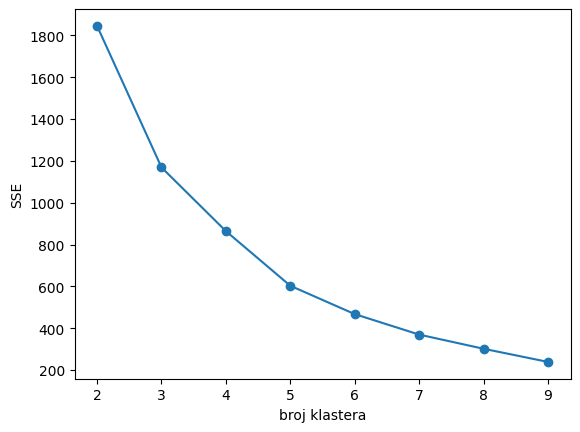
Za KMeans algoritam broj klastera mora biti unapred poznat, što može ujedno da se interpretira i kao mana i kao prednost. Ovo je jednostavna tehnika koja se bazira na pronalaženju reprezentativnih predstavnika, konkretno u ovoj implementaciji to je centroid.



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.

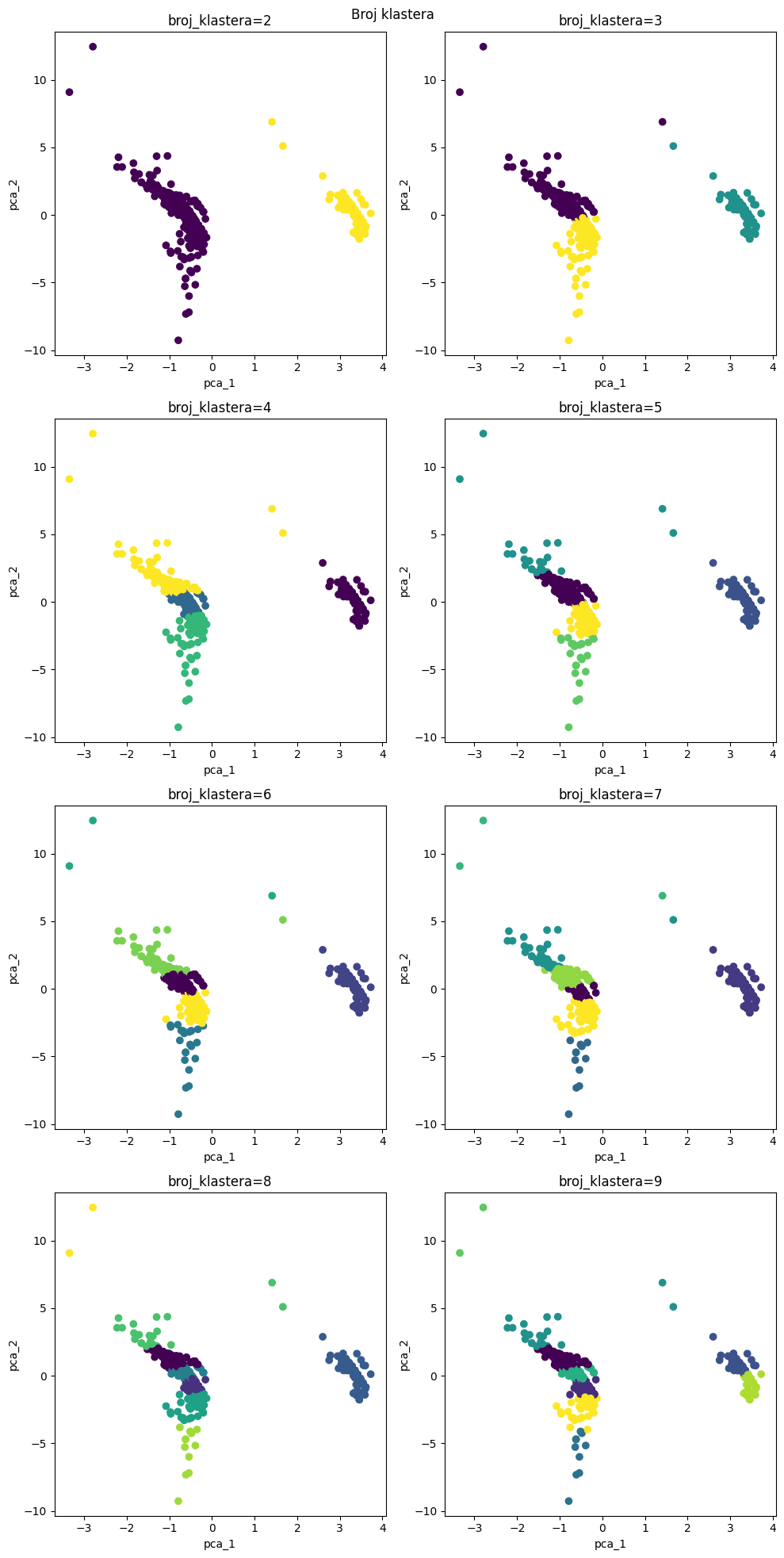


SSE i koeficijent senke sa porastom broja klastera

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.

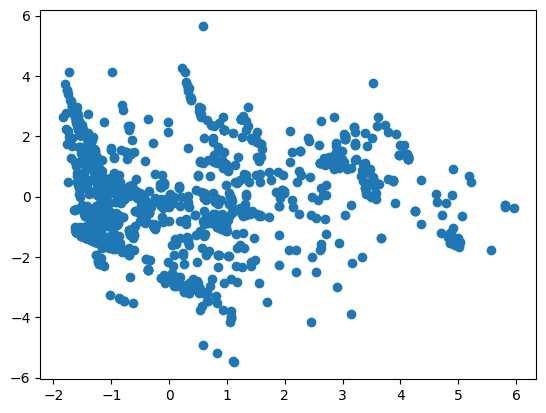
Pravilo lakta je heuristika kojom se traži tačka “na laktu” gde SSE najbrže opadne, da bi se dobila optimalna vrednost parametra K.

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

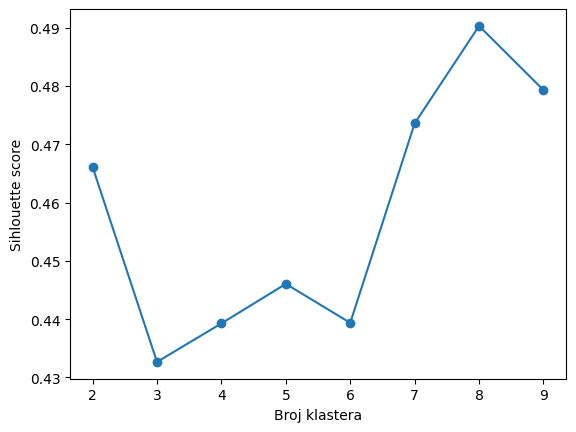
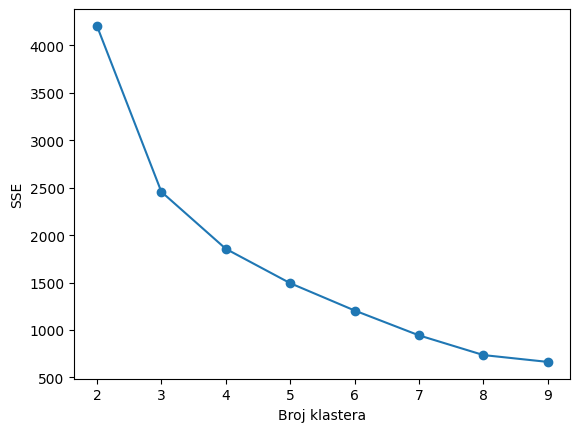


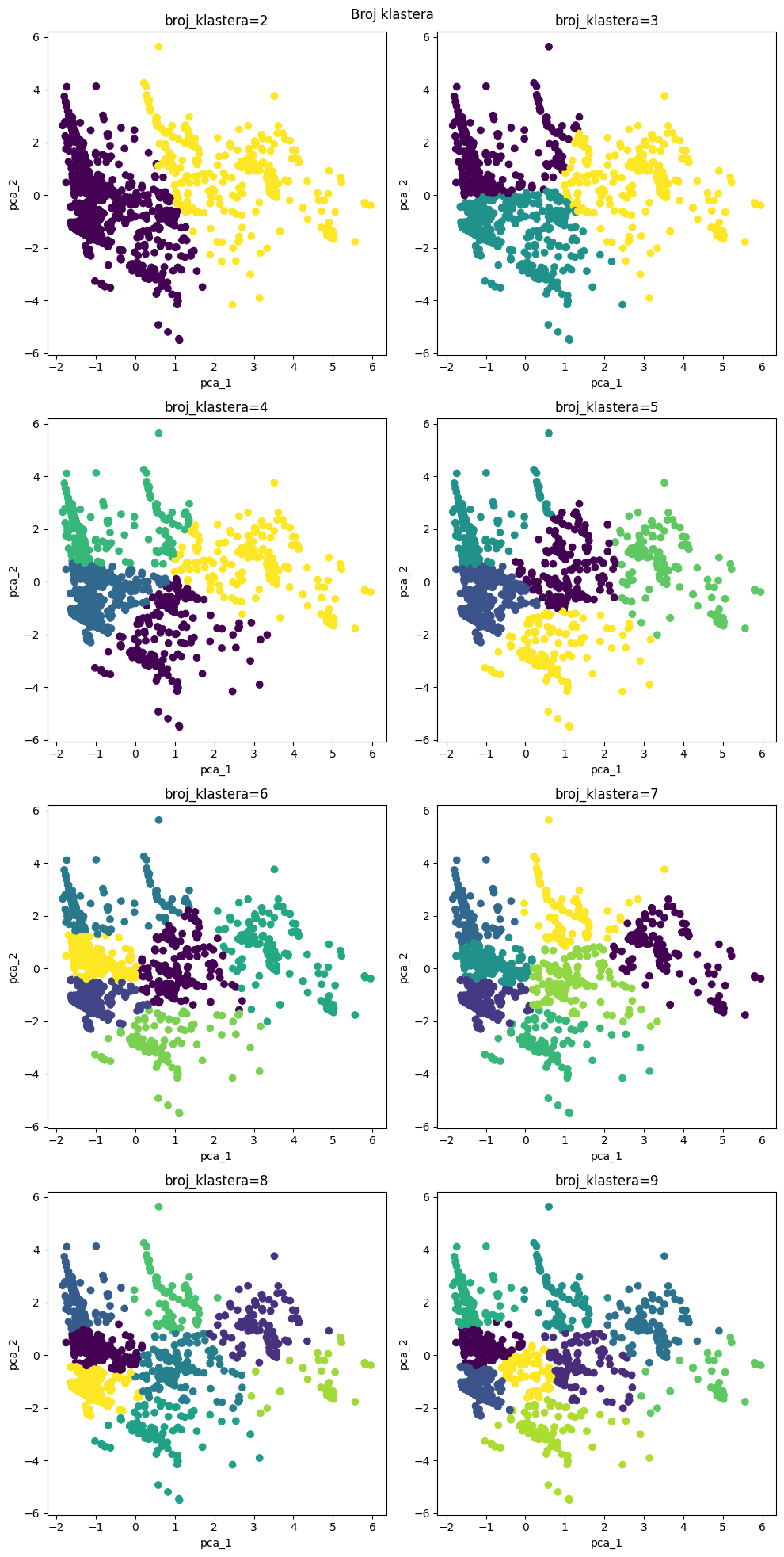
Na osnovu pravila lakta, mogli bismo da kažemo da je optimalna vrednost 3, a može i 4 – promena nije velika i ovaj pristup je jako subjektivan. Nasuprot njemu, na osnovu koeficijenata senke, najveća vrednost se dobija za dva klastera - što–se podudara sa vizuelizovanim podacima.

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.



Vizuelizacija podataka



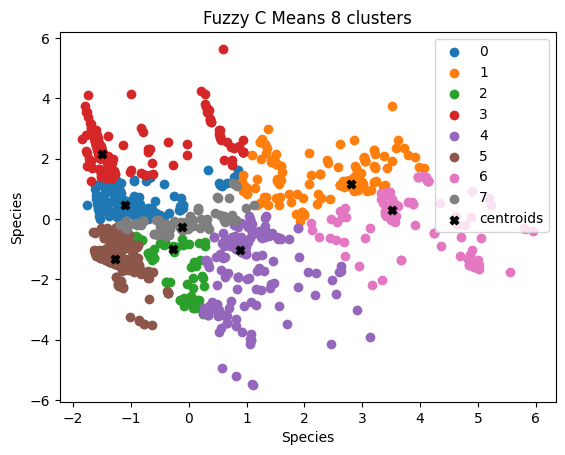


Koeficijent senke se maksimizuje za 8 klastera što izgleda kao razuman izbor.

## fuzzy cmeans

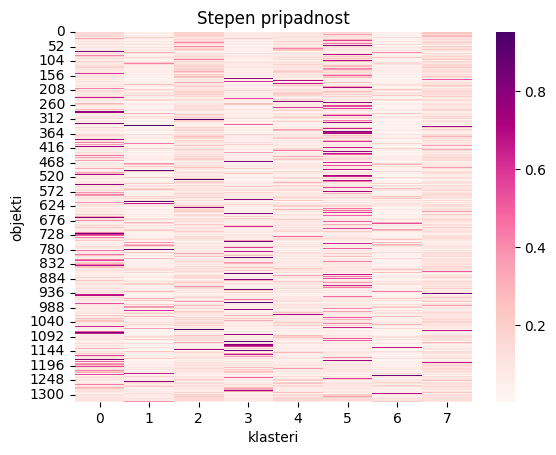
Ovo je predstavnik *soft clustering* algortiama koja omogućava da instance pripada većem broju klastera. Na osnovu težinskih sumi rastojanja se odrežuje pripadnost svakom od C klastera.

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.

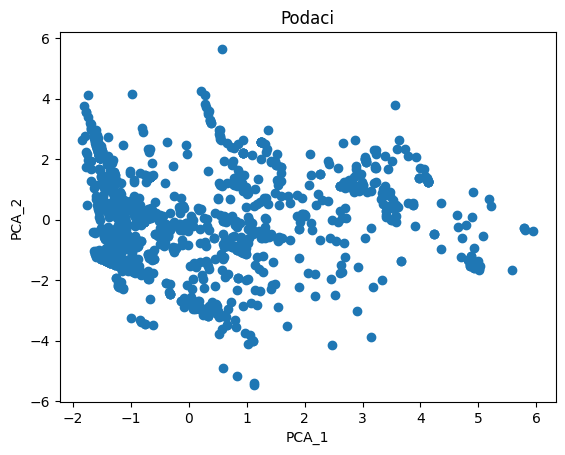


U ovom modelu dobijen je koeficijent senke od 0.65 što je prihvatljivo.

## 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.

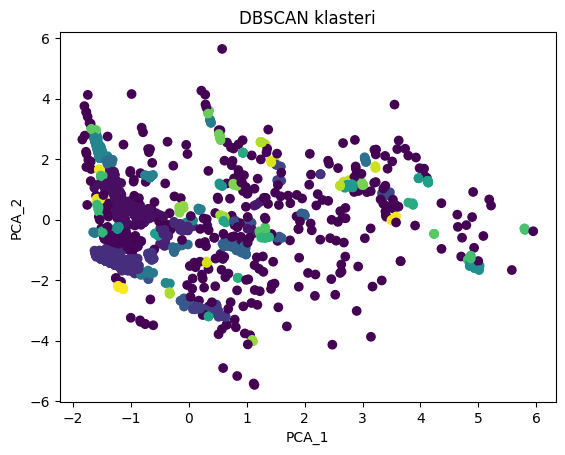
U ovom primeru ćemo klasterovanje primeniti na OneHotEncoded kategoričke atribute, sa smanjenom dimenzionalnošću na 2.



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.



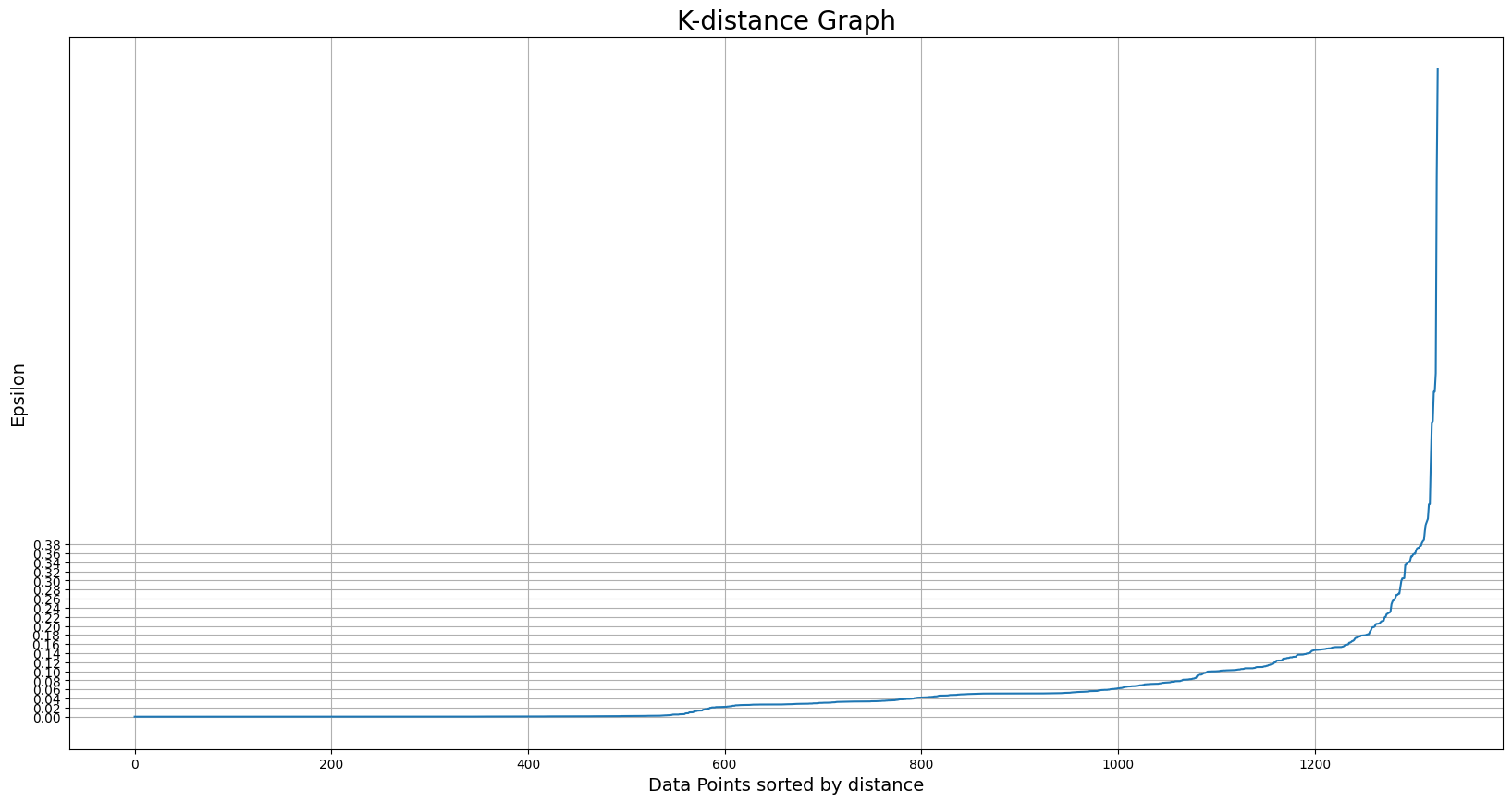
43 Klasteri dobijenim DBSCAN algortimom sa parametrima eps=0.1 i minPoints=3

Imamo dosta klastera koji se preklapaju.

Kako izabrati optimalne parameter?

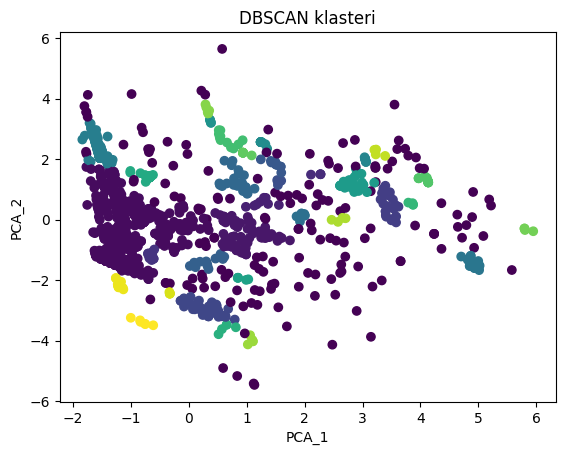
MinPoints bi trebalo da bude bar za jedan veći od broja dimenzija podataka. Često se uzima broj dimenzija puta dva.

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.

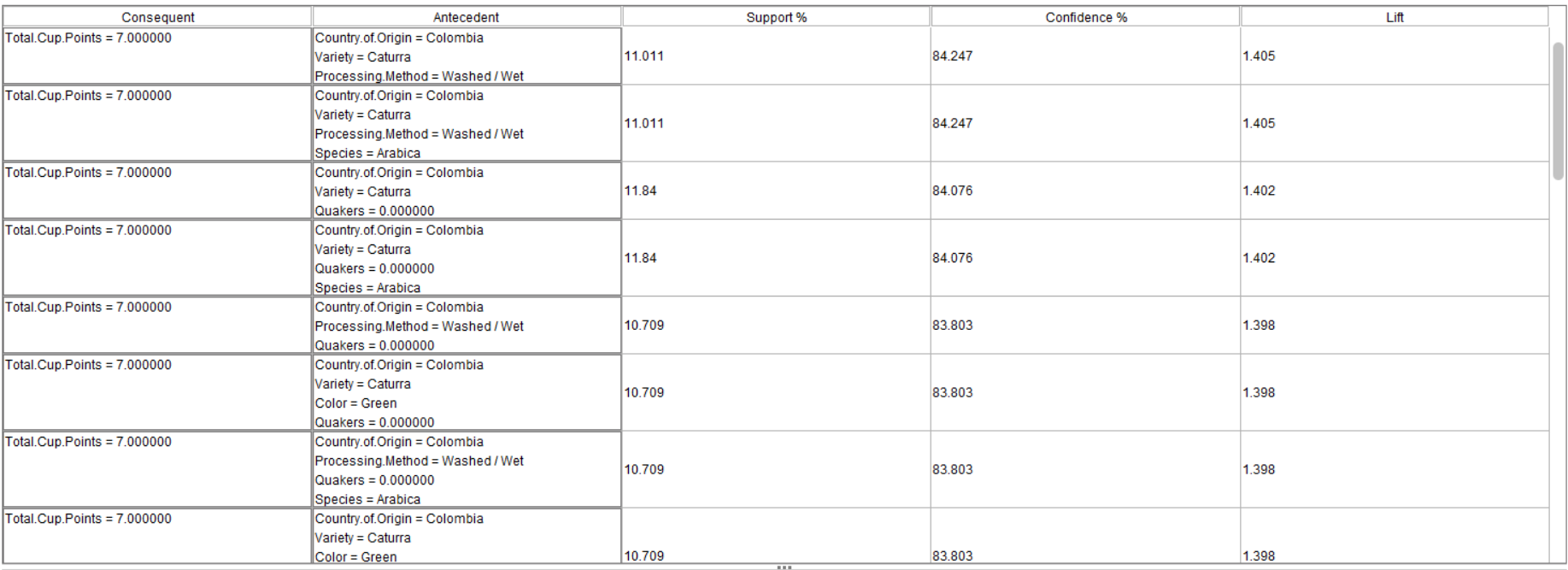
# Pravila pridruživanja

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.

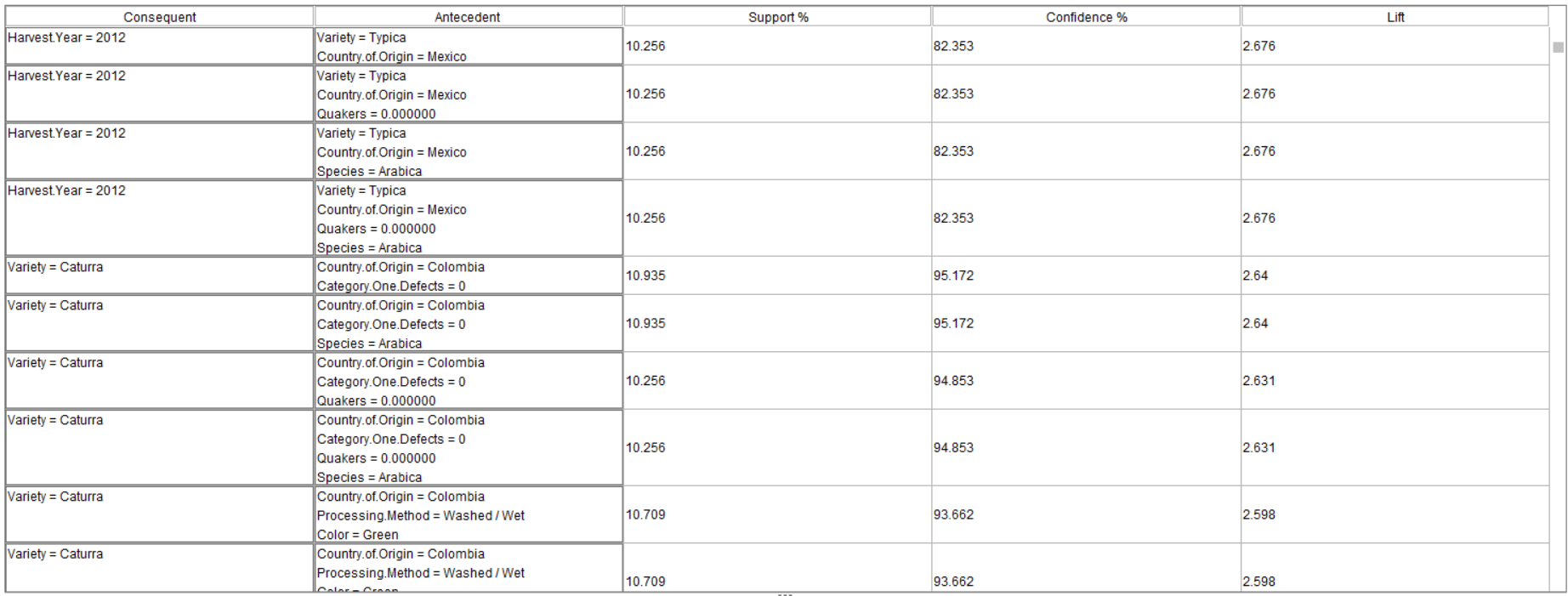


45 Pravila pridruživanja tipa telo -> Total Cup Points

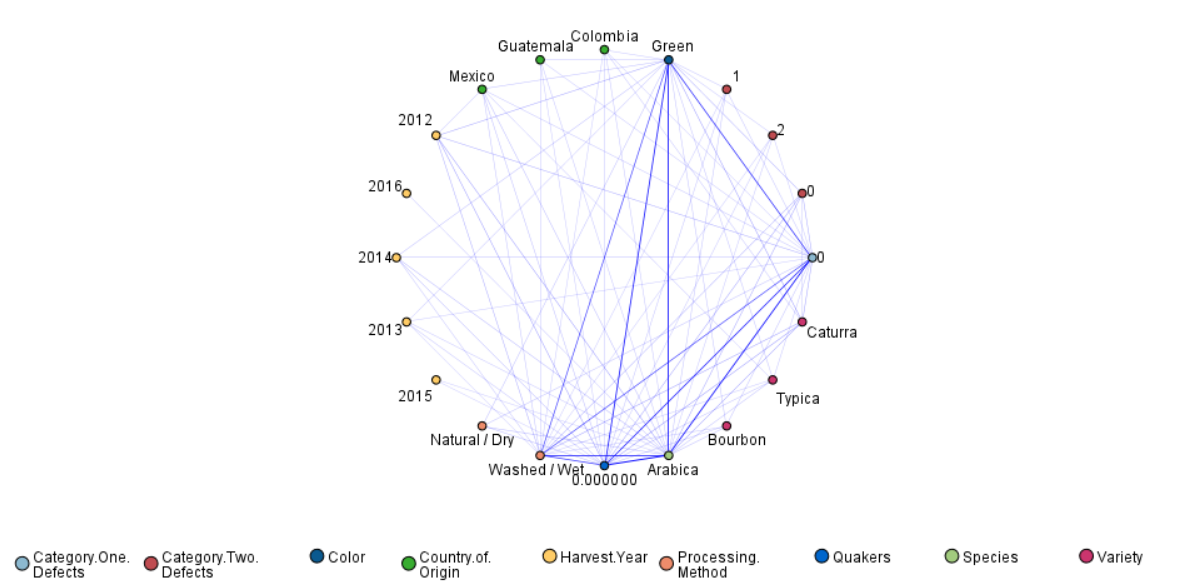
Prona]eno je 32 pravila i sva su vezana za ocenu 7 što ih ne čini naročito interesantnim, mada je lift svuda veći od 1.

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



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.