Dokumentacija Zadaće 1 iz predmeta Mašinsko učenje

Beglerović Vedad Beglerović Vildana Buturović Lejla

Elektrotehnički fakultet Sarajevo
24. april 2022.

Sadržaj

1 Zadatak 1

1.1 Osnovne metode deskriptivne statistike

1.1.1 Upoznavanje sa setom podataka

Set podataka učitavamo naredbom:

```
podaci <- read.csv("customer_data_train.csv", fileEncoding = 'UTF-8')
```

Kako bismo se upoznali sa podacima možemo iskoristiti naredbu head(podaci) koja će nam prikazati prvih nekoliko instanci dataseta iz čega možemo zaključiti koje kolone imamo u setu.

```
head(podaci)
```

```
##   gender Dependents tenure PhoneService    MultipleLines InternetService
## 1   Male         No      8          Yes               No            <NA>
## 2 Female       <NA>      8          Yes             <NA>     Fiber optic
## 3   <NA>         No     21          Yes               No     Fiber optic
## 4   <NA>       <NA>      1           No No phone service             DSL
## 5   <NA>         No     NA         <NA>             <NA>     Fiber optic
## 6   Male         No     69          Yes               No              No
##   StreamingTV StreamingMovies       Contract             PaymentMethod
## 1         Yes             Yes           <NA>                      <NA>
## 2          No              No Month-to-month   Credit card (automatic)
## 3         Yes             Yes       One year              Mailed check
## 4          No              No Month-to-month              Mailed check
## 5          No             Yes Month-to-month          Electronic check
## 6        <NA>            <NA>       Two year Bank transfer (automatic)
##   MonthlyCharges TotalCharges DailyCharges Churn
## 1             NA       832.35           NA   Yes
## 2             NA       548.90           NA    No
## 3         104.55      2239.40        20.91    No
## 4          35.90        35.90         7.18    No
## 5          81.10        81.10        16.22   Yes
## 6          19.30      1447.90         3.86    No
```

Vidimo da imamo 14 atributa i već na prvi pogled uočavamo da imamo neke nedostajuće vrijednosti. Možemo zaključiti da imamo kategoričke i numeričke podatke kroz koje se detaljnije upoznajemo pomoću sljedećih naredbi:

```
cat("Kategorije spolova: ", unique(podaci$gender),
"\nKategorije izdrzavanja (dependents): ", unique(podaci$Dependents),
"\nKategorije PhoneServices: ", unique(podaci$PhoneService),
"\nKategorije (brojevi) nekretnina mu terije koji koriste usluge: ", sort(unique(podaci$
    tenure)),
"\nKategorije MulitpleLines: ", unique(podaci$MultipleLines),
"\nKategorije InternetService: ", unique(podaci$InternetService),
"\nKategorije StreamingTV: ", unique(podaci$StreamingTV),
"\nKategorije StreamingMovies: ", unique(podaci$StreamingMovies),
"\nKategorije ugovora: ", unique(podaci$Contract),
"\nKategorije metode pla anja: ", unique(podaci$PaymentMethod),
"\nKategorije Churn: ", unique(podaci$Churn))
```

Rezultat izvršavanja koda je sljedeći:

Kategorije spolova: Male Female NA
Kategorije izdrzavanja (dependents): No NA Yes Maybe
Kategorije PhoneServices: Yes No NA
Kategorije (brojevi) nekretnina mušterije koji koriste usluge: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72
Kategorije MulitpleLines: No NA No phone service Yes
Kategorije InternetService: NA Fiber optic DSL No
Kategorije StreamingTV: Yes No NA No internet service
Kategorije StreamingMovies: Yes No NA No internet service
Kategorije ugovora: NA Month-to-month One year Two year
Kategorije metode plaćanja: NA Credit card (automatic) Mailed check Electronic check Bank transfer (automatic) abcd
Kategorije Churn: Yes No NA

1.2 Metode procjene lokacije i varijabilnosti podataka

1.2.1 Metode za procjenu lokacije

Da bismo dobili neke osnovne podatke o procjeni lokacije svih atributa iz našeg data seta možemo iskoristiti sljedeću naredbu:

```
summary <- lapply(podaci, summary)
summary
```

Nakon izvršavanja prethodnog isječka dobijamo sljedeći pregled:

```
$gender
   Length     Class      Mode
     2000 character character

$Dependents
   Length     Class      Mode
     2000 character character

$tenure
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    9.00   29.00   32.52   56.00   72.00

$PhoneService
   Length     Class      Mode
     2000 character character

$MultipleLines
   Length     Class      Mode
     2000 character character

$InternetService
   Length     Class      Mode
     2000 character character

$StreamingTV
   Length     Class      Mode
     2000 character character
```

(a)

```
$StreamingMovies
   Length     Class      Mode
     2000 character character

$Contract
   Length     Class      Mode
     2000 character character

$PaymentMethod
   Length     Class      Mode
     2000 character character

$MonthlyCharges
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  -1.22   35.00   70.30   64.41   89.40  118.65     256

$TotalCharges
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   19.1   433.4  1415.4  2280.6  3751.7 10000.0     276

$DailyCharges
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -0.244   6.940  14.050  12.855  17.880  23.730     255

$Churn
   Length     Class      Mode
     2000 character character
```

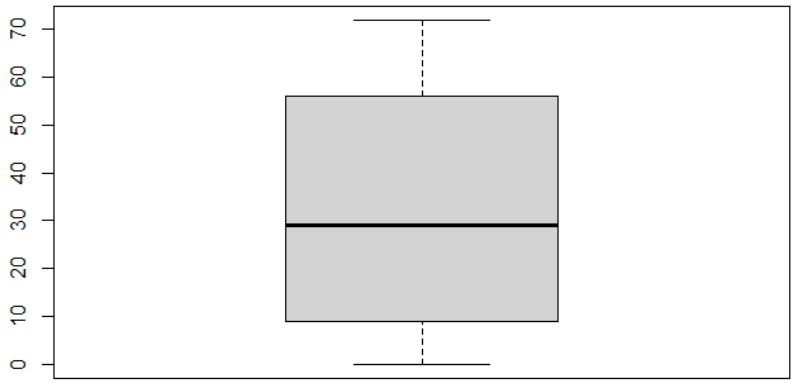
(b)

Slika 1: Statistički podaci o atributima dataseta

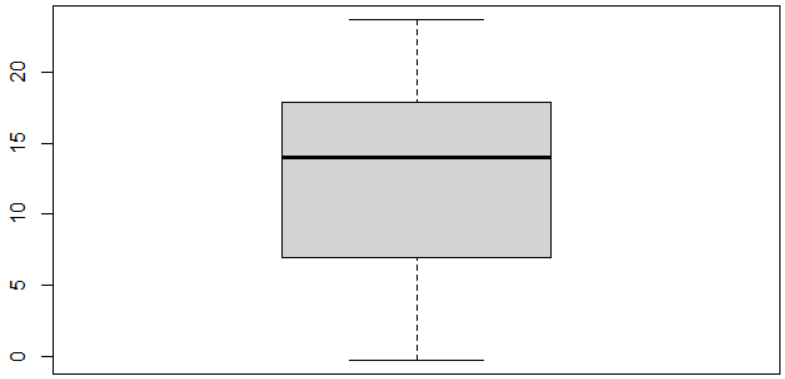
Za numeričke atribute, statističke podatke ćemo vizualizirati pomoću boxplot funkcije. Za tu svrhu možemo izdvojiti subset sa numeričkim vrijednostima:

```
num_cols <- subset(podaci, select = c(tenure,DailyCharges, MonthlyCharges, TotalCharges))
boxplots <- lapply(num_cols, boxplot)
```

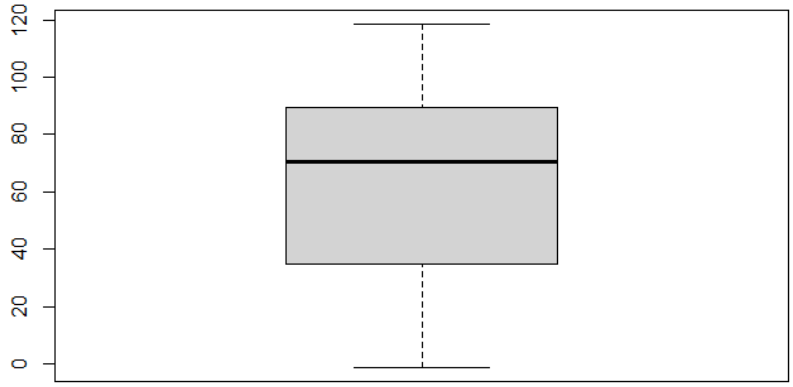
Na osnovu boxplot grafika atributa **'tenure'** (slika 2a) može se zaključiti da je distibucija podataka prilično ravnomjerna, s obzirom da se interkvartalni opseg nalazi nešto ispod sredine. Srednja vrijednost se ne nalazi na sredini opsega vrijednosti, već se nalazi nešto niže što znači da su vrijednosti broja zakupa teže ka nešto manjim vrijednostima. Srednja vrijednost se također ne nalazi na sredini između prvog i trećeg kvartala, već nešto malo ispod, što znači da više klijenata ima broj zakupa od 29 do 56. Za **dnevnu naplatu** (slika 2b) može se zaključiti da je distibucija podataka prilično ravnomjerna, s obzirom da se interkvartalni opseg nalazi skoro na sredini. Srednja vrijednost se ne nalazi na sredini opsega vrijednosti, već se nalazi nešto iznad što znači da vrijednosti dnevne naplate teže ka nešto većim vrijednostima. Srednja vrijednost se također ne nalazi na sredini između prvog i trećeg kvartala, već nešto malo iznad, što znači da više klijenata ima manju dnevnu naplatu (6.940 - 14.050). Za **mjesečnu naplatu** vrijedi ista analiza, ali sa drugačijim opsegom vrijednosti. Ukoliko zanemarimo opsege, možemo uočiti da su njihovi boxplot grafici identični (slike 2b i 2c) što nam ukazuje na potencijalnu jaku korelaciju. **Ukupna naplata** (slika 2d) ima neravnomjernu distribuciju podataka s obzirom da nema mnogo klijenata s ukupnom naplatom iznad interkvartalnog opsega (iznad 3752.175). Srednja vrijednost se ne nalazi na sredini opsega vrijednosti, već se nalazi blizu donje granice što znači da je totalna naplata uglavnom jako niska. Srednja vrijednost se također ne nalazi na sredini između prvog i trećeg kvartala, već blizu prvog kvartala, što znači da se mnogo više klijenata nalazi u donjem kraju interkvrtalnog opsega (433.125 - 1415.425).



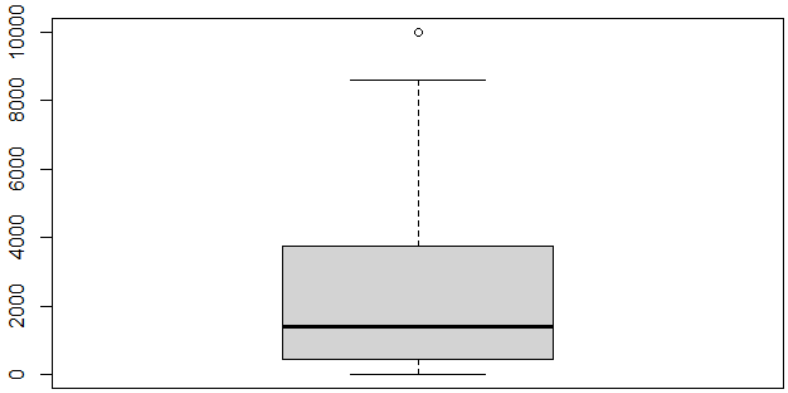
(a) Tenure



(b) Dnevna naplata



(c) Mjesečna naplata



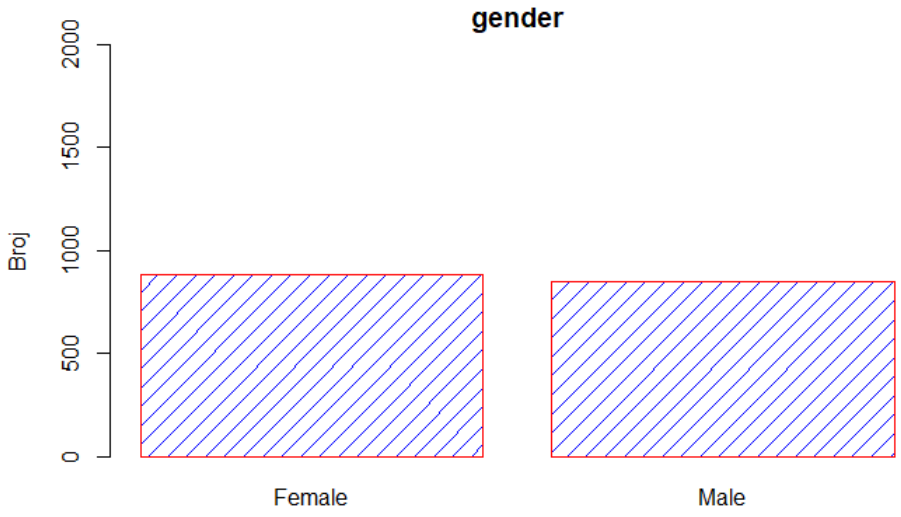
(d) Ukupna naplata

Slika 2: Boxplot grafici numeričkih atributa

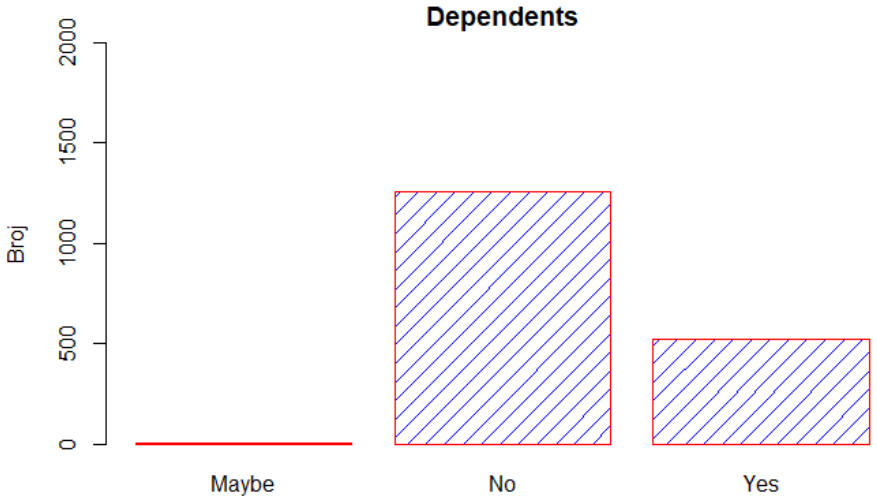
Za statistički prikaz kategoričkih atributa koristit ćemo bar plot:

```
for (i in colnames(podaci)){
  if (class(podaci[[i]]) == "character")
    barplot(table(podaci[[i]]),
    main=colnames(podaci[i]),
    ylim= c(0, length(podaci[[i]])),
    ylab="Broj",
    border="red",
    col="blue",
    density=10
    )
}
```

Navedeni kod rezultira sa 10 grafika:

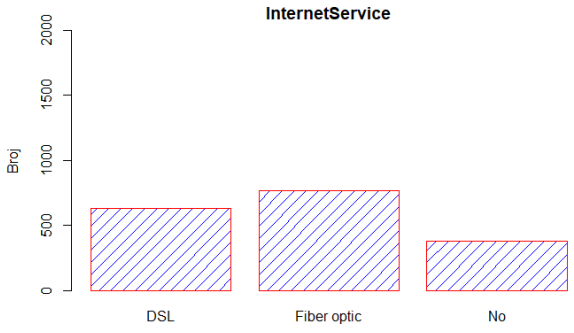


(a) Gender

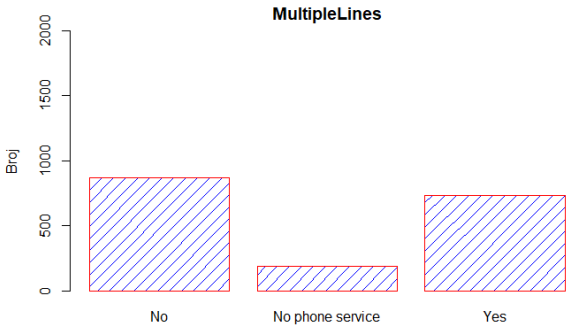


(b) Dependents

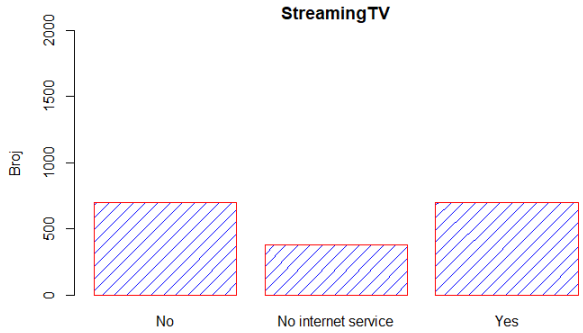
Slika 3: Spol i



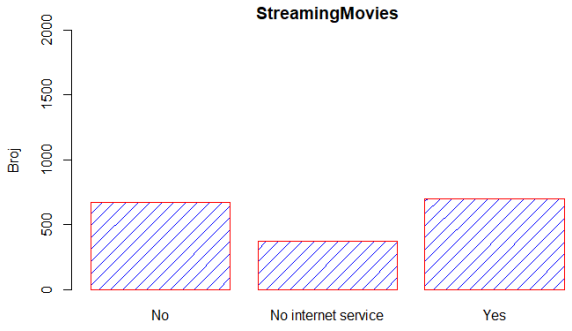
(a) InternetService



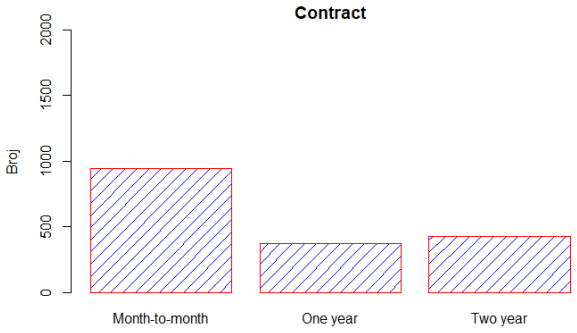
(b) MultipleLines



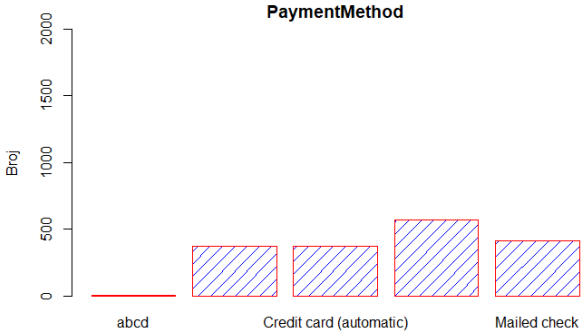
(c) StreamingTV



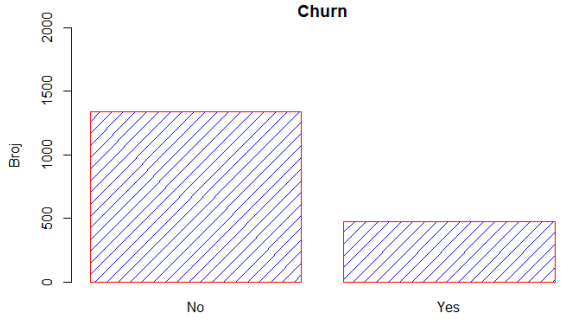
(d) StreamingMovies



(e) Contract



(f) PaymentMethod



(g) Churn

Slika 4: Barplot grafici kategoričkih atributa

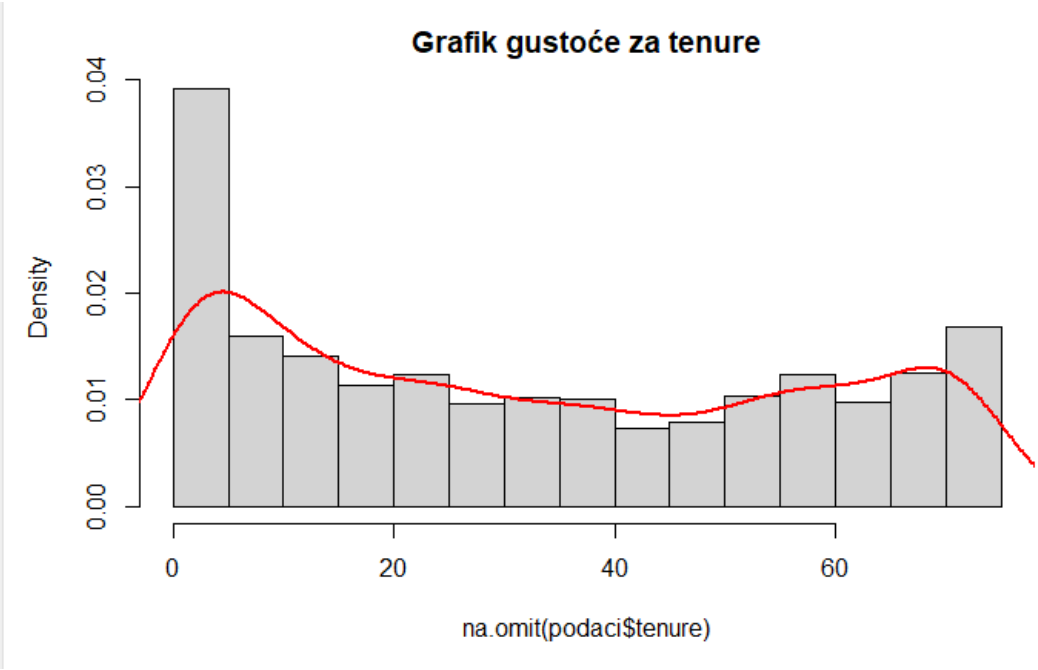
1.2.2 Metode za procjenu varijabilnosti

```
trimmed_mean <- function(vektor,p)
{
 vektor <- sort(vektor)
```

```
 trimmed_vektor <- vektor[(1 + p) : (length(vektor) - p)]
 return (sum(trimmed_vektor) / length(trimmed_vektor))
}

cat("Mean za broj nekretnina koje mu terija posjeduje:", mean(podaci_brojcani_bez_na$tenure),
    "\n")
cat("Trimmed_mean za broj nekretnina koje mu terija posjeduje (p=1):", trimmed_mean(podaci_
    brojcani_bez_na$tenure, 1),"\n")
cat("Trimmed_mean za broj nekretnina koje mu terija posjeduje (p=2):", trimmed_mean(podaci_
    brojcani_bez_na$tenure, 2),"\n")
cat("Trimmed_mean za broj nekretnina koje mu terija posjeduje (p=4):", trimmed_mean(podaci_
    brojcani_bez_na$tenure, 4),"\n")
cat("Median za broj nekretnina koje mu terija posjeduje:", median(podaci$tenure), "\n\n")

cat("Mean za dnevnu naplatu:", mean(podaci_brojcani_bez_na$DailyCharges),"\n")
cat("Trimmed_mean za dnevnu naplatu (p=1):", trimmed_mean(podaci_brojcani_bez_na$DailyCharges,
     1),"\n")
cat("Trimmed_mean za dnevnu naplatu (p=2):", trimmed_mean(podaci_brojcani_bez_na$DailyCharges,
     2),"\n")
cat("Trimmed_mean za dnevnu naplatu (p=4):", trimmed_mean(podaci_brojcani_bez_na$DailyCharges,
     4),"\n")
cat("Median za dnevnu naplatu:", median(podaci_brojcani_bez_na$DailyCharges), "\n\n")

 cat("Mean za mjese nu naplatu:", mean(podaci$MonthlyCharges),"\n")
 cat("Trimmed_mean za mjese nu naplatu (p=1):", trimmed_mean(podaci$MonthlyCharges, 1),"\n")
 cat("Trimmed_mean za mjese nu naplatu (p=2):", trimmed_mean(podaci$MonthlyCharges, 2),"\n")
 cat("Trimmed_mean za mjese nu naplatu (p=4):", trimmed_mean(podaci$MonthlyCharges, 4),"\n")
 cat("Median za mjese nu naplatu:", median(podaci$MonthlyCharges), "\n\n")
```

U navedenom isječku prikazano je izračunavanje mean, median i trimmed_mean vrijednosti brojčanih atributa, ali bez instanci sa na vrijednostima brojčanih atributa, tako da se ovi parametri trebaju uzeti s rezervom, jer ukoliko bi se pojedinačno računalo za svaki brojčani atribut vrlo je moguće da će se podaci razlikovati.

Grafik gustoće za tenure prikazan je na sljedećoj slici:



Slika 5: Grafik gustoće ta tenure

1.3 Metode za procjenu korelacije između varijabli

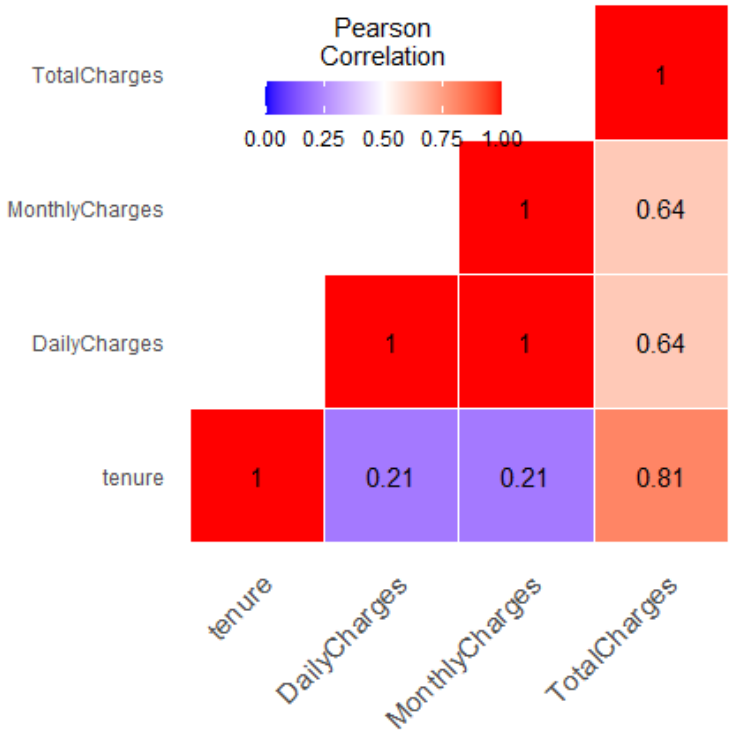
1.3.1 Korelacija između numeričkih varijabli

Za određivanje korelacije između numeričkih varijabli, koristili smo Pearsonov koeficijent korelacije. Odredili smo korelaciju svakog numeričkog atributa sa svakim, odnosno korelacije između atributa: ‘tenure‘, ‘DailyCharges‘, ‘MonthlyCharges‘ i ‘TotalCharges‘.

```
  library(reshape2)
library(ggplot2)
# funkcija za odsijecanje gornjeg dijela matrice
get_upper_tri <- function(cormat){
 cormat[lower.tri(cormat)]<- NA
 return(cormat)
}
# korelacijski koeficijent koriste i NA podatke ima rezultnu vrijednost NA
podaci_brojcani <- subset(podaci, select = c(tenure,DailyCharges,MonthlyCharges,TotalCharges))
podaci_brojcani_bez_na <- na.omit(podaci_brojcani)
# kreiranje korelacijske matrice za sve podatke
cormat <- round(cor(podaci_brojcani_bez_na), 2)
# odsijecanje gornjeg dijela matrice
upper_tri <- get_upper_tri(cormat)
melted_cormat <- melt(upper_tri, na.rm = TRUE)

ggheatmap <- ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
geom_tile(color = "white")+
scale_fill_gradient2(low = "blue", high = "red", mid = "white",
 midpoint = 0.5, limit = c(0,1), space = "Lab",
 name="Pearson\nCorrelation") +
```

```
22  theme_minimal()+
23 theme(axis.text.x = element_text(angle = 45, vjust = 1,
24  size = 12, hjust = 1))+
25 coord_fixed()
26 ggheatmap +
27 geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
28 theme(
29  axis.title.x = element_blank(),
30  axis.title.y = element_blank(),
31  panel.grid.major = element_blank(),
32  panel.border = element_blank(),
33  panel.background = element_blank(),
34  axis.ticks = element_blank(),
35  legend.justification = c(1, 0),
36  legend.position = c(0.6, 0.7),
37  legend.direction = "horizontal")+
38  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
39  title.position = "top", title.hjust = 0.5))
40
41  ggheatmap
```

U liniji 9 smo u varijablu podaci_brojčani izdvojili numeričke atribute, zatim smo u liniji 10 pomoću funkcije na.omit odstranili instance (klijente) koje nemaju definisane vrijednosti jedne ili više navedenih brojčanih atributa i preostale instance pohranili u varijablu podaci_brojcani_bez_na. U linija koje slijede kreirali smo heat mapu koja pokazuje koeficijente korelacije između svakih od atributa. Heat mapa je prikazana na slici 6.



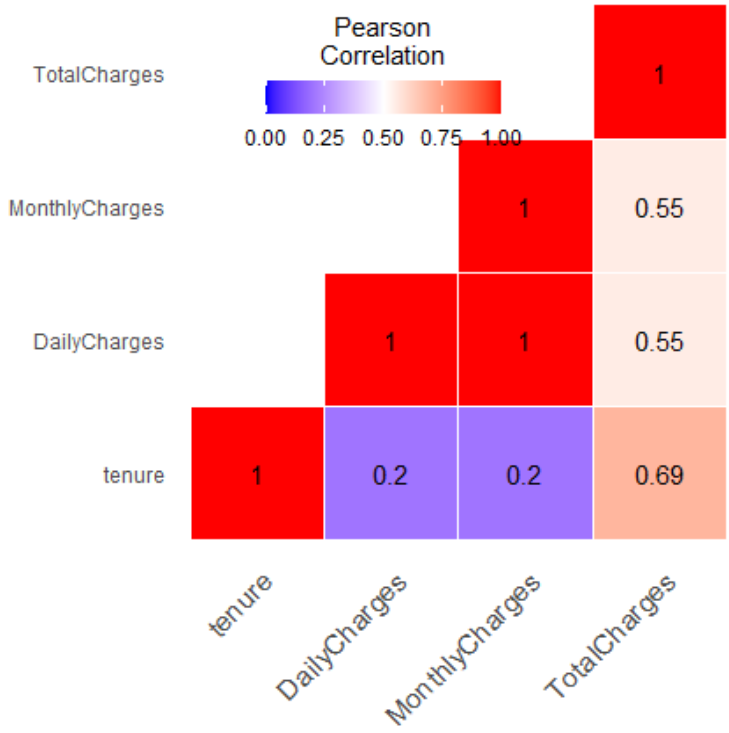
Slika 6: Pearsonov koeficijent korelacije između numeričkih atributa

Svi Pearsonovi koeficijenti korelacija su pozitivni što znači da se varijable pomjeraju u istom smjeru, odnosno da im vrijednosti uzajamno rastu ili uzajamno opadaju. Dnevna naplata i mjesečni troškovi klijenata su u potpunoj korelaciji (koeficijent 1), pa je potrebno izbaciti jedan od navedenih atributa. Smatra se da su u jakoj korelaciji svi atributi čiji Pearsonov koeficijent ima vrijednost veću od 0,7 ukoliko se kreću u istom smijeru, odnosno manju od -0,7 ukoliko se kreću u suprotnim smjerovima. Prema tome, koeficijent koji iznosi 0.81 indicira da su u jakoj korelaciji broj zakupa (tenure) i ukupna naplata (TotalCharges), pa je potrebno izbaciti jedan od ovih atributa.

Možemo provesti i analizu korelacije u slučaju da ne izbacujemo NA vrijednosti. Da bismo to učinili, prvo ćemo popuniti nedostajuće vrijednosti sa medijanima odgovarajućeg atributa. To ćemo učiniti sljedećim kodom:

```
podaci_brojcani <- podaci_brojcani %>%
mutate_if(is.numeric, function(x) ifelse(is.na(x), median(x, na.rm = T), x))
```

Ukoliko sada nad varijablom podaci_brojcani izvršimo analizu koleracije dobit ćemo sljedeći grafik:



Slika 7: Pearsonov koeficijent korelacije između numeričkih atributa (NA popunjene medijanom)

Vidimo da se koeficijent korelacije između atributa TotalCharges i tenure malo smanjio, no i dalje je jak i iznosi približno 0.7. S obzirom da je koeficijent na granici za izbacivanje atributa, možemo još i napraviti subset koji će se sastojati iz atributa 'tenure' i 'TotalCharges' bez instanci koje posjeduju bar jednu NA vrijednost i izračunati njihovu korelaciju. Navedeno je prikazano u sljedećem isječku:

```
total_tenure <- subset(podaci, select = c(tenure,TotalCharges))
total_tenure_bez_na <- na.omit(total_tenure)
cor <- cor.test(total_tenure_bez_na$tenure, total_tenure_bez_na$TotalCharges, method = "
    pearson")
cat("Pearsonov koeficijent korelacije tenure-TotalCharges: ", cor$estimate)
```

Rezultat izvršavanja isječka je:

Pearsonov koeficijent korelacije tenure-TotalCharges: 0.8143779

Dakle, definitvno ćemo izbaciti ili atribut TotalCharges ili atribut tenure.

1.3.2 Korelacija između kategoričkih varijabli

Za određivanje korelacije između kategoričkih varijabli koristili smo chi-square koeficijent korealcije. Odredili smo korelaciju svakog kategoričkog atributa sa svakim i koeficijente smjestili u korelacijsku matricu. Prije određivanja chi-square izbačeni su outlieri, s obzirom da je to jedan od uslova da bi analiza bila uspješna.

```r
chi_matrica <- matrix(0,ncol(podaci), ncol(podaci))
rownames(chi_matrica) <- colnames(podaci)
colnames(chi_matrica) <- colnames(podaci)
critical_matrica <- matrix(0,ncol(podaci), ncol(podaci))
rownames(critical_matrica) <- colnames(podaci)
colnames(critical_matrica) <- colnames(podaci)
razlika_matrica <- matrix(0,ncol(podaci), ncol(podaci))
rownames(razlika_matrica) <- colnames(podaci)
colnames(razlika_matrica) <- colnames(podaci)
for (k in colnames(podaci)){
  atribut1 <- sort(unique(podaci[[k]]))
  for (l in colnames(podaci)){
    atribut2 <- sort(unique(podaci[[l]]))
   if(class(podaci[[k]]) == "character" & class(podaci[[l]]) == "character"){
        data2 <- matrix(0, length(atribut1), length(atribut2))
        rownames(data2) <- atribut1
        colnames(data2) <- atribut2
        cat("\nMatrica korelacije atributa ", colnames(podaci[k]), " i ", colnames(podaci[l]),
            ":\n\n")
           for (i in 1 : length(atribut1)) {

               for (j in 1 : length(atribut2)) {

                   redovi <- subset(podaci, (podaci[[k]] == atribut1[i] & podaci[[l]] ==
                       atribut2[j]))
                   data2[i, j] = length(redovi[[k]])
               }
           }
      print(data2)
      chi <- chisq.test(data2)
      chi_matrica[k,l] <- chi$statistic
      critical <- qchisq(p = chi$p.value, df = chi$parameter)
      critical_matrica[k,l] <- critical
      if(chi$statistic - critical >0){
      razlika_matrica[k,l] <- chi_matrica[k,l] - critical
      }
   }
  }
}
cat("Chi matrica:\n\n")
print(chi_matrica)
cat("Matrica krticnih vrijednosti:\n\n")
print(critical_matrica)
cat("Matrica razlike chi matrice i kriticnih vrijednosti:\n\n")
#print(razlika_matrica)
razlika_matrica <- razlika_matrica[,!colnames(razlika_matrica) %in% c("DailyCharges", "tenure"
    , "MonthlyCharges", "TotalCharges")]
razlika_matrica <- razlika_matrica[!rownames(razlika_matrica) %in% c("DailyCharges", "tenure",
```

```
      "MonthlyCharges", "TotalCharges"),]
print(razlika_matrica)
razlika.df <- as.data.frame(t(razlika_matrica))
print(razlika.df)
normalize <- function(x) {
return (round((x - min(x)) / (max(x) - min(x)),5))
}
for (i in colnames(razlika.df)){
razlika.df[[i]]<-normalize(razlika.df[[i]])
}
print(razlika.df)
razlika_matrica <- as.matrix(razlika.df)
print(razlika_matrica)
```

U linijama 1 - 9 deklarišemo matrice: chi_matrica, critical_matrica i razlika_matrica i postavimo njihove vrijednosti na 0. Zatim u for petlji u liniji 10 prolazimo kroz sve atribute dataframea i u svakoj iteraciji petlje po jedan pohranjujemo u varijablu 'atribut1'. Zatim u liniji 12 otvaramo novu petlju kojom također prolazimo kroz sve atribute i pohranjujemo ih u varijablu atribut 2. Ukoliko su u pitanju kategorički atributi (uslov u liniji 14) formira se korelacijska matrica. Zatim se matrica šalje u funkciju chisq.test u liniji 28 i u chi matricu se dodaje vrijednost chi$statistic. Kritične vrijednosti se dodaju u odgovarajuća polja u crtical matrici. Konačno, u liniji 33 se u matricu razlike dodaje razlika između chi-square vrijednosti i kritične vrijednosti ukoliko je razlika pozitivna, odnosno ukoliko su odgovarajući atributi u korelaciji. Na slici 8 su prikazane neke od korelacijskih matrica. Na slici 10 je prikazana heat mapa korelacije kategoričkih atributa.

```
Matrica korelacije atributa  gender  i  Dependents

        No Yes
Female 543 224
Male   542 229

Matrica korelacije atributa  gender  i  PhoneService

       No Yes
Female 79 685
Male   66 665

Matrica korelacije atributa  gender  i  Multipl

        No No phone service Yes
Female 376               90 331
Male   365               71 312

Matrica korelacije atributa  gender  i  Interne

       DSL Fiber optic  No
Female 285         345 147
Male   253         317 175

Matrica korelacije atributa  gender  i  Streami

        No No internet service Yes
Female 318                 153 314
Male   287                 171 290

Matrica korelacije atributa  gender  i  Streami

        No No internet service Yes
Female 290                 156 320
Male   286                 170 286

Matrica korelacije atributa  gender  i  Contract

       Month-to-month One year Two year
Female            426      167      180
Male              393      154      185
```

(a)

```
Matrica korelacije atributa  Dependents  i  StreamingMovies

     No No internet service Yes
No  452                203 443
Yes 137                136 174

Matrica korelacije atributa  Dependents  i  Contract

    Month-to-month One year Two year
No             673      222      199
Yes            162      116      174

Matrica korelacije atributa  Dependents  i  PaymentMethod

    Bank transfer (automatic) Credit card (automatic) Electronic check Mailed check
No                        205                     208              417          247
Yes                       119                     117               91          122

Matrica korelacije atributa  Dependents  i  Churn

     No Yes
No  774 372
Yes 411  57

Matrica korelacije atributa  PhoneService  i  gender

    Female Male
No      79   66
Yes    685  665

Matrica korelacije atributa  PhoneService  i  Dependents

     No Yes
No   99  52
Yes 994 396

Matrica korelacije atributa  PhoneService  i  PhoneService

     No  Yes
No  169    0
```

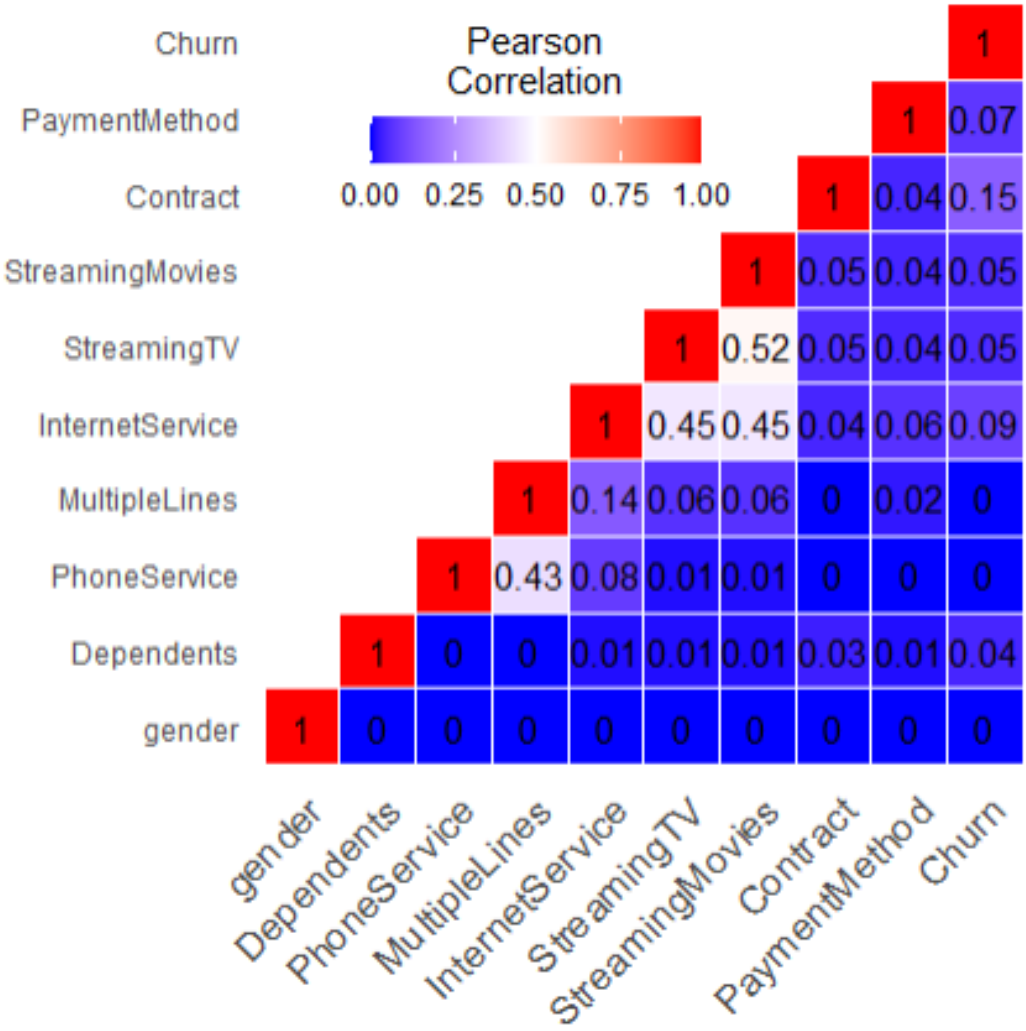
(b)

Slika 8: Neke od korelacijskih matrica

Formirana normalizirana matrica razlike chi-square vrijednosti i kritičnih vrijednosti:

	gender <dbl>	Dependents <dbl>	PhoneService <dbl>	MultipleLines <dbl>	InternetService <dbl>	StreamingTV <dbl>	StreamingMovies <dbl>	Contract <dbl>	PaymentMethod <dbl>
gender	1.00000	0.00000	0.00014	0.00002	0.00000	0.00000	0.00000	0.00000	0.00089
Dependents	0.00000	1.00000	0.00118	0.00000	0.01227	0.00815	0.00823	0.02817	0.00954
PhoneService	0.00014	0.00114	1.00000	0.43350	0.08275	0.01237	0.01291	0.00000	0.00000
MultipleLines	0.00003	0.00000	0.89999	1.00000	0.14012	0.05683	0.05523	0.00404	0.01915
InternetService	0.00271	0.02712	0.17313	0.14031	1.00000	0.45491	0.45180	0.04104	0.06185
StreamingTV	0.00118	0.01744	0.02671	0.05705	0.45501	1.00000	0.52498	0.04746	0.04221
StreamingMovies	0.00077	0.01694	0.02698	0.05440	0.44464	0.51659	1.00000	0.05427	0.03910
Contract	0.00000	0.05529	0.00000	0.00394	0.03903	0.04600	0.05372	1.00000	0.03833
PaymentMethod	0.00301	0.02815	0.00036	0.02789	0.08911	0.06109	0.05765	0.05707	1.00000
Churn	0.00000	0.03896	0.00000	0.00196	0.04206	0.02302	0.02444	0.08022	0.02454

Slika 9: Matrica razlike kategoričkih atributa



Slika 10: Heat mapa korealcije kategoričkih atributa

1.4 Predprocesiranje podataka

1.4.1 Odbacivanje atributa sa visokim stepenom korelacije

Prva dva atributa koja su u korelaciji su 'MonthlyCharges' i 'DailyCharges'. Ova dva atributa imaju iste **koeficijente korelacije sa ostalim atributima** što se vidi na slikama 6 i 7. Da bismo odlučili koji atribut izbaciti, provjerit ćemo koji atribut ima više NA vrijednosti jer smatramo da je bolje izbaciti atribut koji ima više nedostajućih vrijednosti.

```
cat("Broj NA vrijednosti u koloni MonthlyCharges: ",
length(subset(podaci, {is.na(podaci$MonthlyCharges)})$MonthlyCharges),
"\nBroj NA vrijednosti u koloni DailyCharges: ",
length(subset(podaci, {is.na(podaci$DailyCharges)})$DailyCharges))
```

Nakon izvršavanja isječka dobijamo sljedeći rezultat:

Broj NA vrijednosti u koloni MonthlyCharges: 256
Broj NA vrijednosti u koloni DailyCharges: 255

S obzirom da imaju približan broj NA vrijednosti, izbacit ćemo atribut dnevne naplate jer mjesečna naplata ima približnije vrijednosti ukupnoj naplati nego dnevna naplata. Ovo ćemo uraditi pomoću sljedećeg koda:

```
podaci <- subset(podaci, select = -c(DailyCharges))
```

Analognu analizu radimo i za atribute 'tenure' i 'TotalCharges':

```
cat("Broj NA vrijednosti u koloni TotalCharges: ",
length(subset(podaci, {is.na(podaci$TotalCharges)})$TotalCharges),
"\nBroj NA vrijednosti u koloni tenure: ",
length(subset(podaci, {is.na(podaci$tenre)})$tenure))
```

Nakon izvršavanja isječka dobijamo sljedeći rezultat:

Broj NA vrijednosti u koloni TotalCharges: 276
Broj NA vrijednosti u koloni tenure: 0

Najbitniju ulogu u izbacivanju atributa igra **stepen korelacije sa ostalim atributima**, pa tako vidimo na slici 6 da atribut 'TotalCharges' ima visok stepen korelacije sa 'MonthlyCharges' i 'DailyCharges' (0,64), dok atribut 'tenure' ima nešto niže koeficijente (0.21). Također, atribut 'TotalCharges' ima više nedostajućih vrijednosti. Iz navedenih razloga izbacujemo atribut 'TotalCharges'.

```
podaci <- subset(podaci, select = -c(TotalCharges))
```

Svi parovi kategoričkih atributa čija je vrijednost nula nisu u korelaciji. Možemo primijetiti da atributi **phoneService** i **MultipleLines** imaju nešto viši stepen korelacije (0,43) što ima smisla jer više linija ne može imati neko ko ne koristi telefonske usluge. Također, parovi **InternetService** i **StreamingTV** (0,45), **InternetService** i **StreamingMovies** (0,45), **StreamingTV** i **StreamingMovies** (0,52). Možemo izbaciti kategoriju **MultipleLines** s obzirom da ima veću korelaciju i sa ostalim atributima u odnosu na phoneService. Sasvim je svejedno da li ćemo odbaciti **StreamingMovies** ili **StreamingTV**, s obzirom da su im korelacije sa ostalim atributima jednake. Izbacit ćemo kategoriju koja ima više NA vrijednosti.

```
cat("Broj NA vrijednosti u koloni StreamingMovies: ",length(subset(podaci, {is.na(podaci$
    StreamingMovies)})$StreamingMovies), "\nBroj NA vrijednosti u StreamingTV: ",length(subset
    (podaci, {is.na(podaci$StreamingTV)})$StreamingTV))
```

Kod iznad daje rezultat:

Broj NA vrijednosti u koloni StreamingMovies: 252
Broj NA vrijednosti u StreamingTV: 224

Dakle, izbacit ćemo **StreamingMovies**. Možemo još izbaciti ili **StreamingTV** ili **InternetService**. InternetService ima nešto veću korelaciju sa ostalim varijablama, pa ćemo nju izbaciti.

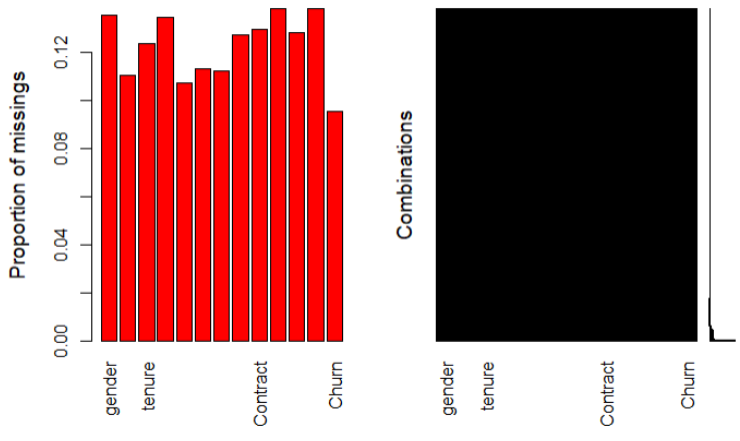
Izbacivanje dovršavamo sljedećim kodom:

```
podaci <- subset(podaci, select = -c(MultipleLines, StreamingMovies, InternetService))
```

1.4.2 Popunjavanje nedostajućih vrijednosti

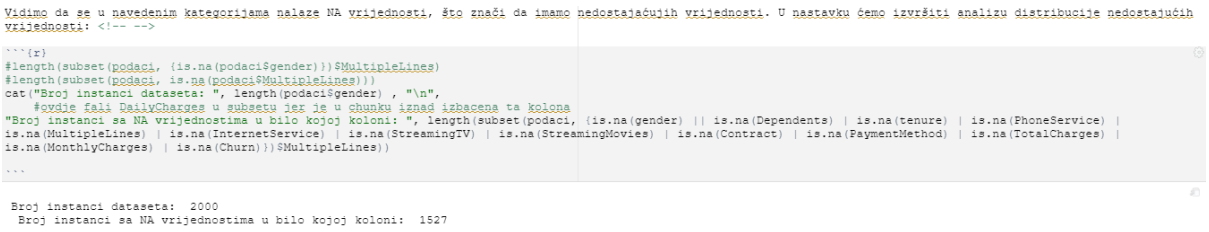
Analiza na vrijednosti pokazuje sljedeću situaciju:

```
library(VIM)
aggr(podaci)
podaci_stari <- podaci
#svjesni smo da drugi grafik Combinations nije adekvatan, ali to je tehnicki nedostatak funkcije
nekoliko kolona
```



Slika 11: Analiza NA vrijednosti

Vidimo da u svim atributima imamo nedostajućih vrijednosti, stoga ih je potrebno popuniti. Oko tri četvrtine instanci ima bar jednu nedostajuću vrijednost što može biti problem za naše modele:



```
Vidimo da se u navedenim kategorijama nalaze NA vrijednosti, što znači da imamo nedostajaćujih vrijednosti. U nastavku ćemo izvršiti analizu distribucije nedostajućih
vrijednosti: <!-- -->

```{r}
#length(subset(podaci, (is.na(podaci$gender))$MultipleLines)
#length(subset(podaci, is.na(podaci$MultipleLines)))
cat("Broj instanci dataseta: ", length(podaci$gender) , "\n",
    #vidje fali DailyCharges u subsetu jer je u chunku iznad izbacena ta kolona
"Broj instanci sa NA vrijednostima u bilo kojoj koloni: ", length(subset(podaci, (is.na(gender) || is.na(Dependents) | is.na(tenure) | is.na(PhoneService) |
is.na(MultipleLines) | is.na(InternetService) | is.na(StreamingTV) | is.na(StreamingMovies) | is.na(Contract) | is.na(PaymentMethod) | is.na(TotalCharges) |
is.na(MonthlyCharges) | is.na(Churn))$MultipleLines))

```

 Broj instanci dataseta: 2000
  Broj instanci sa NA vrijednostima u bilo kojoj koloni: 1527
```

Slika 12: Broj instanci sa bar jednim nedostajućim podatkom

Brojčane atribute koje nećemo odbaciti probali smo popuniti na primitivniji način (samo medijanom). To smo uradili pomoću sljedećeg isječka koda:

```
  library(tidyverse)
 podaci <- podaci %>%
 mutate_if(is.numeric, function(x) ifelse(is.na(x), median(x, na.rm = T), x))
```

Drugi način koji smo pokušali je da atribut tenure popunimo medijanom, a da atribut MonthlyCharges popunimo prosječnom vrijednosti po jednoj kategoriji atributa tenure (s obzirom da ima diskretne vrijednosti). Naši modeli su oba slučaja davali slične rezultate.

```
 podaci_stari <- podaci
not_na <- subset(podaci, is.na(tenure) == FALSE)$tenure
median <- median(not_na)
for (i in 1 : length(podaci$tenure))
{
 if (is.na(podaci$tenure[i]) == TRUE)
 {
 podaci$tenure[i] <- median
 }
}

podaci_kategorija <- podaci
podaci_kategorija$tenure <- factor(podaci_kategorija$tenure)
not_na <- subset(podaci_kategorija, is.na(MonthlyCharges) == FALSE)
srednje_vrijednosti <- c()
svi_ratinzi <- levels(podaci_kategorija$tenure)
for (i in 1 : length(svi_ratinzi))
{
 srednja_vrijednost <- median(subset(not_na, tenure ==
svi_ratinzi[i])$MonthlyCharges)
 srednje_vrijednosti <- append(srednje_vrijednosti, srednja_vrijednost)
}
for (i in 1 : length(podaci_kategorija$MonthlyCharges))
{
 if (is.na(podaci_kategorija$MonthlyCharges[i]) == TRUE)
 {
 index <- podaci_kategorija$tenure[i]
 podaci_kategorija$MonthlyCharges[i] <- srednje_vrijednosti[index]
```

```
 }
}
podaci_kategorija$tenure <- as.numeric(as.character(podaci_kategorija$tenure))
podaci <- podaci_kategorija
```

Vrijednosti kategoričkih varijabli smo popunili na sljedeći način:

```
male<-0
dependent<-0
for (i in 1 : length(podaci$gender))
{

  #popunjavanje StreamingTV na No internet service ako je InternetService No
  if (is.na(podaci$StreamingTV[i]) == TRUE & is.na(podaci$InternetService[i]) == FALSE &
      podaci$InternetService[i]=="No")
  {
    podaci$StreamingTV[i] <- "No internet service"
  }
  #popunjavanje PhoneService na Yes ako je MultipleLines yes
  if (is.na(podaci$PhoneService[i]) == TRUE & is.na(podaci$MultipleLines[i]) == FALSE & podaci
      $MultipleLines[i]=="Yes")
  {
    podaci$PhoneService[i] <- "Yes"
  }

   if (is.na(podaci$gender[i]) == TRUE)
  {
     if (male == 1){
      podaci$gender[i] <- "Female"
      male<-0
     } else {
        podaci$gender[i] <- "Male"
        male<-1
     }
   }
   if (is.na(podaci$Dependents[i]) == TRUE)
  {
     if (dependent == 1){
      podaci$Dependents[i] <- "No"
      dependent<-0
     } else {
        podaci$Dependents[i] <- "Yes"
        dependent<-1
     }
   }
  #ako su PhoneService i MultipleLines oba Na
  if (is.na(podaci$PhoneService[i]) == TRUE & is.na(podaci$MultipleLines[i])==TRUE){
    podaci$PhoneService[i]<-'No'
```

```
    podaci$MultipleLines[i]<-'No phone service'
  }
  #ako su InternetService i StreamingTV i StreamingMovies svi Na
  if (is.na(podaci$InternetService[i]) == TRUE & is.na(podaci$StreamingTV[i])==TRUE & is.na(
      podaci$StreamingMovies[i])==TRUE){
    podaci$InternetService[i]<-'No'
    podaci$StreamingTV[i]<-'No internet service'
    podaci$StreamingMovies[i]<-'No internet service'
  }
  #ako su InternetService i StreamingTV Na i StreamingMovies No
  if (is.na(podaci$InternetService[i]) == TRUE & is.na(podaci$StreamingTV[i])==TRUE & podaci$
      StreamingMovies[i]=='No'){
    podaci$InternetService[i]<-'No'
    podaci$StreamingTV[i]<-'No internet service'
  }
  #ako su InternetService i StreamingMovies Na i StreamingTV No
  if (is.na(podaci$InternetService[i]) == TRUE & is.na(podaci$StreamingMovies[i])==TRUE &
      podaci$StreamingTV[i]=='No'){
    podaci$InternetService[i]<-'No'
    podaci$StreamingMovies[i]<-'No internet service'
  }
  if (is.na(podaci$InternetService[i]) == TRUE){
    podaci$InternetService[i]<-'DSL'
  }
  if (is.na(podaci$PhoneService[i]) == TRUE){
    podaci$PhoneService[i]<-'Yes'
  }
  if (is.na(podaci$MultipleLines[i]) == TRUE){
    podaci$MultipleLines[i]<-'Yes'
  }
  if (is.na(podaci$StreamingTV[i]) == TRUE){
    podaci$StreamingTV[i]<-'Yes'
  }
   if (is.na(podaci$StreamingMovies[i]) == TRUE){
    podaci$StreamingMovies[i]<-'Yes'
   }
  if (is.na(podaci$Contract[i]) == TRUE){
    podaci$Contract[i]<-'Month-to-month'
  }
  if (is.na(podaci$PaymentMethod[i]) == TRUE){
    podaci$PaymentMethod[i]<-'Electronic check'
  }
}
```

1.4.3 Odbacivanje pronađenih outliera

Postoji mnogo tehnika pomoću kojih možemo uočiti outliere. Jedna od tehnika za kategoričke atribute ili atribute sa diskretnim vrijednostima jeste pomoću analize barplot grafika. Pa tako na grafiku 3b možemo uočiti kako imamo zanemariv broj klijenata čija je vrijednost atributa 'Depentents' 'Maybe'. U sljedećem isječku pronalazimo te instance:

```
print(subset(podaci, Dependents == "Maybe"))
```

	gender <chr>	Dependents <chr>	tenure <int>	PhoneService <chr>	MultipleLines <chr>	InternetService <chr>
550	Female	Maybe	61	Yes	No	DSL

Slika 13: Outlier atributa 'Dependents'

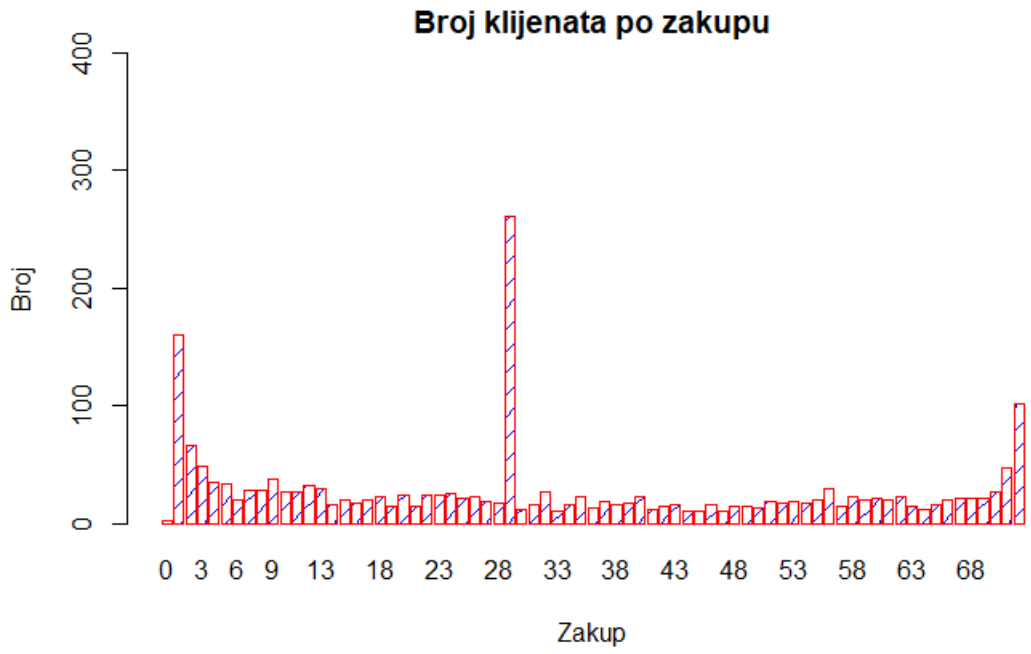
Na slici 4f uočavamo da postoji outlier kategorije 'PaymentMethod' koji ima vrijednost 'abcd':

```
print(subset(podaci, PaymentMethod == "abcd"))
```

	gender <chr>	Dependents <chr>	tenure <int>	PhoneService <chr>	MultipleLines <chr>	InternetService <chr>	Streamin <chr>
70	*NA*	No	1	Yes	No	No	No intern

Slika 14: Outlier atributa 'PaymentMethod'

Preostaje nam da ispitamo da li ima outliera u atributu 'tenure' i za tu svrhu pravimo barplot dijagram.



Slika 15: Barplot atributa 'tenure'

Možemo primijetiti da outlier ima vrijednost atributa tenure 0. U pitanju su dvije instance:

	gender <chr>	Dependents <chr>	tenure <int>	PhoneService <chr>	MultipleLines <chr>	InternetService <chr>
76	Female	Yes	0	Yes	Yes	DSL
1632	Female	*NA*	0	*NA*	No	DSL

Slika 16: Outlieri atributa 'tenure'

Pronađene outliere ćemo ukloniti iz dataframea sljedećom naredbom:

```
podaci <- podaci[!rownames(podaci) %in% c("550", "70", "76","1632"), ]
```

Isto je bilo moguće i koristeći subset funkciju sa kriterijima, ali s obzirom da nema puno instanci koje treba ukloniti navedeni način je prihvatljiv.

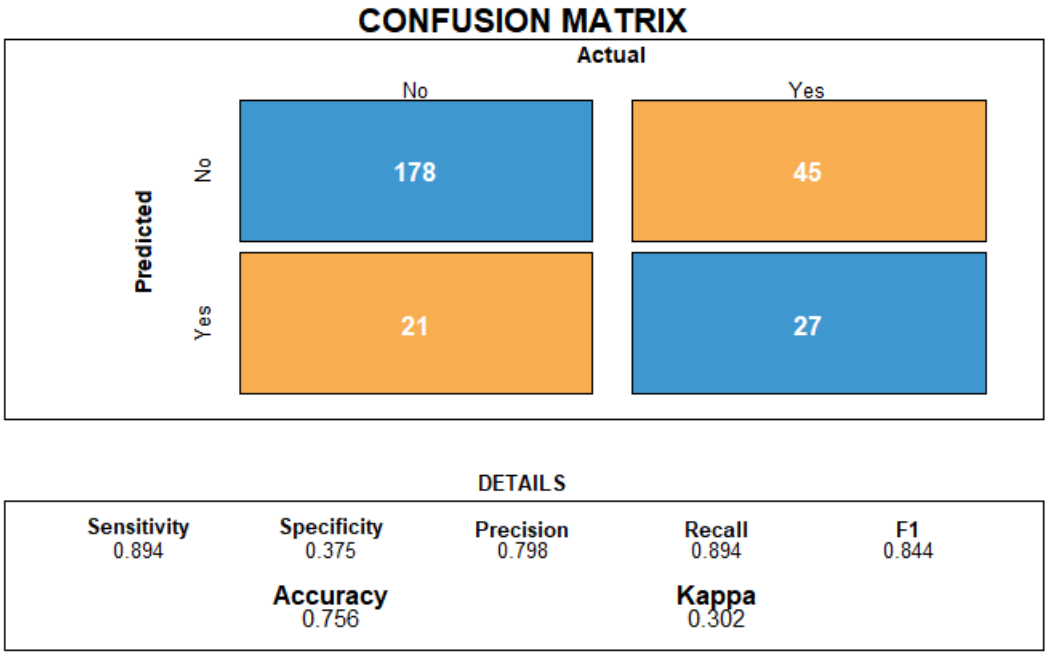
2 Zadatak 2

Prije kreiranja modela uradili smo faktorizaciju kategoričkih varijabli i podjeli skup podataka na trening i testni podskup. Zbog broja instanci u skupu, podjelu smo uradili tako da se 85% podataka nalazi u trening podskupu, a 15% u testnom podskupu.

2.1 Izgradnja modela klasifikacije

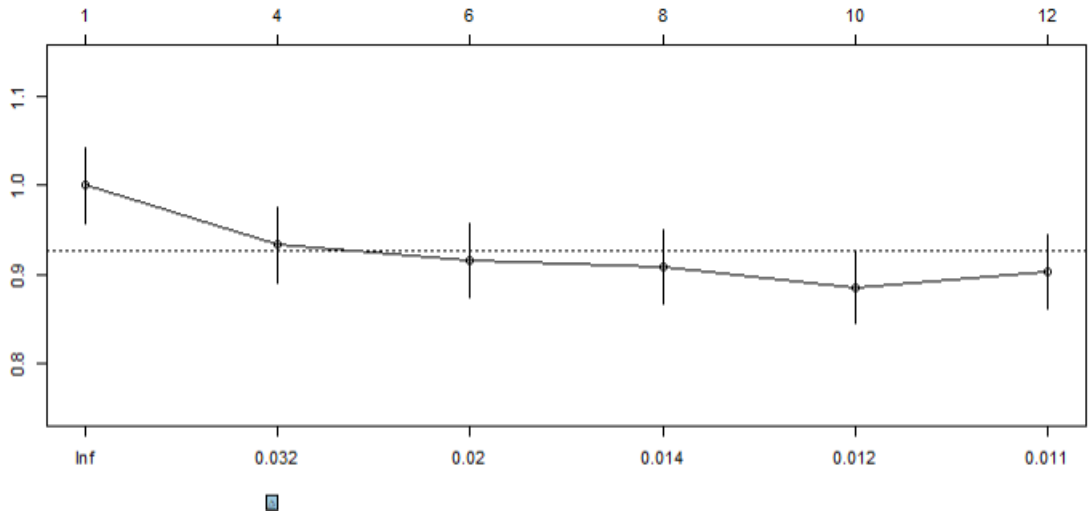
*Napomena: Vrijednosti Sensivity i Specifity su zamijenjene.

2.1.1 Drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit



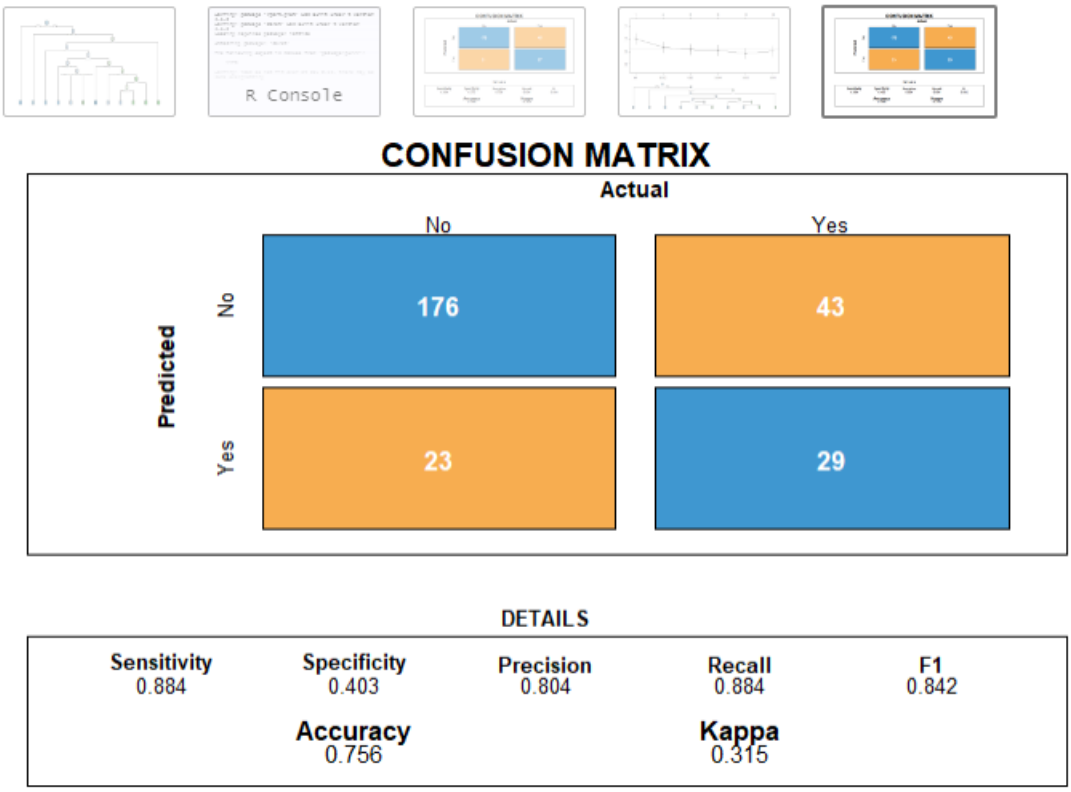
Slika 17: Konfuzijska matrica za drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit

Kao što možemo vidjeti, Sensitivity je niska što znači da klasifikator ne klasifikuje ispravno stvarno pozitivne instance.



Slika 18: Roc kriva za drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit

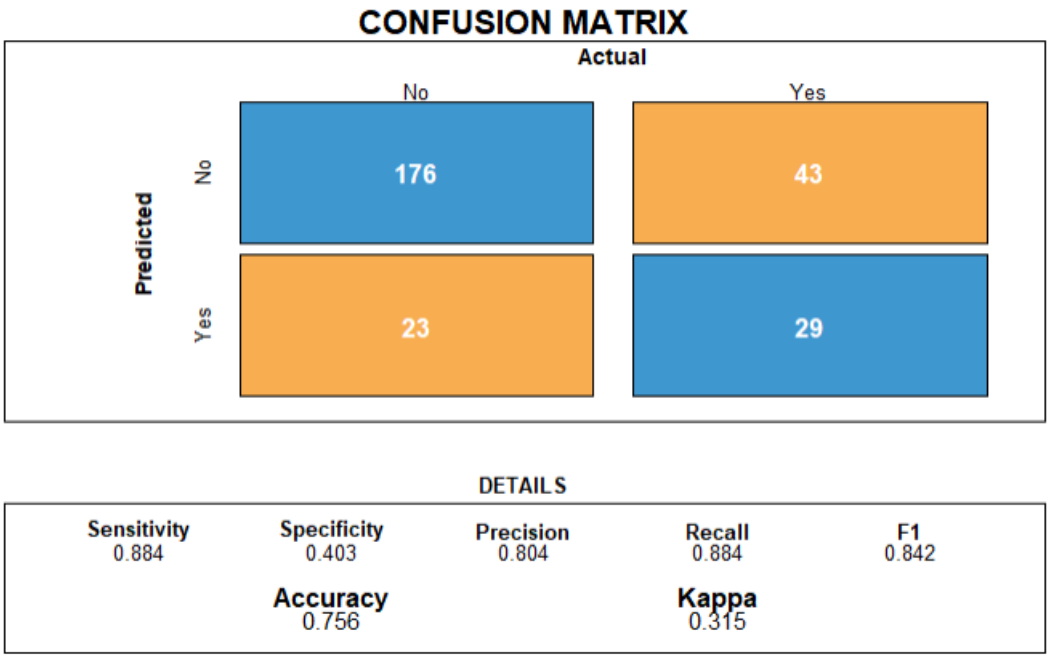
Na ovaj model smo primijeili čišćenje stabla i dobili smo bolje rezultate za Sensivity.



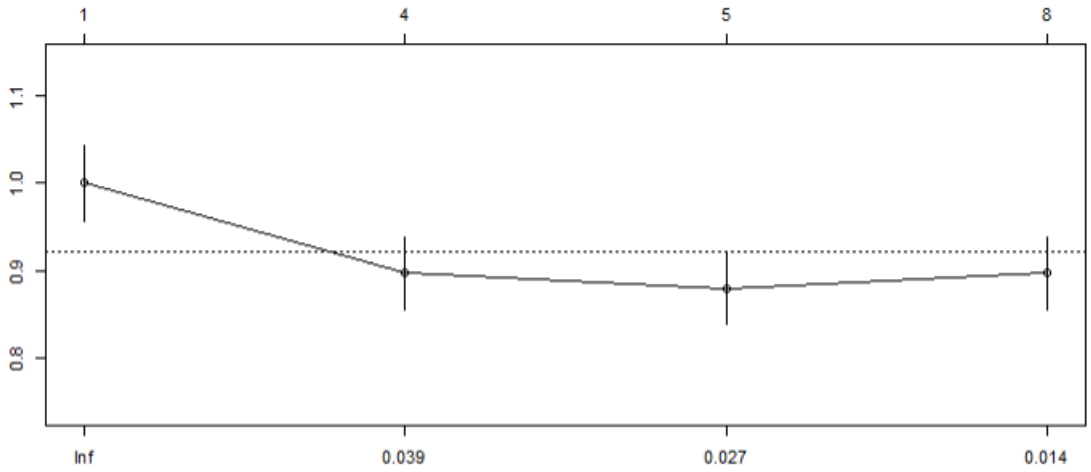
Slika 19: Konfuzijska matrica za drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit

2.1.2 Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks

Za model koji koristi gini indeks smo dobili slične rezultate kao kod modela koji koristi informacijsku dobit.

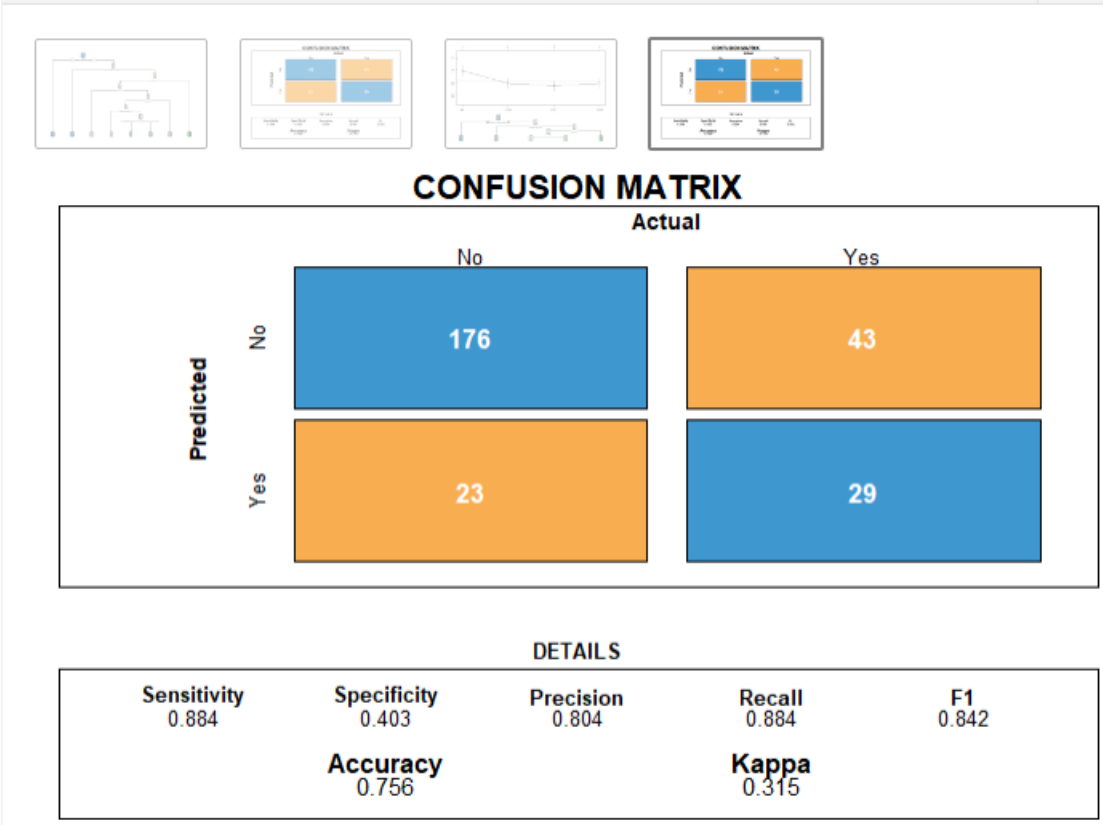


Slika 20: Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks



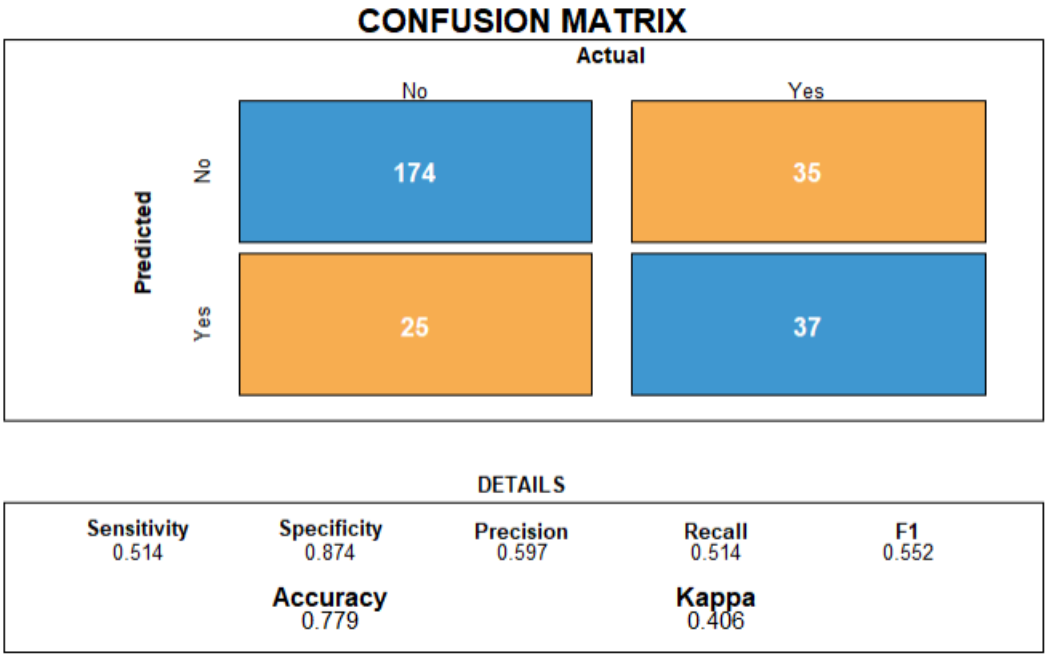
Slika 21: Roc kriva za drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks

Čišćenje stabla kod modela koji koristi gini indeks nije poboljšalo rezultate.



Slika 22: Konfuzijska matrica za drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks

2.1.3 C5.0 model klasifikacije

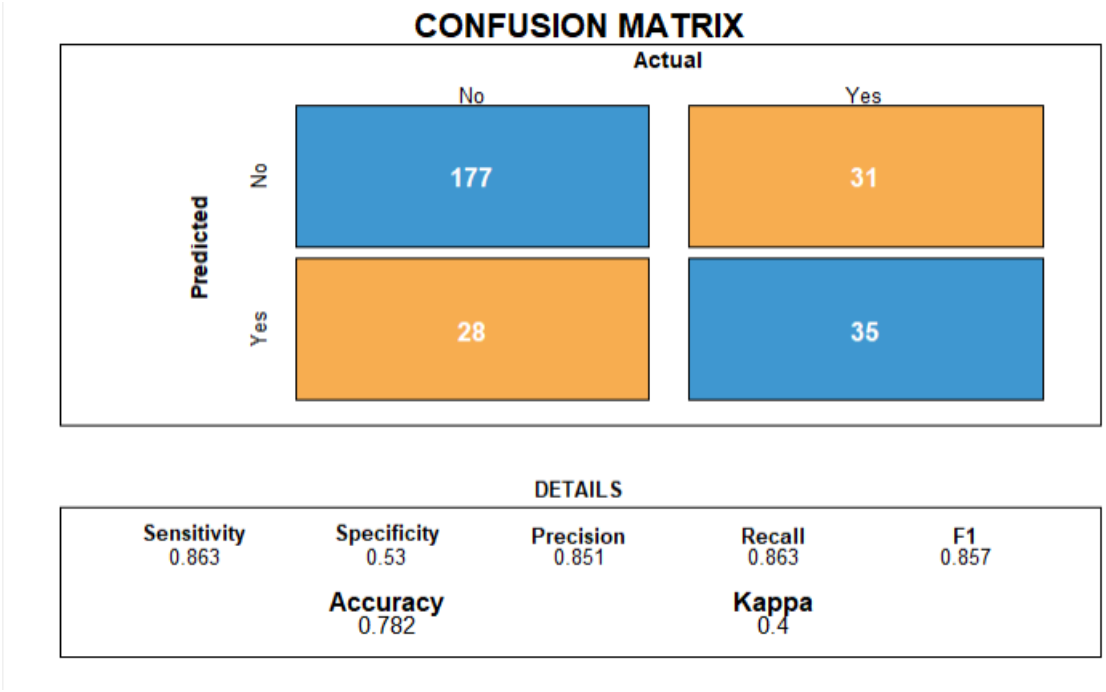


Slika 23: C5.0 model klasifikacije

2.2 Predikcijski modeli sa metodom holdouta

2.2.1 Drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit sa metodom holdouta

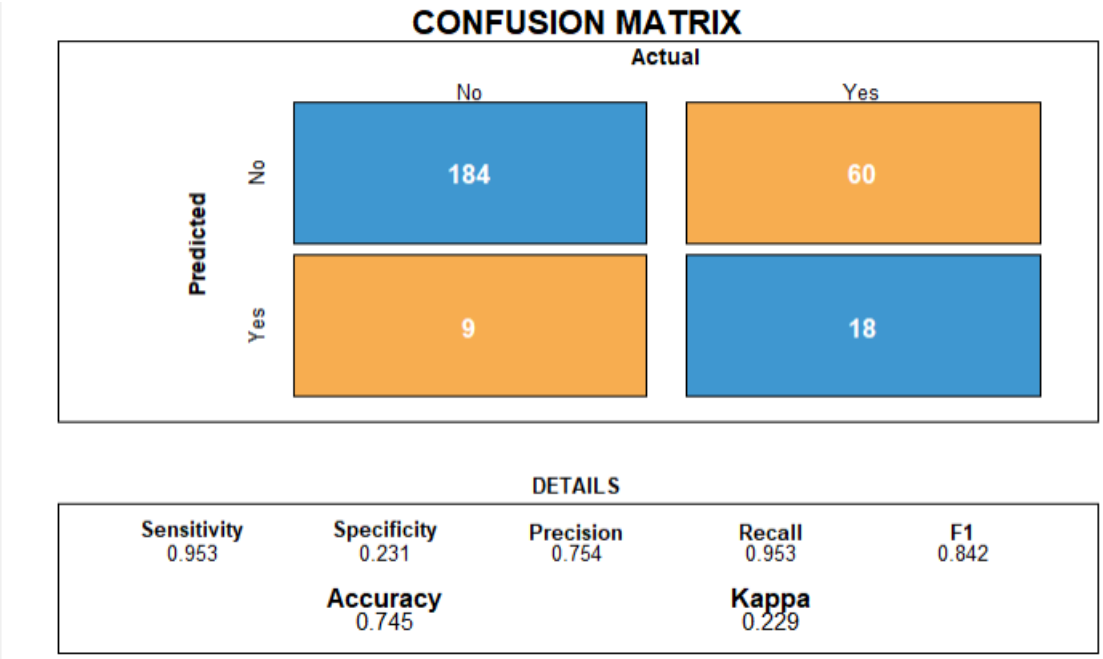
Metoda holdouta pokazuje bolju vrijednosti svih parametara osim parametra Sensivity.



Slika 24: Konfuzijska matrica za drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit sa metodom holdouta

2.2.2 Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks sa metodom holdouta

Metodom holdouta smo dobili lošiji Sensivity za model koji koristi gini indeks.



Slika 25: Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks sa metodom holdouta

2.2.3 C5.0 model klasifikacije sa metodom holdouta

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  484 126
       Yes  50  62

               Accuracy : 0.7562
                 95% CI : (0.7232, 0.7871)
    No Information Rate : 0.7396
    P-Value [Acc > NIR] : 0.1648

                  Kappa : 0.2717

 Mcnemar's Test P-Value : 1.574e-08

            Sensitivity : 0.32979
            Specificity : 0.90637
         Pos Pred Value : 0.55357
         Neg Pred Value : 0.79344
             Prevalence : 0.26039
         Detection Rate : 0.08587
   Detection Prevalence : 0.15512
      Balanced Accuracy : 0.61808

       'Positive' Class : Yes
```

Slika 26: C5.0 model klasifikacije sa metodom holdouta

2.3 Predikcijski modeli sa metodom k-fold unakrsne validacije

2.3.1 Drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit sa metodom k-fold unakrsne validacije

Za k-fold unakrsnu validaciju smo napravili funkciju koju smo pozivali sa vrijednostima parametra k = 5 i k = 10. Parametar k pokazuje koliko puta će se ponoviti klasifikacija. Za obje vrijednosti parametara dobili smo slične rezultate.

```
10-fold validacija
Najveća tačnost: 0.8166667 , fold: 2, najveća kappa: 0.441906 , fold: 6
Najmanja tačnost: 0.7333333 , fold: 5, najmanja kappa: 0.2270992 , fold: 9
Srednja tačnost: 0.7815838, srednja kappa: 0.3214565

5-fold validacija
Najveća tačnost: 0.8199446 , fold: 1, najveća kappa: 0.4414558 , fold: 1
Najmanja tačnost: 0.7479224 , fold: 2, najmanja kappa: 0.2521456 , fold: 4
Srednja tačnost: 0.781602, srednja kappa: 0.3324738
```

Slika 27: Konfuzijska matrica za drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit sa metodom k-fold unakrsne validacije

2.3.2 Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks sa metodom k-fold unakrsne validacije

I za model koji koristi gini indeks smo dobili slične vrijednosti srednje tačnosti za vrijednosti parametara k=5 i k=10.

```
10-fold validacija
Najveća tačnost: 0.8287293 , fold: 4, najveća kappa: 0.4925387 , fold: 4
Najmanja tačnost: 0.7222222 , fold: 6, najmanja kappa: 0.1680532 , fold: 6
Srednja tačnost: 0.7815316, srednja kappa: 0.3342524

5-fold validacija
Najveća tačnost: 0.800554 , fold: 5, najveća kappa: 0.3701047 , fold: 5
Najmanja tačnost: 0.7534626 , fold: 2, najmanja kappa: 0.2905983 , fold: 3
Srednja tačnost: 0.7815897, srednja kappa: 0.333075
```

Slika 28: Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks sa metodom k-fold unakrsne validacije

2.3.3 C5.0 model klasifikacije sa metodom k-fold unakrsne validacije

Kod C5.0 modela smo dobili bolju srednju tačnost kappe za vrijednost parametra k = 10.

```
10-fold validacija
Najveća tačnost: 0.8277778 , fold: 3, najveća kappa: 0.5735249 , fold: 3
Najmanja tačnost: 0.7333333 , fold: 2, najmanja kappa: 0.2785571 , fold: 2
Srednja tačnost: 0.7788091, srednja kappa: 0.4160015

5-fold validacija
Najveća tačnost: 0.7950139 , fold: 1, najveća kappa: 0.4490596 , fold: 1
Najmanja tačnost: 0.7229917 , fold: 5, najmanja kappa: 0.185556 , fold: 2
Srednja tačnost: 0.7633102, srednja kappa: 0.3356188
```

Slika 29: C5.0 model klasifikacije sa metodom k-fold unakrsne validacije

2.4 Predikcijski modeli sa metodom k-fold bootrstaping validacije

2.4.1 Drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit sa metodom k-fold bootrstaping validacije

Kao i kod k-fold unakrsne validacije i kod k-fold bootstraping validacije smo pozivali funkciju za različite vrijednosti parametra k.

```
10-fold bootstrap
Najveća tačnost: 0.8277778 , fold: 3, najveća kappa: 0.4149196 , fold: 8
Najmanja tačnost: 0.7444444 , fold: 4, najmanja kappa: 0.200237 , fold: 9
Srednja tačnost: 0.7872222, srednja kappa: 0.3165408

5-fold bootstrap
Najveća tačnost: 0.7833333 , fold: 2, najveća kappa: 0.3129771 , fold: 4
Najmanja tačnost: 0.7611111 , fold: 5, najmanja kappa: 0.2271978 , fold: 5
Srednja tačnost: 0.7705556, srednja kappa: 0.282632
```

Slika 30: Konfuzijska matrica za drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit sa metodom k-fold bootrstaping validacije

2.4.2 Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks sa metodom k-fold bootrstaping validacije

```
10-fold bootstrap
Najveća tačnost: 0.8055556 , fold: 1, najveća kappa: 0.3697479 , fold: 4
Najmanja tačnost: 0.7388889 , fold: 3, najmanja kappa: 0.2634762 , fold: 5
Srednja tačnost: 0.775, srednja kappa: 0.3176853

5-fold bootstrap
Najveća tačnost: 0.8055556 , fold: 3, najveća kappa: 0.3435791 , fold: 3
Najmanja tačnost: 0.7611111 , fold: 5, najmanja kappa: 0.2461832 , fold: 2
Srednja tačnost: 0.7805556, srednja kappa: 0.3059102
```

Slika 31: Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks sa metodom k-fold bootrstaping validacije

2.4.3 C5.0 model klasifikacije sa metodom k-fold bootrstaping validacije

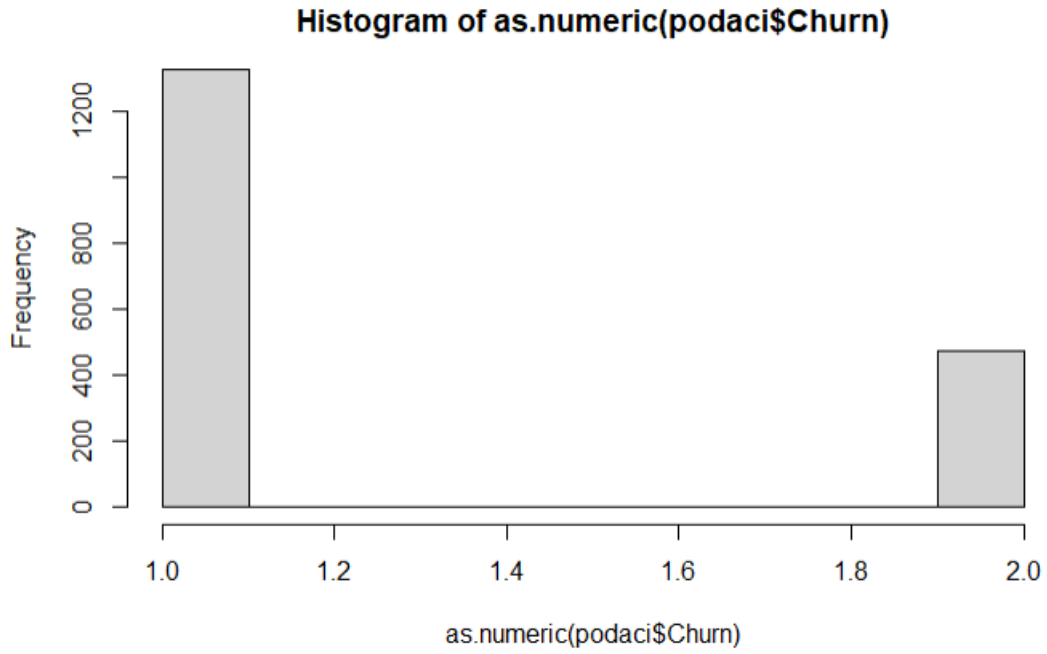
```
10-fold bootstrap
Najveća tačnost: 0.8444444 , fold: 6, najveća kappa: 0.542068 , fold: 6
Najmanja tačnost: 0.7777778 , fold: 1, najmanja kappa: 0.3939394 , fold: 3
Srednja tačnost: 0.8005556, srednja kappa: 0.471508

5-fold bootstrap
Najveća tačnost: 0.8416667 , fold: 4, najveća kappa: 0.5823836 , fold: 4
Najmanja tačnost: 0.7861111 , fold: 3, najmanja kappa: 0.4340547 , fold: 3
Srednja tačnost: 0.8105556, srednja kappa: 0.497575
```

Slika 32: C5.0 model klasifikacije sa metodom k-fold bootrstaping validacije

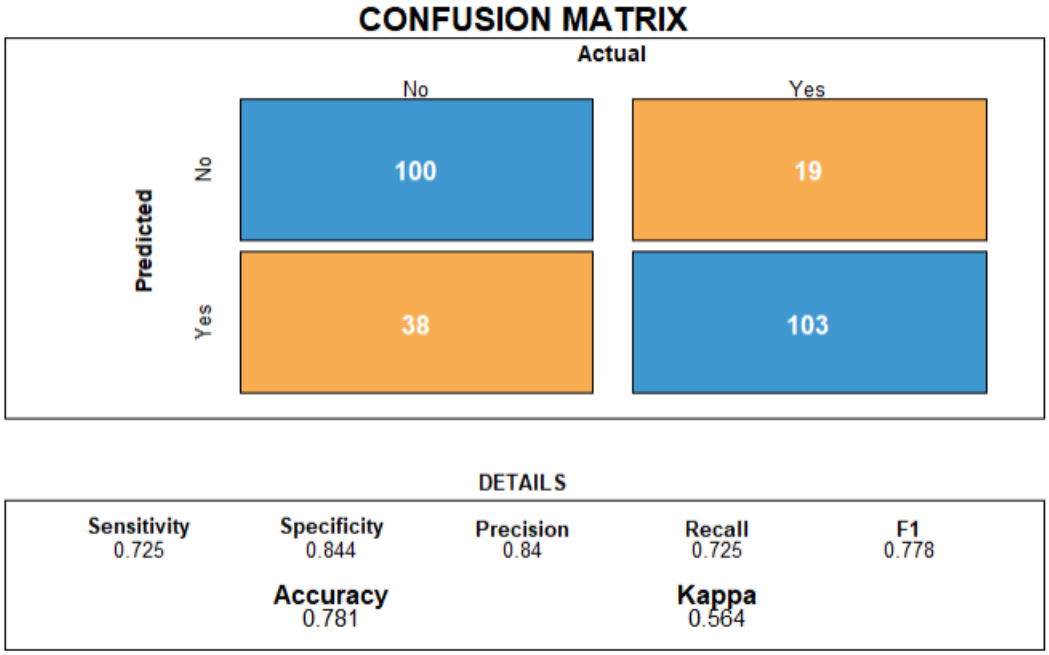
2.5 Balansiranje podataka

Na histogramu možemo uočiti da u datasetu postoji više instanci koje imaju No vrijednost za varijablu Churn. Stoga je potrebno izvršiti balansiranje dataseta. Za vrijednost parametra N smo uzeli 2600 kako bi se izbalansirao set.



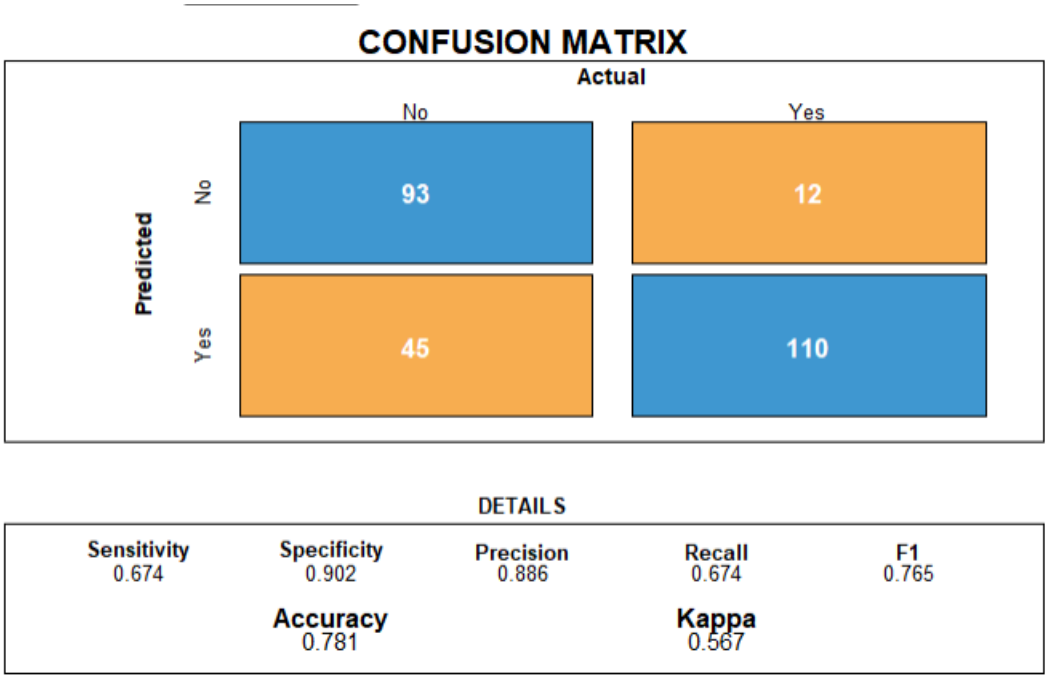
Slika 33: Histogram prije balansiranja

2.5.1 Drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit nakon balansiranja podataka



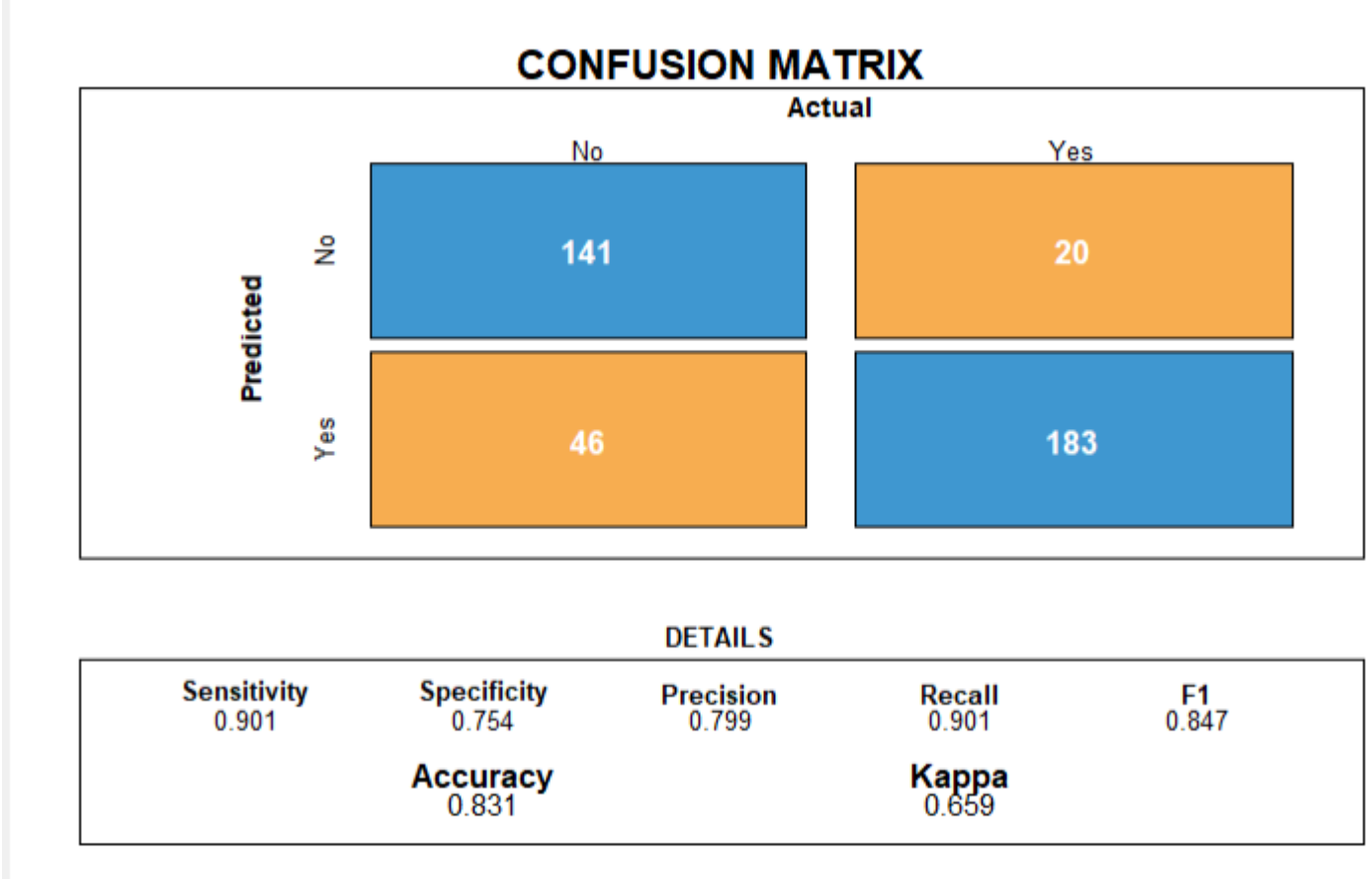
Slika 34: Drvo odlučivanja koje kao mjeru atributa selekcije koristi informacijsku dobit nakon balansiranja podataka

2.5.2 Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks nakon balansiranja podataka



Slika 35: Drvo odlučivanja koje kao mjeru atributa selekcije koristi gini indeks nakon balansiranja podataka

2.5.3 C5.0 model nakon balansiranja



Slika 36: C5.0 model nakon balansiranja

Za sve modele smo dobili bolje vrijednosti parametara nakon treniranja nad balansiranim podacima.

2.6 Ansambl tehnike za unaprjeđenje tačnosti klasifikacije

Za parametar nbagg smo uzeli vrijednost 50 što znači da će se kreirati 50 modela, a za parametar minsplit smo uzeli vrijednost 4 što označava minimalni broj instanci za granjanje.

2.6.1 Bagging model

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  127   3
       Yes  11 119

               Accuracy : 0.9462
                 95% CI : (0.9113, 0.9703)
    No Information Rate : 0.5308
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8923

 Mcnemar's Test P-Value : 0.06137

            Sensitivity : 0.9203
            Specificity : 0.9754
         Pos Pred Value : 0.9769
         Neg Pred Value : 0.9154
             Prevalence : 0.5308
         Detection Rate : 0.4885
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.9478

       'Positive' Class : No
```

Slika 37: Bagging model

2.6.2 Boosting model uz korišenje AdaBoost metode

Na slici su prikazani parametri za boosting model koji koristi AdaBoost metodu. Vrijednost parametra mfinal je 50.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  110  20
       Yes  28 102

               Accuracy : 0.8154
                 95% CI : (0.7628, 0.8606)
    No Information Rate : 0.5308
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6308

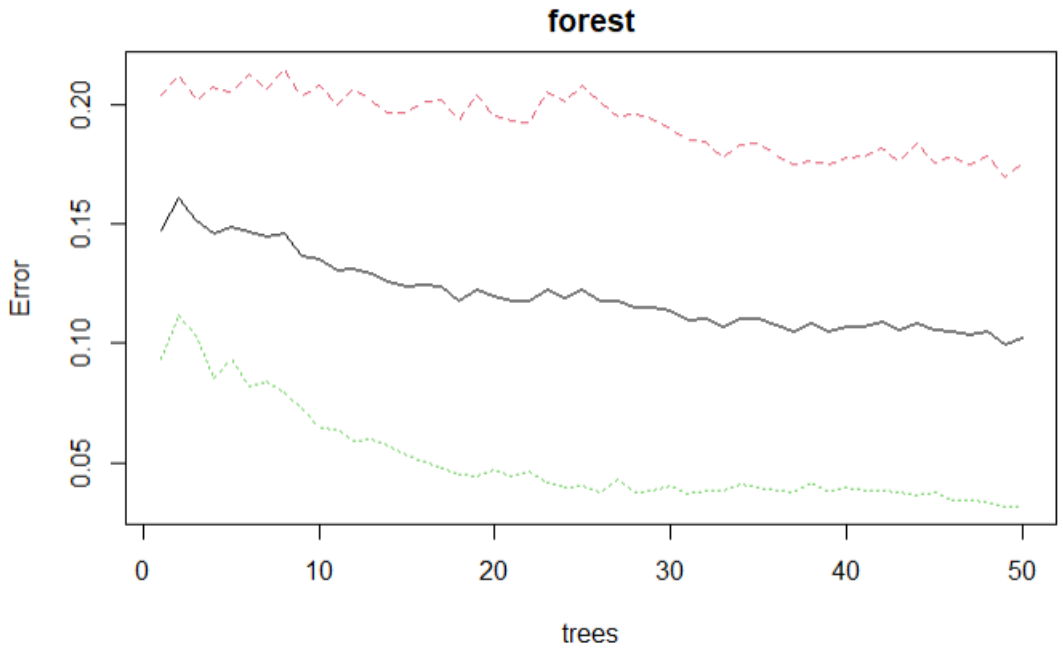
 Mcnemar's Test P-Value : 0.3123

            Sensitivity : 0.7971
            Specificity : 0.8361
         Pos Pred Value : 0.8462
         Neg Pred Value : 0.7846
             Prevalence : 0.5308
         Detection Rate : 0.4231
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.8166
```

Slika 38: Boosting model uz korišenje AdaBoost metode

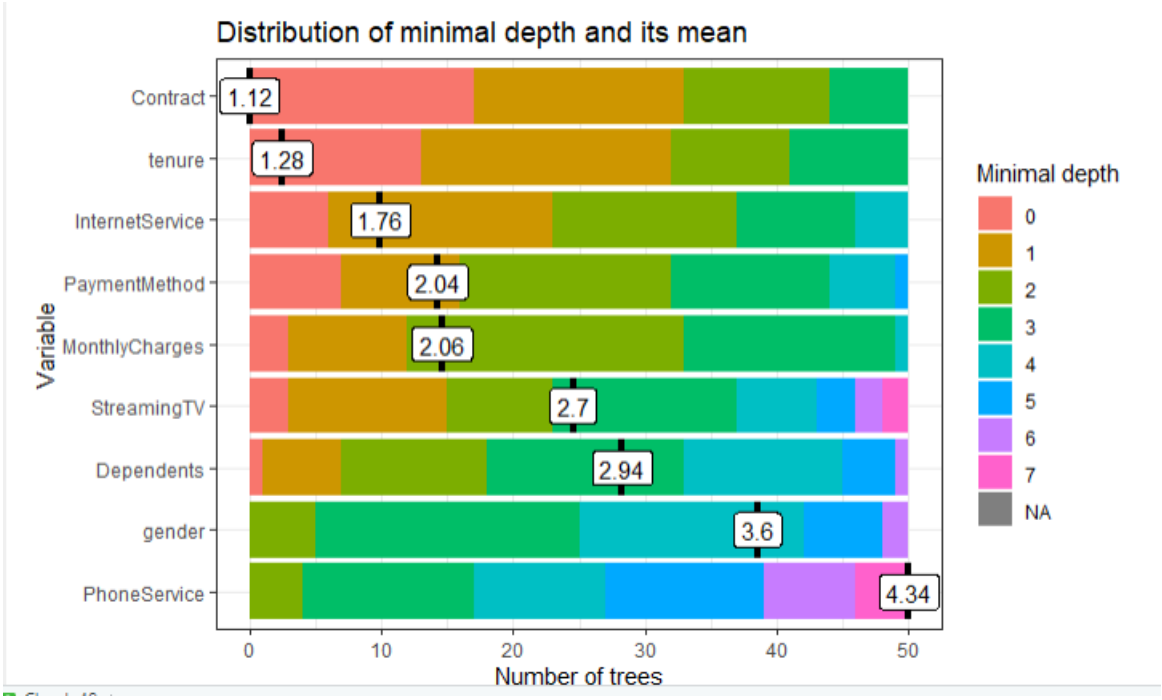
2.6.3 Random forest model

Za ovaj model smo kkoristi vrijednost parametra mtree = 50.



Slika 39: Prikaz ovisnosti greške klasifikacije o broju modela

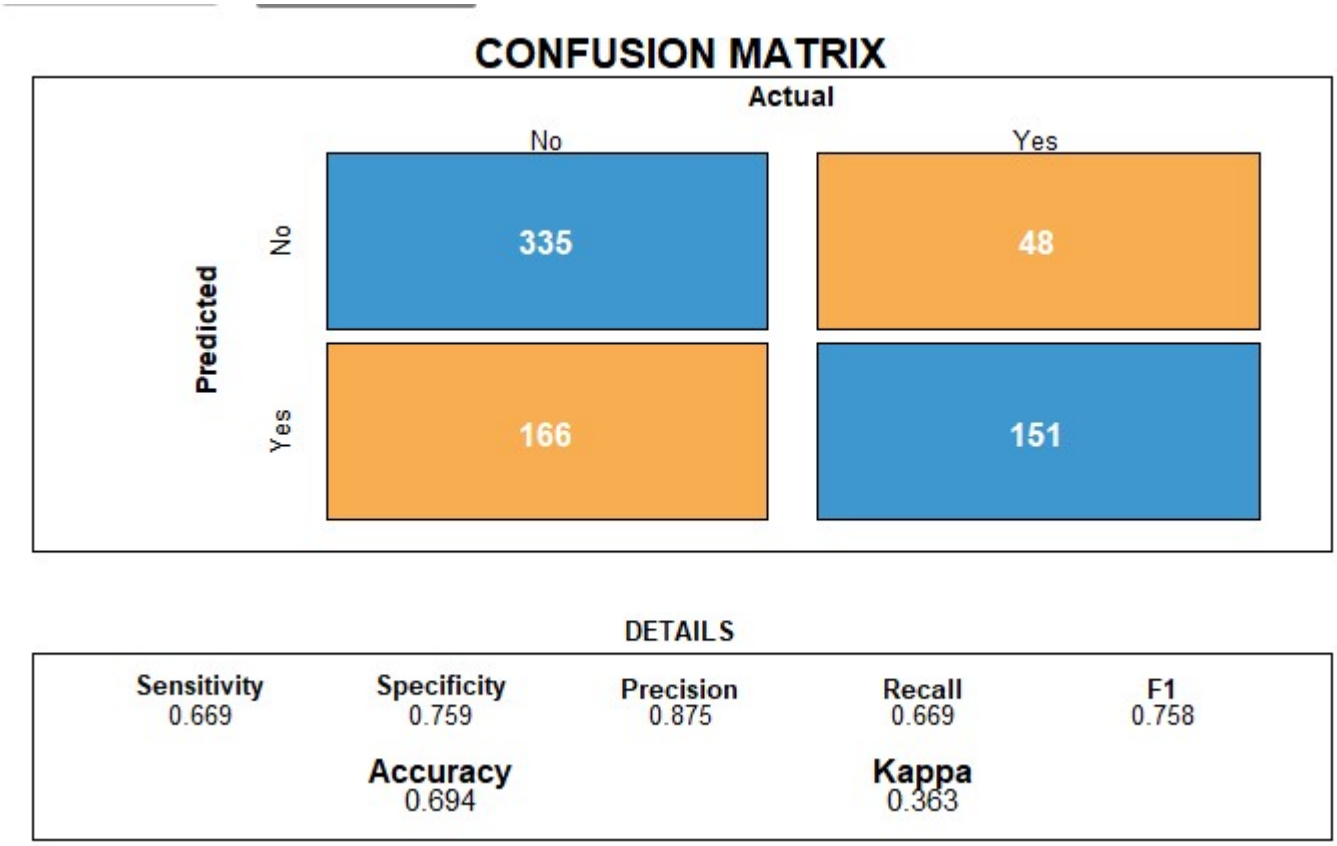
Na osnovu stepena značaja svih atributa možemo zaključiti da su Contract i tenure najvažniji atributi.



Slika 40: Prikaz stpena značaja svih atributa u random forest modelu

3 Zadatak 3

Najbolje rezultate u drugom zadatku nam je dao bagging model. Međutim nad testnim skupom bolje rezultate nam je dao C5.0 pa smo prikazali njegove rezultate.



Slika 41: Primjena C5.0 modela nad testnim skupom

4 Zadatak 4


```
Call:
OneR.data.frame(x = podaci_train)

Rules:
If Contract = Month-to-month then Churn = Yes
If Contract = One year       then Churn = No
If Contract = Two year       then Churn = No

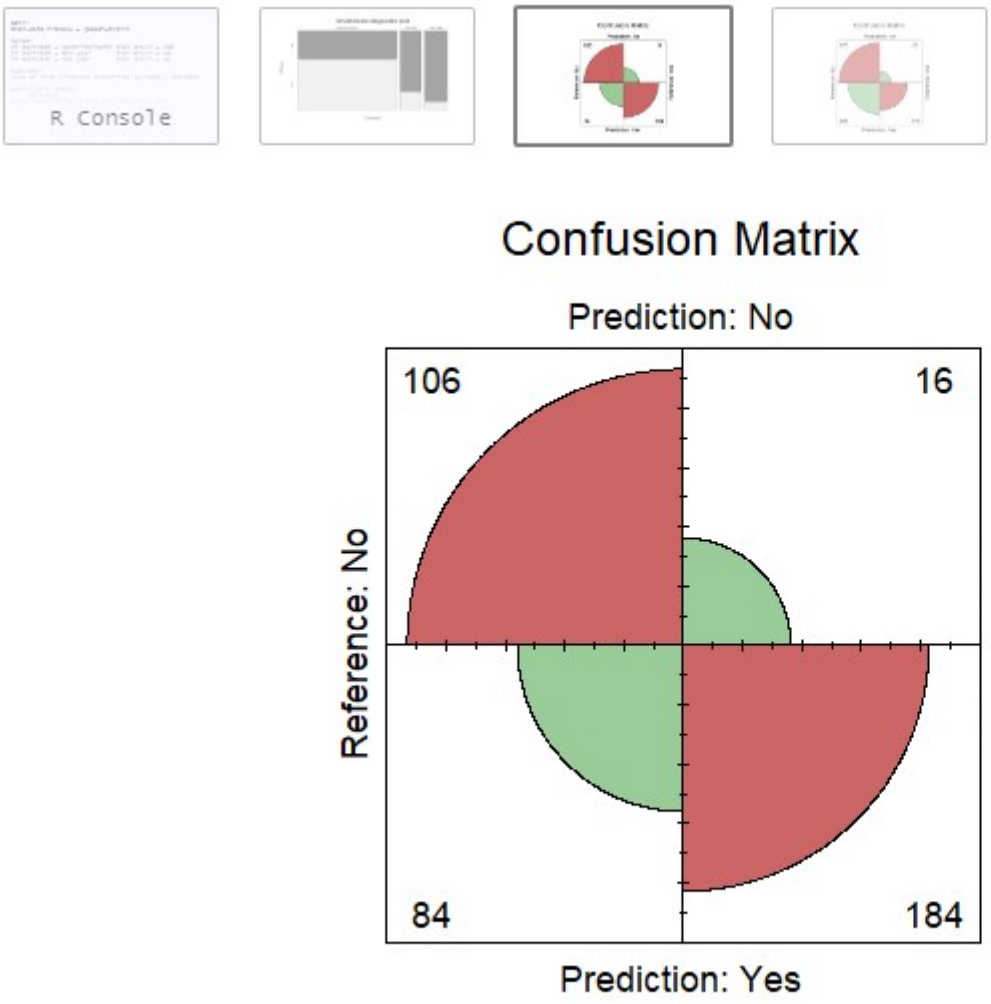
Accuracy:
1540 of 2210 instances classified correctly (69.68%)

Contingency table:
     Contract
Churn Month-to-month One year Two year  Sum
  No              572    * 259    * 325 1156
  Yes           * 956       69       29 1054
  Sum            1528      328      354 2210
---
Maximum in each column: '*'

Pearson's Chi-squared test:
X-squared = 450.32, df = 2, p-value < 2.2e-16

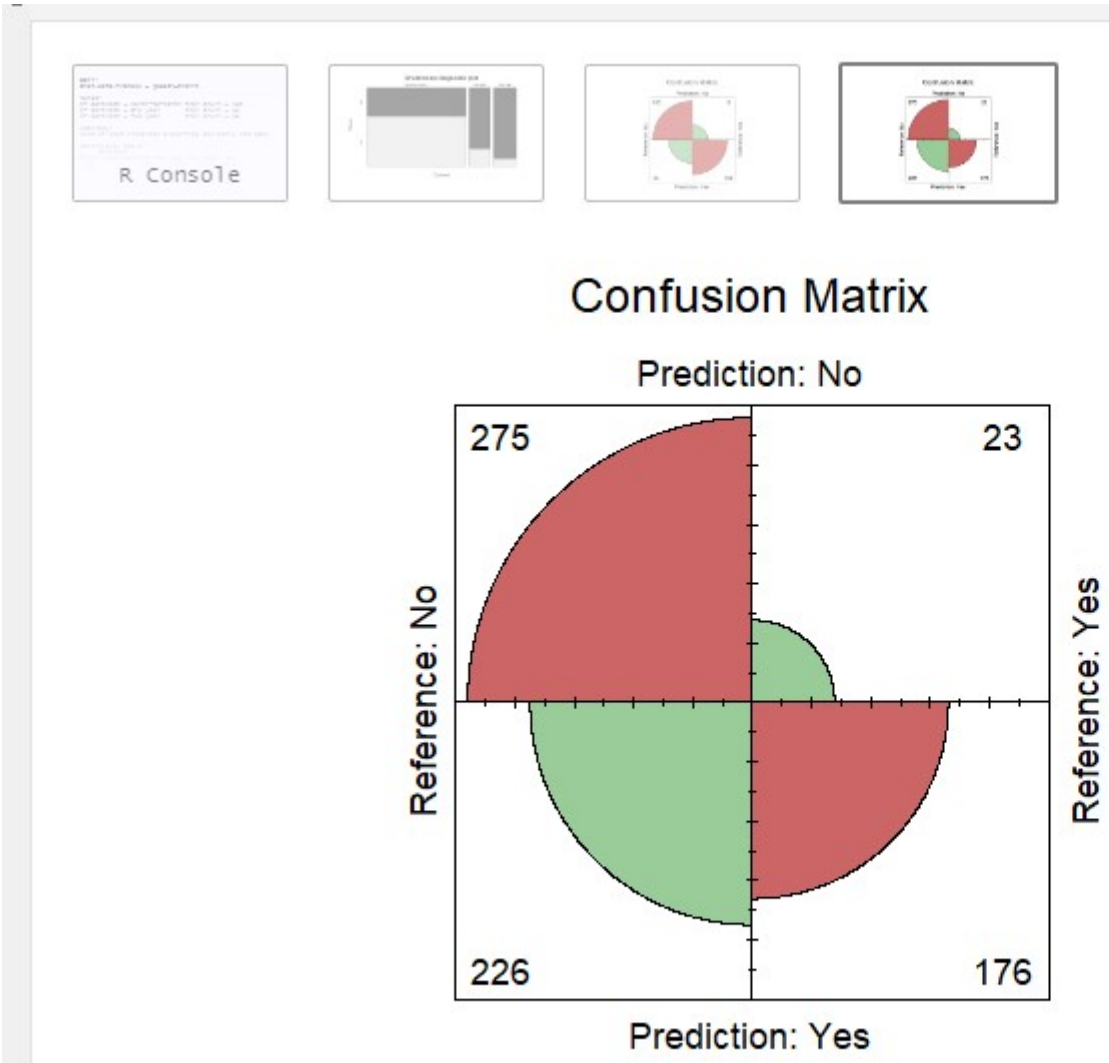
Pravilo: IF Month-to-month THEN Yes , pokrivenost: 0.6914027 , tačnost: 0.6256545
Pravilo: IF One year THEN No , pokrivenost: 0.1484163 , tačnost: 0.7896341
Pravilo: IF Two year THEN No , pokrivenost: 0.160181 , tačnost: 0.9180791
```

Slika 42: Model OneR klasifikatora



1538

Slika 43: Prikaz konfuzijske matrice sa proporcijama tačnih i netačnih klasifikacija



Slika 44: Prikaz konfuzijske matrice sa proporcijama tačnih i netačnih klasifikacija za testni skup