SVEUČILIŠTE U RIJECI

ODJEL ZA INFORMATIKU

RIJEKA

Nola Čumlievski

Ivan Šimičić

PROJEKT IZ KOLEGIJA INTELIGENTNI SUSTAVI 2

PROJEKTNI ZADATAK

Rijeka, 2020.

Sadržaj

Uvod

U ovom projektom radu ćemo obraditi neke od odabranih tema koje smo obradili u sklopu kolegija Inteligentni sustavi 2. Za projekt smo odabrali zadatak 1.1.3 koji obuhvaća vizualizaciju i istraživanje podataka, zadatak 1.2.b koji obuhvaća metriku važnosti značajke te zadatak 2.b koji obuhvaća analizu podataka primjenom logističke regresije.

Zadatak 1.1.3.

Prilikom vizualizacije i istraživanja podataka koristit ćemo biblioteku *ggplot2* koja je najpoznatija biblioteka za R programski jezik koja omogućuje vizualizaciju podataka. Biblioteka *ggplot2* nam omogućuje visoku razinu apstrakcije i modifikacije željenih grafikona što je poželjno prilikom vizualizacije kompleksnijih podataka.

U projektu ćemo koristiti .csv datoteke *data1 i data2* koje sadrže mjerenja nekih parametara plodova breskve. Zbog jednostavnosti, iz oba skupa izostavljene su varijable koje nije potrebno koristiti prilikom analize i skupovi su spojeni u jedan skup podataka kako bi imali više podataka u nadi da modeli budu što kvalitetniji i precizniji.

Skupove učitavamo kroz mogućnost Import Dataset unutar RStudio razvojnog okruženja, važno je napomenuti da opciju heading treba postaviti na True budući da prvi red sadržava nazive varijabli.

Nakon što smo učitali skupove, potrebno je izbaciti varijable koje neće biti korištene u projektu kako bi mogli spojiti oba skupa u jedan veći skup.

```
data1 = data1[,-c(1,2,3,4,5,6,7,8,9,11,13,14,15,16,18, 19,23)]

data2 = data2[, -c(1,2,3,4,5,7,9,10,11,12,14,15)]

podaci=rbind(data1,data2)
```

Budući da podatci nemaju konzistentan način zapisa u oba skupa, potrebno je zaokružiti decimalne vrijednosti na dvije decimale , a to postižemo uz pomoć funkcije `round`.

```
'data.frame':   200 obs. of  6 variables:
$ volume.cm3   : num  122 120 127 127 122 ...
$ density.g.cm3: num  1.23 1.24 1.2 1.22 1.19 1.45 1.32 1.21 1.2 1.26 ...
$ FirmnessAv   : num  1.63 3.23 0.6 1.24 1.03 0.68 0.78 1.45 3.29 3.65 ...
$ SSC.TA       : num  9.82 14.82 15.54 14.98 12.57 ...
$ Zs           : num  1.18 1.4 1.05 1.03 0.91 0.85 0.86 1.38 1.31 1.34 ...
$ θ            : num  -41.7 -46.9 -36.7 -40.5 -33.6 ...
```

Na prethodnom prikazu vidljiva je struktura spojenog dataset-a *podaci*. Vidimo da se skup podataka sastoji od 200 opservacija i 6 numeričkih varijabli (izabrane varijable za analizu).

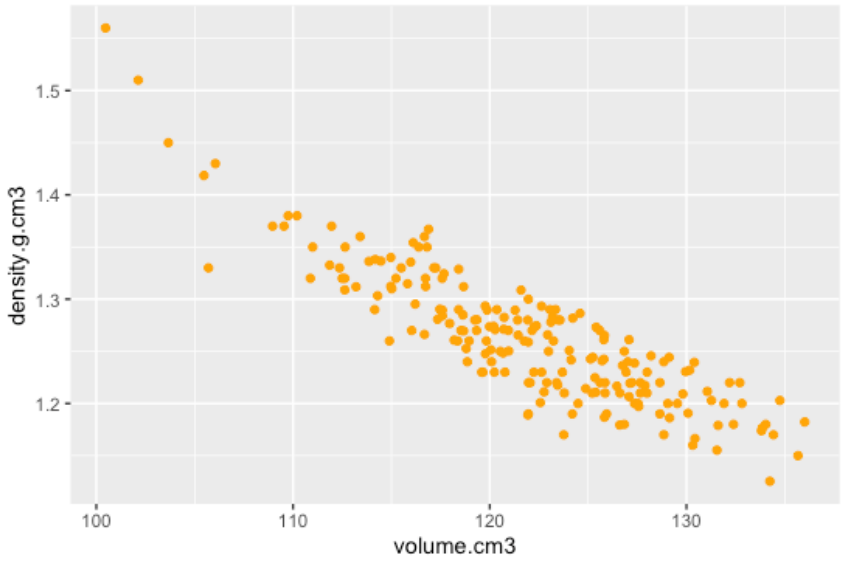
Odnos između varijabli

Uz pomoć funkcije *cor* ispitujemo korelaciju varijabli kako bi saznali koje varijable možemo odabrati za analizu. Ispis funkcije `cor(podaci)`:

```
              volume.cm3   density.g.cm3 FirmnessAv    SSC.TA     Zs
volume.cm3     1.00000000    -0.8767210  0.22462244 -0.07751461  0.2240924
density.g.cm3 -0.87672105     1.0000000 -0.23775786  0.16304566 -0.2256604
FirmnessAv     0.22462244    -0.2377579  1.00000000 -0.01522054  0.7232339
SSC.TA        -0.07751461     0.1630457 -0.01522054  1.00000000  0.1128358
Zs             0.22409236    -0.2256604  0.72323393  0.11283582  1.0000000
θ             -0.20467564     0.1781956 -0.72144587 -0.05636513 -0.8163714
                         θ
volume.cm3    -0.20467564
density.g.cm3  0.17819560
FirmnessAv    -0.72144587
SSC.TA        -0.05636513
Zs            -0.81637141
θ              1.00000000
```

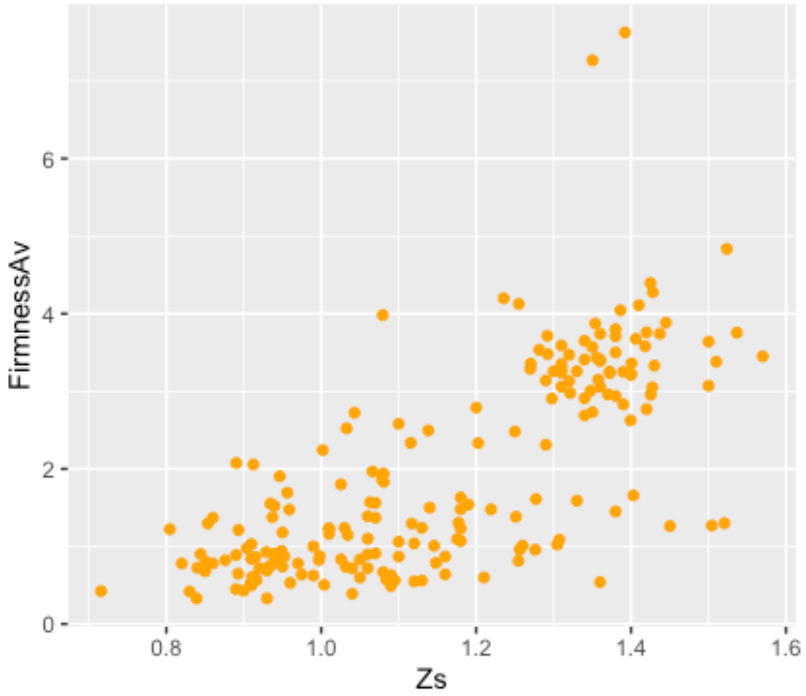
Iz priloženog outputa vidljive su korelacije između varijabli dataset-a. Vidimo da najveću korelaciju s varijablom *FirmnessAv* imaju varijable *Zs* (impendanca) i θ (imaginarni dio otpora breskve) na način da impendanca ima visoku, pozitivnu korelaciju sa zrelosti breskve dok kut theta (imaginarni dio otpora breskve) ima visoku negativnu korelaciju s navedenom varijablom. Ostale varijable imaku znatno nisku korelaciju u odnosu na zrelost breskve te je fokus za daljnju analizu postavljen na impendancu i imaginarni dio otpora breskve.

Uz pomoć funckije *ggplot* na temelju podataka i odabranih varijabli možemo vizualizirati njihov odnos.



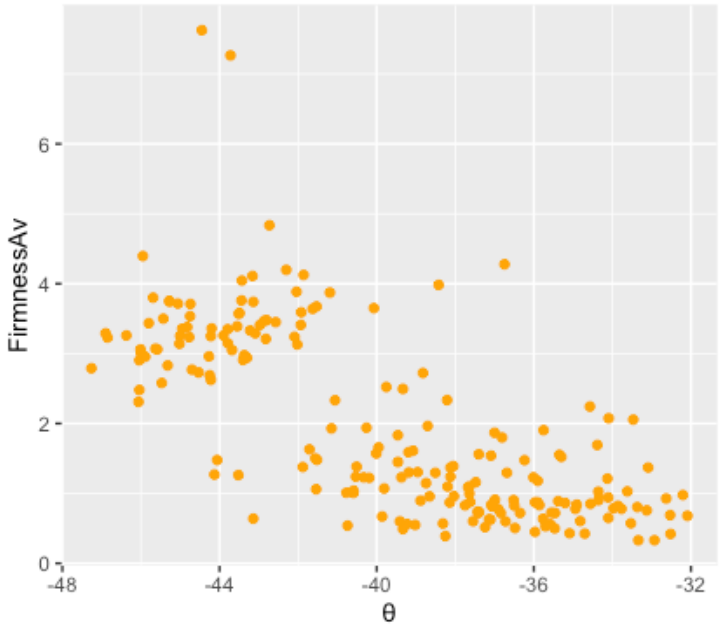
Slika 1. odnos između varijabli volume i density

Na prethodnom grafu (slika 1.) možemo vidjeti odnos između varijabli *volume.cm*^{3} (volumen) i *density.g.cm*^{3} (gustoća). Možemo uočiti kako je gustoća breskve manja s porastom volumena što se objašnjava time da gustoća voća postaje manja kako ono sazrijeva i dobiva na veličini (točan primjer negativne korelacije).



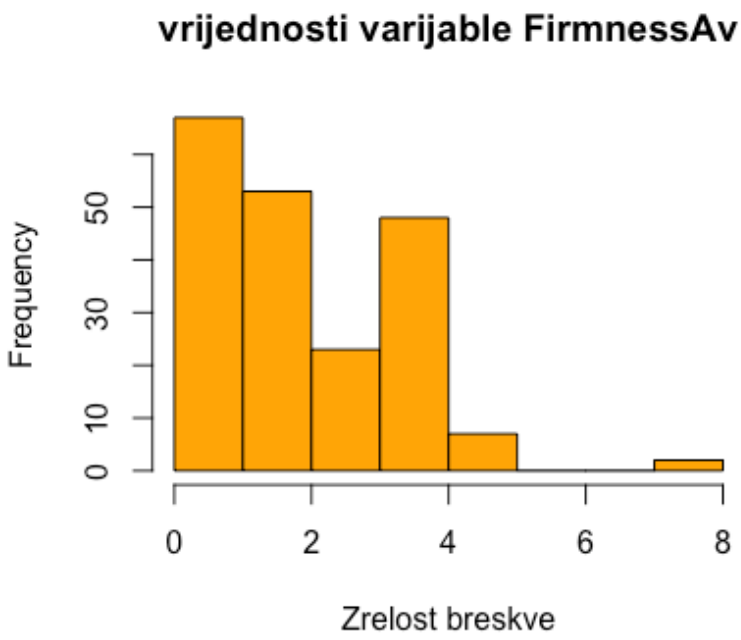
Slika 2. odnos između varijable FirmnessAv i Zs

Na prethodnom grafu (slika 2.) možemo vidjeti odnos između varijabli *FirmnessAv* (prosječna „zrelost“) i *Zs* (otpor električne struje na breskvu). Možemo uočiti pozitivnu korelaciju između navedene dvije varijable tj. kako otpor električne struje postaje sve veći što je voćka zrelija.



Slika 3. odnos između varijabli FirmnessAv i kuta theta

Na prethodnom grafu (slika 3.) možemo vidjeti odnos između varijabli FirmnessAv (prosječna „zrelost“) i kuta theta (imaginarni dio otpora breskve). Možemo uočiti kako je imaginarni otpor breskve veći kod manje zrelih voćki te sukladno tome, kako je otpor manji što je voćka zrelija tj. da varijable negativno koreliraju.

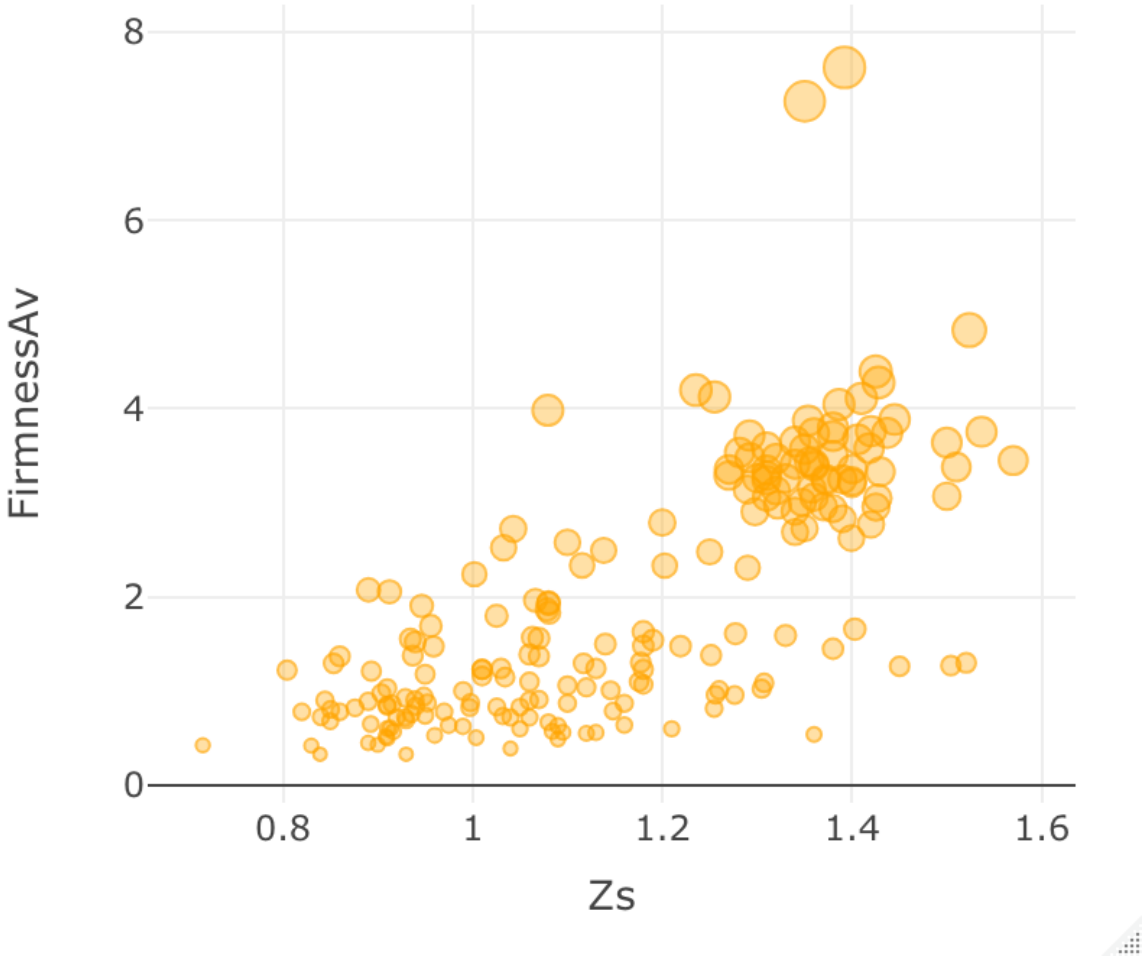


Slika 4. histogram vrijednosti FirmnessAv

Histogram je dobiven pomoću funkcije *hist.* Možemo vidjeti (slika 4.) histogram s frekvencijama varijable *FirmnessAv* gdje uočavamo da najveći broj promatranih voćaka ima vrijednost varijable *FirmnessAv* između 0 i 4.

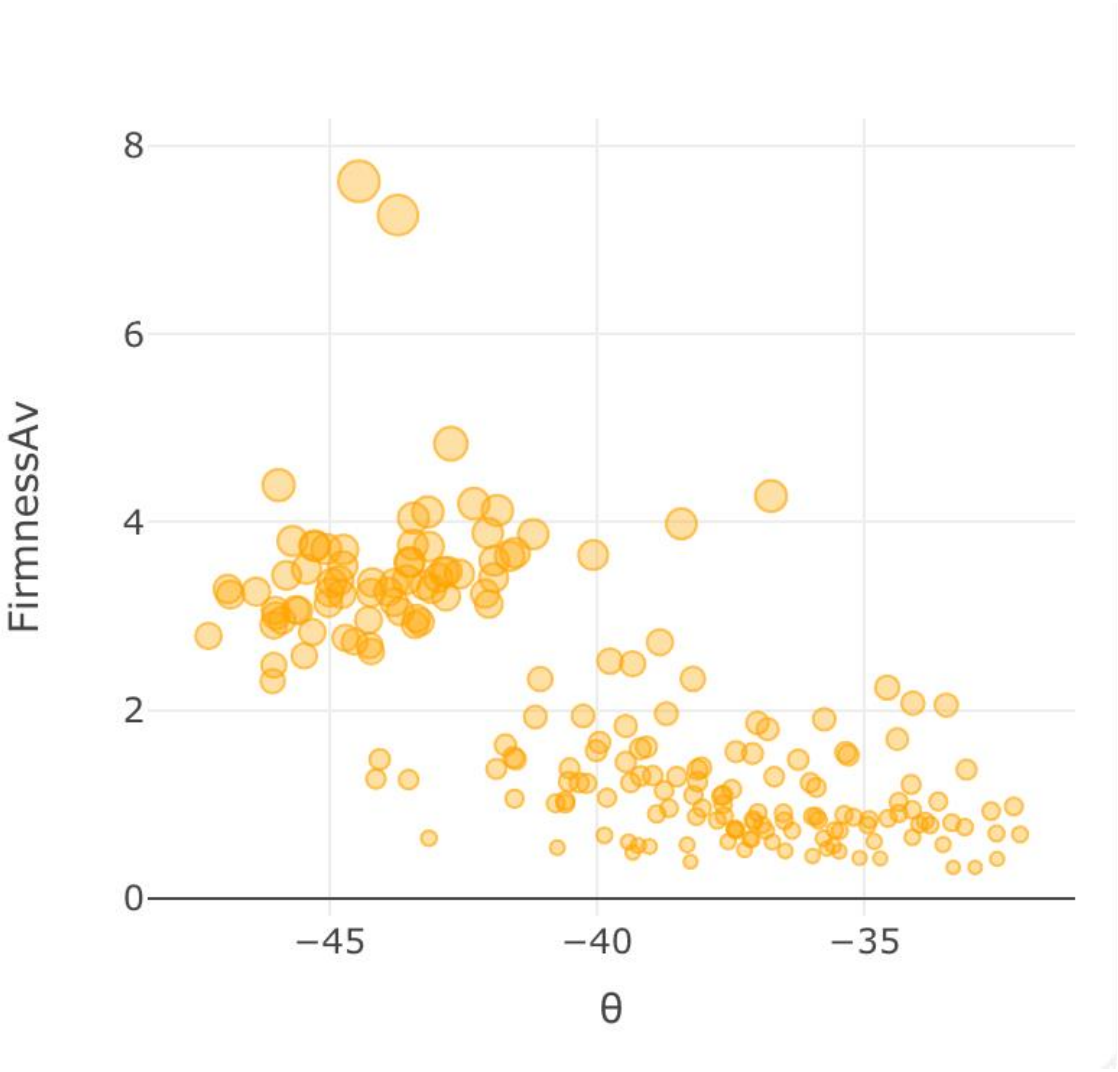
Slika 5. density plot

Ovaj grafikon (slika 5.) dobiven je pomoću funkcije *densityplot.* Sa grafikona vidljiv je density plot koji vizualizira podatke slično kao histogram gdje dobivamo detaljniji prikaz frekvencije neke varijable, u ovom slučaju zrelosti breskve.



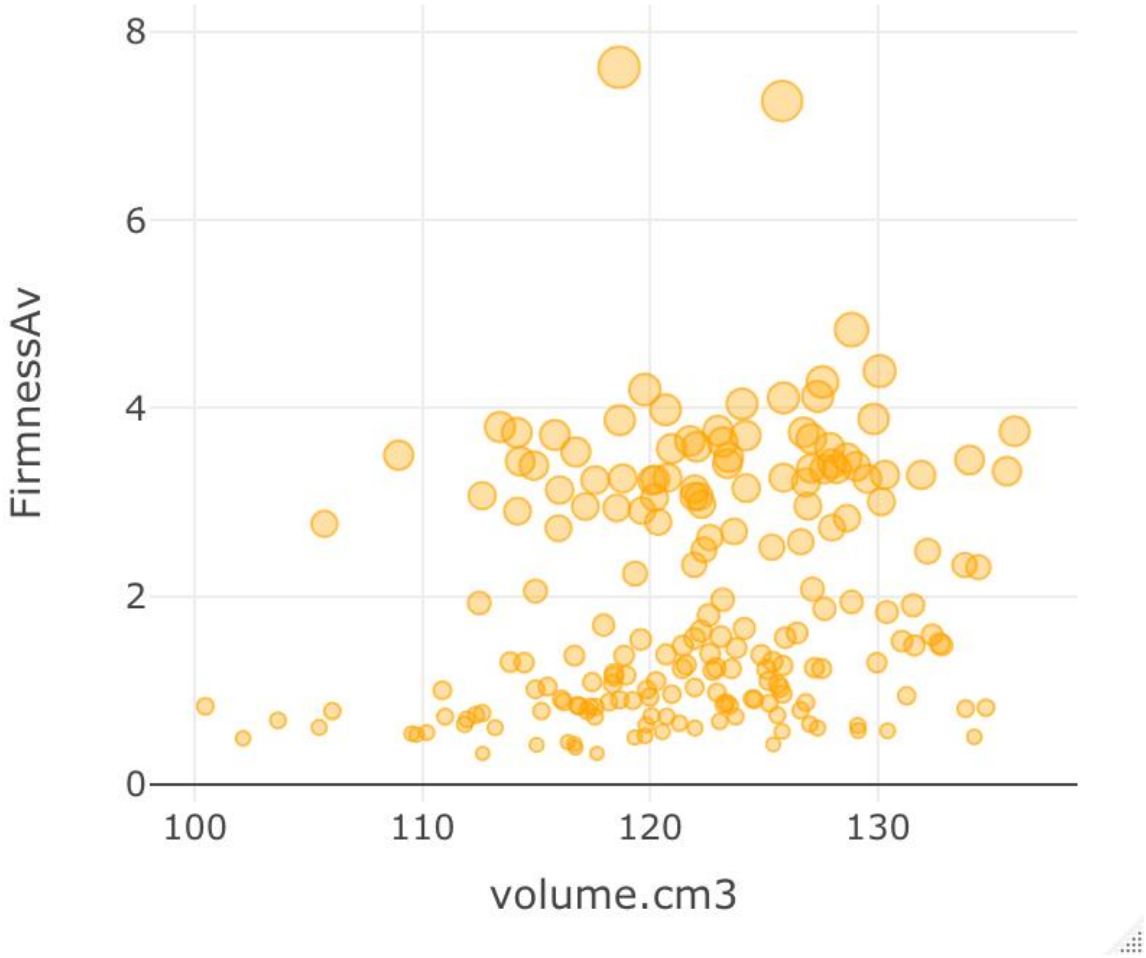
Slika 6. scatter plot varijabli FirmnessAv i Zs

Na prethodnoj slici (slika 6.) možemo vidjeti scatter plot za varijable *FirmnessAv i Zs* koji je sličan grafu na slici 2. Ovaj grafikon dobiven je pomoću funkcije *plot_ly.* Za razliku od običnog scatterplota, na ovom grafikonu veličina točaka prilagođena je vrijednosti varijable zrelosti breskve te možemo uočiti gdje se nalazi najveći indeks zrelosti u odnosu na vrijednosti varijable impendance. Kao što je vidljivo sa grafikona, najveći indeks zrelosti breskve nalazimo u rasponu vrijednosti impendance od 1.2 do 1.6.



Slika 7. scatter plot varijabli FirmnessAv i kuta theta

Na prethodnoj slici možemo vidjeti scatter plot za varijable *FirmnessAv* i kuta theta koji je sličan grafu na slici 3 (funkcija *plot_ly*). Kao što vidimo, najveće vrijednosti indeksa zrelosti breskve nalazimo u rasponu vrijednosti imaginarnog kuta theta od ~ -48 do -40, dok je kod ostalih vrijednosti kuta theta indeks zrelosti breskve relativno malen.



Slika 8. scatter plot varijabli FirmnessAv i volume

Na prethodnoj slici (slika 8.) možemo vidjeti scatter plot za varijable *FirmnessAv* i *volumen.cm*^{3} koji nam prikazuje kako zrelost ne utječe nužno i na volumen i da razlika između zrelih i nezrelih voćki po volumenu nije velika. Može se jedino primjetiti da su voćke sa većim indeksom zrelosti volimena između ~115 do 130 (uz iznimke).

Uz pomoć funkcije summary možemo dobiti sažete statističke informacije o varijablama u skupu koji koristimo (minimalna vrijednost, maksimalna vrijednost, kvartali, median, aritmetičku srednju vrijednost).

Summary podataka:

```
  volume.cm3    density.g.cm3     FirmnessAv         SSC.TA            Zs
θ
 Min.   :100.5   Min.   :1.125   Min.   :0.3300   Min.   : 8.94   Min.
:0.7156   Min.   :-47.26
 1st Qu.:117.6   1st Qu.:1.219   1st Qu.:0.8475   1st Qu.:11.64   1st
Qu.:0.9736   1st Qu.:-43.38
```

```
 Median :122.2   Median :1.260   Median :1.4763   Median :12.83   Median
:1.1391   Median :-39.33
 Mean   :121.9   Mean   :1.262   Mean   :1.9547   Mean   :13.07   Mean
:1.1575   Mean   :-39.67
 3rd Qu.:126.8   3rd Qu.:1.291   3rd Qu.:3.1425   3rd Qu.:14.14   3rd
Qu.:1.3419   3rd Qu.:-36.49
 Max.   :136.0   Max.   :1.560   Max.   :7.6250   Max.   :21.60   Max.
:1.5700   Max.   :-32.09
```

Zadatak 1.2.a.

Slučajne šume

Slučajne šume (random forest) predstavlja algoritam u kojem se gradi određen broj stabala odluke na uzorcima za treniranje dobivenim metodom *bootstrap* – metoda uzrokovanja koja, za razliku od metode uzrokovanja unakrsne validacije bez zamjene, koristi uzrokovanje sa zamjenom da bi se formirao skup za učenje).

Za vrijeme gradnje pojedinog stabla odluke unutar slučajne šume odabran je **slučajan uzorak** od *n* prediktora iz skupa prediktora skupa, te se prilikom svakog grananja preuzima novi uzorak od *n* prediktora (obično se bira *n = sqrt(p)*). Za razliku od bagging metode gdje većina stabla koristi jake prediktore za grananje (što ne dovodi do značajnog smanjenja varijance), algoritam slučajne šume ne koristi jake prediktore pri grananju tako da slabi prediktori imaju bolju šansu. Slučajne šume koriste se s ciljem redukcije pogreške testiranja kao i OOB pogreške (*out-of-bag* pogreške – pogreška testiranja modela dobivenog pomoću metode bagging).

Koraci djelovanja algoritma:

1. Odabir *n* slučajnih podskupova iz skupa za treniranje

2. Treniranje *n* stabla odluka (1 nasumični skup iz prethodnog koraka ide jednom stablu odluke)

3. Svako stablo odluke **neovisno** radi predikcije skupa za testiranje

4. Konačno predviđanje izvodi se **glasovanjem –** svako stablo odluke glasa svojom predikcijom te slučajna šuma kao konačno predviđanje koristi onu klasu s **najviše glasova.** Kod regresije koristi se uprosječenje rezultata stabla odluke.

Analiza

Nakon što smo napravili vizualizaciju, možemo kreirati skupove za treniranje i testiranje modela. Za ovu analizu nismo diskretizirali ciljnu varijablu s obzirom da se analiza može provoditi nad numeričkim vrijednostima.

```
set.seed(35)
treniranje = sample(1:nrow(podaci), 0.8*nrow(podaci))
testiranje = podaci[-treniranje,"FirmnessAv"]
```

Kao što je vidljivo iz priloga, skup za treniranje sastoji se od 80% opservacija, dok skup za testiranje sadrži 20% opservacija. Nakon kreiranja skupova, možemo kreirati modele za slučajne šume. Koristit ćemo funkciju *randomForest* iz istoimene biblioteke.

```
set.seed(35)
suma2=randomForest(FirmnessAv~Zs,data=podaci,subset=treniranje,
importance=TRUE, ntree=994, sampsize=7, nPerm=8, nodesize=1, maxnodes=8)
```

Prilikom unosa parametara, potrebno je navesti formulu (*FirmnessAv ~ Zs*) koja sadrži ciljnu varijablu i prediktor te skup i podskup podataka. Ostali parametri (*ntree, mtry, importance, nPerm, nodesize, maxnodes, sampize,* itd. su proizvoljni i u ovom primjeru nekoliko je korišteno za optimizaciju stabla - smanjivanje MSE što predstavlja srednju vrijednosti kvadrata reziduala). Slijedi opis **korištenih** parametara:

- **ntree**: broj stabla odluke unutar slučajne šume
- **sampsize**: veličina uzoraka
- **nPerm**: broj permutacije OOB podataka unutar svakog stabla (za postavljanje važnosti varijable)
- **nodesize**: minimalna veličina terminalnih čvorova (veči broj terminalnih čvorova uzrokuje izgradnju stabala manje dubine)
- **maxnodes**: maksimalan broj terminalnih čvorova koje pojedino stablo odluke može imati (ako nije navedeno, grade se stabla maksimalne moguće veličine).

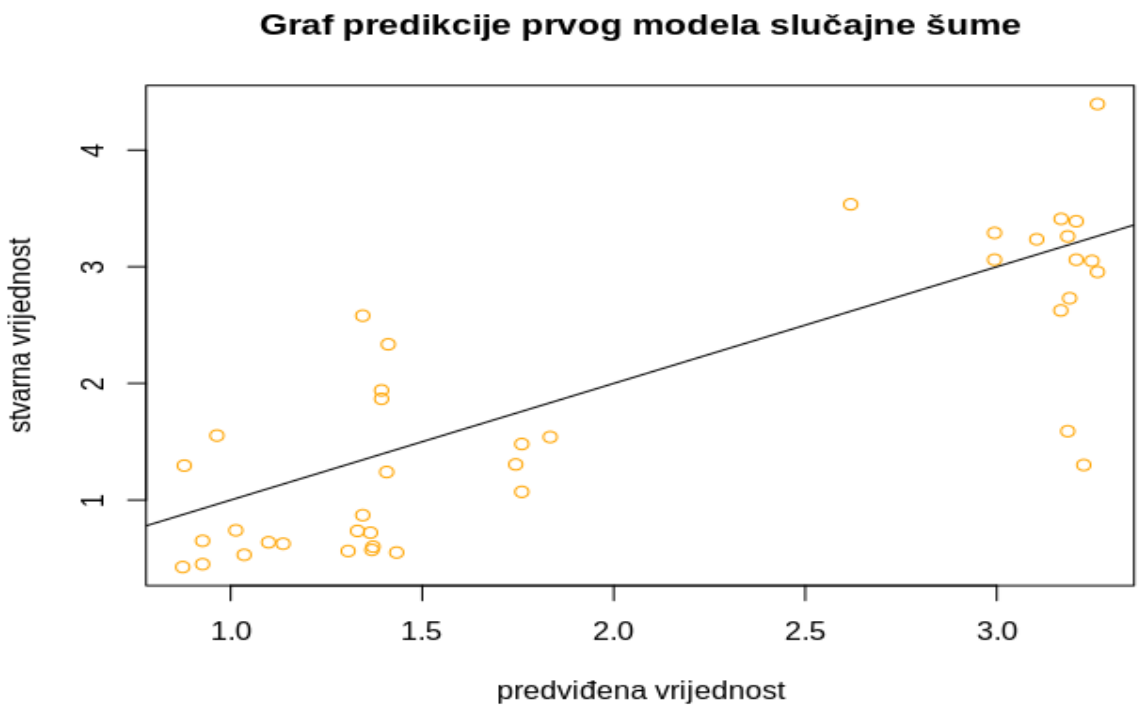
Isprobano je više kombinacija parametara te na kraju dobiveno slijedeće:

sity.g.cm3: num 1.23 1.24 1.2 1.22 1.19 1.45 1.32 1.21 1.2 1.26 ...
$ FiAv Call:
randomForest(formula = FirmnessAv ~ Zs, data = podaci, importance = TRUE, ntree = 994, sampsize = 7, nPerm = 8, nodesize = 1, maxnodes = 8, subset = treniranje)
Type of random forest: regression
Number of trees: 994
No. of variables tried at each split: 1

Mean of squared residuals: 0.8563256
% Var explained: 53.17
: num 1.63 3.23

Nakon što smo kreirali model za slučajnu šumu, možemo izvesti predikciju slučajne šume koju ćemo kasnije koristiti prilikom prikazivanja grafa.

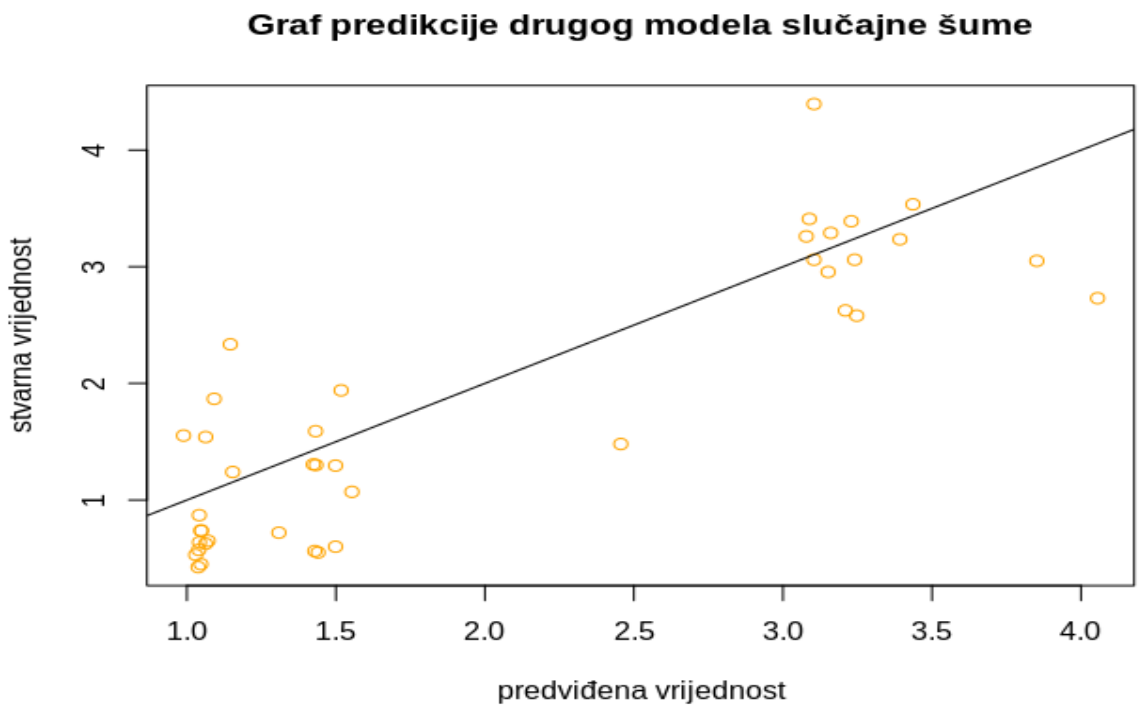
```
predikcija3 = predict(suma2,newdata=podaci[-treniranje,])
```



Slika 9. graf predikcije

MSE (*mean squared error*) predstavlja prosjek kvadrata pogreške tj. prosječnu kvadratnu razliku između predviđenih i stvarnih vrijednosti. Što je MSE bliži nuli, to je model bolji. MSE u R-u možemo izračunati uz pomoć funkcije *MSE*, a za ovaj model ona iznosi 0.4577372. Za potrebe analize i testiranja, kreiramo novi model s drugom varijablom kako bi testirali kvalitetu i usporedili rezultate.

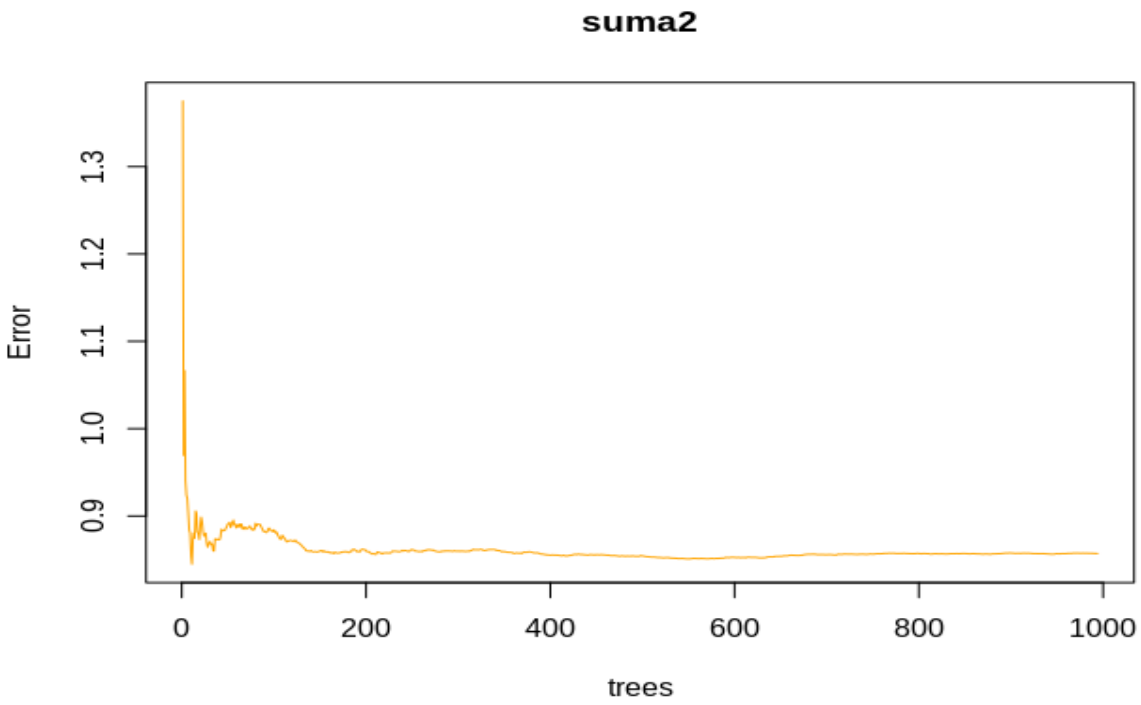
```
suma4=randomForest(FirmnessAv~0,data=podaci,subset=treniranje,importa
nce=TRUE, ntree=1510, sampsize=5, nodesize=1, maxnodes=8)
```



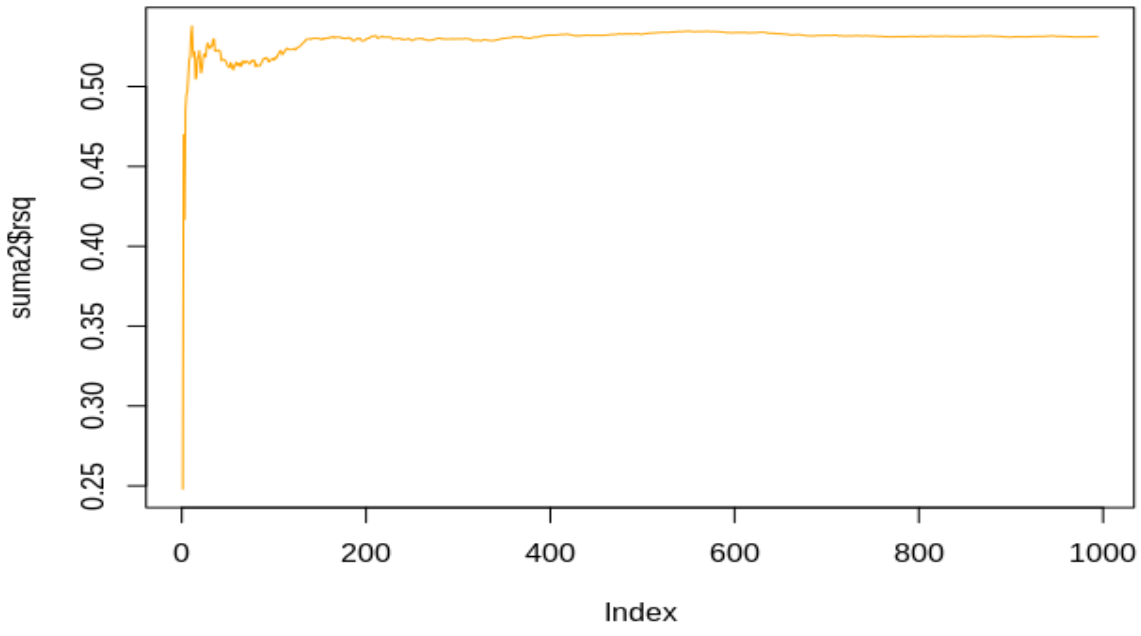
Slika 10. graf predikcije s drugom varijablom

MSE za ovaj model iznosi 0.3246449 te prema samom grafikonu vidljiva je manja raspršenost točaka (linija bolje opisuje podatke).

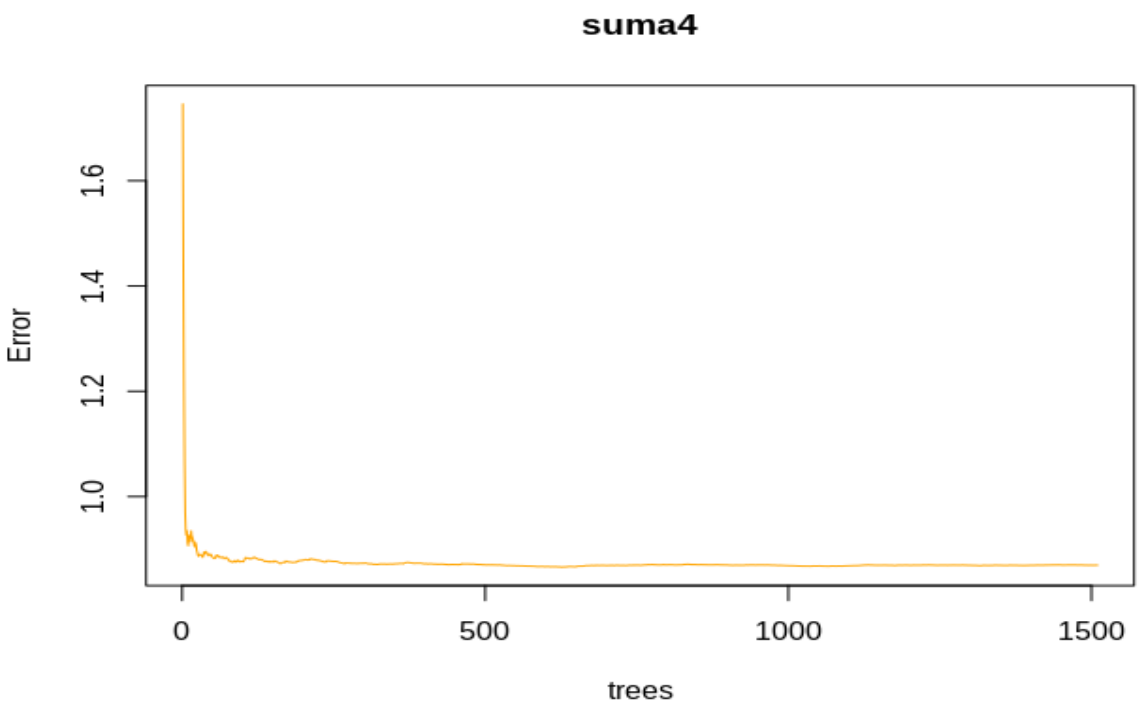
Na slijedeća 4 grafikona prikazani su grafikoni koji prikazuju MSE oba modela prema broju stabala te R^2 oba modela.



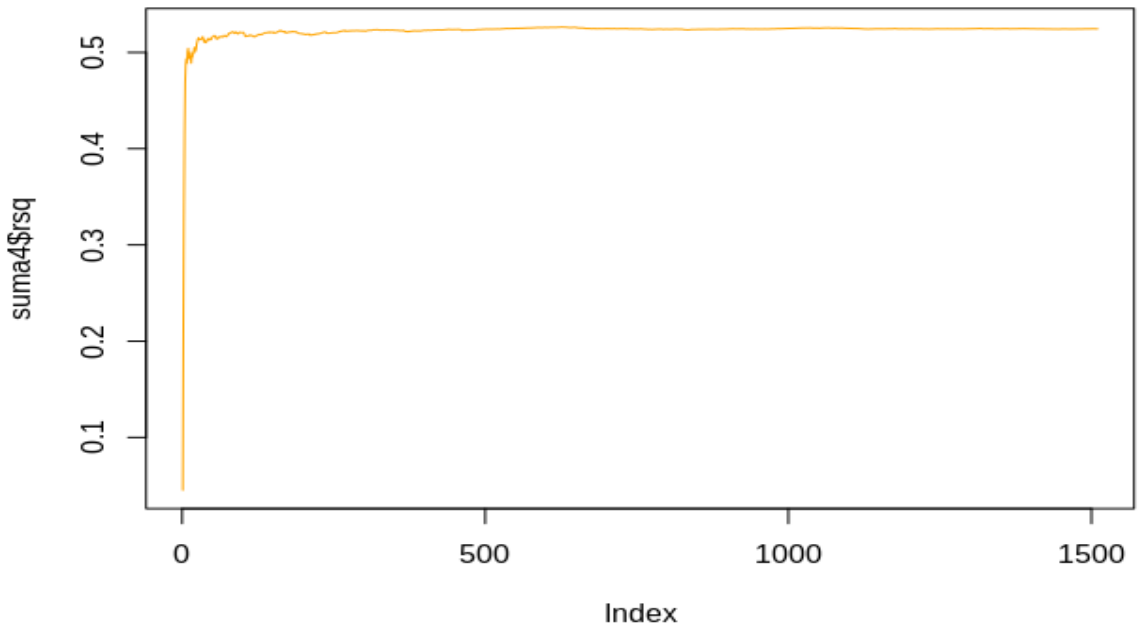
Slika 11. graf MSE modela suma2 prema broju stabala



Slika 12. graf R2 za model suma2



Slika 13. graf MSE modela suma4 prema broju stabala

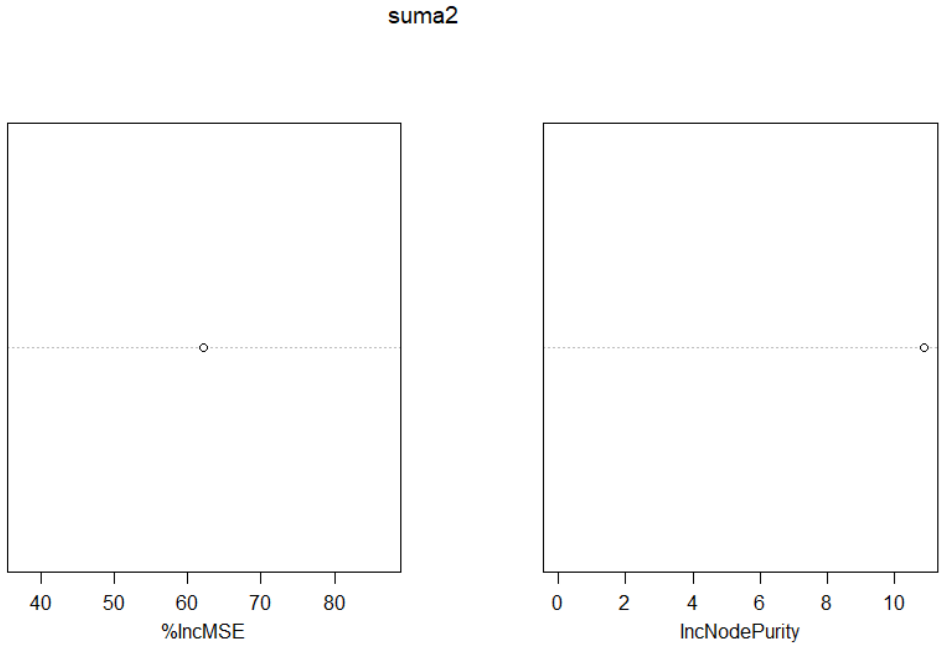


Slika 14. graf R2 za model suma4

Zadatak 1.2.

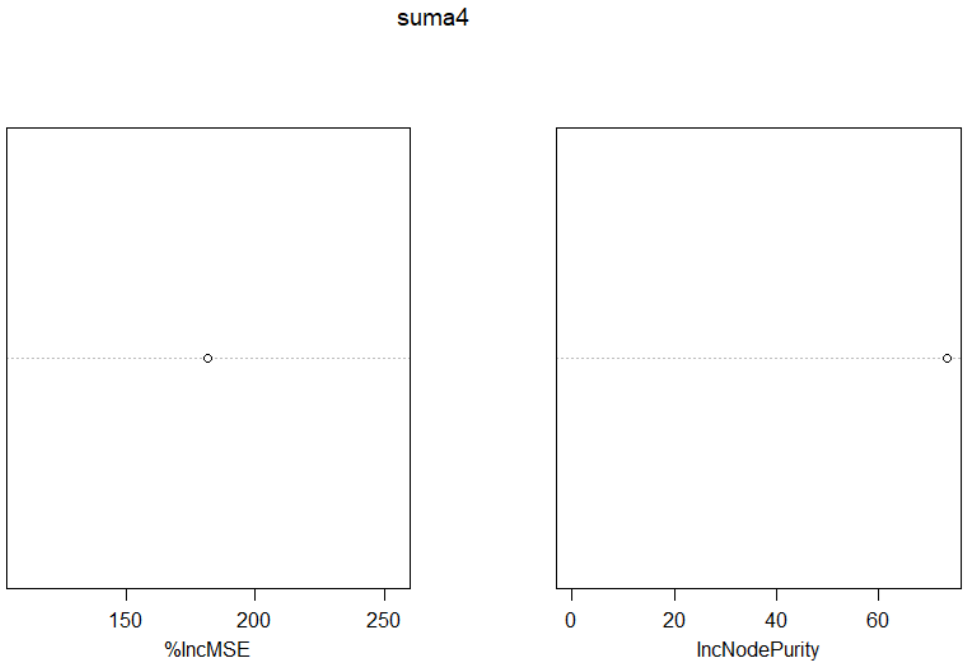
Kako bi izračunali značaj varijabli, koristit ćemo funkcije *importance, varImpPlot i filterVarImp.*

```
 importance(suma2):
    %IncMSE IncNodePurity
Zs 62.21377      10.86432
```



Slika 13. važnost varijable Zs u modelu suma2

```
importance(suma4):
    %IncMSE IncNodePurity
θ    181.7456      73.28007
```



Slika 14. važnost varijable theta u modelu suma4

Vidimo da veći značaj za predviđanje zrelosti breskve ima kut theta breskve. Uz pomoć funkcije filterVarImp možemo saznati koje su varijable najvažnije za predikciju zrelosti.

```
filterVarImp(podaci[1:6], podaci$FirmnessAv, nonpara = TRUE)

                Overall
volume.cm3    0.05863383
density.g.cm3 0.09222641
FirmnessAv    1.00000000
SSC.TA        0.04621618
Zs            0.57501981
θ             0.59301223
```

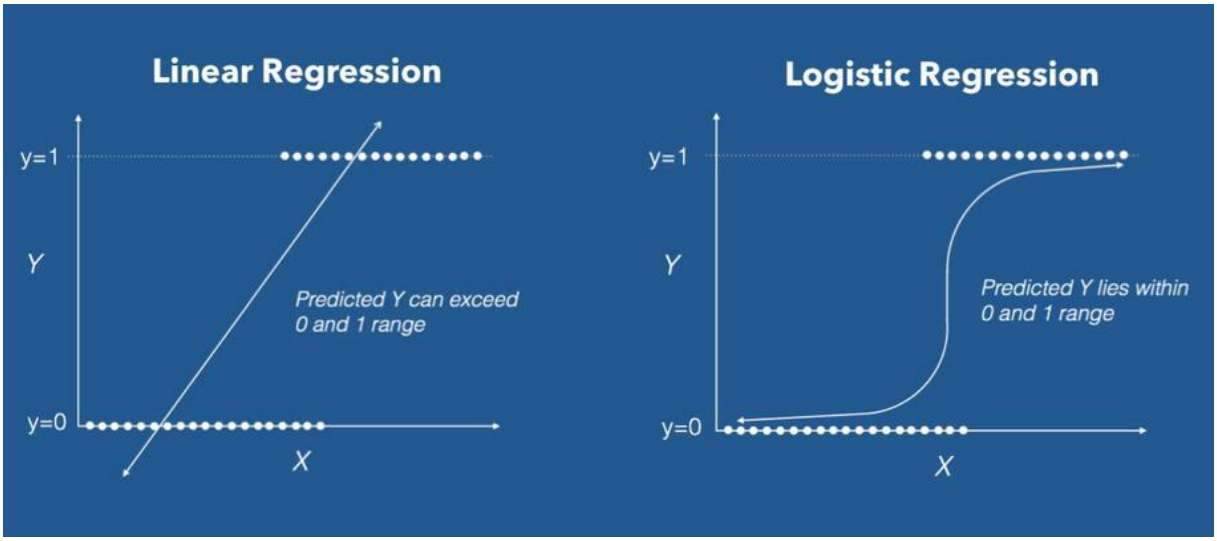
U slučaju ovih podataka, možemo uočiti kako varijable θ i Zs imaju najveći značaj za predikciju (0.58 i 0.59).

Zadatak 2.b.

Logistička regresija

Modelom logističke regresije opisujemo veze između prediktora tj. vjerojatnosti pripadanja svakoj kategoriji za dan skup prediktora. Ova metoda predstavlja vjerodostojnu statističku tehniku kada je zavisna varijabla (ciljna varijabla) kategoričkog tipa i kada su nezavisne varijable metričke ili nemetričke.

Logistička regresija se, za razliku od linearne, koristi kada je zavisna (ciljna varijabla) kateogričkog tipa.



Slika 15: Razlika između linearnog i logističkog modela

Kako bi ustanovili koje varijable koristiti prilikom izrada modela, s obzirom na prethodnu analizu, izabrali smo dvije prethodno navedene varijable za koje je ustanovljeno da imaju najveći značaj/utjecaj na zrelost breskve te jedna od preostale 3 varijable koja nema velik utjecaj na zrelost breskve kako bi se mogla izvršiti usporedba logističkih modela.

Izrada modela

Prije izrade modela bilo je potrebno drugačije postaviti skupove za treniranje i testiranje te diskretizirati ciljnu varijablu (zrelost breskve).

```
      podaci$FirmnessAv=discretize(FirmnessAv, method = "frequency", breaks
= 2,labels =c("nezrela", "zrela"))
      head(podaci)
      #podjela podataka
      train = sample(1:nrow(podaci), 0.8*nrow(podaci))
      treniranje = podaci[train,]
      testiranje = podaci[-train,]
      test_y = podaci$FirmnessAv[-train]
```

Model logističke regresije u R-u kreiramo uz pomoć funkcije *glm* koja se nalazi unutar biblioteke *ISLR*:

```
      logModelθ <- glm(formula=FirmnessAv ~ θ, data=treniranje, family =
"binomial")
      summary(logModelθ)
```

Glm funkcija za parametre uzima formulu gdje specificiramo odnos ciljne varijable i prediktora, podatke koji se koriste pri logističkog regresiji te funkciju koja će se koristiti u modelu. U ovom modelu smo koristili kut theta, a informacije o modelu možemo vidjeti uz pomoć funkcije *summary*.

```
      Call:
glm(formula = FirmnessAv ~ θ, family = "binomial", data = treniranje)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1379  -0.6948  -0.2188   0.5733   2.4963

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -19.54731    2.86971  -6.812 9.65e-12 ***
θ            -0.49230    0.07213  -6.825 8.76e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221.78  on 159  degrees of freedom
Residual deviance: 137.29  on 158  degrees of freedom
```

```
AIC: 141.29

Number of Fisher Scoring iterations: 5
```

Iz ovog ispisa vidljivo je da je naš model statistički značajan (mala *p* vrijednost). ß$_1$ je negativan (-0.49230) što znači da kada je kut theta veći, to je indeks zrelosti breskve manji i obrnuto.

Nakon što smo kreirali model, uz pomoć funkcije predict izvodimo predikciju modela nad skupom za testiranje.

```
logpredikcija = predict(logModelθ, testiranje, type = "response")
logpredY = rep("nezrela", length(test_y))
logpredY[logpredikcija > 0.5] = "zrela"
```

Kreiran je novi vektor *logpredY* i inicijalizirane su vrijednost svih elemenata na „nezrela”, a zatim je isti vektor ažuriran tako da je na pozicije za koje je predviđena vjerojatnost veća od 0.5 postavljena vrijednost „zrela”.

```
table(logpredY, test_y)

#stopa pogreške klasifikacije
mean(logpredY != test_y)
```

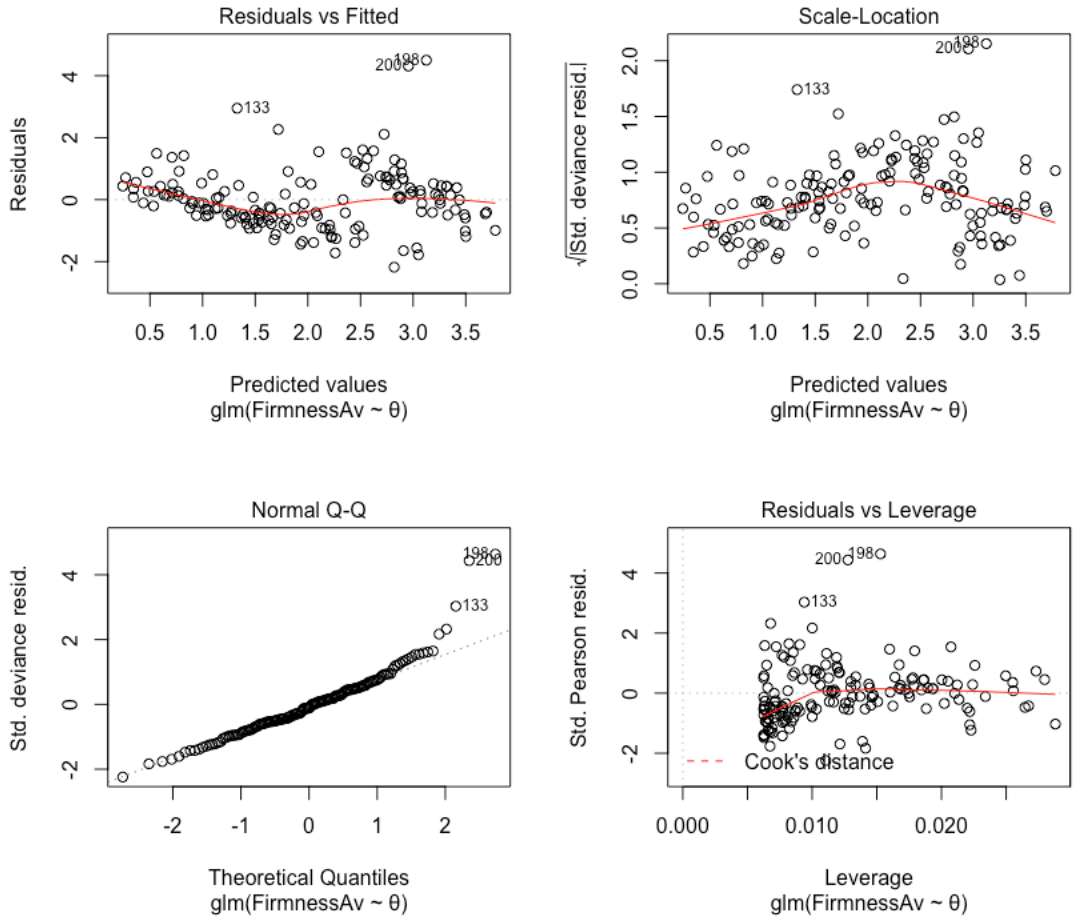
Nakon predikcije i ažuriranja vrijednosti, pomoću funkcije *table* kreirana je matrica konfuzije modela te pomoću funkcije *mean* izračunali smo stopu pogreške modela koja iznosi 0.15 ili ~15%.

```
         test_y
logpredY  nezrela zrela
  nezrela      18     5
  zrela         1    16
```

Vidimo da je model netočno predvidio 5 instanci klase *zrela* kao *nezrela* te samo jednu instancu klase *nezrela* je netočno klasificirao kao *zrela.* Na dijagonali se nalaze točno predviđene vrijednosti.

Pozivanjem funkcije plot nad *glm* modelom dobijemo grafove: odnosa između predviđenih vrijednosti i reziduala, odnosa između predviđenih vrijednosti i standardne devijacije reziduala, Normal QQ plot koji prikazuje distribucije reziduala te scale location plot koji nam pomaže u identifikaciji outliera. Za prikaz grafikona izabran je logistički model koji

predviđa zrelost na temelju kuta theta breskve, s obzirom da smo pomoću tog modela dobili najbolje rezultate.



Slika 16. grafovi modela logističke regresije

Za potrebe analize kreirana su još dva logistička modela kako bi se mogli usporediti rezultati. Za drugi model odabrana je također varijabla visoke korelacije Zs (impendanca), a za treći model odabrana je varijabla koja nema visoku korelaciju s varijablom *FirmnessAv* (zrelosti breskve).

Za izradu drugog modela također je korištena funkcija *glm* uz navođenje pripadajućeg prediktora. Statističke informacije modela možemo vidjeti pomoću funkcije *summary.*

```
      Call:
glm(formula = FirmnessAv ~ Zs, family = "binomial", data = treniranje)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5970  -0.7329   0.3267   0.6118   2.1412
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.990      1.510  -6.618 3.64e-11 ***
Zs             8.766      1.303   6.725 1.76e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 220.91  on 159  degrees of freedom
Residual deviance: 148.13  on 158  degrees of freedom
AIC: 152.13

Number of Fisher Scoring iterations: 4
```

Iz ovog ispisa vidljivo je da je drugi model također statistički značajan (mala *p* vrijednost). $ß_1$ je pozitivan (-0.49230) što znači da kada je kut theta veći, to je indeks zrelosti breskve također veći i obrnuto. Mala *p* vrijednost i pozitivan $ß_1$ ukazuju nam da se radi o značajnom koeficijentu (varijabla promatranja ima višu vjerojatnost za ciljnu varijablu).

Nakon što smo kreirali model, uz pomoć funkcije *predict* izvodimo predikciju modela nad skupom za testiranje.

```
#predikcija drugog logističkog modela

> logpredikcija2 = predict(logModelZs, testiranje, type = "response")
> logpredY2 = rep("nezrela", length(test_y))
> logpredY2[logpredikcija2 > 0.5] = "zrela"
> table(logpredY2, test_y)
         test_y
logpredY2 nezrela zrela
  nezrela      16     5
  zrela         3    16
> #stopa pogreške klasifikacije
> mean(logpredY2 != test_y)
[1] 0.2
```

Iz ovog izvještaja vidljiva je matrica konfuzije dobivena pomoću funkcije *table* te stopa pogreške modela dobivena pomoću funkcije *mean*. Vidimo da je drugi model netočno

klasificirao 5 instanci klase *zrela* te 3 instance klase *nezrela* uz nešto veću stopu pogreške od prvog modela koja iznosi 0.2 što je otprilike 20%.

Za izradu trećeg modela odabrana je varijabla *volume.cm*^{3} (volumen breskve) koja, prema prethodno odrađenim analizama, ima manji utjecaj na zrelost breskve u odnosu na imepndancu i kut theta.

```
Call:
glm(formula = FirmnessAv ~ volume.cm3, family = "binomial", data =
treniranje)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7443  -1.0886  -0.4867   1.0964   1.9040

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.21983    3.35519  -3.642  0.00027 ***
volume.cm3    0.10015    0.02749   3.643  0.00027 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 221.78  on 159  degrees of freedom
Residual deviance: 206.25  on 158  degrees of freedom
AIC: 210.25
```

Iz ovog ispisa vidljivo je da je treći model također statistički značajan – *p* vrijednost jest malena, no u usporedbi s *p* vrijednostima prethodnih modela je veća. ß$_1$ je pozitivan (0.10015) što znači da što je volumen breskve veći, to je i indeks zrelosti veći. ß$_1$ jest pozitivan broj, ali je neznatno veći od nule što ukazuje da navedena varijabla nema velik utjecaj na ciljnu varijablu.

```
 #predikcija trećeg logističkog modela
> logpredikcija3 = predict(logModelV, testiranje, type = "response")
> logpredY3 = rep("nezrela", length(test_y))
> logpredY3[logpredikcija3 > 0.5] = "zrela"
> table(logpredY3, test_y)
         test_y
logpredY3 nezrela zrela
  nezrela       8     9
  zrela        11    12
> #stopa pogreške klasifikacije
> mean(logpredY3 != test_y)
[1] 0.5
```

Iz priloženog ispisa vidimo da je treći model netočno klasificirao 11 instanci klase *nezrela* te 9 instanci klase *zrela* uz stopu pogreške od 50%.

Izvješće o timskom radu

Timski rad je bio dobro organiziran. Kolegica Nola i ja smo već ranije radili zajedno na nekim projektima pa smo već navikli na zajedničke radne navike i način rada te zbog toga nije bilo nikakvih problema tijekom izrade ovog projekta. Zajedno smo radili gotovo cijeli projekt, izuzev neki osobnih dorada i preinaka za koje je netko od nas smatrao da bi bilo dobro promijeniti. Sve je bilo na vrijeme i po dogovoru.

- Ivan Šimičić

Kolega Ivan Šimičić i ja projekt smo organizirali na način da ravnopravno podjelimo zadatke. Oboje smo imali udjela u pisanju koda, kao i pisanju ovog seminara te smo zajedničkim razmišljanjem došli do zaključaka i interpretacija navedenih u seminaru.

- Nola Čumlievski