

Istraživanje skupa podataka “Anuran Calls”

02.06.2023

—
Student: Stefanija Marković

Profesor: dr Nenad Mitić

Asistent: Stefan Kapunac

Matematički fakultet, Univerzitet u Beogradu

Uvod

Svrha rada je demonstracija metoda klasifikacije i klasterovanja različitih vrsta žaba koristeći snimljene audio zapise njihovih oglašavanja.

Skup podataka je kreiran segmentacijom 60 audio zapisa zvukova žaba iz četiri različite porodice, osam rodova i deset vrsti. Svaki zapis odgovara jedinki, i njegov redni broj je zabeležen u zasebnoj koloni. Nakon segmentacije dobijeno je 7195 slogova, koji predstavljaju instance u skupu podataka. Za svaki slog izračunato je 22 kepstralnih koeficijenata mel skale (MFCC) koji su, zatim, normalizovani.

Skup ukupno sadrži 26 kolona, od kojih 22 predstavljaju odgovarajuće MFC koeficijente i numeričkog su tipa, jedna predstavlja pripadnost odgovarajućem audio zapisu i uzima vrednosti od 1 do 60, i tri kolone koje govore o porodici, rodu i vrsti jedinke, redom.

Potencijalni problem je interpretacija atributa i, radi boljeg razumevanja i korišćenja skupa podataka, treba da se osvrnemo na njihovo značenje. Kepstralne karakteristike sadrže informacije o brzini promene u različitim opsezima spektra. Možemo razlikovati uticaj glasnih žica i vokalnog trakta jer se niskofrekventna ekscitacija i formantno filtriranje vokalnog trakta nalaze u različitim delovima kepstralnog domena. Koeficijenti prvog reda predstavljaju raspodelu spektralne energije između niskih i visokih frekvencija (obično niske frekvencije predstavljaju sonorantan zvuk, a visoke frikativan). Koeficijenti nižeg reda sadrže većinu informacija o ukupnom spektralnom obliku funkcije prenosa izvor-filter, dok koeficijenti višeg nivoa predstavljaju povećane nivoe spektralnih detalja. U proseku je optimalno koristiti između 12 i 20 kepstralnih koeficijenata za analizu zvukova. Iz tog razloga, ne možemo odmah odlučiti koji su atributi najvažniji.

Klasifikacija

Uvod

Cilj ovog dela istraživanja je poređenje performansi različitih klasifikacionih modela. Istraživanje se sastoji iz više koraka: analiza skupa podataka, preprocesiranja podataka, korišćenje modela i poređenje modela.

Za potrebe ovog istraživanja odabrani su sledeći modeli: stablo odlučivanja (Decision Tree), K najbližih suseda (K Nearest Neighbors) i naivni Bajes (Naive Bayes), kao i dva ansambla: slučajna šuma (Random Forest) i pakovanje (Bagging).

Modeli su izabrani u skladu sa prirodom problema koji rešavamo, specifikacijama skupa podataka i rezultatima koje želimo. Motivacije za odabir modela svakog ponaosob će biti navedene u odgovarajućem odeljku.

Metodologija

Analiza skupa podataka

Skup se sastoji od 7195 instanci i 26 atributa. Prvih 22 atributa, redom imenovani MFCC_1, MFCC_2, ..., MFCC_21, MFCC_22, su neprekidni i razmerni. Naredna tri atributa predstavljaju redom porodicu, rod i vrstu i sva tri atributa su diskretna i imenska. Poslednji atribut je redni broj jedinke i takođe je diskretan imenski atribut koji je numeričkog tipa sa vrednostima od 1 do 60.

	MFCCs_1	MFCCs_2	MFCCs_3	MFCCs_4	MFCCs_5	MFCCs_6	MFCCs_7	MFCCs_8	MFCCs_9	MFCCs_10	...	MFCCs_17	MFCCs_18	MFCCs_19	MFCCs_20	MFCCs_21	MFCCs_22	Family	Genus	Species	RecordID
0	1.0	0.152936	-0.105586	0.200722	0.317201	0.260764	0.100945	-0.150063	-0.171128	0.124676	...	-0.108351	-0.077623	-0.009568	0.057684	0.118680	0.014038	Leptodactylidae	Adenomera	AdenomeraAndre	1
1	1.0	0.171534	-0.098975	0.268425	0.338672	0.268353	0.060835	-0.222475	-0.207693	0.170883	...	-0.090974	-0.056510	-0.035303	0.020140	0.082263	0.029056	Leptodactylidae	Adenomera	AdenomeraAndre	1
2	1.0	0.152317	-0.082973	0.287128	0.276014	0.189867	0.008714	-0.242234	-0.219153	0.232538	...	-0.050691	-0.023590	-0.066722	-0.025083	0.099108	0.077162	Leptodactylidae	Adenomera	AdenomeraAndre	1
3	1.0	0.224392	0.118985	0.329432	0.372088	0.361005	0.015501	-0.194347	-0.098181	0.270375	...	-0.136009	-0.177037	-0.130498	-0.054766	-0.018691	0.023954	Leptodactylidae	Adenomera	AdenomeraAndre	1
4	1.0	0.087817	-0.068345	0.306967	0.330923	0.249144	0.006884	-0.265423	-0.172700	0.266434	...	-0.048885	-0.053074	-0.088550	-0.031346	0.108610	0.079244	Leptodactylidae	Adenomera	AdenomeraAndre	1

5 rows × 26 columns

Kolone ne sadrže nedostajuće vrednosti.

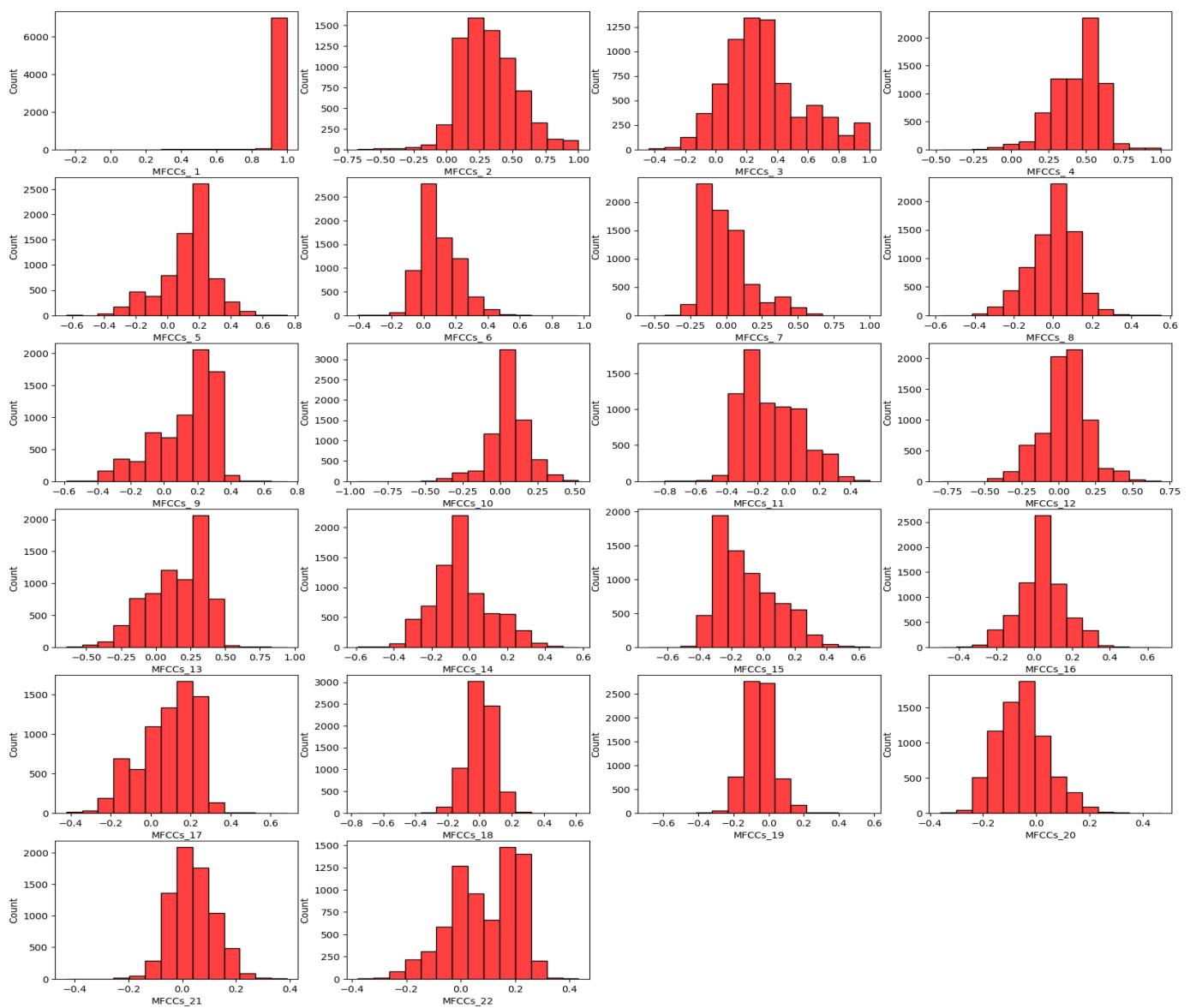
Atributi koji predstavljaju koeficijente su normalizovani i uzimaju vrednosti od -1 do 1. Skup obuhvata podatke o 4 različite porodice: Leptodactylidae, Hylidae, Dendrobatidae, Bufonidae; 8 rodova: Adenomera, Ameerega, Dendropsophus, Hypsiboas, Leptodactylus, Osteocephalus, Rhinella, Scinax; i 10 vrsta: AdenomeraAndre, AdenomeraHylaedactylus,



Ameeregatrivittata, HylaMinuta, HypsiboasCinerascens, HypsiboasCordobae,
LeptodactylusFuscus, OsteocephalusOophagus, Rhinellagranulosa, ScinaxRuber.

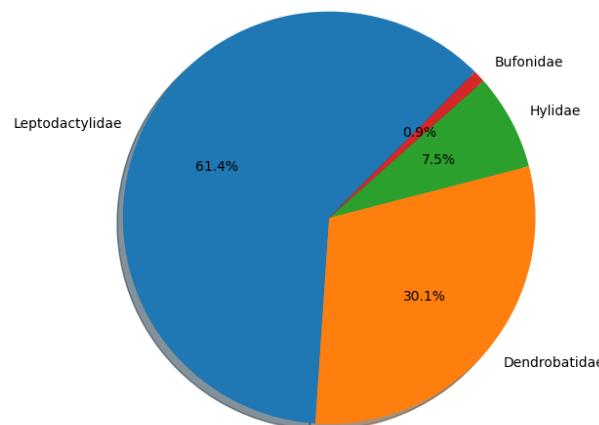
Pregled statistika i raspodele atributa.

	MFCCs_1	MFCCs_2	MFCCs_3	MFCCs_4	MFCCs_5	MFCCs_6	MFCCs_7	MFCCs_8	MFCCs_9	MFCCs_10	...	MFCCs_14	MFCCs_15	MFCCs_16	MFCCs_17	MFCCs_18	MFCCs_19	MFCCs_20	MFCCs_21	MFCCs_22
count	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	...	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000	7195.000000
mean	0.989885	0.323584	0.311224	0.445997	0.127046	0.097939	-0.001397	-0.000370	0.128213	0.055998	...	-0.039244	-0.101748	0.042062	0.088680	0.007755	-0.049474	-0.053244	0.037313	0.087567
std	0.069016	0.218653	0.263527	0.160328	0.162722	0.120412	0.171404	0.116302	0.179008	0.127099	...	0.152515	0.187618	0.119915	0.138055	0.084733	0.082546	0.094181	0.079470	0.123442
min	-0.251179	-0.673025	-0.436028	-0.472676	-0.636012	-0.410417	-0.538982	-0.576506	-0.587313	-0.952266	...	-0.590380	-0.717156	-0.498675	-0.421480	-0.759322	-0.680745	-0.361649	-0.430812	-0.379304
25%	1.000000	0.165945	0.138445	0.336737	0.051717	0.012581	-0.125737	-0.063109	0.004648	-0.001132	...	-0.132980	-0.255929	-0.019549	-0.001764	-0.042122	-0.106079	-0.120971	-0.017620	0.000533
50%	1.000000	0.302184	0.274626	0.481463	0.161361	0.072079	-0.052630	0.013265	0.189317	0.063478	...	-0.050715	-0.143259	0.041081	0.112769	0.011820	-0.052626	-0.055180	0.031274	0.105373
75%	1.000000	0.466566	0.430695	0.559861	0.222592	0.175957	0.085580	0.075108	0.265395	0.117725	...	0.039157	0.017348	0.107046	0.201932	0.061889	0.006321	0.001342	0.089619	0.194819
max	1.000000	1.000000	1.000000	1.000000	0.752246	0.964240	1.000000	0.551762	0.738033	0.522768	...	0.575749	0.668924	0.670700	0.681157	0.614064	0.574209	0.467831	0.389797	0.432207

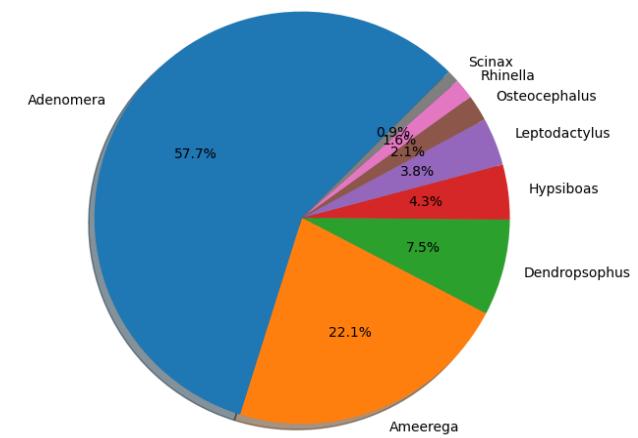


Žabe nisu ravnomerno raspoređene u porodice, odnosno, rodove i vrste. Više od 60% žaba pripada porodici Leptodactylidae, dok porodica Bufonidae sadrži manje od 1% žaba.

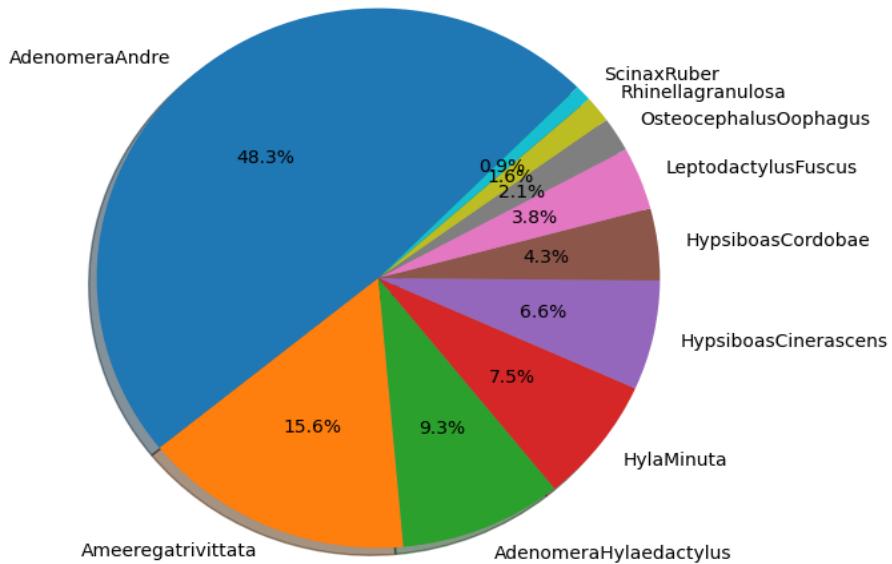
Family Distribution



Genus Distribution

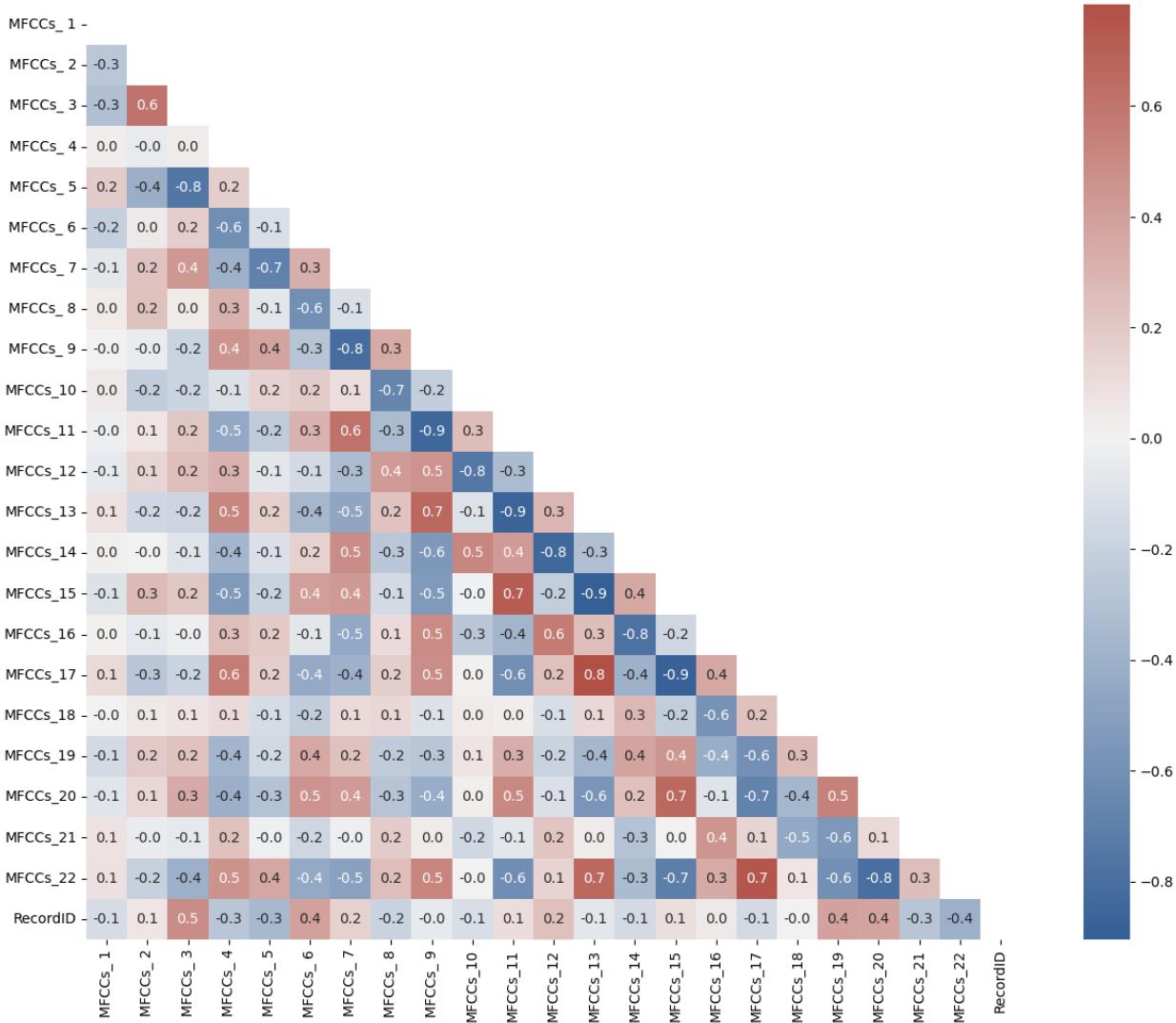


Species Distribution





Koristimo matricu korelacija da predstavimo zavisnosti između atributa. Postoji negativna korelisanost atributa MFCC_i sa atributima MFCC_{i-2} i MFCC_{i+2}, kao i pozitivna korelisanost sa atributima MFCC_{i-4} i MFCC_{i+4}. Najmanja korelisanost iznosi -0.9, dok je najveća 0.8.



Klasifikacija može da se vrši na tri nivoa, tj. na porodice, rodove i vrste. U okviru ovog rada klasifikacija će se vršiti na vrste. Motivacija za ovakav odabir je deskriptivnost vrste, informacija o pripadnosti nekoj vrsti automatski nosi i informaciju o pripadnosti rodu, tj. porodici.

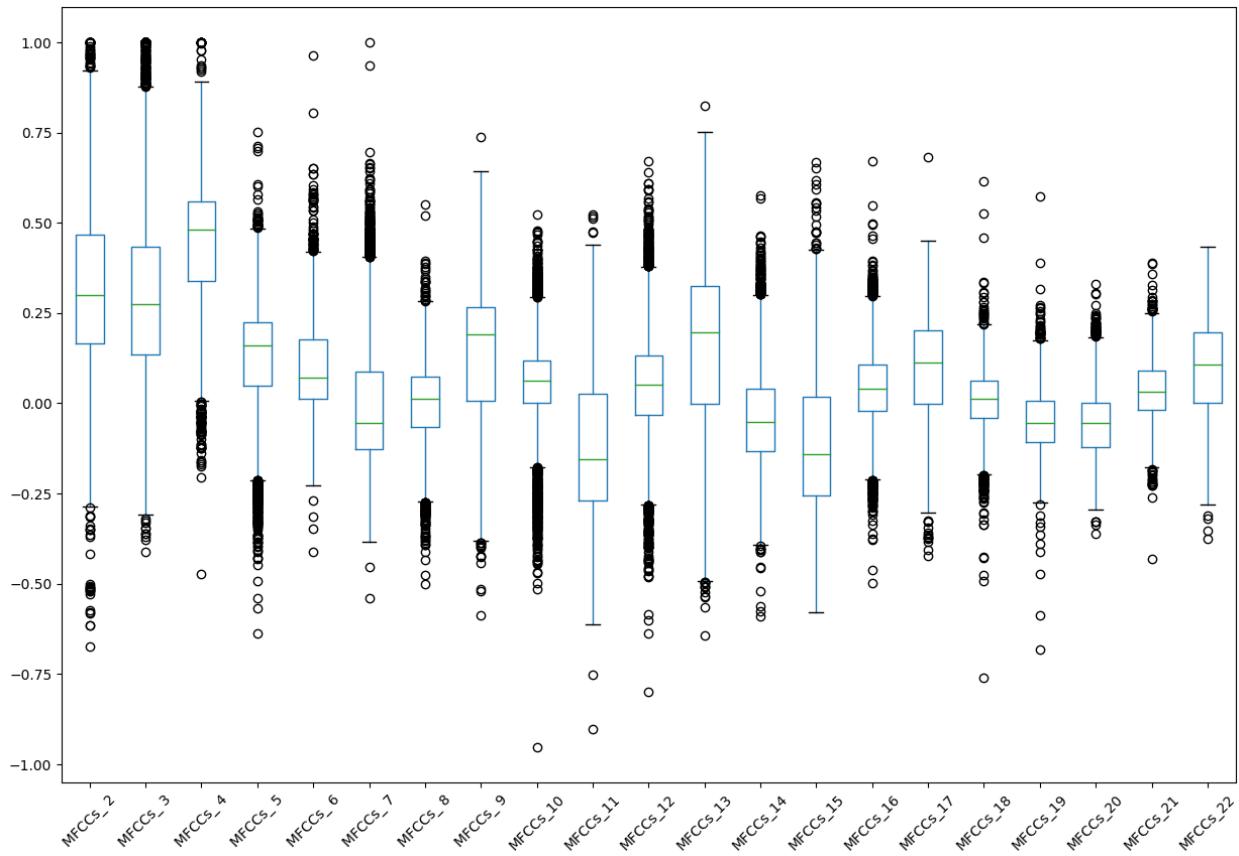
Preprocesiranje

Prvi korak u radu sa podacima je podela na dva podskupa, jedan koji će služiti za treniranje modela i drugi koji će služiti testiranju. Instance se dele tako da se održi postojeća raspodela instanci po klasama. Iz prvobitnog skupa podataka izdvajamo 70% instanci za trening skup, a 30% instanci za test skup. U trening skupu se nalazi 5036 instanci, dok se u test skupu nalazi 2159 instanci.

Preprocesiranje nastavljamo na trening skupu, izuzev PCA koji će biti primenjen i na test skup.

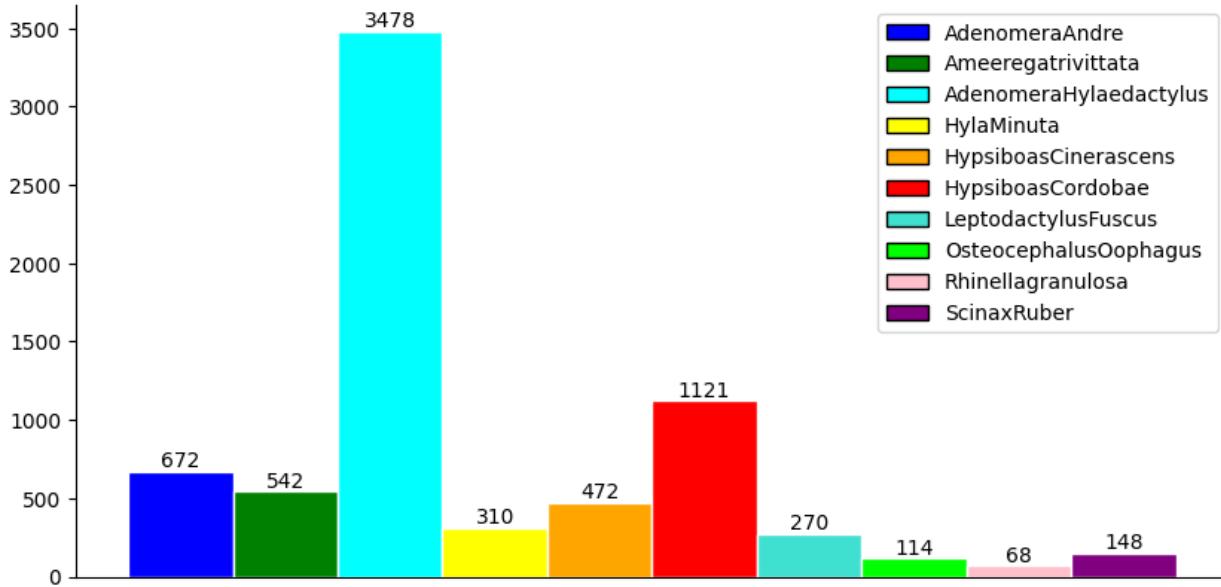
Rad sa elementima van granica

Elemente van granica detektujemo i vizuelno predstavljamo boxplot tehnikom.

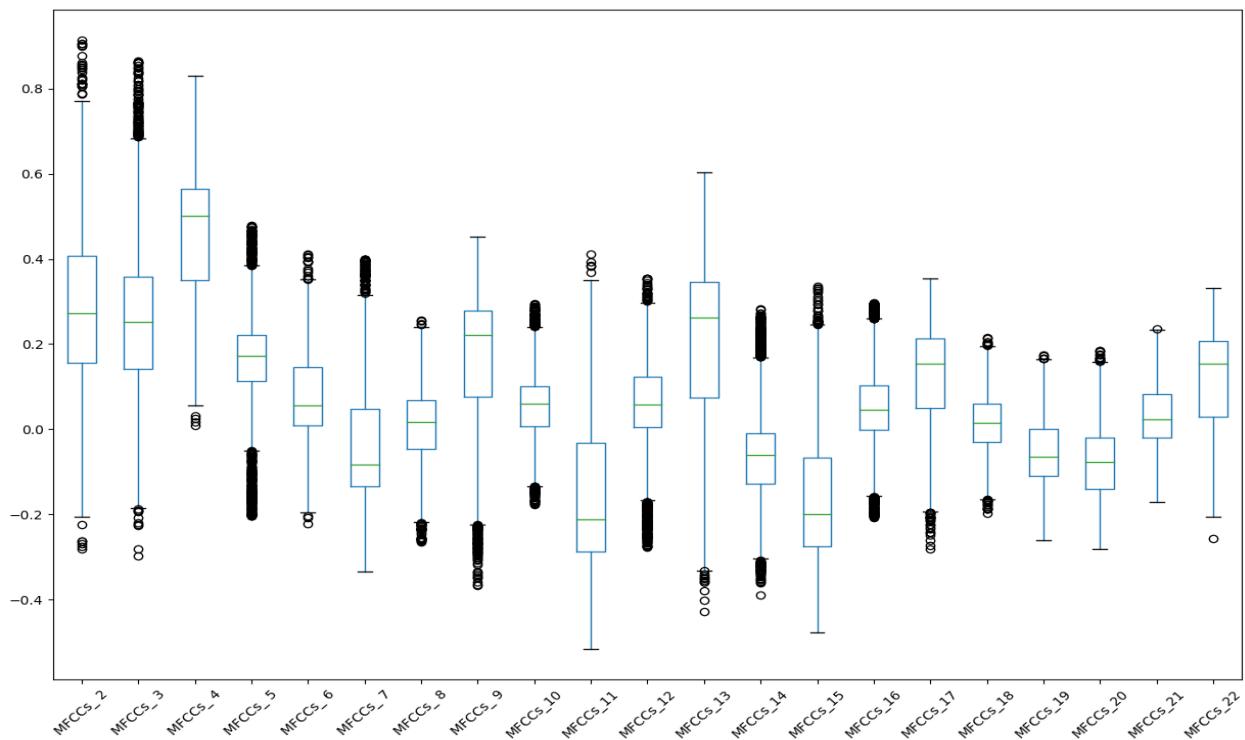




Raspoređenost instanci po vrstama pre eliminisanja elemenata van granica.

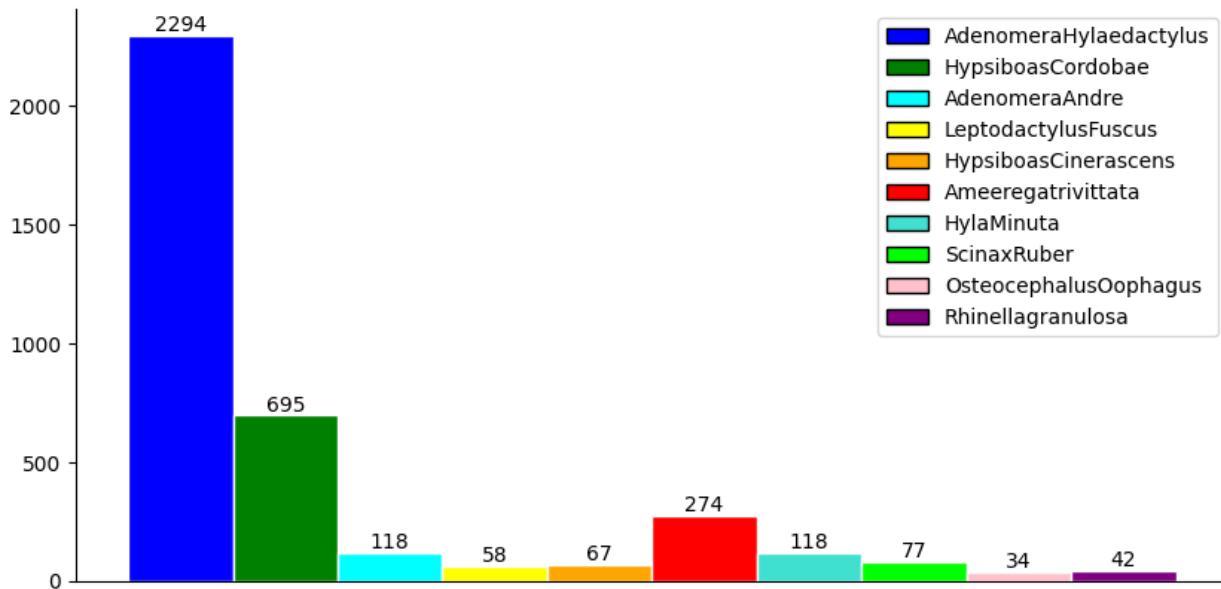


Pomoću IQR (interkvartilni raspon) metoda uočavamo 1259 elemenata van granica. Pošto je, između ostalih, odabran i metod k najbližih suseda za klasifikaciju koji je osetljiv na elemente van granica, eliminišemo svih 1259 elemenata iz trening skupa, čime ga svodimo na 3777 instanci.



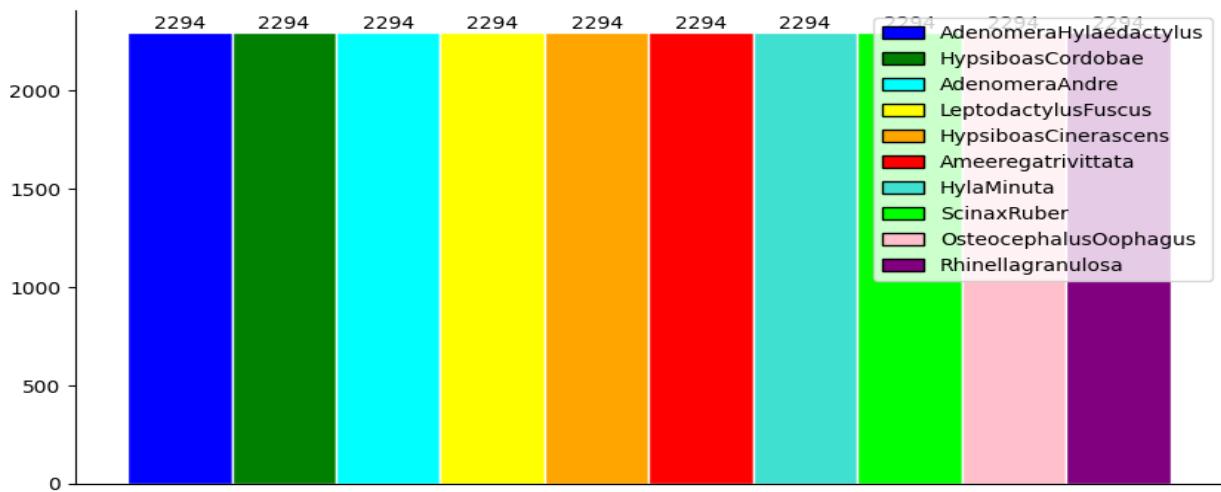


Nakon eliminacije elemenata van granica, raspoređenost instanci po klasama se ne menja značajno.

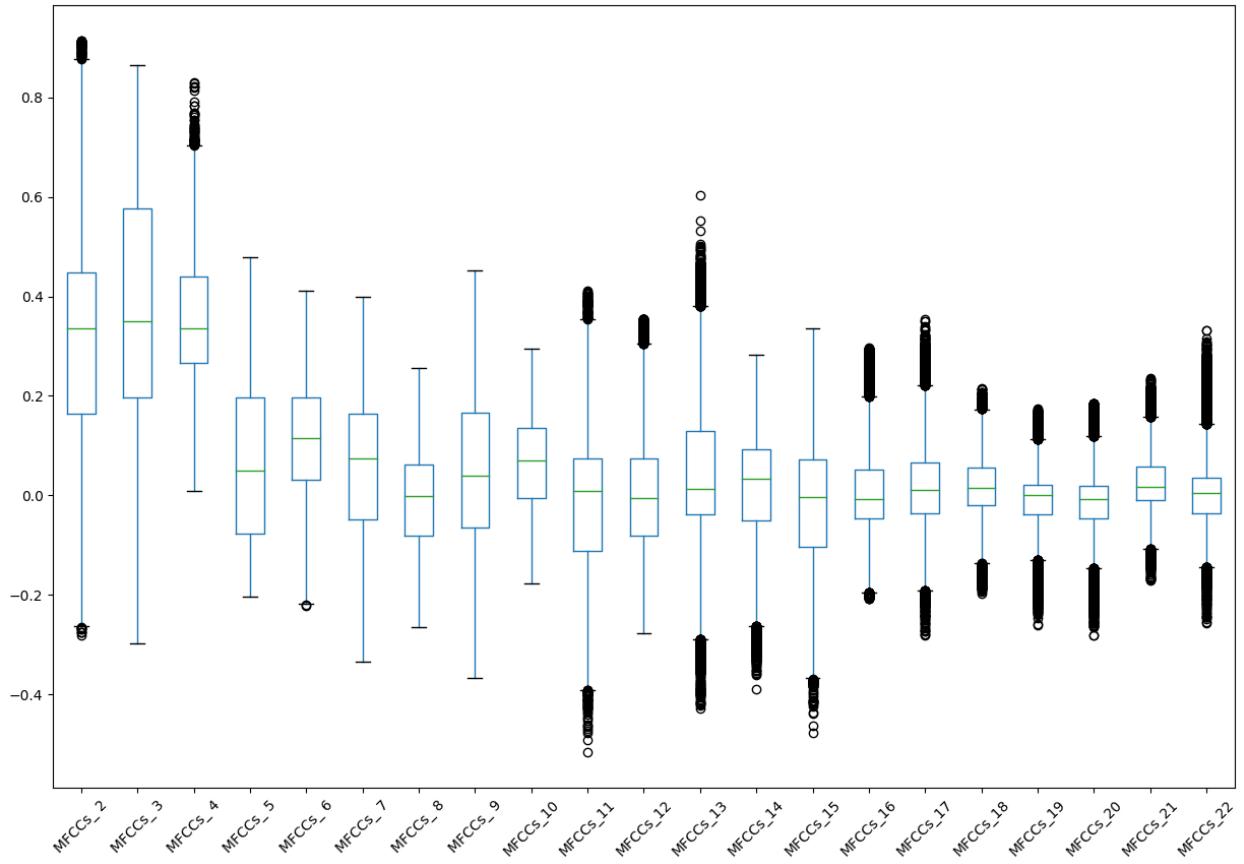


Balansiranje

Skup podataka je nebalansiran, naime, oko 60% instanci pripada vrsti *AdenomeraHylaedactylus*, dok manje od 1% instanci pripada vrsti *OsteocephalusOophagus*. Da bismo izbegli favorizovanje većinske klase, vršimo balansiranje podataka u trening skupu. Jedna od popularnih i učinkovitih metoda balansiranja je takozvana tehnika sintetičkog preuzorkovanja manjinskih uzoraka, odnosno algoritam SMOTE. SMOTE generiše nove instance interpolacijom između postojećih instanci manjinske klase.



SMOTE potencijalno može da unese šum u podatke. Koristeći boxplot tehniku detektujemo i vizualizujemo elemente van granica nakon balansiranja podataka i primećujemo da ne postoji značajan porast šuma, iako su uočljive promene na poslednjih 10 atributa.



Redukcija dimenzionalnosti

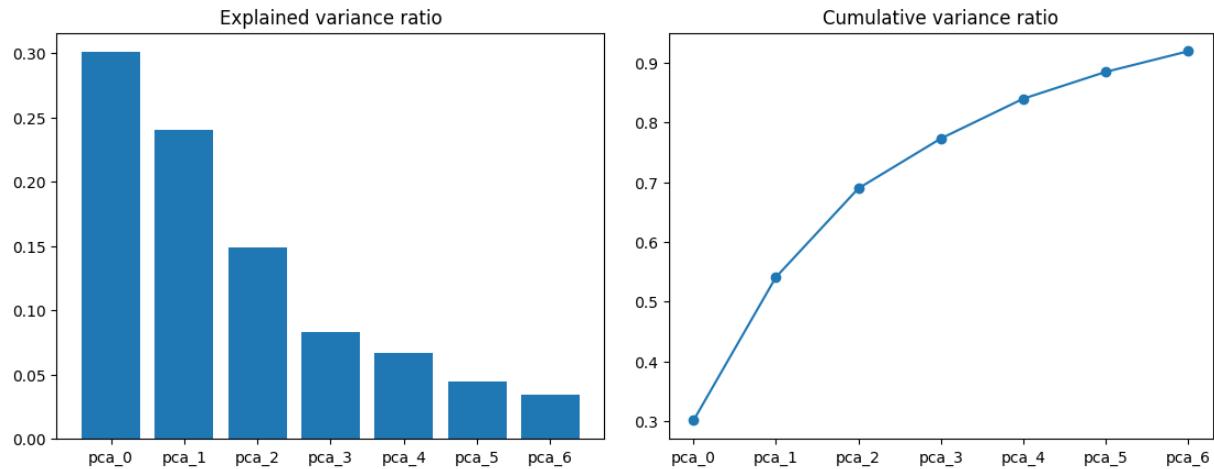
Poslednji atribut nosi informaciju o rednom broju jedinke i ne doprinosi klasifikaciji na vrste, zbog čega ga brišemo iz skupa podataka.

Prvi atribut, MFCC_1, u najvećem broju instanci ima vrednost 1. Ne donosi nam mnogo informacija i zato ga brišemo iz skupa podataka.

Tehnike izdvajanja atributa se koriste za smanjenje dimenzionalnosti skupa podataka radi dostizanja boljih performansi i bržeg vremena izvršavanja. Za potrebe ovog istraživanja odabrana je tehnika analize glavnih komponenti, odnosno PCA (Principal component analysis). PCA transformiše korelisane attribute u nekorelisane, koje zovemo glavne komponente. U daljem radu koristićemo samo onoliko komponenti koliko je dovoljno da se objasni bar 90% varijanse podataka. U našem slučaju odabранo je prvih 6 glavnih komponenti, s kojima nastavljamo dalji rad.



PCA koji smo prilagodili trening skupu, koristimo i za transformaciju test skupa.



Na prvom grafiku je predstavljeno koliko varijanse svaki atribut opisuje, dok je na drugom predstavljena njihova kumulativna suma.

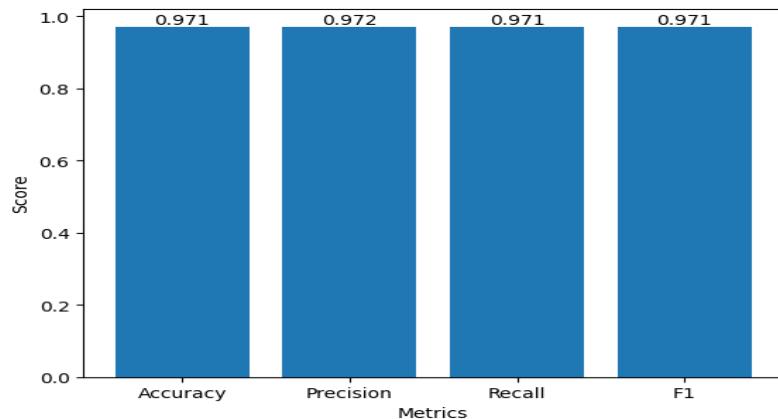
Stablo odlučivanja

Motivacije za odabir ove metode u našem radu su:

1. Interpretabilnost: Drvo odlučivanja pruža jasan i intuitivan pregled procesa odlučivanja koji je lako razumeti.
2. Rad sa numeričkim atributima: Drvo odlučivanja radi sa različitim tipovima atributa bez dodatne potrebe za preprocesiranjem.

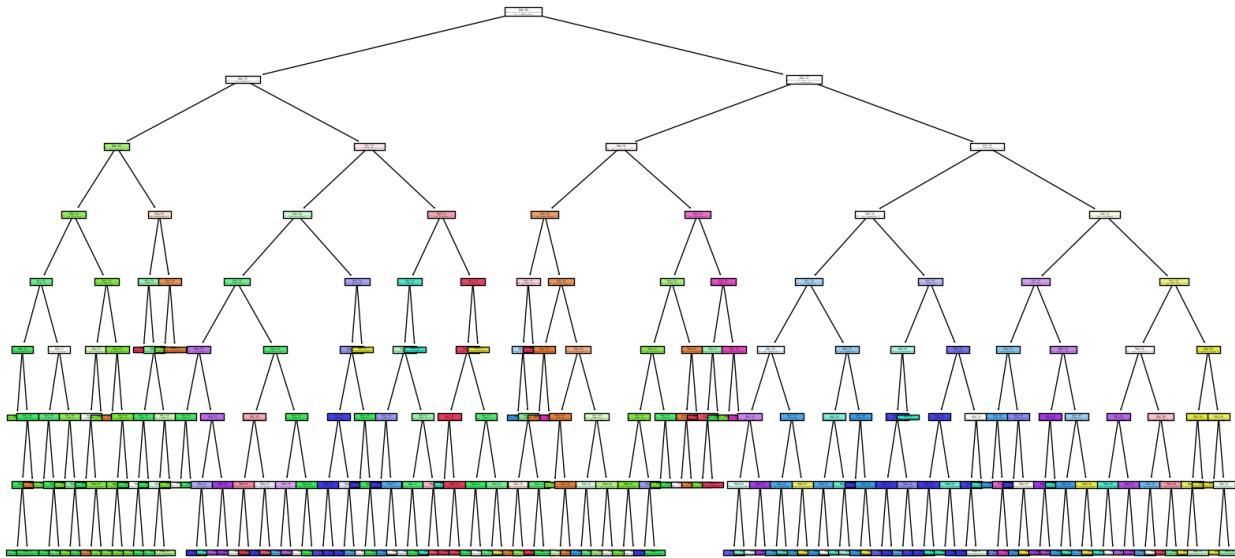
Za pronalaženje optimalnih hiperparametara koristimo GridSearch. Prosleđujemo vrednosti parametara čije će se kombinacije dalje pretraživati. Testiramo različite mere (Ginijev indeks, entropija), dubine stabla i kriterijume za podelu unutrašnjih čvorova i određivanje listova.

Koristimo različite mere ocena performansi unakrsno validiranog modela. Primećujemo da sve četiri mere kvaliteta modela: tačnost, preciznost, odziv i F1 imaju vrednost oko 97%.



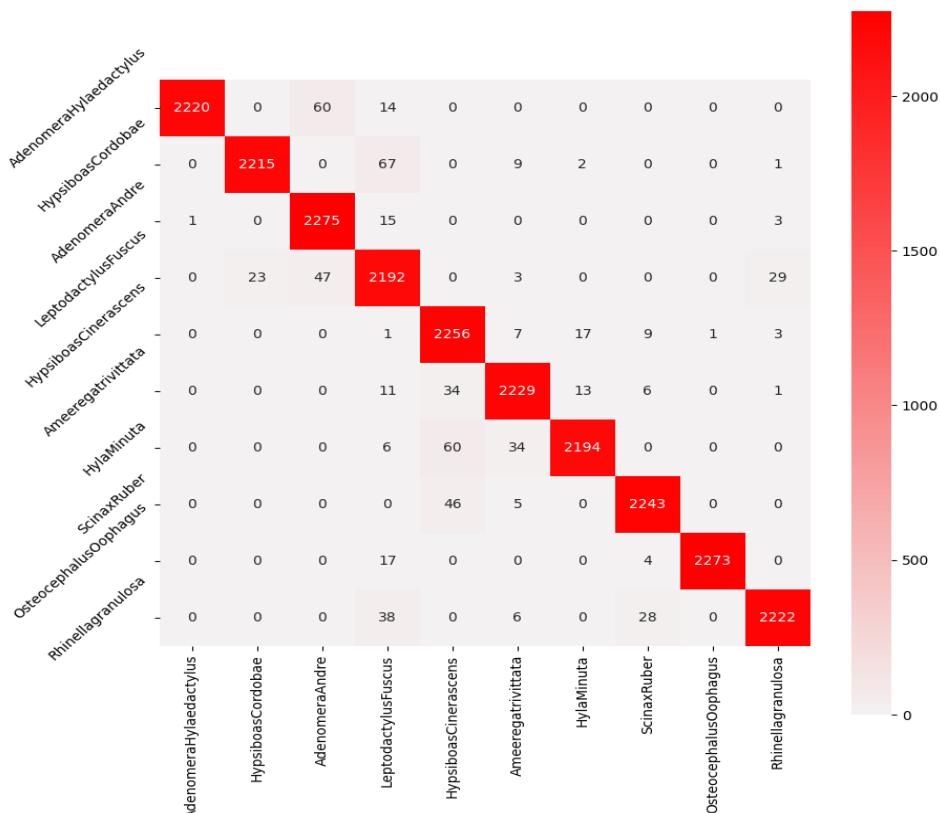


Stablo sa najboljim hiperparametrima možemo lako vizualizovati zahvaljujući interpretabilnosti modela stabla odlučivanja.



Predviđanja klase na trening skupu daju tačnost od oko 97%.

Pomoću matrice konfuzije možemo lakše predstaviti kvalitet klasifikovanja modela.

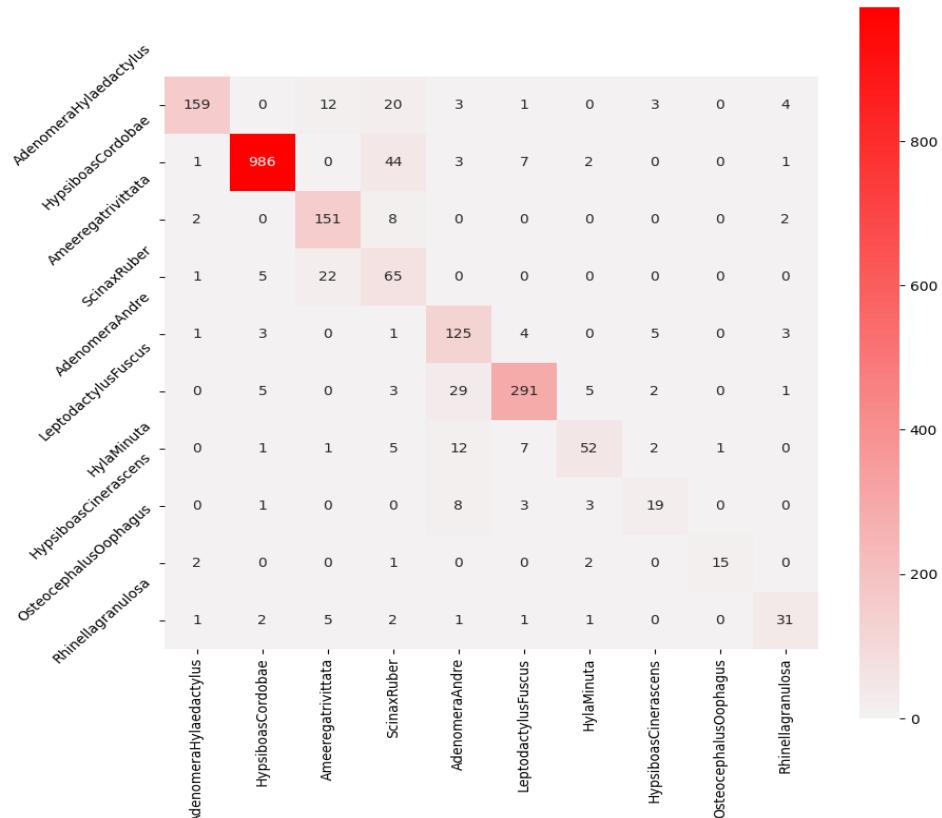




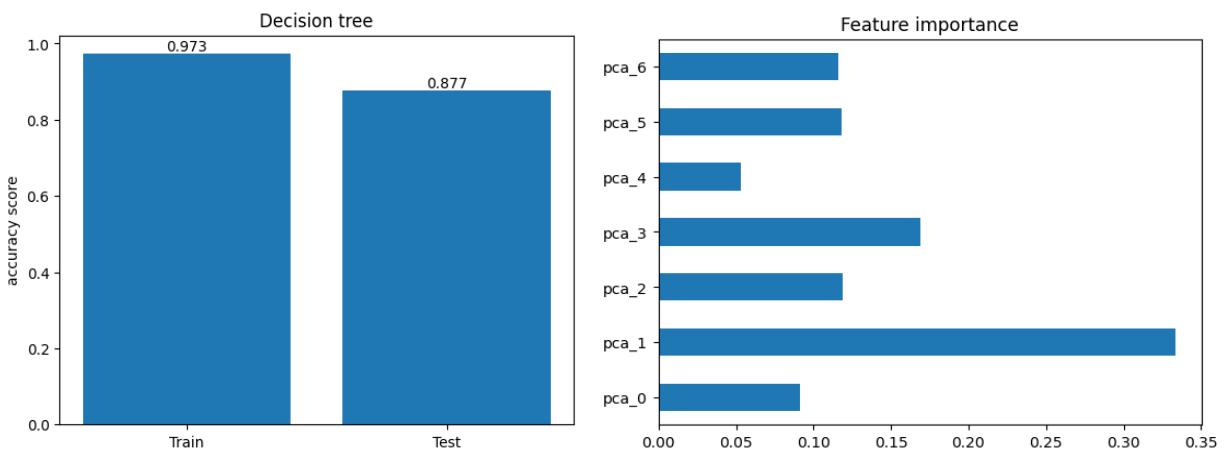
...analiza...

Predviđanja klase na test skupu daju tačnost od 88%.

Pomoću matrice konfuzije možemo lakše predstaviti kvalitet klasifikovanja modela.



Feature importance grafik predstavlja važnost atributa u procesu klasifikacije, dok *decision tree* grafik poredi tačnost predviđanja na trening i test skupu. Uočljiv je stepen preprilagođavanja modela.



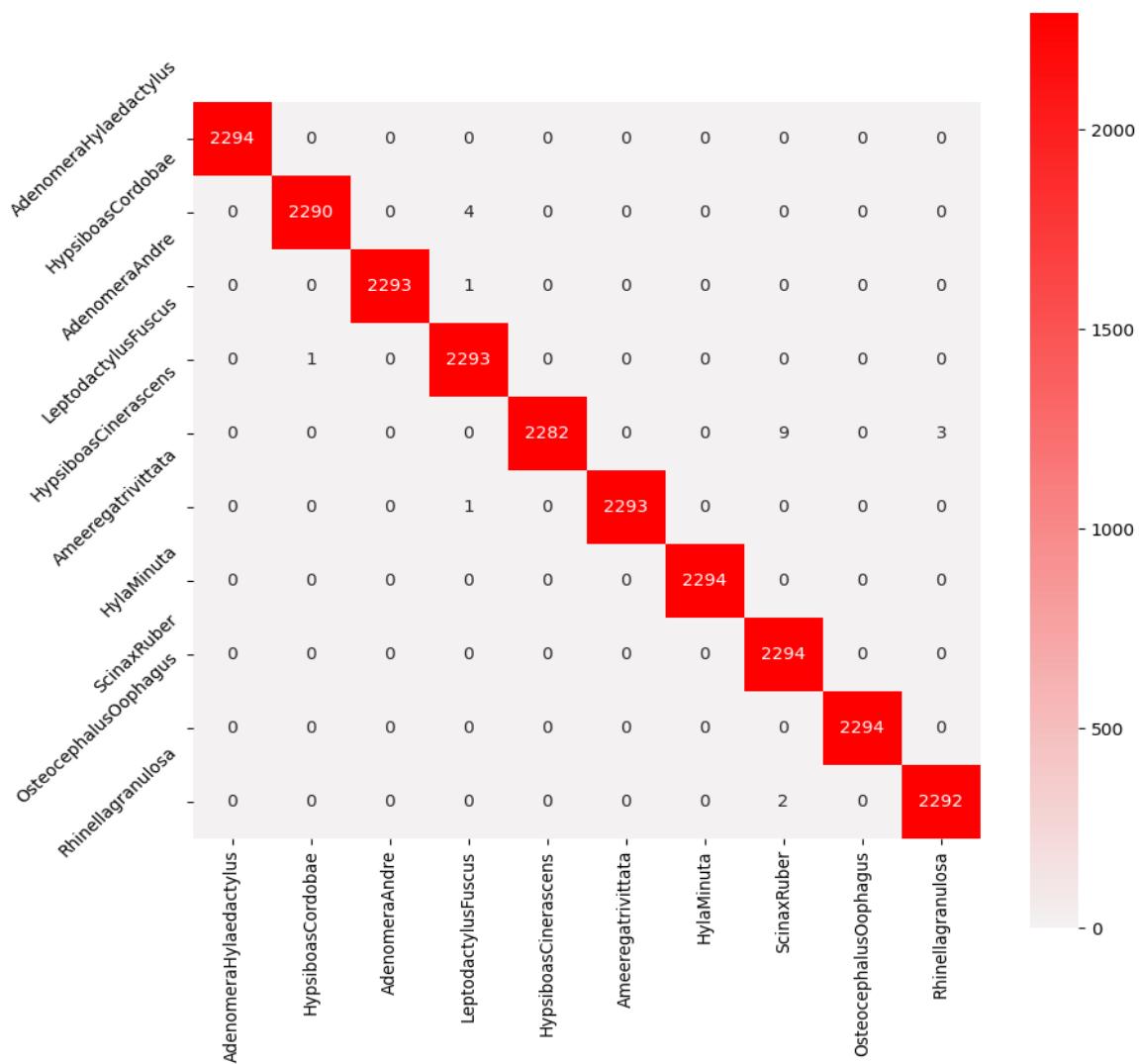


Slučajna šuma

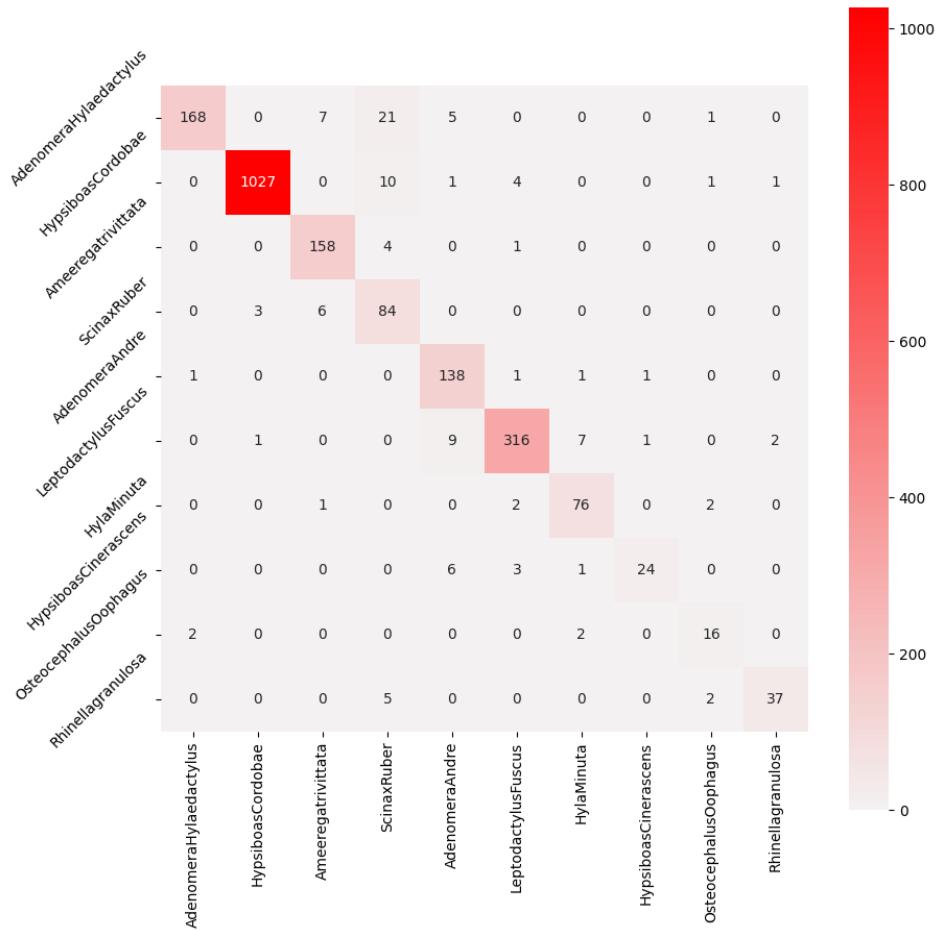
Slučajna šuma je ansambl koji se sastoji od m stabala tretiranih na različitim podskupovima skupa za treniranje. Motivacija za korišćenje je manja šansa za preprilagođavanje od jednog stabla odlučivanja. Potencijalni problemi su zahtevniji račun i manja interpretabilnost.

Ansambl slučajna šuma, koji je korišćen u ovom istraživanju, sastoji se od 100 stabala odlučivanja dubine 8.

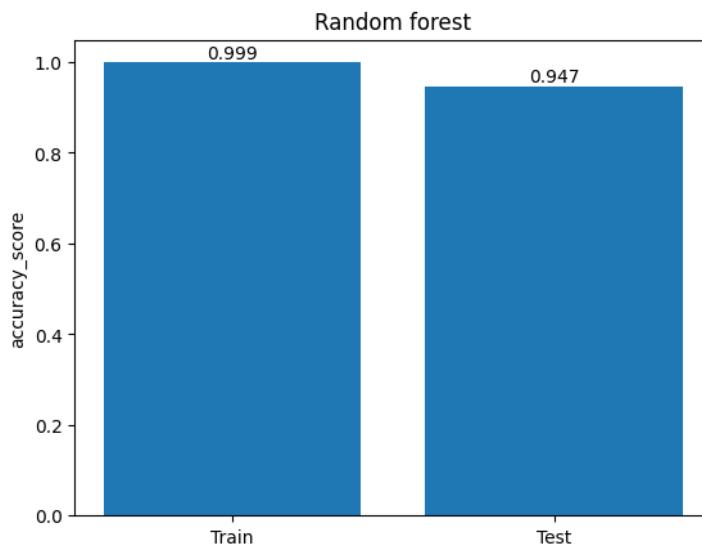
Predviđanja klasa na trening skupu daju nešto manje od 100% tačnosti, sa sledećom matricom konfuzije.



Predviđanja klase na test skupu daju tačnost od 95%, sa sledećom matricom konfuzije.



Očekivano, ansambl je dao bolje rezultate od jednog stabla odlučivanja, kao i manji stepen preprilagođavanja.



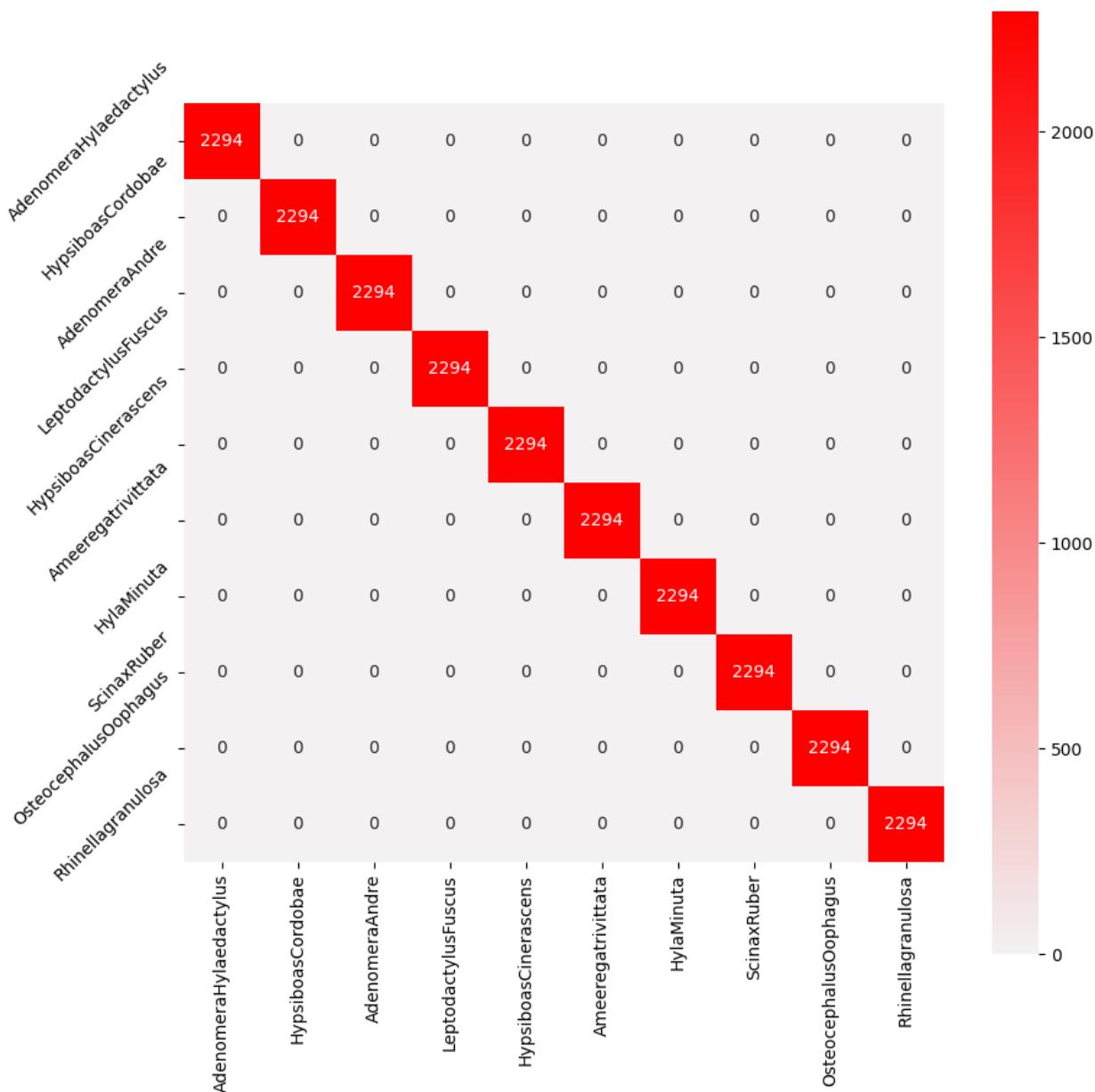


K najbližih suseda

Motivacije za korišćenje modela K najbližih suseda su jednostavan rad sa numeričkim atributima bez dodatnog preprocesiranja, manji broj hiperparametara za testiranje i brzina izvršavanja.

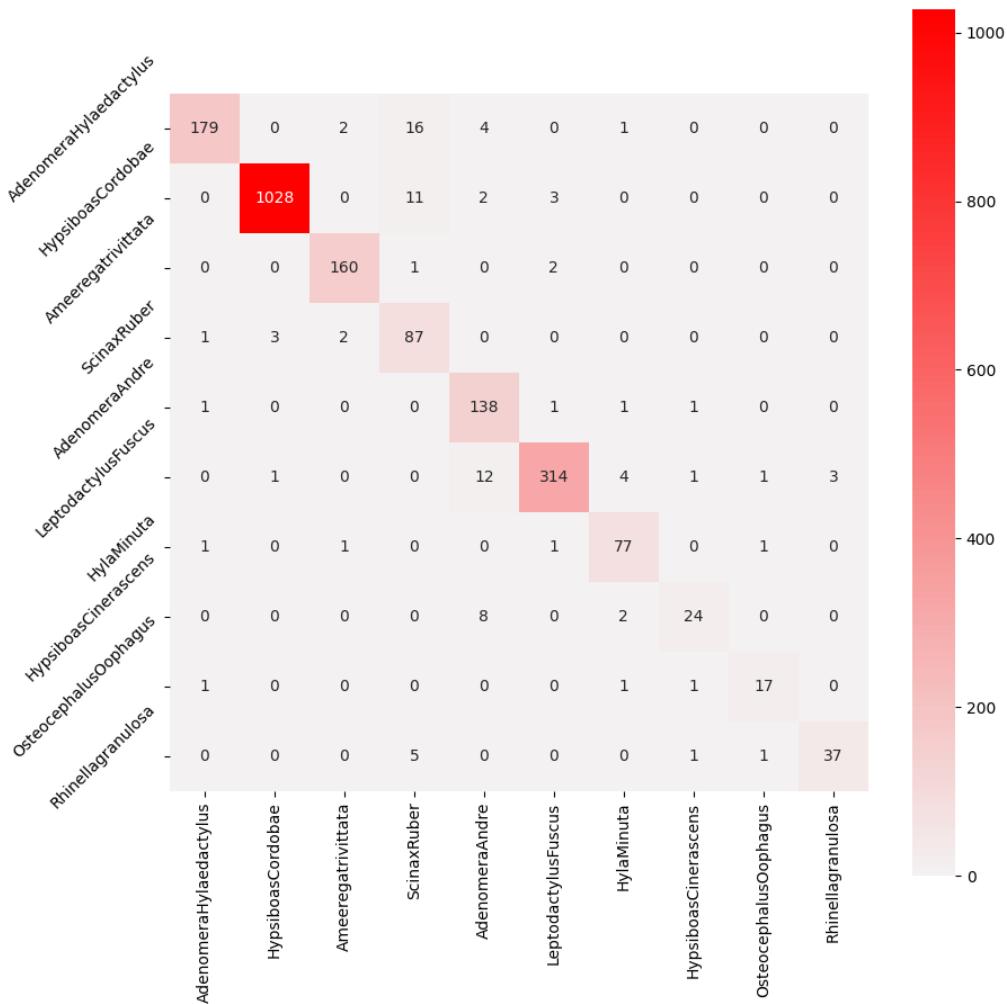
Da bismo dobili model sa optimalnim vrednostima hiperparametara, koristimo GridSearch.

Predloženi model ima vrednost $k=10$, predviđa klase na trening skupu sa tačnošću od 100% i sledećom matricom konfuzije.

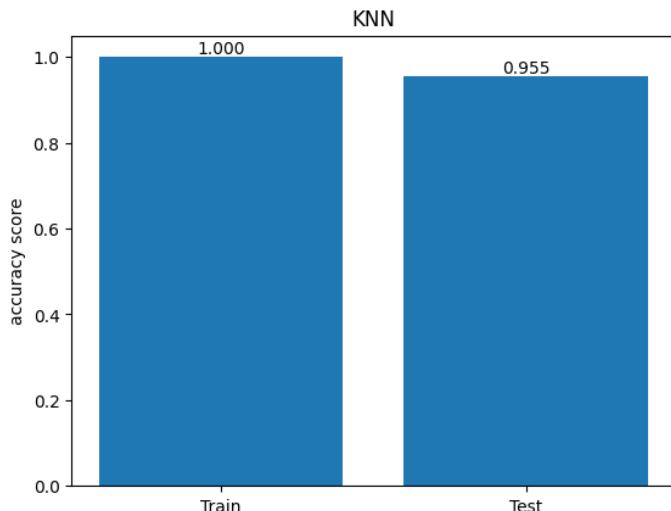




Na test skupu, model predviđa klase sa tačnošću od 95% i sledećom matricom konfuzije.



Pregled dobijenih tačnosti na trening i test skupu.



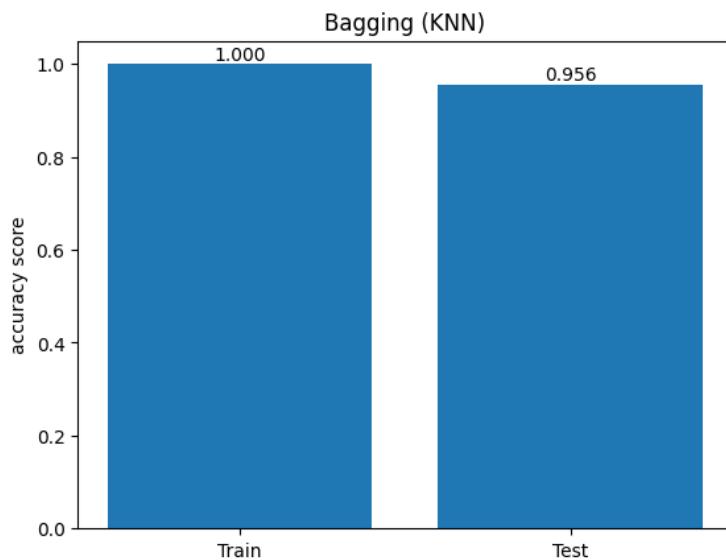


Pakovanje

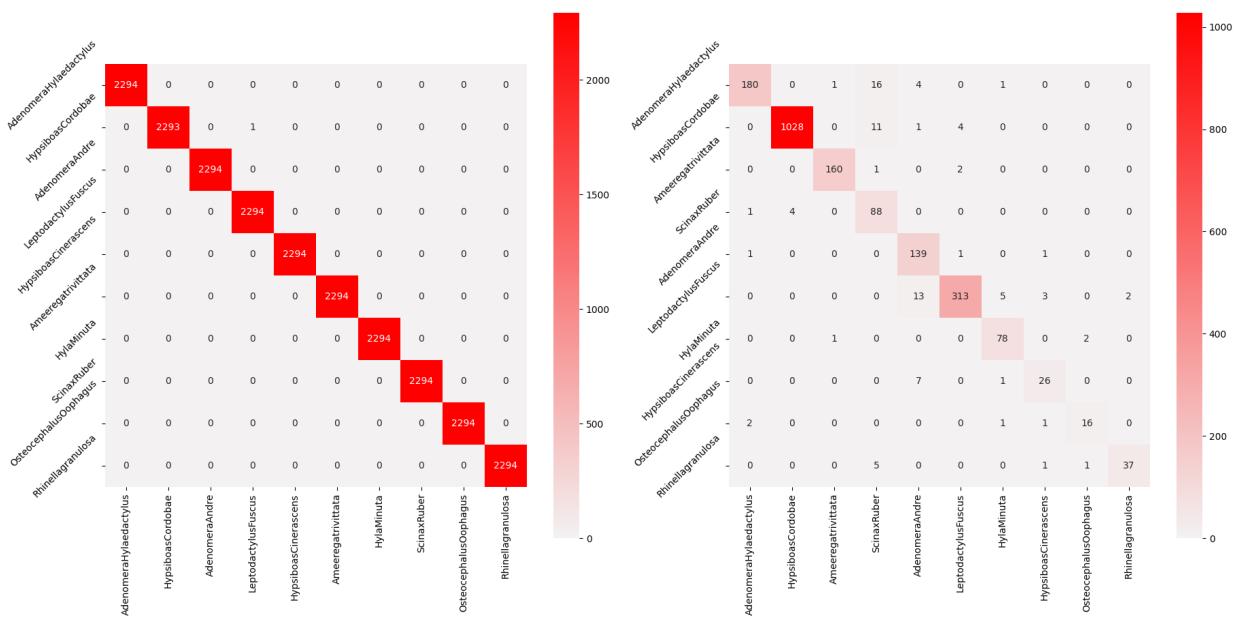
Pakovanje (Bagging ensemble) ima jednostavnu ideju: uklapamo nekoliko nezavisnih modela i prosečimo njihova predviđanja kako bismo dobili model sa manjom varijansom. Model koji će ansambl pakovanja korisiti je model k najbližih suseda.

Parametri koji su korišćeni za model u okviru pakovanja, su najbolji parametri iz prethodno treniranog modela k najbližih suseda.

Slično, na trening skupu tačnost iznosi 100%, dok je na test skupu oko 96%.



Dobijene su sledeće matrice konfuzije, prva se odnosi na trening, a druga na test skup.





Naivni Bajes

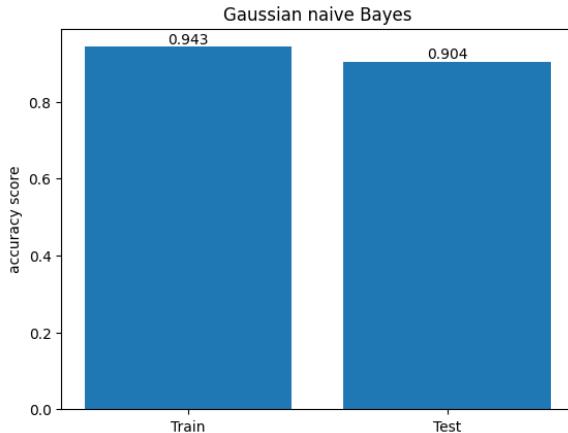
Poslednji klasifikacioni model koji ćemo koristiti je naivni Bajes.

Motivacije za njegovu upotrebu su:

1. Jednostavnost i efikasnost: Naivni Bajes je jednostavan i računski efikasan algoritam, što ga čini pogodnim za velike skupove podataka.
2. Prepostavka nezavisnosti: Naivni Bajes prepostavlja da su atributi uslovno nezavisni u odnosu na klasnu oznaku, što pojednostavljuje model i smanjuje rizik od preprilagođavanja.

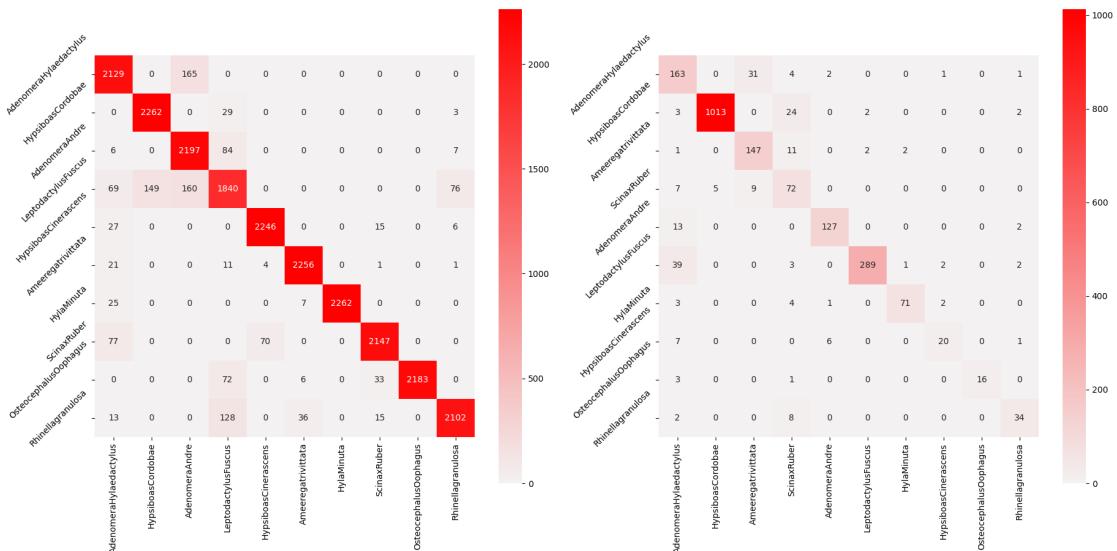
Pošto su svi atributi neprekidni, koristićemo Gausov naivni Bajesov klasifikator.

Model nema hiperparametre koje bismo testirali, tako da nije korišćen GridSearch.



Na trening skupu, model predviđa klase sa tačnošću od 94%, dok je na test skupu ta vrednost 90%.

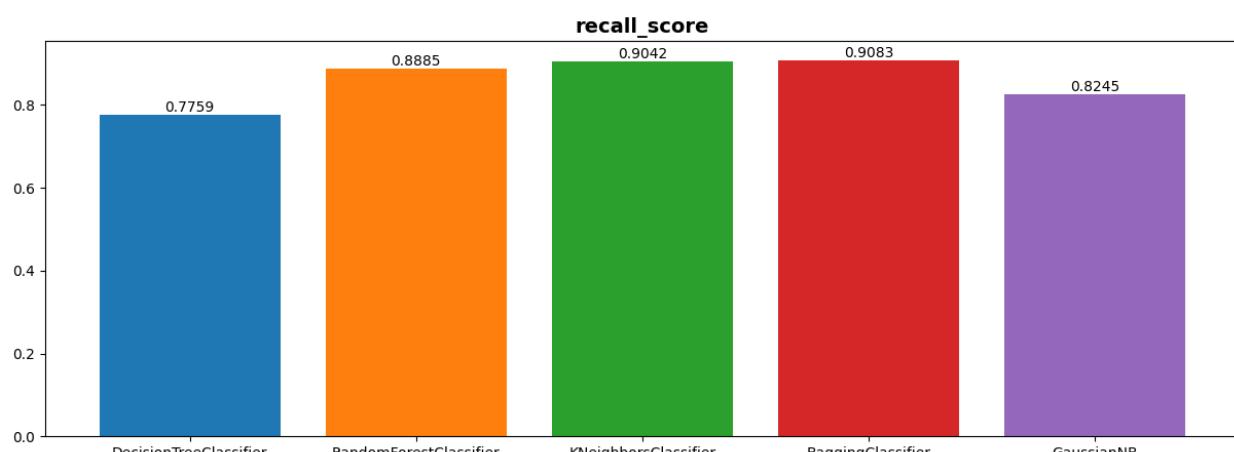
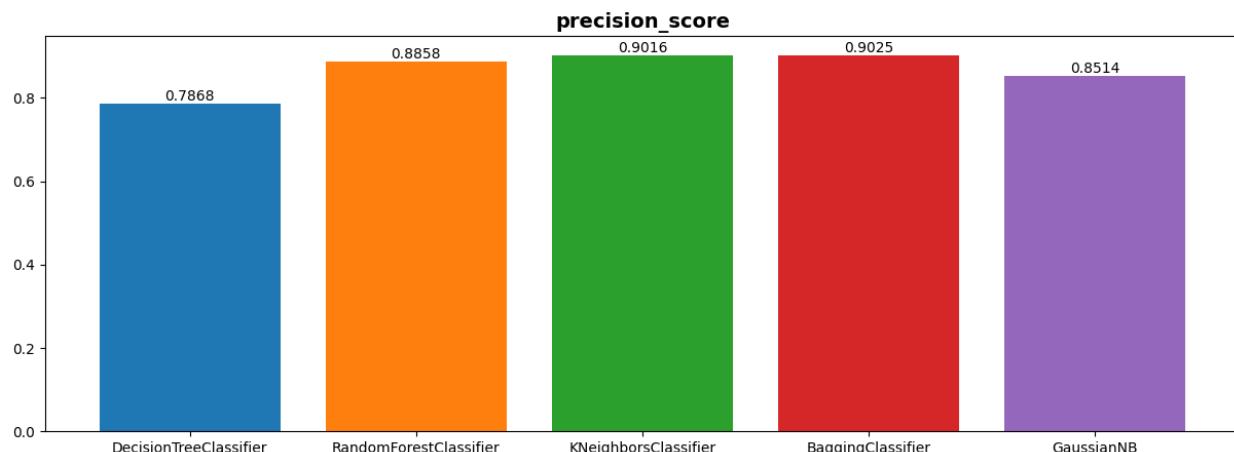
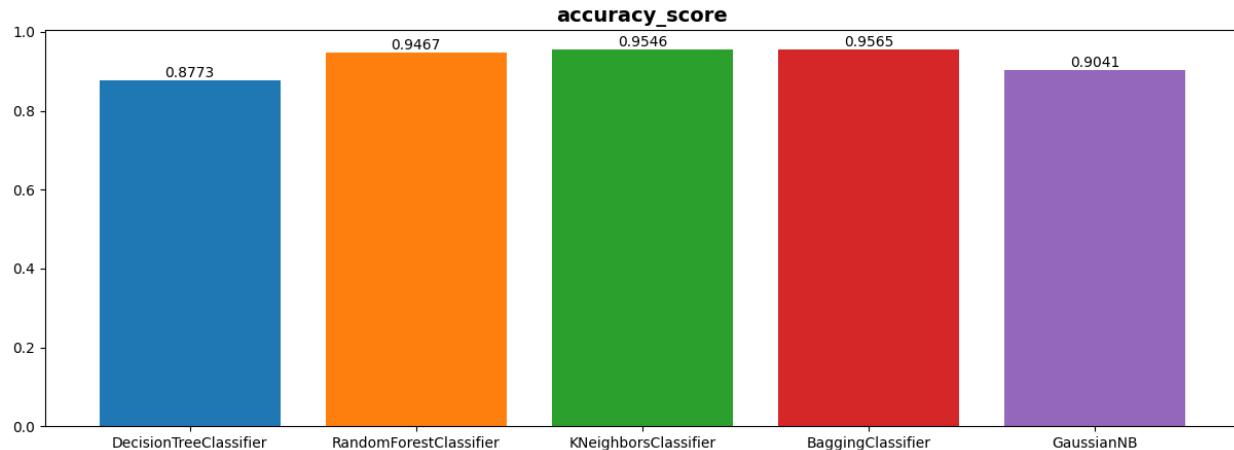
Dobijene su sledeće matrice konfuzije, prva se odnosi na trening, a druga na test skup.

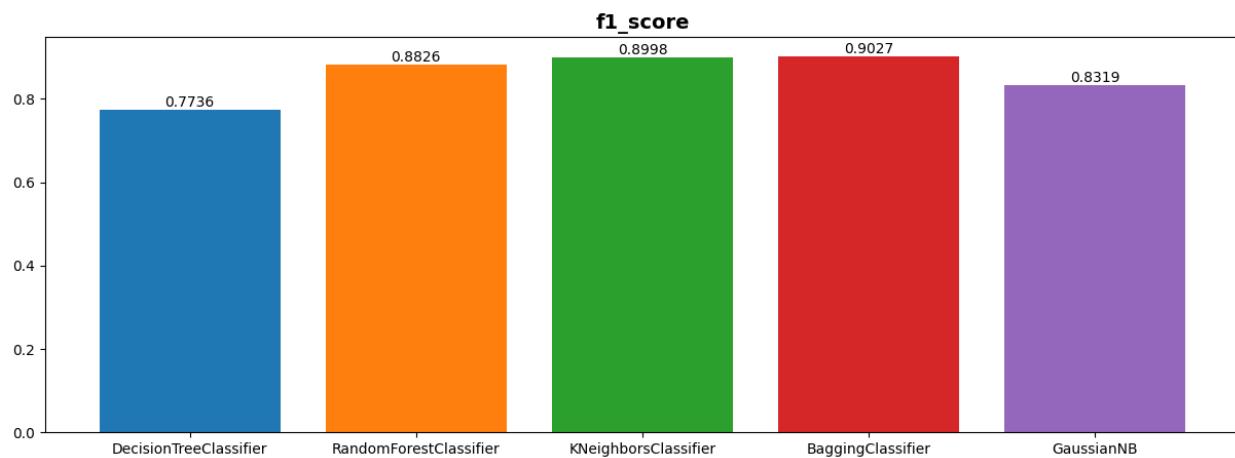




Poređenje modela

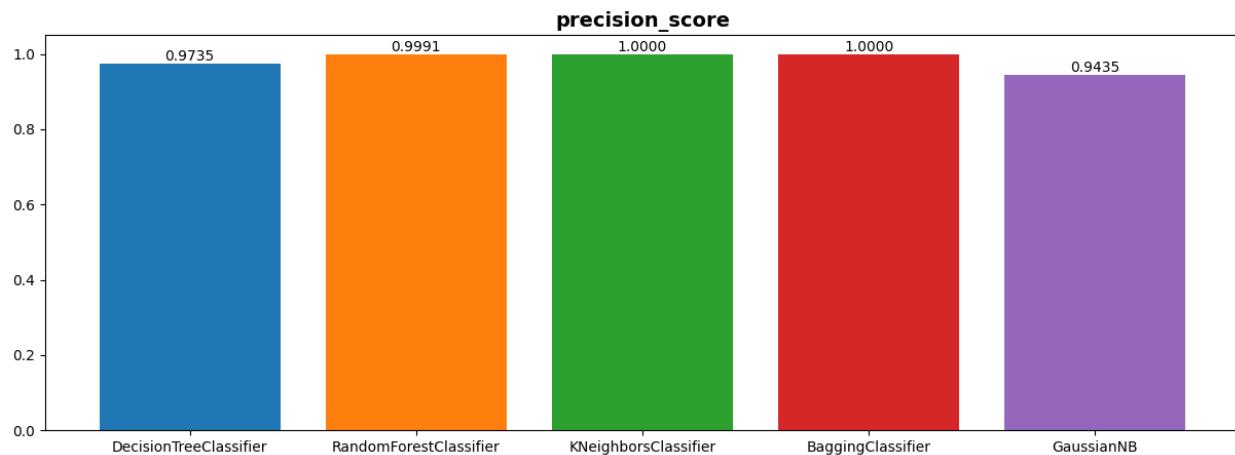
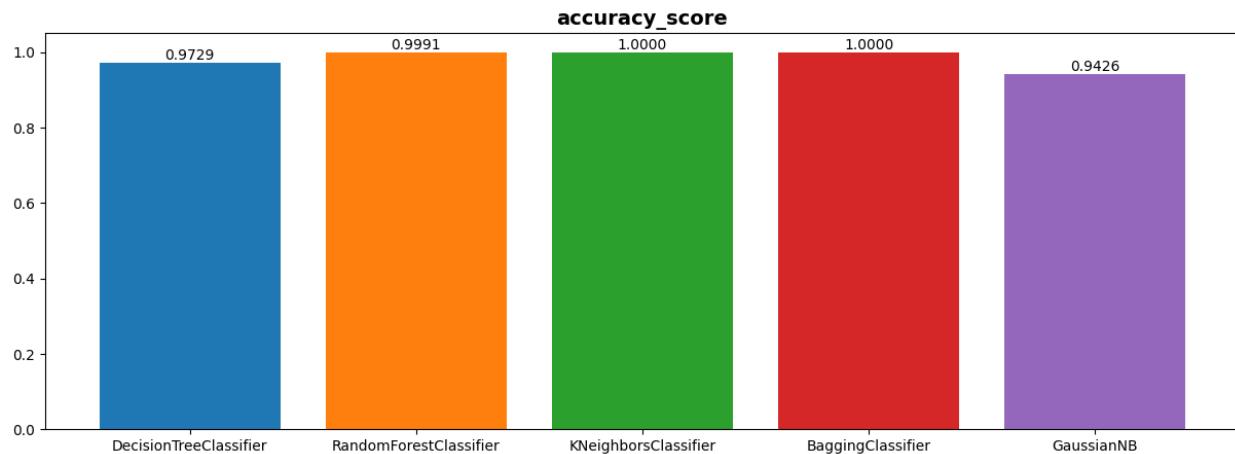
Predstavićemo vrednosti različitih mera kvaliteta za sve klasifikacione modele koji su korišćeni u istraživanju. Mere su izračunate nakon predviđanja klase na **test** skupu. Sem redukcije dimenzionalnosti, test skup nije dodatno preprocesiran.

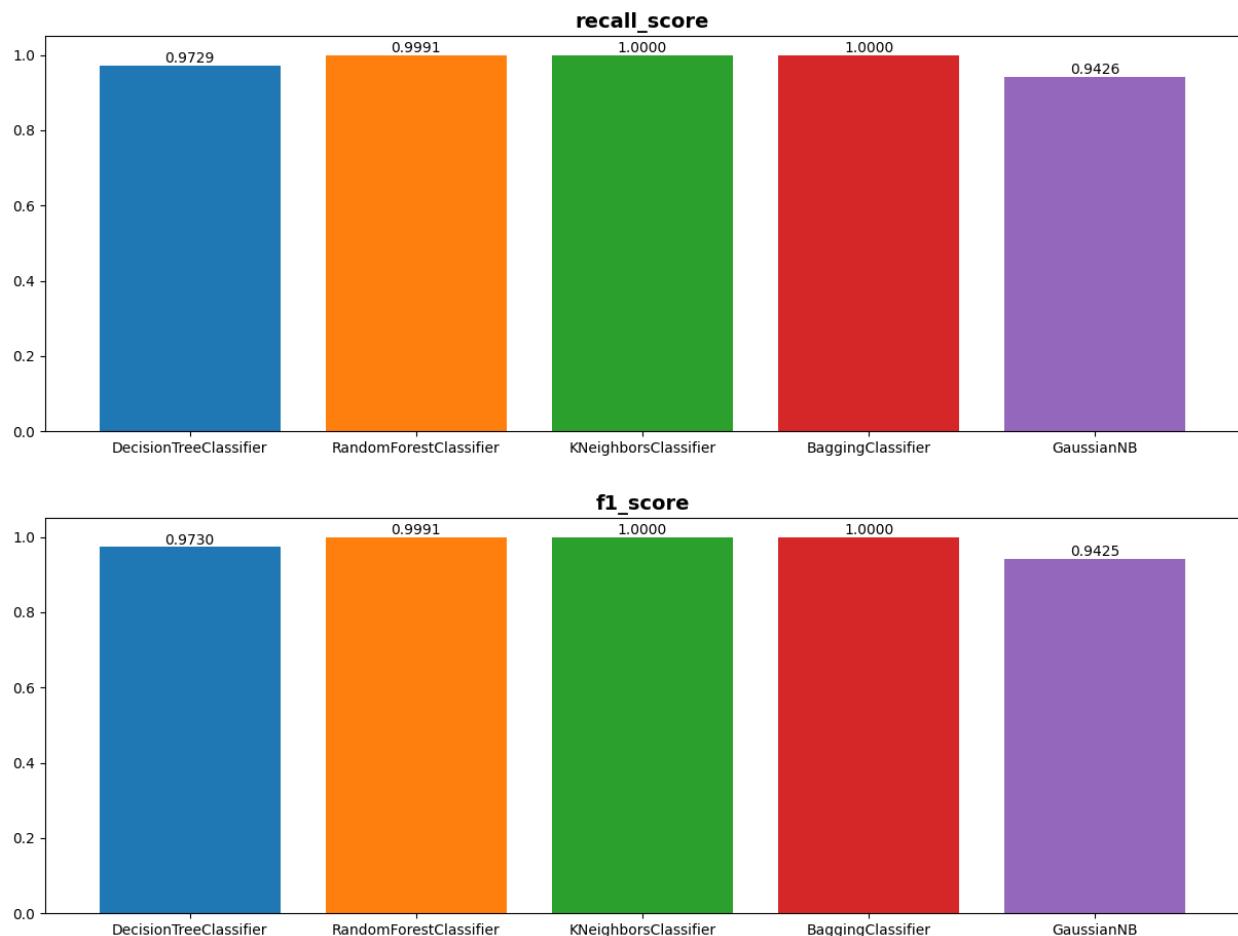




Najbolju tačnost, preciznost, odziv i f1 ima ansambl pakovanja koji koristi model k najbližih suseda, dok najslabije rezultate daje stablo odlučivanja.

Možemo porebiti mere tačnosti, preciznosti, odziva i f1 i na **trening** skupu.





Primećujemo da na trening skupu stablo odlučivanja ne daje najslabije rezultate i vrednosti su više u odnosu na test skup, što nam zapravo govori da ono ima najveći stepen preprilagođavanja.

Slučajna šuma, kao ansambl koji okuplja više stabala odlučivanja, daje bolje rezultate od pojedinačnog stabla, samim tim i manji stepen preprilagođavanja.

Gausov naivni Bajes, takođe daje nešto lošije rezultate, međutim u njegovom slučaju stepen preprilagođavanja nije velik kao kod stabla odlučivanja. S obzirom na to da algoritam prepostavlja normalnu raspodelu atributa, koja u našem istraživanju nije slučaj, očekivani su slabiji rezultati. Tehnika normalizacije raspodele primenjena na svaki atribut bi poboljšala kvalitet modela.

Najbolje rezultate generišu model *k* najbližih suseda i odgovarajući ansambl pakovanja, s tim što ansambl daje nešto bolje ocene, kako bismo i očekivali.

Klasterovanje

Uvod

Cilj ovog dela istraživanja je poređenje performansi različitih modela za klasterovanje. Koraci su slični onima za klasifikaciju, s tim što je neophodno izvršavati ponovno preprocesiranje jer klasterovanje zahteva drugačiji pristup obradi podataka i više ne vršimo podelu na skupove za obuku i testiranje. Svaki odabrani model biće obrađen zasebno, nakon čega će biti prikazano i njihovo poređenje.

Za potrebe ovog istraživanja odabrani su sledeći modeli: *K* sredina (*K*-Means), hijerarhijsko klasterovanje (Agglomerative) i DBSCAN.

Modeli su izabrani u skladu sa prirodom problema koji rešavamo, specifikacijama skupa podataka i rezultatima koje želimo. Motivacije za odabir modela svakog ponaosob će biti navedene u odgovarajućem odeljku.

Metodologija

Analiza skupa podataka

Oslonićemo se na [analizu podataka](#) izvršenu u okviru klasifikacije.

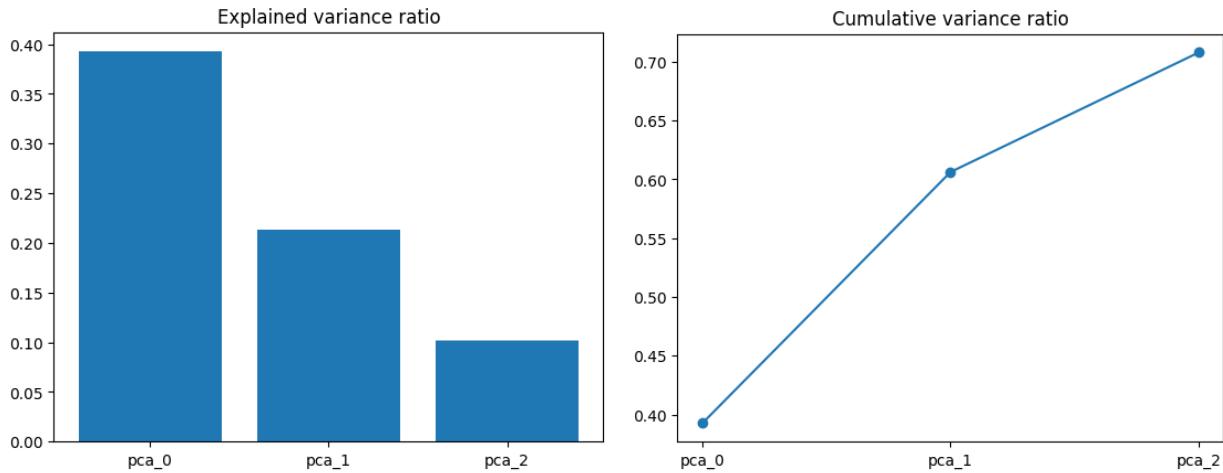
Preprocesiranje

Iako analiza pokazuje da postoje elementi van granica, nisu uklonjeni pre klasterovanja. Ideja je da se zadrži celokupnu sliku skupa podataka. Kao što je prethodno rečeno, skup je nebalansiran, međutim, to ne utiče na klasterovanje i zato je taj korak izuzet iz ovog dela istraživanja.

Redukcija dimenzionalnosti

Iz originalnog skupa podataka izuzeti su atributi koji predstavljaju klase, kao i dva atributa, identifikator jedinke i MFCC_1, koji ne doprinose informativnosti.

Na preostalom skupu je primenjena tehnika analize glavnih komponenti (PCA), pomoću koje smo sveli broj atributa sa 21 na 3. Iako tri dimenzije opisuju oko 70% varijanse podataka, praktičnija je vizualizacija klastera i efikasnije je izvršavanje algoritama.

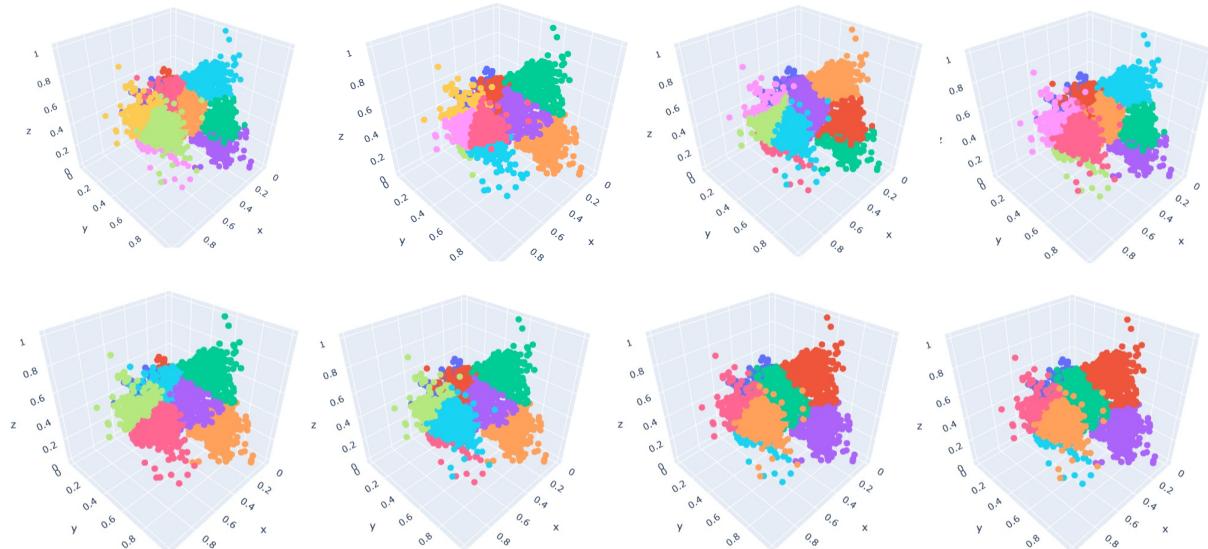


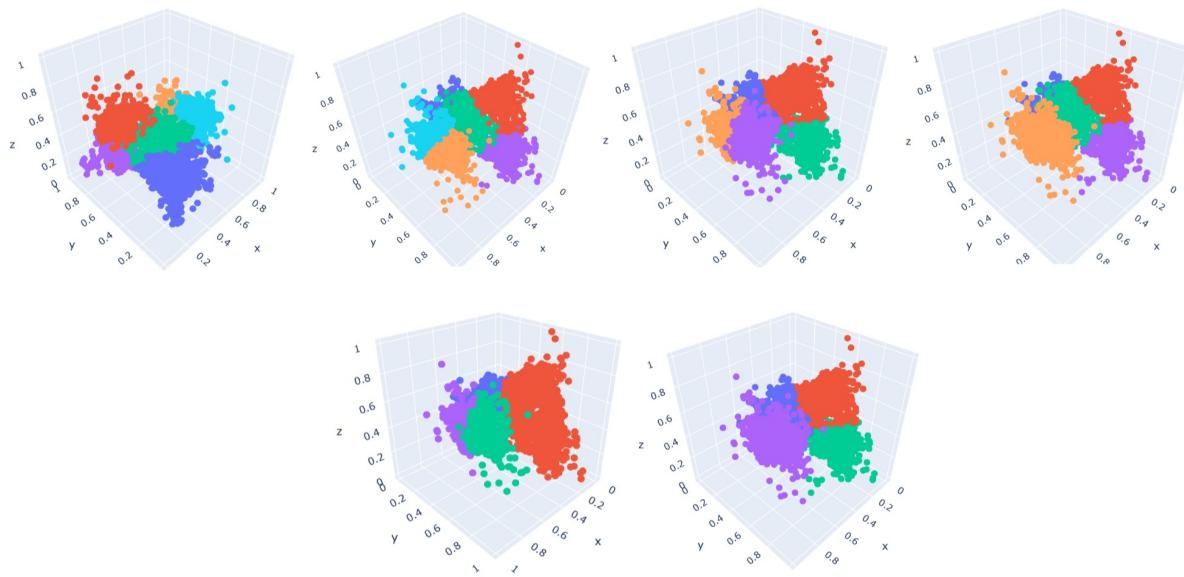
K sredina

K -means je iterativni algoritam koji deli podatke u K klastera, pri čemu se broj klastera unapred definiše. Algoritam računa centroid, i klastere deli u zavisnosti od udaljenosti od centroida. Suma kvadrata rastojanja između tačaka iz klastera i centroide treba da bude što manja. K sredina je jednostavan i efikasan algoritam koji dobro radi sa skupovima podataka sličnim našem, što je motivacija za njegovo korišćenje.

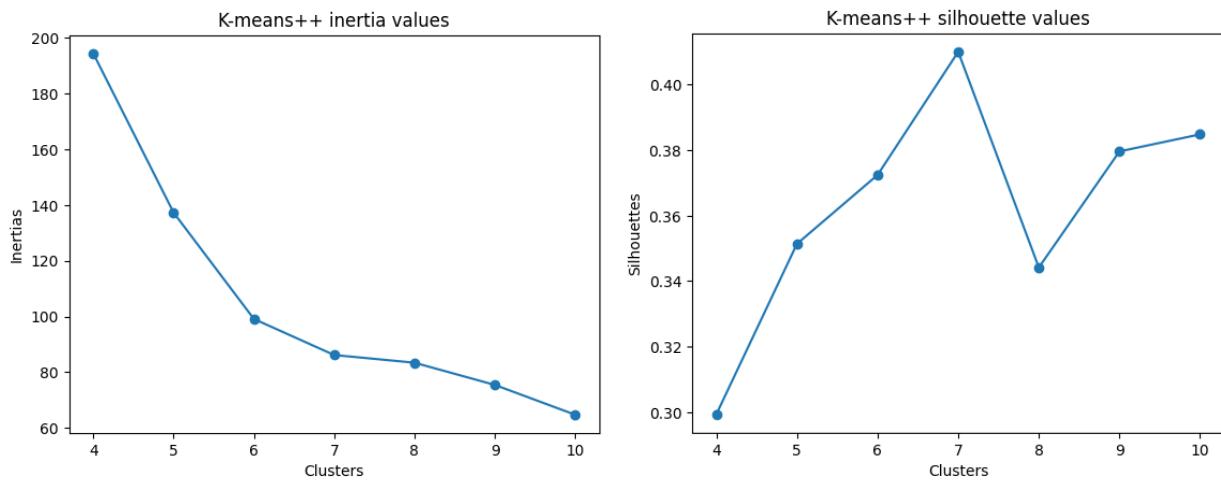
Definišemo broj klastera u rasponu od 4 do 10 i testiramo algoritam. Prirodno, skup podataka se deli na 4 porodice, 8 rodova i 10 vrsti, zbog čega je izabran ovaj raspon. Takođe, testiramo dva načina za inicijalizaciju centroida: *k-means++* i *random*.

Svake dve slike predstavljaju grafike za fiksani broj klastera (redom od 10 do 4), s tim što leva predstavlja korišćenje *k-means++*, a desna *random* parametra.





Mere koje koristimo za ocenu rada algoritma su inercija i silueta.



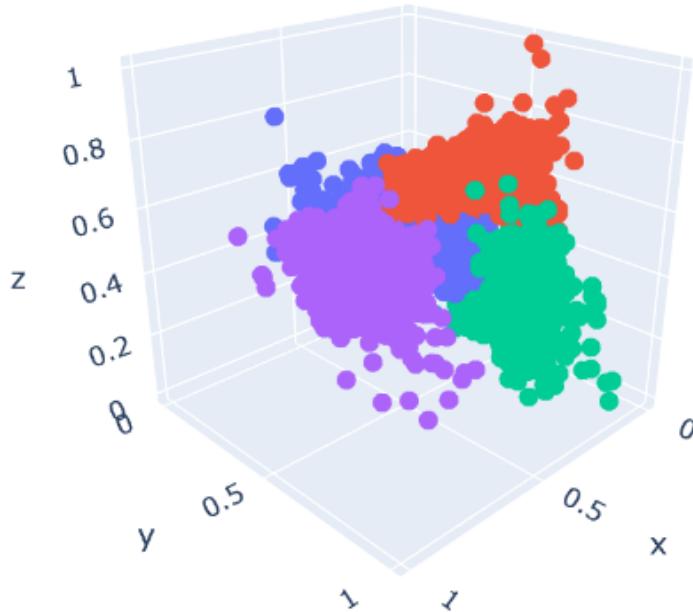
Cilj je da minimizujemo inerciju, za šta se najčešće koristi pravilo lakti. Na našem grafiku nije potpuno uočljiv lakt, tako da uzimamo u obzir još jednu važnu meru za ocenjivanje klastera, siluetu. Silueta je najveća za 7 i 10 klastera, ali je inercija približnja "laktu" u vrednosti 7 za broj klastera, tako da možemo da izdvojimo taj model kao najbolji.

Hijerarhijsko klasterovanje

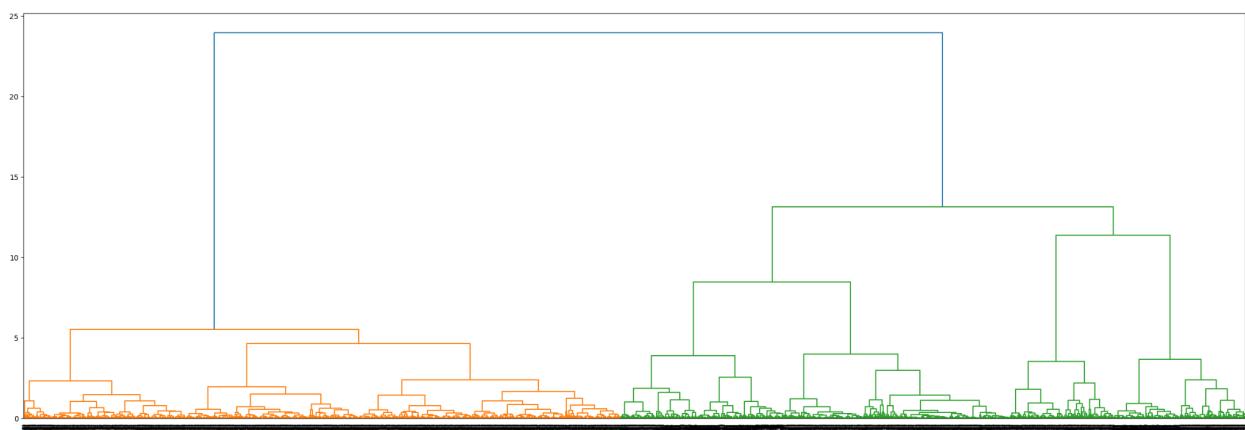
Motivacija za korišćenje hijerarhijskog klasterovanja je, pre svega, njegova interpretabilnost. Algoritam proizvodi hijerarhijsku strukturu klastera, koja se često predstavlja dendrogramom. Postavljanjem željenog broja klastera ili određivanjem praga, možemo izdvojiti klastere različitih veličina i gustina, u skladu sa specifičnim potrebama. Dodatno, hijerarhijsko klasterovanje je otpornije na elemente van granica, što je pogodno za skup podataka kakav je naš.

Testiramo ponašanje algoritma za različite vrednosti parametara. Definišemo klastere u rasponu od 4 do 10, slično kao za algoritam K sredina. Isprobavamo sve moguće vrednosti za parametar vezivanja, tj. linkage (*average*, *ward*, *complete*, *single*).

Izdvajamo model sa najboljom ocenom siluete, koja iznosi 0.52. Model definiše 4 različita klastera.



Da bismo dobili vizuelnu reprezentaciju hijerarhijskog klasterovanja koristimo biblioteku *scipy*, pomoću koje kreiramo dendrogram. Parametar *linkage* imaće vrednost *ward*.

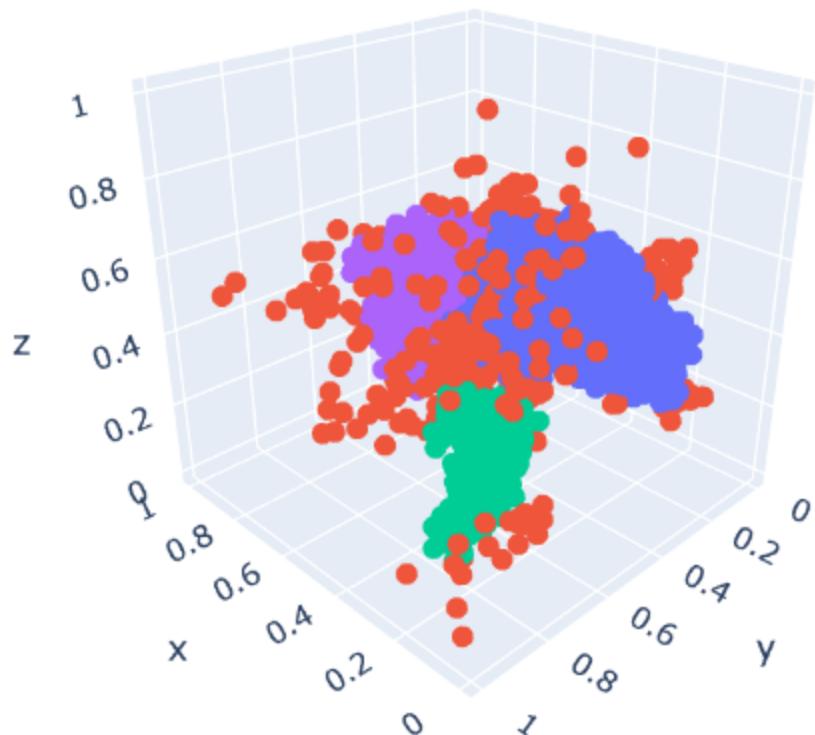


Zbog velikog broja instanci dendrogram gubi na deskriptivnosti, ali i dalje je moguće videti proces hijerahiskog klasterovanja.

DBSCAN

Poslednji model koji ćemo koristiti je DBSCAN. Prednost modela je što može da prepozna i obeleži elemente van granica u zasebnu grupu. Nije potrebno unapred definisati broj klastera. Jedine parametre koje testiramo su vrednost za epsilon i minimalan broj tačaka koje treba da se nađu u susedstvu neke tačke da bi se ona smatrала jezgrom klastera (*min_samples*). Pošto su podaci normalizovani i gusto raspoređeni, epsilon vrednost je najmanje 0.03, a najviše 0.1. Parametar *min_samples* će uzimati vrednosti od 10 do 50, sa korakom 10.

Model sa najboljom siluetom (0.47) definiše 3 klastera i elemente van granica.

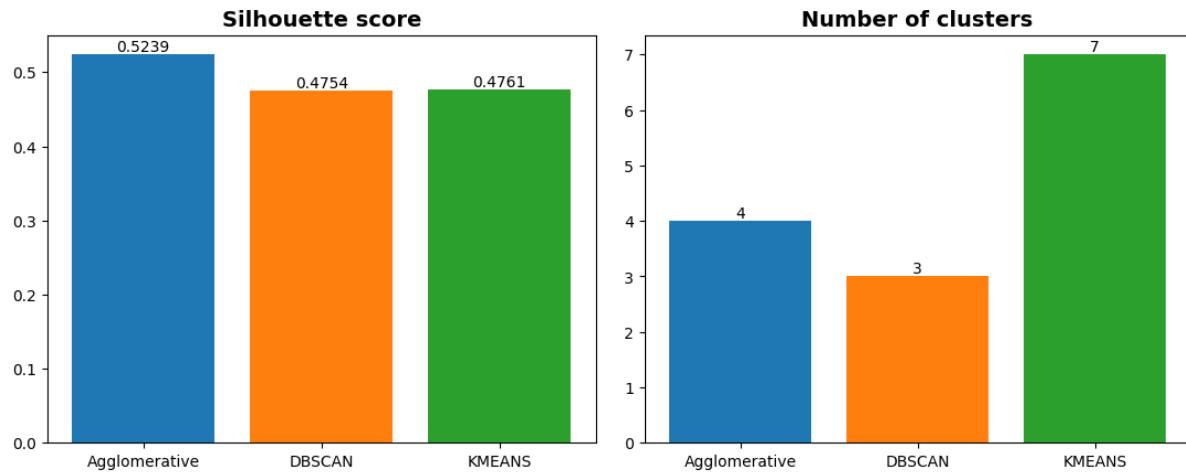


Plavom, zelenom i ljubičastom bojom su obeleženi klasteri. Crvena boja predstavlja elemente van granica.



Poređenje modela

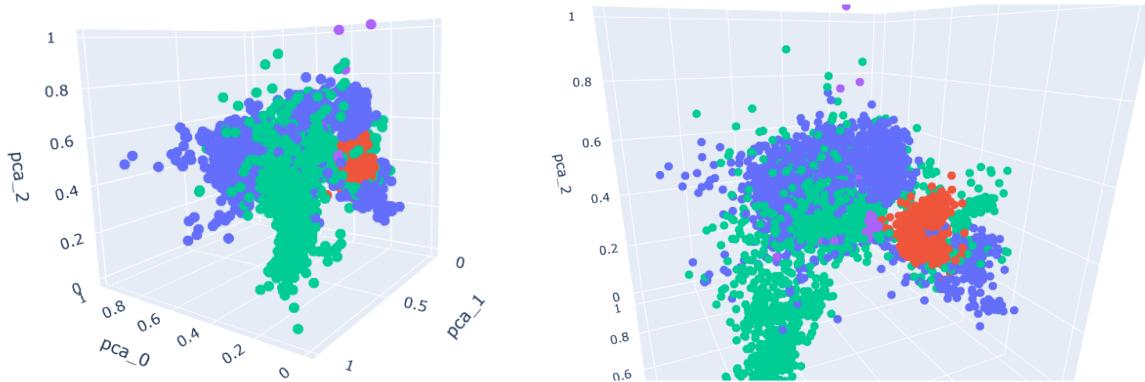
Poredili smo modele sa najboljim ocenama siluete, izuzev modela za K sredina koji je izabran koriteći metodu "lakta" zajedno sa najboljom ocenom siluete.



Hijerahijsko klasterovanje generiše model sa najboljom siluetom. Vrednost siluete za DBSCAN i K sredina je približna.

Najbolji model hijerahijskog klasterovanja definiše 4, a K sredina 7 klastera, dok DBSCAN definiše 3 sa dodatkom jednog koji predstavlja elemente van granica.

Ako uporedimo rezultate klasterovanja sa prirodnim grupama kojima pripadaju žabe, zaključujemo da nijedna metoda nije oslikala prirodnu podelu. Naime, ako pogledamo raspodelu po porodicama, primećujemo da su instance koje pripadaju različitim porodicama međusobno blizu i da je takvo grupisanje teško oslikati ovim modelima.



Pravila pridruživanja

Uvod

Pravila pridruživanja su tehnika mašinskog učenja zasnovana na pravilima koja se koristi za otkrivanje zanimljivih veza ili asocijacije između stavki u velikim skupovima podataka. Ova pravila identifikuju obrasce ili korelacije između različitih stavki na osnovu njihovog zajedničkog pojavljivanja u skupu podataka.

Koristeći pravila pridruživanja želimo da vidimo da li postoji međusobna zavisnost između kepstralnih koeficijenata mel skale, ili zavisnost vrste, odnosno roda i porodice od jednog ili više koeficijenata.

Proces će se izvršavati u SPSS modeleru.

Metodologija će podrazumevati učitavanje i pripremu podataka i upotrebu dva algoritma za pravila pridruživanja, *Apriori* i *FP-Growth*.

Apriori algoritam radi na osnovu "apriori" principa, koji kaže da bilo koji podskup čestog skupa stavki takođe mora biti čest. Algoritam počinje identifikacijom čestih pojedinačnih stavki, a zatim iterativno generiše veće skupove stavki spajanjem čestih skupova stavki iz prethodne iteracije. Primenjuje princip smanjenja pretrage primenom *downward closure* svojstva i koristi pragove podrške kako bi filtrirao retke skupove stavki.

FP-Growth (*Frequent Pattern-Growth*) algoritam je efikasnija alternativa *Apriori* algoritmu. On konstruiše kompaktnu strukturu podataka nazvanu FP-stablo (*FP-tree*), koja predstavlja česte skupove stavki i njihove veze. Algoritam izvodi dva prolaza kroz podatke: prvo, gradi FP-stablo, a zatim pronalazi česte skupove stavki rekursivnim pretraživanjem stabla. *FP-Growth* izbegava eksplicitno generisanje kandidatskih skupova stavki, koje može biti računski zahtevno kod *Apriori* algoritma, što rezultira bržim izvršavanjem.

Metodologija

Preprocesiranje

Podaci su učitani u originalnom formatu.

Da bismo izbegli kreiranje očiglednih pravila između vrsta, roda i porodice, eliminišemo atribut *Genus* i *Family*. Takođe, eliminišemo atribut koji predstavlja identifikaciju žabe, jer nam ne nosi važnu informaciju. S obizrom na to da atribut *MFCC_1* skoro uvek iznosi 1, uklanjamo i njega iz skupa, jer ne želimo da nam se kreiraju beznačajna pravila.

Atributi koji predstavljaju kepstralne koeficijente mel skale su numerički i kao takvi nisu nam pogodni za primenu algoritama za pravila pridruživanja. Nad njima primenjujemo

metodu diskretizacije (*binning*) i za svaki koeficijent kreiramo četiri grupe. Pošto smo attribute transformisali tako da su kategoričkog tipa, stare eliminisemo iz skupa podataka. Atribut *Species* je kategorički, tako da njega ostavljamo u originalnom formatu.

Za svaki atribut smo definisali da je ujedno i ulazni i ciljni, jer želimo da proverimo sve vrste zavisnosti.

Kao meru kvaliteta pravila posmatramo podršku (*support*), pouzdanost (*confidence*) i meru podizanja (*lift*).

Apriori

Nakon preprocesiranja, koristeći apriori čvor u SPSS modeleru, dobijamo skup generisanih pravila.

Number of Rules: 36,130	
Number of Valid Transactions: 7,195	
Minimum Support: 10.007%	
Maximum Support: 58.277%	
Minimum Confidence: 80.0%	
Maximum Confidence: 100.0%	
Minimum Lift: 1.655%	Minimalna vrednost za podršku je 10%, a maksimalni broj prethodnika je 5. U obzir uzimamo pravila čija je mera pouzdanosti bar 80%.
Maximum Lift: 6.378%	
Minimum Deployability: 0.0%	
Maximum Deployability: 11.647%	
Minimum Rule Support: 8.019%	
Maximum Rule Support: 47.477%	

Prikaz pet pravila sa naivćom merom podrške.

Consequent	Antecedent	Support %	Confidence %	Lift
Species = AdenomeraHylaedactylus	MFCCs_9_BIN = 3 MFCCs_10_BIN = 3	58.277	80.014	1.655
Species = AdenomeraHylaedactylus	MFCCs_9_BIN = 3 MFCCs_12_BIN = 3	56.511	82.612	1.709
Species = AdenomeraHylaedactylus	MFCCs_15_BIN = 2 MFCCs_5_BIN = 3 MFCCs_12_BIN = 3	55.942	80.547	1.666
Species = AdenomeraHylaedactylus	MFCCs_15_BIN = 2 MFCCs_5_BIN = 3 MFCCs_12_BIN = 3 MFCCs_10_BIN = 3	54.454	80.807	1.672
Species = AdenomeraHylaedactylus	MFCCs_9_BIN = 3 MFCCs_10_BIN = 3 MFCCs_6_BIN = 2	54.149	82.161	1.7

Primećujemo da su pravila prisutna u nešto više od pola instanci, imaju pouzdanost od oko 80% da će ciljna vrsta biti AdenomeraHylaedactylus i podizanje od oko 1.7 koje sugerije da je vrsta AdenomeraHylaedactylus 1.7 puta verovatnija kada se javljaju navedne kombinacije vrednosti.

Prikaz pet pravila sa najvećom merom pouzdanosti.

Consequent	Antecedent	Support %	Confidence %	Lift
Species = AdenomeraHylaedactylus	MFCCs_20_BIN = 1 MFCCs_11_BIN = 2 MFCCs_17_BIN = 3 MFCCs_3_BIN = 2	12.203	100.0	2.069
Species = AdenomeraHylaedactylus	MFCCs_20_BIN = 1 MFCCs_11_BIN = 2 MFCCs_17_BIN = 3 MFCCs_19_BIN = 2	12.064	100.0	2.069
Species = AdenomeraHylaedactylus	MFCCs_20_BIN = 1 MFCCs_11_BIN = 2 MFCCs_17_BIN = 3 MFCCs_13_BIN = 3	13.468	100.0	2.069
Species = AdenomeraHylaedactylus	MFCCs_20_BIN = 1 MFCCs_11_BIN = 2 MFCCs_17_BIN = 3 MFCCs_14_BIN = 2	11.383	100.0	2.069
Species = AdenomeraHylaedactylus	MFCCs_20_BIN = 1 MFCCs_11_BIN = 2 MFCCs_17_BIN = 3 MFCCs_16_BIN = 2	11.744	100.0	2.069

Vrednosti mera sugeriju da se navedena pravila javljaju samo u malom broju slučajeva, ali kada se javje, imaju visoku tačnost ili pouzdanost. Ovo može biti rezultat raspodele specifičnih atributa.

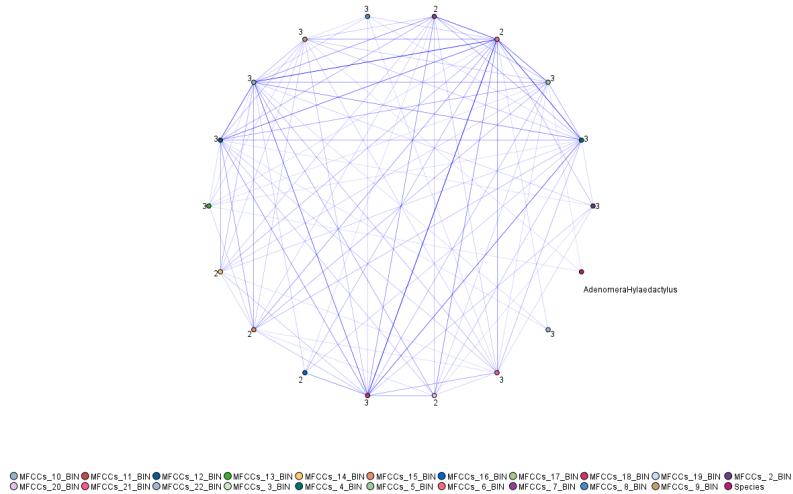
Prikaz pet pravila sa najvećom merom podizanja.

Consequent	Antecedent	Support %	Confidence %	Lift
Species = HypsiboasCordobae	MFCCs_3_BIN = 3 MFCCs_8_BIN = 2 MFCCs_11_BIN = 3 MFCCs_14_BIN = 2 MFCCs_12_BIN = 3	11.049	99.371	6.378
Species = HypsiboasCordobae	MFCCs_3_BIN = 3 MFCCs_8_BIN = 2 MFCCs_13_BIN = 2 MFCCs_14_BIN = 2 MFCCs_12_BIN = 3	10.716	99.351	6.377
Species = HypsiboasCordobae	MFCCs_3_BIN = 3 MFCCs_8_BIN = 2 MFCCs_13_BIN = 2 MFCCs_17_BIN = 2 MFCCs_14_BIN = 2	10.021	99.307	6.374
Species = HypsiboasCordobae	MFCCs_3_BIN = 3 MFCCs_8_BIN = 2 MFCCs_11_BIN = 3 MFCCs_17_BIN = 2 MFCCs_14_BIN = 2	10.354	99.195	6.367
Species = HypsiboasCordobae	MFCCs_3_BIN = 3 MFCCs_8_BIN = 2 MFCCs_13_BIN = 2 MFCCs_14_BIN = 2 MFCCs_20_BIN = 2	10.688	99.09	6.36

Visoka vrednost podizanja ukazuje na veoma jaku povezanost između uslova i posledice. Podrška od oko 10% govori o relativno retkom pojavljivanju uslova.

Ovi rezultati sugeruju da je pravilo ili uzorak veoma pouzdan i da je veoma verovatno da će se posledica desiti kada je uslov ispunjen.

Prikaz mreže koja predstavlja dobijena pravila.



Među pravilima sa najboljim rezultatima, u ulozi posledice, izdvojile su se samo dve vrste, *AdenomeraHylaedactylus* i *HypsiboasCordobae*, koje, redom, pripadaju porodicama *Leptodactylidae* i *Hylidae*.

Pogledaćemo kako se ponaša *Apriori* algoritam, ako vrstu zamenimo porodicom.

Consequent	Antecedent	Support %	Confidence %	Lift	Consequent	Antecedent	Support %	Confidence %	Lift
Family = Leptodactylidae	MFCCs_9_BIN = 3 MFCCs_10_BIN = 3	58.277	81.588	1.328	Family = Leptodactylidae	MFCCs_20_BIN... MFCCs_3_BIN...	12.495	100.0	1.628
Family = Leptodactylidae	MFCCs_9_BIN = 3 MFCCs_12_BIN = 3	56.511	83.571	1.36	Family = Leptodactylidae	MFCCs_20_BIN... MFCCs_11_BIN... MFCCs_19_BIN...	12.245	100.0	1.628
Family = Leptodactylidae	MFCCs_15_BIN = 2 MFCCs_5_BIN = 3 MFCCs_10_BIN = 3	56.275	80.02	1.303	Family = Leptodactylidae	MFCCs_20_BIN... MFCCs_11_BIN... MFCCs_17_BIN... MFCCs_3_BIN...	12.203	100.0	1.628
Family = Leptodactylidae	MFCCs_15_BIN = 2 MFCCs_5_BIN = 3 MFCCs_12_BIN = 3	55.942	80.671	1.313	Family = Leptodactylidae	MFCCs_20_BIN... MFCCs_11_BIN... MFCCs_17_BIN... MFCCs_19_BIN...	12.064	100.0	1.628
Family = Leptodactylidae	MFCCs_15_BIN = 2 MFCCs_5_BIN = 3 MFCCs_12_BIN = 3 MFCCs_10_BIN = 3	54.454	80.909	1.317	Family = Leptodactylidae	MFCCs_20_BIN... MFCCs_11_BIN... MFCCs_17_BIN... MFCCs_13_BIN...	13.468	100.0	1.628

Consequent	Antecedent	Support %	Confidence %	Lift
Family = Hylidae	MFCCs_3_BIN = 3 MFCCs_13_BIN = 2 MFCCs_19_BIN = 3 MFCCs_14_BIN = 2	13.259	99.895	3.32
Family = Hylidae	MFCCs_3_BIN = 3 MFCCs_13_BIN = 2 MFCCs_11_BIN = 3 MFCCs_19_BIN = 3 MFCCs_14_BIN = 2	13.204	99.895	3.32
Family = Hylidae	MFCCs_3_BIN = 3 MFCCs_13_BIN = 2 MFCCs_19_BIN = 3 MFCCs_14_BIN = 2 MFCCs_7_BIN = 2	13.176	99.895	3.32
Family = Hylidae	MFCCs_3_BIN = 3 MFCCs_13_BIN = 2 MFCCs_14_BIN = 2 MFCCs_18_BIN = 3	12.995	99.893	3.32
Family = Hylidae	MFCCs_3_BIN = 3 MFCCs_13_BIN = 2 MFCCs_19_BIN = 3 MFCCs_14_BIN = 2 MFCCs_6_BIN = 2	12.995	99.893	3.32

Dobijeni su veoma slični rezultati. Izdvajaju se porodice *Leptodactylidae* i *Hylidae*.

FP-Growth

Nakon preprocesiranja, koristeći *association rules* čvor u SPSS modeleru, dobijamo skup generisanih pravila. Parametri su slični onima u *Apriori* algoritmu.

Build Settings		Rule Statistics				
Maximum Number of Rules	1,000					
Minimum Condition Support	0.05					
Minimum Confidence	0.80					
Minimum Rule Support	0.10	Measurements	Minimum	Maximum	Mean	Standard Deviation
Minimum Lift	2.00	Condition Support (%)	10.22	18.28	13.00	1.27
Maximum Number of Items in a Rule	6	Confidence (%)	80.00	99.37	87.50	5.12
Maximum Number of Items in a Condition	5	Rule Support (%)	10.01	15.82	11.33	0.81
Maximum number of Items in a Prediction	1	Lift	2.77	6.38	5.34	0.88
Use only True Value for Flag Fields	True	Deployability (%)	0.07	3.24	1.67	0.77
Allow Rules without Conditions	False					
Evaluation Measure Sorting the Rules	Lift					

Kao meru kvaliteta pravila posmatramo podršku (*support*), pouzdanost (*confidence*) i meru podizanja (*lift*).

Izdvajamo pravila sa najvećom merom podrške.

Rank	Rule ID	Condition	Prediction	Other Evaluation Statistics				
				Sorted By Rule Support(%)	Condition Support (%)	Confidence (%)	Lift	Deployability (%)
1	984	MFCCs_16_BIN= 2 MFCCs_12_BIN = 2	MFCCs_14_BIN = 3	15.82	18.28	86.54	2.83	2.46
2	990	MFCCs_4_BIN = 3 MFCCs_12_BIN = 2	MFCCs_14_BIN = 3	14.64	17.08	85.68	2.80	2.45
3	970	MFCCs_18_BIN = 3 MFCCs_12_BIN = 2	MFCCs_14_BIN = 3	14.50	16.50	87.87	2.88	2.00
4	952	MFCCs_17_BIN= 2 MFCCs_12_BIN = 2	MFCCs_14_BIN = 3	14.32	15.98	89.57	2.93	1.67
5	954	MFCCs_18_BIN = 3 MFCCs_16_BIN = 2 MFCCs_12_BIN = 2	MFCCs_14_BIN = 3	13.50	15.08	89.49	2.93	1.58

Ne izdvajaju se pravila koja kao posledicu imaju vrstu. Pravila sa najvećom podrškom, i za posledicu i za vrstu imaju kepstralni koeficijent mel skale. Primećujemo da je podrška mala, što znači da ova pravila nisu česta, međutim pouzdanost i podizanje su visoki, tako da su pravila verodostojna.

Za razliku od onih pravila kojima je posledica neka od vrsta, ova su interesantna jer pokazuju zavisnost između određenih vrednosti kepstralnih koeficijenata mel skale.

Izdvajamo pravila sa najvećom merom pouzdanosti.

Most Interesting Rules by Confidence

Rank	Rule ID	Condition	Prediction	Sorted By Confidence(%)	Other Evaluation Statistics			
					Condition Support (%)	Rule Support (%)	Lift	Deployability (%)
1	1	MFCCs_12_BIN = 3 MFCCs_14_BIN = 2 MFCCs_11_BIN = 3 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	99.37	11.05	10.98	6.38	0.07
2	2	MFCCs_12_BIN = 3 MFCCs_14_BIN = 2 MFCCs_13_BIN = 2 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	99.35	10.72	10.65	6.38	0.07
3	3	MFCCs_14_BIN = 2 MFCCs_17_BIN = 2 MFCCs_11_BIN = 3 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	99.19	10.35	10.27	6.37	0.08
4	4	MFCCs_20_BIN = 2 MFCCs_14_BIN = 2 MFCCs_13_BIN = 2 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	99.09	10.69	10.59	6.36	0.10
5	5	MFCCs_20_BIN = 2 MFCCs_14_BIN = 2 MFCCs_11_BIN = 3 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	98.99	11.04	10.92	6.35	0.11

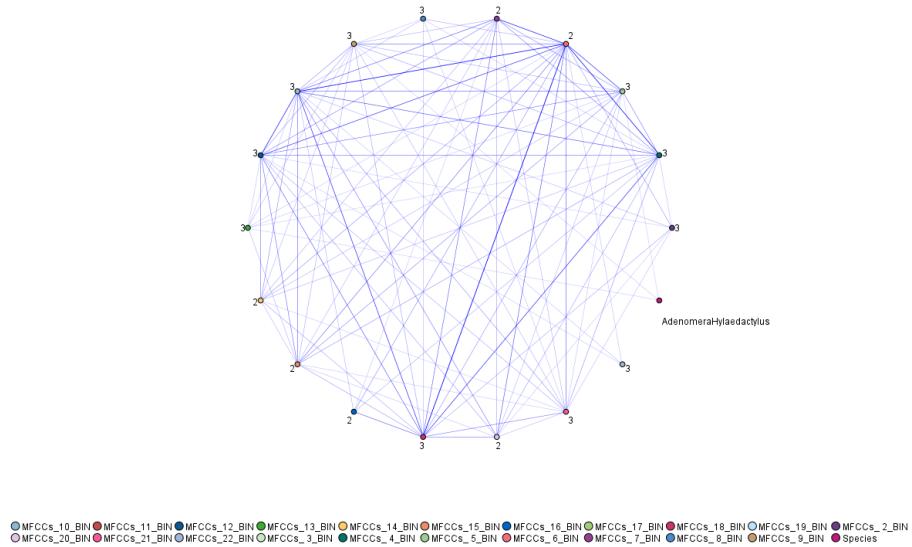
Izdvajamo pravila sa najvećom merom podizanja.

Most Interesting Rules by Lift

Rank	Rule ID	Condition	Prediction	Sorted By Lift	Condition Support (%)	Other Evaluation Statistics		
						Confidence (%)	Rule Support (%)	Deployability (%)
1	1	MFCCs_12_BIN = 3 MFCCs_14_BIN = 2 MFCCs_11_BIN = 3 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	6.38	11.05	99.37	10.98	0.07
2	2	MFCCs_12_BIN = 3 MFCCs_14_BIN = 2 MFCCs_13_BIN = 2 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	6.38	10.72	99.35	10.65	0.07
3	3	MFCCs_14_BIN = 2 MFCCs_17_BIN = 2 MFCCs_11_BIN = 3 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	6.37	10.35	99.19	10.27	0.08
4	4	MFCCs_20_BIN = 2 MFCCs_14_BIN = 2 MFCCs_13_BIN = 2 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	6.36	10.69	99.09	10.59	0.10
5	5	MFCCs_20_BIN = 2 MFCCs_14_BIN = 2 MFCCs_11_BIN = 3 MFCCs_8_BIN = 2 MFCCs_3_BIN = 3	Species = HypsiboasCordobae	6.35	11.04	98.99	10.92	0.11

Pravila sa najvećom merom pouzdanosti su ujedno i ona sa najvećom merom podizanja, to ukazuje na snažan i pouzdan odnos između uslova i posledice, što označava značajnu povezanost u skupu podataka. Uslov je najčešće vrsta *HypsiboasCordobae*.

Prikaz mreže koja predstavlja dobijene rezultate.



Možemo da pogledamo kako se ponaša algoritam sa izmenjenim parametrima. Ovaj put ograničavamo ukupan broj prethodnika po pravilu na 3.

Build Settings

Maximum Number of Rules	1,000
Minimum Condition Support	0.05
Minimum Confidence	0.70
Minimum Rule Support	0.05
Minimum Lift	2.00
Maximum Number of Items in a Rule	6
Maximum Number of Items in a Condition	3
Maximum number of Items in a Prediction	1
Use only True Value for Flag Fields	True
Allow Rules without Conditions	False
Evaluation Measure Sorting the Rules	Lift

Rule Statistics					
	Measurements	Minimum	Maximum	Mean	Standard Deviation
	Condition Support (%)	5.05	18.68	8.44	2.51
	Confidence (%)	70.02	99.45	83.69	7.05
	Rule Support (%)	5.00	13.45	6.99	1.82
	Lift	3.03	9.88	4.33	1.49
	Deployability (%)	0.03	5.56	1.45	0.94

Pogledaćemo pravila sa najboljim ocenama podrške, pouzdanosti i podizanja, redom.

Most Interesting Rules by Rule Support

Rank	Rule ID	Condition	Prediction	Other Evaluation Statistics				
				Sorted By Rule Support(%)	Condition Support (%)	Confidence (%)	Lift	Deployability (%)
1	246	MFCCs_12_BIN = 3 MFCCs_19_BIN = 3 MFCCs_8_BIN = 2	Species = <i>HypsiboasCordobae</i>	13.45	18.05	74.52	4.78	4.60
2	556	MFCCs_10_BIN = 4	MFCCs_12_BIN = 2	13.38	16.53	80.99	3.69	3.14
3	290	MFCCs_11_BIN = 3 MFCCs_3_BIN = 3	Species = <i>HypsiboasCordobae</i>	13.19	18.44	71.51	4.59	5.25
4	224	MFCCs_7_BIN = 2 MFCCs_11_BIN = 3 MFCCs_3_BIN = 3	Species = <i>HypsiboasCordobae</i>	13.18	17.33	76.02	4.88	4.16
5	844	MFCCs_13_BIN = 2 MFCCs_14_BIN = 3	MFCCs_12_BIN = 2	13.12	18.68	70.24	3.20	5.56

Most Interesting Rules by Confidence

Rank	Rule ID	Condition	Prediction	Other Evaluation Statistics				
				Sorted By Confidence(%)	Condition Support (%)	Rule Support (%)	Lift	Deployability (%)
1	819	MFCCs_14_BIN = 2 MFCCs_17_BIN = 3 MFCCs_18_BIN = 2	MFCCs_16_BIN = 3	99.45	5.05	5.02	3.23	0.03
2	827	MFCCs_22_BIN = 3 MFCCs_17_BIN = 3 MFCCs_18_BIN = 2	MFCCs_16_BIN = 3	99.11	6.25	6.20	3.22	0.06
3	829	MFCCs_10_BIN = 3 MFCCs_17_BIN = 3 MFCCs_18_BIN = 2	MFCCs_16_BIN = 3	99.02	7.12	7.05	3.22	0.07
4	830	MFCCs_6_BIN = 2 MFCCs_17_BIN = 3 MFCCs_18_BIN = 2	MFCCs_16_BIN = 3	99.02	7.10	7.03	3.22	0.07
5	832	MFCCs_3_BIN = 2 MFCCs_17_BIN = 3 MFCCs_18_BIN = 2	MFCCs_16_BIN = 3	98.99	5.49	5.43	3.22	0.06

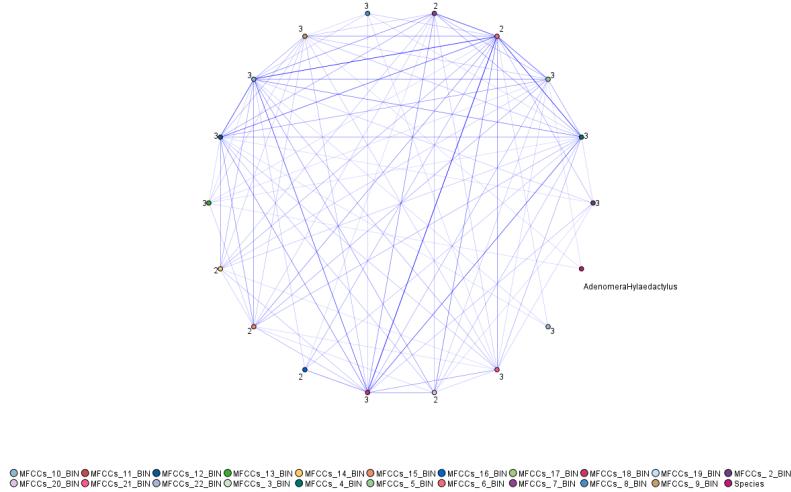
Most Interesting Rules by Lift

Rank	Rule ID	Condition	Prediction	Other Evaluation Statistics				
				Sorted By Lift	Condition Support (%)	Confidence (%)	Rule Support (%)	Deployability (%)
1	1	MFCCs_17_BIN = 2 MFCCs_12_BIN = 2 MFCCs_11_BIN = 4	Species = <i>AdenomeraAndre</i>	9.88	7.00	92.26	6.46	0.54
2	2	MFCCs_21_BIN = 3 MFCCs_12_BIN = 2 MFCCs_11_BIN = 4	Species = <i>AdenomeraAndre</i>	9.82	6.88	91.72	6.31	0.57
3	3	MFCCs_20_BIN = 2 MFCCs_12_BIN = 2 MFCCs_11_BIN = 4	Species = <i>AdenomeraAndre</i>	9.79	6.78	91.39	6.20	0.58
4	4	MFCCs_19_BIN = 2 MFCCs_17_BIN = 2 MFCCs_12_BIN = 2	Species = <i>AdenomeraAndre</i>	9.73	6.25	90.89	5.68	0.57
5	5	MFCCs_13_BIN = 2 MFCCs_12_BIN = 2 MFCCs_11_BIN = 4	Species = <i>AdenomeraAndre</i>	9.67	5.62	90.35	5.07	0.54

Dobili smo nova pravila koja za posledicu imaju vrstu *AdenomeraAndre* i atribut MFCC_16. Pravila sa najvećom podrškom imaju zanimljive posledice, vrstu *HypsiboasCordobae*, koja se javlja i ranija, ali i atribut MFCC_12. Podizanje je visoko, međutim pouzdanost je između 70% i 80%.

Pravila koja imaju visoke vrednosti za podizanje i pouzdanost, imaju manju vrednost podrške.

Prikaz mreže koja predstavlja dobijene rezultate.



Poređenje modela

Iako su oba modela pružila zanimljive rezultate, dajemo prednost algoritmu FP-Growth. Razlog tome je što su pravila koja je generisao ovaj algoritam korisnija i raznolikija, a mere podrške, pouzdanosti i podizanja su obično više.

Uočljivo je da su tri vrste usko vezane za konkretnе raslove vrednosti različitih atributa, pre svega *HypsiboasCordobae*, *AdenomeraHylaedactylus* i *AdenomeraAndre*. Takav rezultat je smislen, ako uzmemo u obzir to da *AdenomeraHylaedactylus* i *AdenomeraAndre* pripadaju porodici *Leptodactylidae* koja obuhvata značajan deo instanci u skupu podataka.

Sa druge strane, interesantna je povezanost vrste *HypsiboasCordobae* sa određenim atributima. Naime, broj instanci te vrste predstavlja samo oko 4% ukupnog broja instanci. Možemo da zaključimo da ima snažan uticaj na druge promenljive ili je povezana sa specifičnim obrascima u podacima. Visoke vrednosti podrške, pouzdanosti i podizanja ukazuju na to da su pravila koja sadrže tu klasu relevantna i često se pojavljuju u podacima.

U ovakvim situacijama, ova retka klasa može biti posebno zanimljiva za dalju analizu i istraživanje. Može se smatrati ključnom promenljivom u pravilima pridruživanja i može pružiti važne uvide o odnosima i uzrocima među promenljivima.

Mreže pokazuju najsnažniju povezanost između vrednosti atributa MFCC_4 i MFCC_6, MFCC_6 i MFCC_10, kao i MFCC_6 i MFCC_18.

Zaključak

U zaključku istraživanja podataka na skupu žaba i njihovih oglašavanja, sprovedene su analize u tri ključna područja: klasifikacija, klasterovanje i asocijativna pravila. Kroz ovo istraživanje, stekli smo dublje razumevanje karakteristika i veza u oglašavanjima žaba.

Prvo, kroz klasifikaciju smo izgradili modeli koji su sposobni da klasifikuju žabe i njihove glasove na različite vrste. Koristili smo modeli kao što su stablo odlučivanja, k najbližih suseda i naivni Bajes. Ovi modeli su pokazali visoku preciznost i uspešnost u klasifikaciji vrsta na osnovu karakteristika glasova.

Drugo, klasterovanje nam je omogućilo identifikaciju sličnosti i grupisanje glasova prema njihovim zajedničkim karakteristikama. Koristili smo hijerarhijsko klasterovanje i metrike poput linkage kako bismo stvorili klastere koji odražavaju prirodne grupe u podacima. Primenili smo k sredina i DBSCAN algoritme kako bismo stvorili klastere na osnovu prostornih i gustinskih karakteristika glasova. Kroz ove metode klasterovanja, identifikovali smo grupe sličnih glasova, omogućujući nam bolje razumevanje varijacija i klasifikaciju nepoznatih glasova. Ovaj pristup nam je omogućio otkrivanje skrivenih obrazaca i grupisanje glasova na osnove njihove strukture i sličnosti.

Treće, primenom pravila pridruživanja na skup podataka, otkrili smo značajne veze između različitih karakteristika glasova. Identifikovali smo pravila koja opisuju veze između određenih glasova žaba i njihovih karakteristika. Ova otkrića mogu biti korisna za bolje razumevanje i interpretaciju glasova, kao i za dalje istraživanje u ovom području.

U celini, ovo istraživanje na skupu podataka o žabama i njihovim oglašavanjima pružilo nam je korisne uvide u klasifikaciju, klasterovanje i asocijativna pravila. Ovi rezultati mogu biti od pomoći u boljem razumevanju anuranskih poziva, njihove raznolikosti i međusobnih veza. Nadalje, rezultati ovog istraživanja mogu poslužiti kao osnova za dalje istraživanje i primenu u biološkim studijama, ekologiji ili zaštiti prirode.