

Univerzitet Singidunum

Tehnički fakultet

Izrada projekta u R programskom jeziku

Profesor: Angelina Njeguš Student:

Asistent: Petar Biševac Maja Ercegovčević 2018/201902

Beograd, 2022.godina

**Sadržaj**

1. Analiza seta podataka **3**
2. Čišćenje seta podataka**4**
3. Smanjenje dimenzionalnosti I vizualizacija podataka **5**
4. Primena algoritma za mašinsko učenje**9**

**4.1**.Logistic Regression**9**

***4.2.****Random Forest****11***

1. Zaključak**13**

**1.Analiza seta podataka**

Za izradu ovog projekta, odlučila sam se za set podataka koji se odnose na crvene varijante portugalskog vina „Vinho Verde“. Zbog problema privatnosti i logistike, dostupne su samo fizičko-hemijske (inputi) i senzorne (output) varijable (npr. nema podataka o vrsti grožđa, brendu vina, prodajnoj ceni vina itd.). Ovi skupovi podataka se mogu posmatrati kao zadaci klasifikacije ili regresije. Časovi su raspoređeni i nisu izbalansirani (npr. ima mnogo više normalnih vina nego odličnih ili loših). Algoritmi za otkrivanje odstupanja mogu se koristiti za otkrivanje nekoliko odličnih ili loših vina.

Dati set podataka „winequality-red“ je preuzet sa Keggle sajta i nalazi se na sledećem linku: <https://www.kaggle.com/code/vishalyo990/prediction-of-quality-of-wine/>

Za potrebe ovog projekta biće pokušano predviđanje kvaliteta vina na osnovu ostalih parametara iz date baze podataka. Biće korišćeni svi parametri jer je dati set relativno mali pa se neće raditi postupak uklanjanja atributa.

Sastoji se iz sledećih 12 atributa(varijabli):

1 - fixed acidity (vinska kiselina - g / dm^3)

2 - volatile acidity (sirćetna kiselina - g / dm^3)

3 - citric acid (g / dm^3)

4 - residual sugar (g / dm^3)

5 - chlorides (natrijum hlorid - g / dm^3

6 - free sulfur dioxide (mg / dm^3)

7 - total sulfur dioxide (mg / dm^3)

8 - density (g / cm^3)

9 - pH

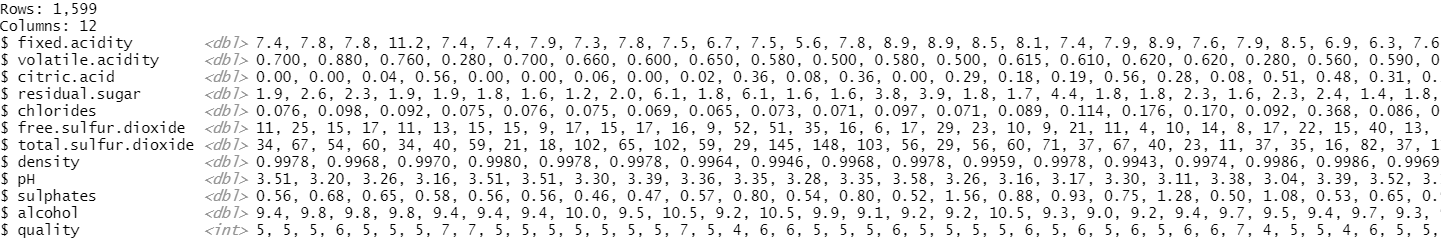
10 - sulphates (kalijum sulfat - g / dm3)

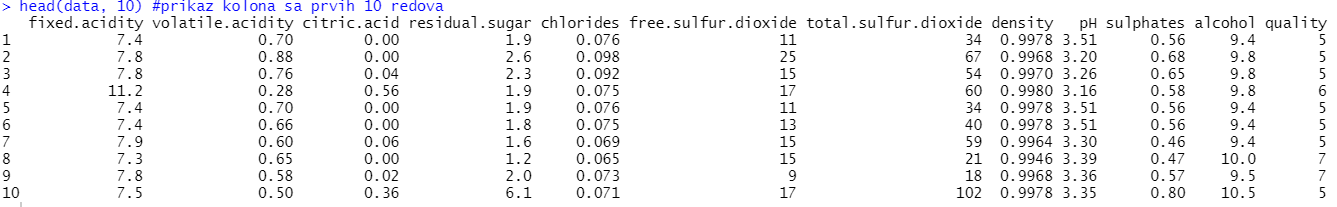
11 - alcohol (% po zapremini)

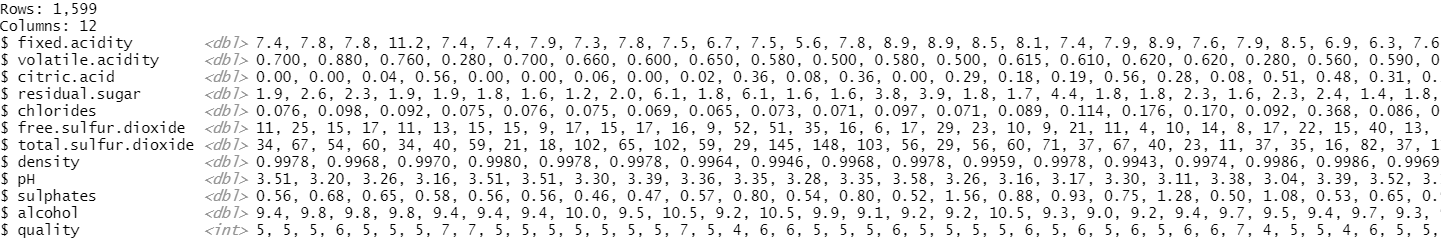
12 - quality (ocena između 0 i 10)

U nastavku je prikazana struktura podataka kao i izgled prvih 10 redova

Naredba kojom se prikazuje struktura: glimpse(data)

*Slika1: Prikaz strukture seta podataka*



*Slika2: Prikaz prvih 10 redova seta podataka*

Na osnovu ovih podataka biće izvršeno predviđanje upotrebom algoritama Logistic Regression I Random Forest koji će klasifikovati naše podatke po tome da li ispunjavaju traženi uslov ili ne.

**2.Čišćenje seta podataka**

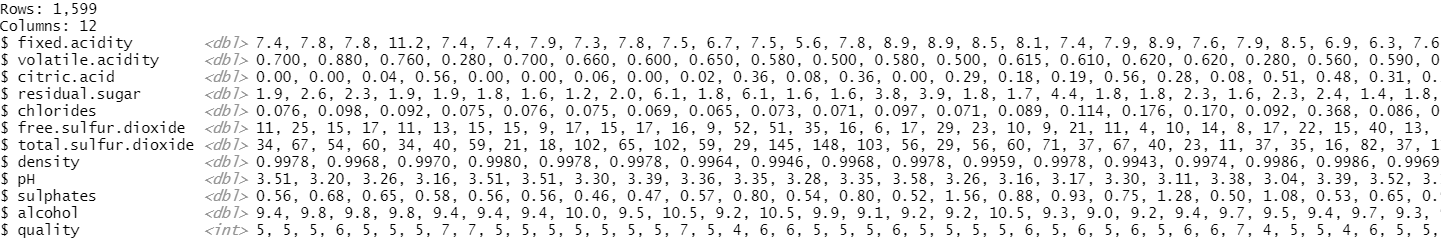
Zbog malog obima seta podataka, nije bilo mogućnosti da se neki atributi uklone, ali su izdvojene jedinstvene vrednosti I izbačene NA vrednosti.



*Slika3: Prikaz primera kako bi se uklonili nepotrebni atributi iz seta podataka*

Nakon ove naredbe set podataka je manjen na 1395 redova

****

*Slika4: Prikaz izbacivanja duplikata iz seta podataka*

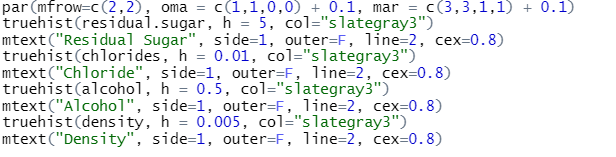
Takođe, proverili smo dali nasa tabela sadrži NA vrednosti koje nam nisu potrebe u daljem radu sa podacima i došli smo do zaključka da u našem setu podataka nema takvih vrednosti.

****

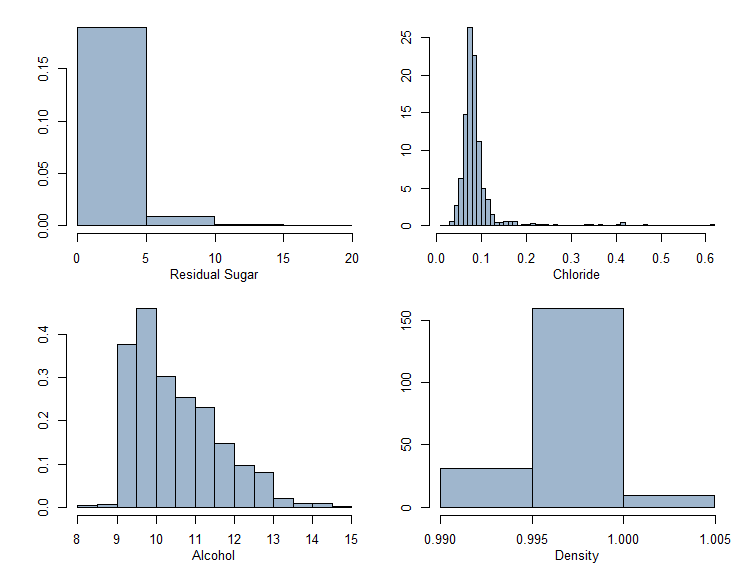
*Slika5: Prikaz provere postojanja Na vrednosti u setu podataka*

**3.Smanjenje dimenzionalnosti I vizualizacija podataka**

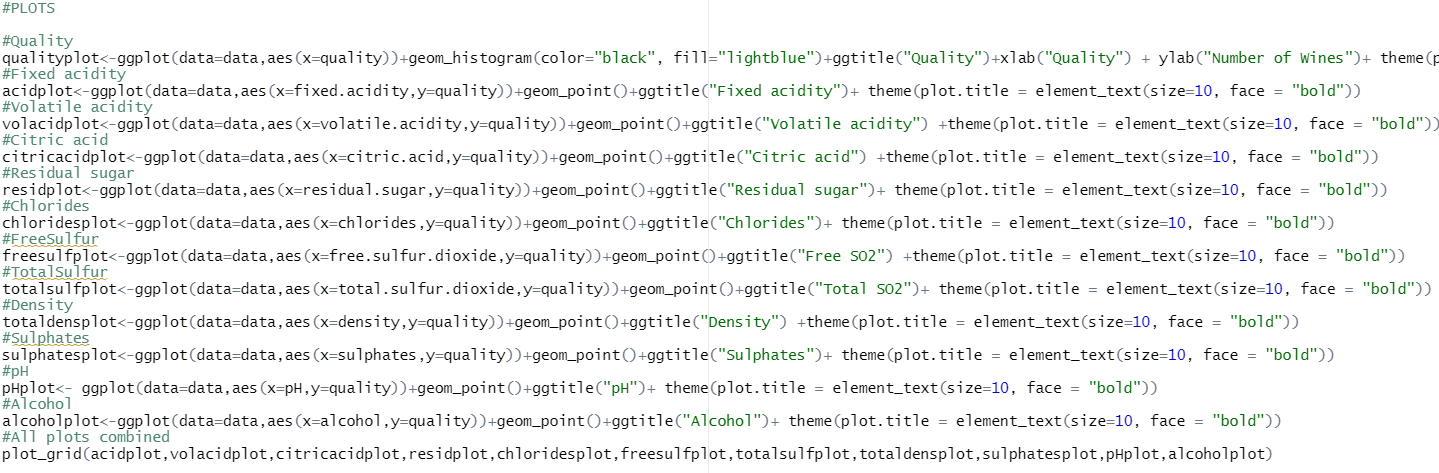
Nakon sto smo malo sredili podatke, sada ćemo videte histograme koji prikazuju količinu alkohola, gustinu, sulfat I preostali šećera. Takodje, biće prikazani I diagrami nezavisnih promenljivih u odnosu na zavisnu promenljivu (quality).

****

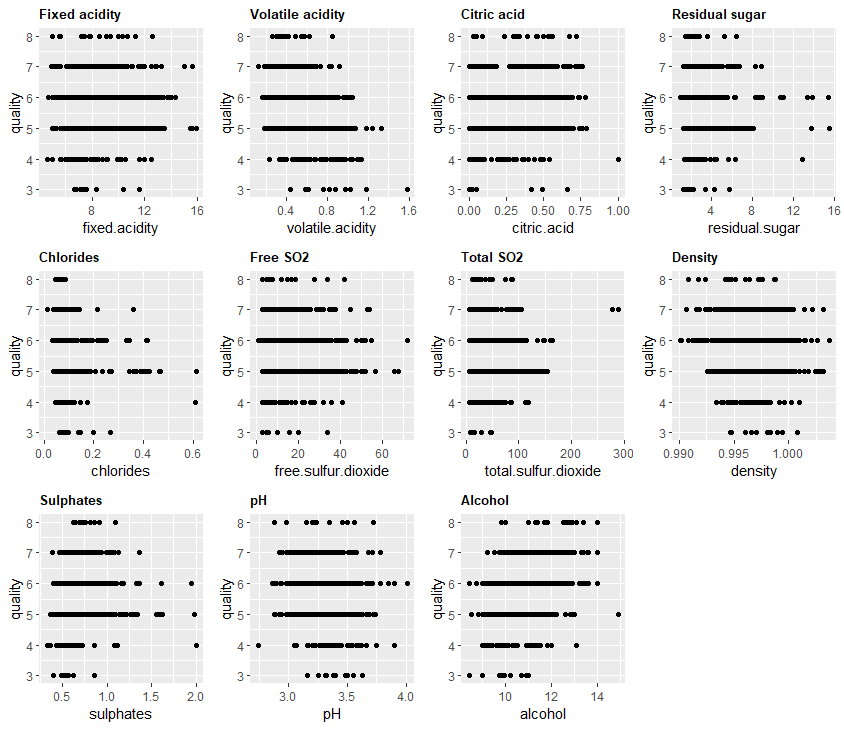
*Slika6: Prikaz koda za crtanje histograma*

****

*Slika7: Prikaz histograma*

**

*Slika8: Prikaz koda za crtanje odnosa nezavisnih promenljivih I zavisne promenljive*

****

*Slika9: Kombinovani prikaz svih diagrama zavisnosti*

Možemo primetiti da postoji linearna veza izmedju kvaliteta I ostalih promenljivih.

Kvalitet je imao najviše vrednosti u rasponu od 5 do 7, pri čemu nijedno vino nije rangirano od 1-2 ili 9-10.

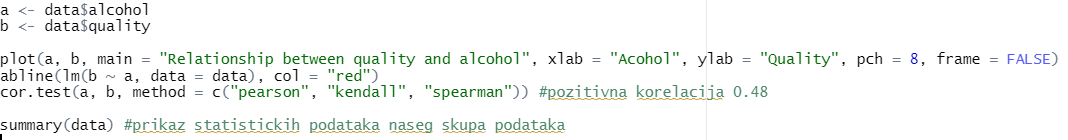
Fiksna I isparljiva kiselost imaju distribuciju koja je približna normalnoj , ali kao I limunska kiselina, ona je nagnuta udesno.

Gustina I pH imaju normalnu distribuciju sa prosečnim pH 3.31 što ukazuje na visoku koncentraciju vina koja su kiselija.

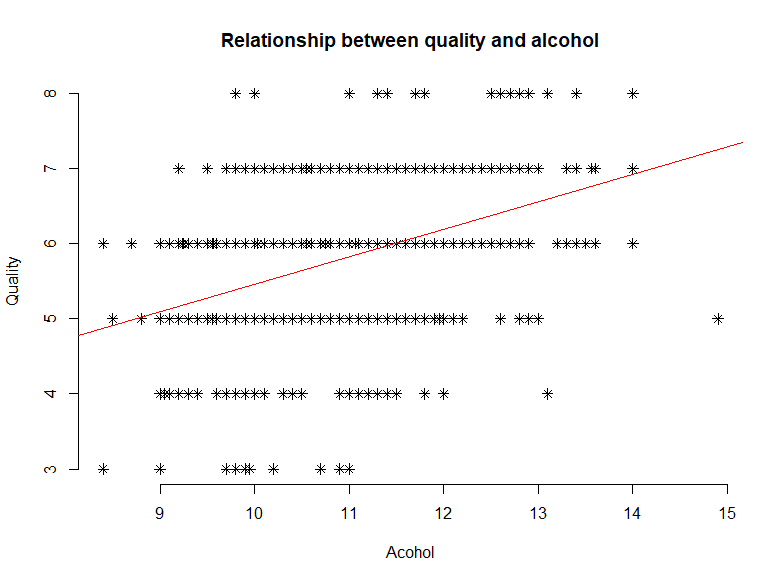
Ukupni sumpor-dioksid i slobodni sumpor-dioksid su takođe nagnuti udesno i takođe se sastoje od graničnih vrednosti.

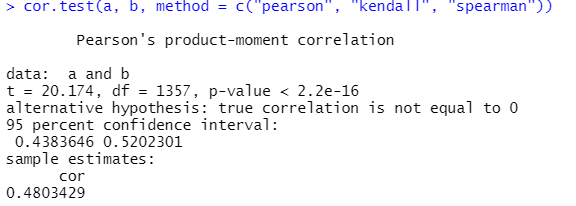
Prosečan procenat alkohola je 10,44%, pri čemu je maksimalni 14,9, a minimalni 8,4%.

Sledeće što smo uradili je bilo utvrđivanje **da li postoji zavisnosti** izmedju alkohola I kvaliteta



*Slika10: Prikaz koda za određivanje korelacije*

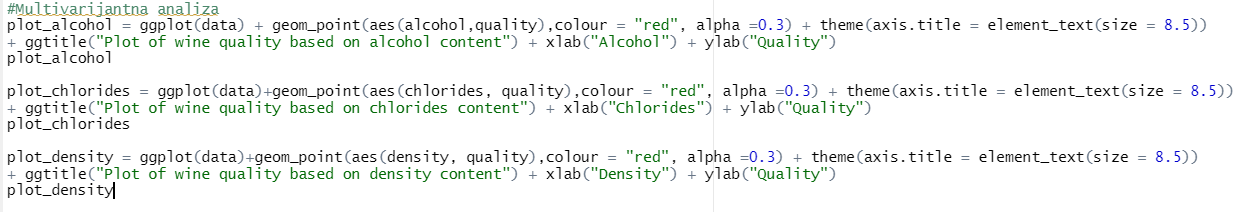
****Navedenim kodom smo došli do pozitivnog rezultata jer nam je koeficijent korelacije pozitivan pa samim tim postoji I pozitivna zavisnost izmedju alkohola I kvaliteta vina.



*Slika11: Prikaz rezultata Pearsonovog testa korelacije*

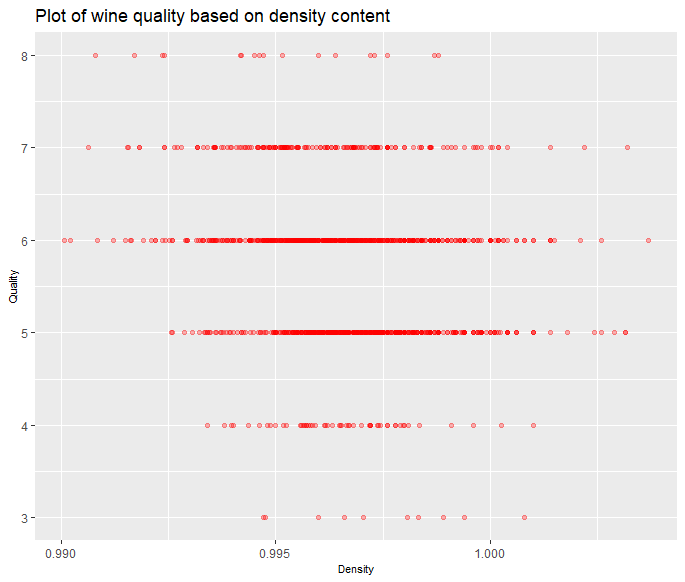
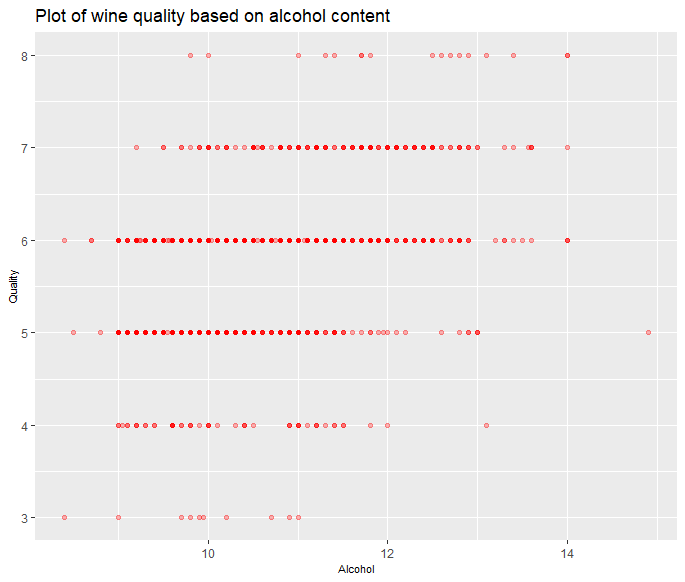
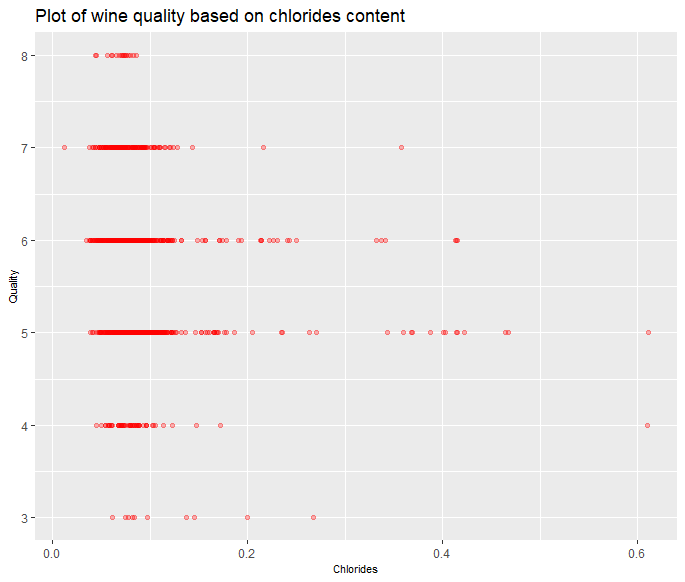
*Slika12: Prikaz odnosa izmedju kvaliteta I alkohola*

Zatim smo primenili multivarijantnu analizu nad atributima koji su prikazani na histogramima I koji predstavljaju odnos kvaliteta vina I alkohola, gustine I sulfata.



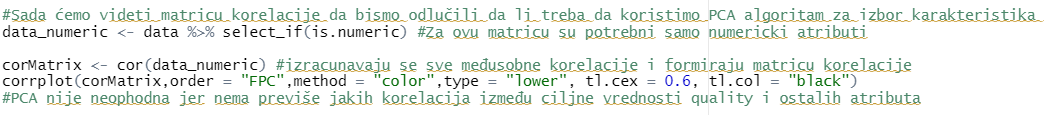
*Slika13: Prikaz koda za mulitarijantnu analizui*

Na osnovu grafikona možemo videti da je prosečan procenat alkohola 10,44%, pri čemu je maksimalni 14,9, a minimalni 8,4% I da se ne može stvoriti jasna predstava o vezi između kvaliteta vina I kolicina hlorida.

**

*Slika14: Prikaz grafikona za mulitarijantnu analizui*

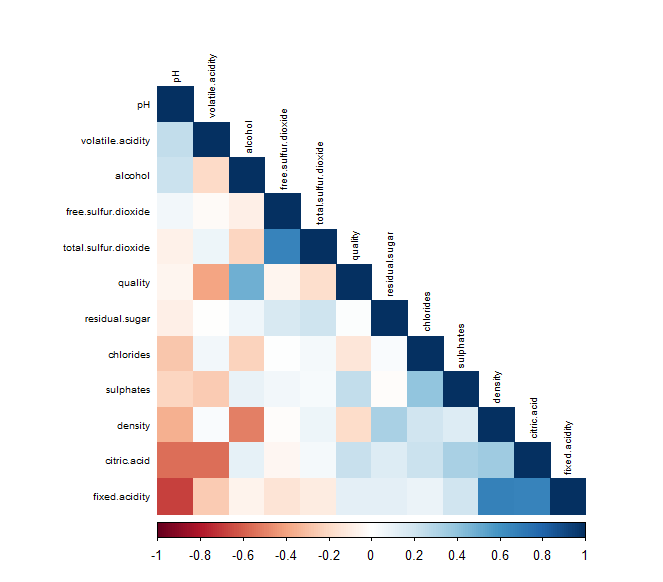
Nakon ovoga smo odredili matricu korelacije da bismo videli da li je potrebno da radimo PCA(Principal Component Analysis) ali pre toga smo morali da izdvojimo samo numeričke podatke

****

*Slika15: Prikaz koda za korelacionu matricu*

Na osnovu grafičkog prikaza vidimo da je mali broj blokova obojenih u tamne boje što znači da nema jakih korelacija izmedju ciljne vrednosti(quality) I ostalih atributa I samim tim nije potrebno da se radi PCA

Analiza glavnih komponenti (PCA) je široko korišćena tehnika koja pomaže u analizi velikih skupova podataka sa velikim brojem dimenzija/obeležja za posmatranju. To je statistička tehnika za smanjenje dimenzionalnosti skupa podataka, kako bi se povećala interpretabilnost uz očuvanje maksimalnih informacija



*Slika16: Prikaz grafikona korelacione matrice*

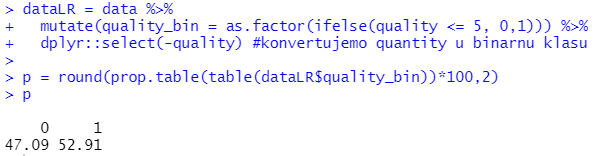
**4.Primena algoritma za mašinsko učenje**

**4.1.Logical regression**

Logistička regresija u R programiranju je klasifikacioni algoritam koji se koristi za pronalaženje verovatnoće uspeha i neuspeha događaja. Koristi se kada je zavisna promenljiva binarna (0/1, Tačno/Netačno, Da/Ne) po prirodi. Logistička regresija je takođe poznata kao binomna logistička regresija. Zasnovan je na sigmoidnoj funkciji gde je izlaz verovatnoća, a ulaz može biti od -beskonačno do +beskonačno.

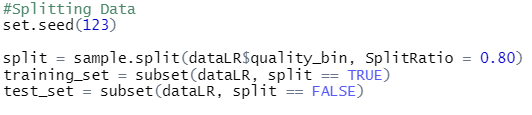
Pre nego što primenimo sam algoritam mašinskog učenja, potrebno je da uradimo još neke pripreme.

Prvo što smo uradili bilo je da konvertujemo našu ciljnu vrednost u binarni zapis I klasifikujemo vino na dobro I lose prema datom uslovu I dobili smo da imamo 47.09% lošeg I 52.91% dobrog vina.

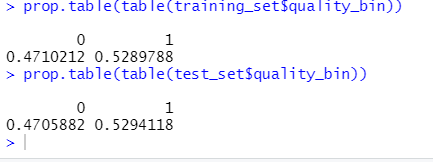


*Slika17: Prikaz koda za pretvaranje u binarni zapis I klasifikaciju*

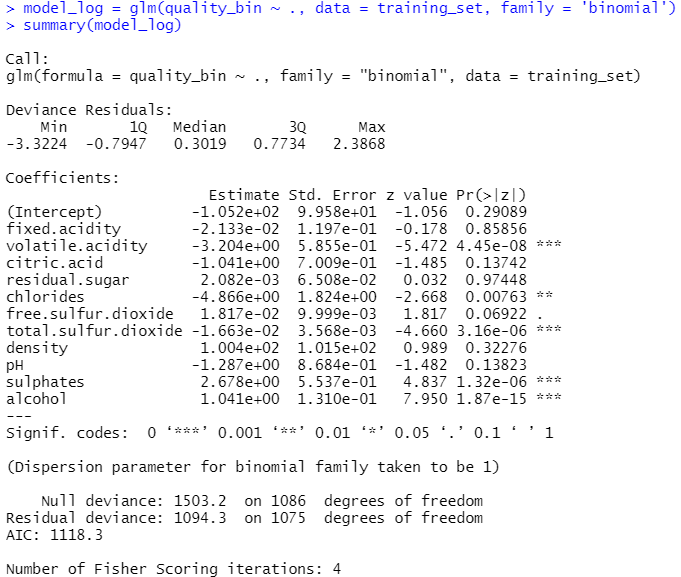
Nakon ovoga smo radili razdvajanje podataka(splitting) na trening I test skup u odnosu 80/20.

**

*Slika18: Prikaz koda za razdvajanje podataka*

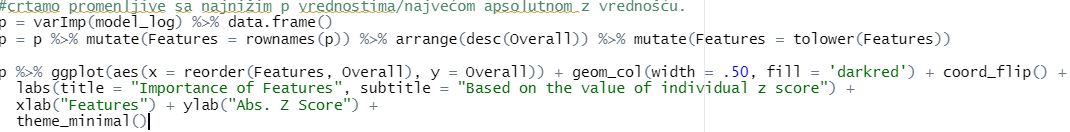
**

*Slika18: Prikaz koda provere ravnoteže* *podataka u podacima o obuci i testiranju.*

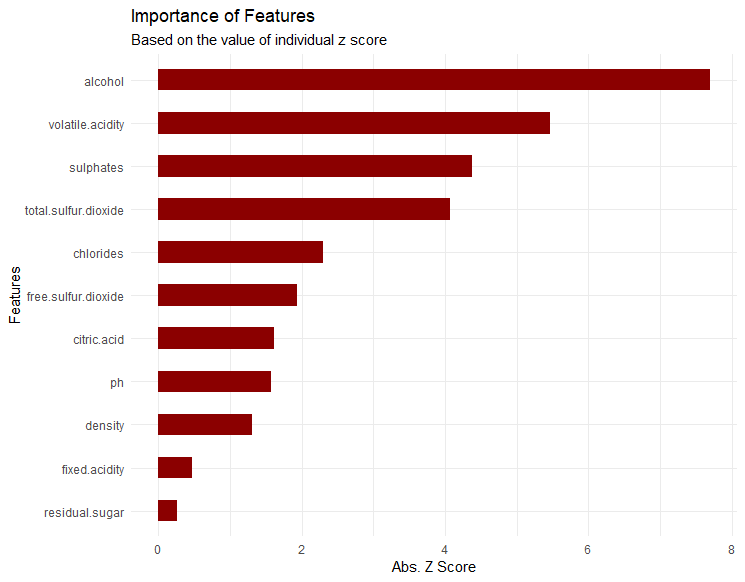
**

*Slika19: Prikaz koda I rezultata za treniranje modela linearnom regresijom.*

Takodje, prikazali smo grafikon sa najnižim p vrednostima/najvećom apsolutnom z vrednosti

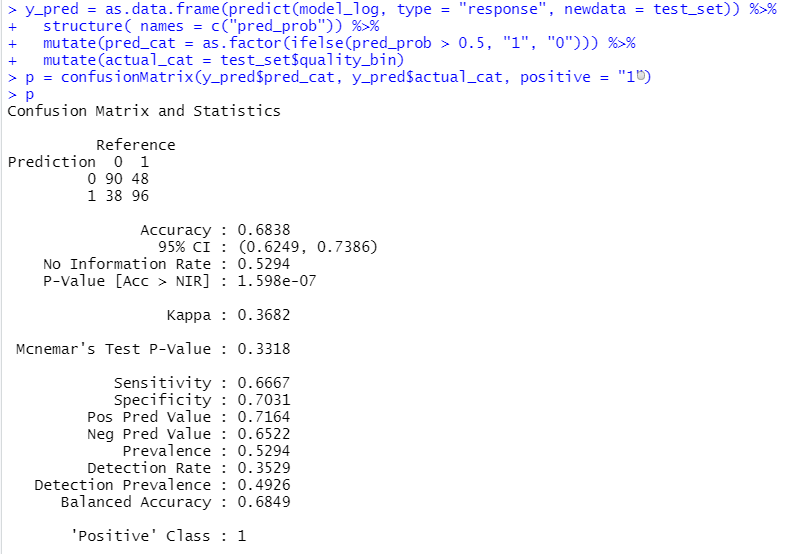


*Slika20: Prikaz koda za crtanje grafikona za prikaz* *najniže p vrednosti/najveće apsolutne z vrednosti*

**

*Slika20: Prikaz grafikona najniže p vrednosti/najveće apsolutne z vrednosti*

I na kraju,proverićemo kako naš model radi tako što ćemo ga pokrenuti na našim prethodno nevidljivim testnim podacima. Uporedićemo predviđeni ishod sa stvarnim ishodom i izračunati neke uobičajeno korišćene metrike merenja performansi modela binarne klasifikacije.

****

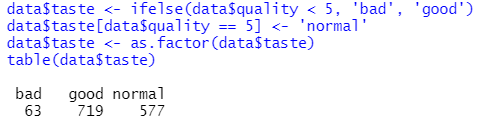
*Slika21: Prikaz koda I statističkih rezultata*

Dobili smo da je 68,38% uzoraka vina tačno(Accuracy) klasifikovano, da je 66.67% stvarnih uzoraka dobrog vina tačno klasifikovano(Sensitivity) I da je **71.64% ukupnih predviđanja dobrih vina zapravo dobra vina**( Pos Pred Value).

**4.2.Random Forest**

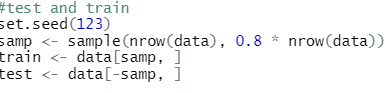
Slučajna šuma u ​​R programiranju je grupa stabala odlučivanja. On gradi i kombinuje više stabala odlučivanja da bi se dobila tačnija predviđanja. To je nelinearni klasifikacioni algoritam. Svaki model stabla odlučivanja se koristi kada se koristi samostalno. Zovu se nasumični jer nasumično biraju prediktore u vreme treninga. Oni se nazivaju šumama jer uzimaju rezultate više stabala da bi doneli odluku. Slučajna šuma nadmašuje stabla odlučivanja jer će veliki broj nekoreliranih stabala (modela) koji rade kao komitet uvek nadmašiti pojedinačne konstitutivne modele.

Prvo smo pripremili podatke da bi mogli kasnije da primenimo Random Forest algoritam I ono što smo prvo uradili jeste da smo klasifikovali ciljnu vrednost(quality) prema datom uslovu na bad, good I normal

****

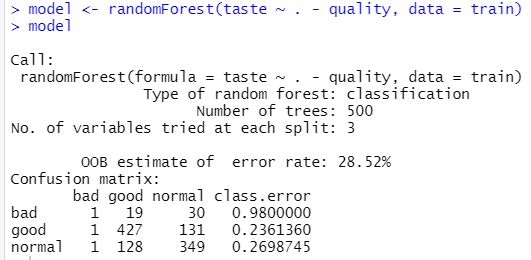
*Slika22: Prikaz koda za klasifikaciju*

Nakon ovoga smo radili razdvajanje podataka(splitting) na trening I test skup u odnosu 80/20.

**

*Slika23: Prikaz koda za razdvajanje podataka na test I train*

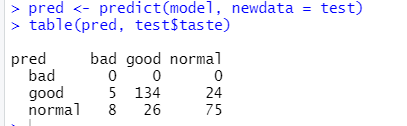
Nakon ovoga smo napravili model za koji nam je bila neophodna randomForest biblioteka



*Slika24: Prikaz koda za model*

Možemo videti da je izgrađeno 500 stabala, a model je nasumično uzorkovao 3 prediktora pri svakom razdvajanju. Takođe prikazuje matricu koja sadrži predviđanje u odnosu na stvarnu, kao i grešku u klasifikaciji za svaku klasu.

Sada ćemo da testiramo model na skupu testnih podataka.

**

*Slika25: Prikaz koda za testiranje modela na skupu testiranih podataka*

I na kraju smo testirali tačnost na sledeći način I dobili rezultat da smo postigli tačnost od **~76,38%**

****

*Slika26: Prikaz koda za testiranje taćnosti*

**5.Zaključak**

Ovim projektom smo videli praktičnu primenu R programskog jezika za obradu, proveru i testiranje podataka, kao I primenu nekih od algoritama mašinskog učenja. Sve u svemu, vidimo da je najvažnija varijabla u određivanju kvaliteta sadržaj alkohola. Iako sulfati značajno doprinose kvalitetu vina, postoji ograničenje količine sulfata uključenih u vino. Dakle, kada pojedinac želi da pronađe vino visokog kvaliteta, najvažniji faktor bi bio sadržaj alkohola.