Univerzitet u Sarajevu
Elektrotehnički fakultet
Mašinsko učenje

IZGRADNJA I EVALUACIJA KLASIFIKACIJSKOG MODELA

ZADAĆA BR. 1

Studenti: Amina Alagić, Nejra Rovčanin, Ema Rudalija
Datum: 28.11.2021.
Predmetni profesor: Prof.dr. Dženana Đonko
Predmetni asistent: Ehlimana Krupalija

S A D R Ž A J

1. ISTRAŽIVANJE PODATAKA

U ovoj zadaći korišten je set podataka *customer data train.csv*. Na osnovu pomenutog seta podataka izgrađen je klasifikacijski model drveta odlučivanja koji utvrđuje da li će mušterija kompanije mrežnih usluga otkazati pretplatu.

Skup podataka se sastoji od ukupno 14 kolona, a prisutno je 2000 instanci.

	gender	Dependents	tenure	PhoneService	MultipleLines	InternetService	StreamingTV	StreamingMovies	Contract	PaymentMethod	MonthlyCharges	TotalCharges	DailyCharges	Churn
1	Male	No	8	Yes	No	*NA*	Yes	Yes	*NA*	*NA*	*NA*	832.35	*NA*	Yes
2	Female	*NA*	8	Yes	*NA*	Fiber optic	No	No	Month-to-month	Credit card (automatic)	*NA*	548.90	*NA*	No
3	*NA*	No	21	Yes	No	Fiber optic	Yes	Yes	One year	Mailed check	104.55	2239.40	20.91	No
4	*NA*	*NA*	1	No	No phone service	DSL	No	No	Month-to-month	Mailed check	35.90	35.90	7.18	No
5	*NA*	No	*NA*	*NA*	*NA*	Fiber optic	No	Yes	Month-to-month	Electronic check	81.10	81.10	16.22	Yes
6	Male	No	69	Yes	No	No	*NA*	*NA*	Two year	Bank transfer (automatic)	19.30	1447.90	3.86	No
7	Female	Yes	9	Yes	No	DSL	Yes	Yes	Month-to-month	Credit card (automatic)	69.40	571.45	13.88	No
8	Male	No	28	Yes	No	No	No internet service	No internet service	One year	*NA*	18.25	*NA*	3.65	No
9	Male	Yes	60	Yes	No	DSL	Yes	Yes	Two year	Bank transfer (automatic)	76.95	4543.95	15.39	No
10	Female	No	5	*NA*	No	Fiber optic	No	Yes	Month-to-month	Electronic check	80.00	412.50	16.00	Yes
11	Male	No	68	*NA*	*NA*	DSL	No	Yes	One year	Credit card (automatic)	76.90	5023.00	15.38	*NA*
12	Female	No	72	*NA*	Yes	Fiber optic	Yes	Yes	Two year	Mailed check	109.75	8075.35	21.95	No
13	*NA*	No	71	Yes	No	Fiber optic	Yes	Yes	Two year	*NA*	99.20	*NA*	19.84	No
14	Male	No	*NA*	*NA*	*NA*	DSL	Yes	*NA*	One year	*NA*	77.40	4155.95	15.48	*NA*
15	Female	No	9	No	No phone service	DSL	Yes	Yes	Month-to-month	Mailed check	58.50	539.85	11.70	Yes
16	Male	No	72	Yes	Yes	DSL	No	Yes	Two year	Bank transfer (automatic)	82.30	5980.55	16.46	*NA*
17	Female	No	5	Yes	No	No	No internet service	No internet service	*NA*	Mailed check	20.15	117.95	4.03	No
18	Female	Yes	27	Yes	Yes	No	*NA*	No internet service	Two year	Electronic check	24.50	761.95	4.90	No
19	Female	*NA*	18	*NA*	No	Fiber optic	No	*NA*	Month-to-month	Electronic check	69.80	*NA*	13.96	No
20	*NA*	No	1	Yes	No	No	No internet service	No internet service	*NA*	Mailed check	20.45	20.45	4.09	No
21	Female	*NA*	*NA*	*NA*	No	No	No internet service	No internet service	Two year	Mailed check	19.40	*NA*	3.88	No
22	Female	No	60	Yes	Yes	Fiber optic	Yes	Yes	Two year	Electronic check	104.70	6333.80	20.94	No
23	Male	Yes	63	Yes	*NA*	No	No internet service	No internet service	Two year	Credit card (automatic)	24.65	1574.50	4.93	No
24	Female	Yes	28	Yes	No	*NA*	No internet service	No internet service	Month-to-month	*NA*	*NA*	543.80	*NA*	No
25	Male	No	72	*NA*	No phone service	DSL	Yes	Yes	Two year	Bank transfer (automatic)	63.70	4464.80	12.74	No
26	Female	No	24	Yes	No	No	No internet service	*NA*	Month-to-month	Mailed check	20.30	459.95	4.06	No
27	*NA*	No	19	Yes	No	DSL	No	No	Month-to-month	Credit card (automatic)	*NA*	1046.50	*NA*	Yes
28	Female	Yes	*NA*	Yes	Yes	No	No internet service	No internet service	Two year	Mailed check	23.85	625.65	4.77	No
29	Male	Yes	*NA*	Yes	No	No	No internet service	No internet service	Month-to-month	Mailed check	19.45	*NA*	3.89	No
30	Female	No	46	Yes	No	Fiber optic	No	*NA*	Month-to-month	Credit card (automatic)	74.80	3548.30	14.96	No

Figure 1 - Prikaz prvih 30 redova data seta

Kao što je vidljivo sa slike iznad u data setu su prisutni:

- Numerički atributi (4 atributa – tenure, MonthlyCharges, TotalCharges, DailyCharges)
- Kategorijski atributi (10 atributa – gender, PhoneService, MultipleLines, InternetService, StreamingTV, StreamingMovies, PaymentMethod, Churn)

Kategorija Churn je posebna kategorija koja predstavlja finalni rezultat za drvo odlučivanja koje je potrebno kreirati.

1.1. UPOZNAVANJE SA SETOM PODATAKA

U cilju boljeg upoznavanja sa setom podataka korišteno je nekoliko metoda. Neke od njih su:

- Osnovne metode deskriptivne statistike
- Metode procjene lokacije i varijabilnosti podataka
- Metode za procjenu korelacije između varijabli

1.1.1. METODE DESKRIPTIVNE STATISTIKE

Unutar primjene metoda deskriptivne statistike uglavnom su razmatrane kategoričke varijable.

Atribut *Gender* ima dvije moguće vrijednosti: Female i Male. Frekventnost ovih varijabli je prikazana na barplotu ispod.

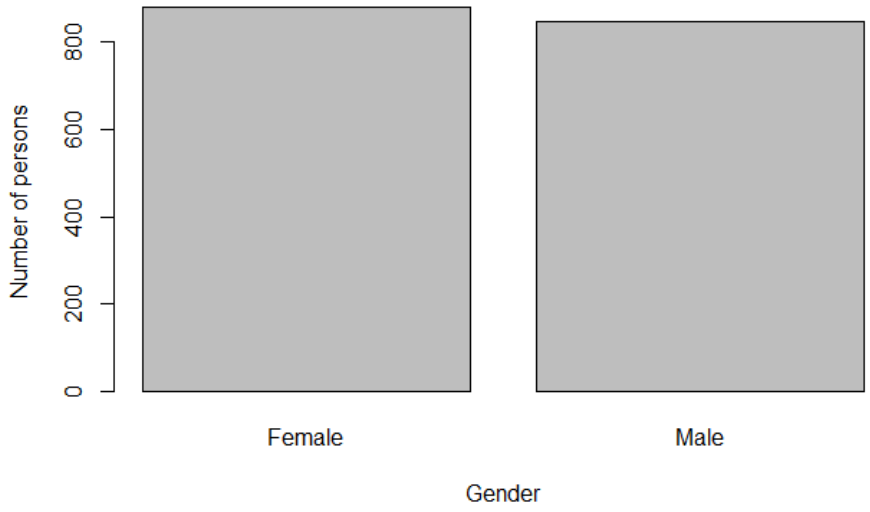


Abbildung 1 - Atribut Gender

Atribut *Dependents* ima tri moguće vrijednosti: Maybe, No i Yes. Frekventnost ovih varijabli je prekazana barplotom ispod.

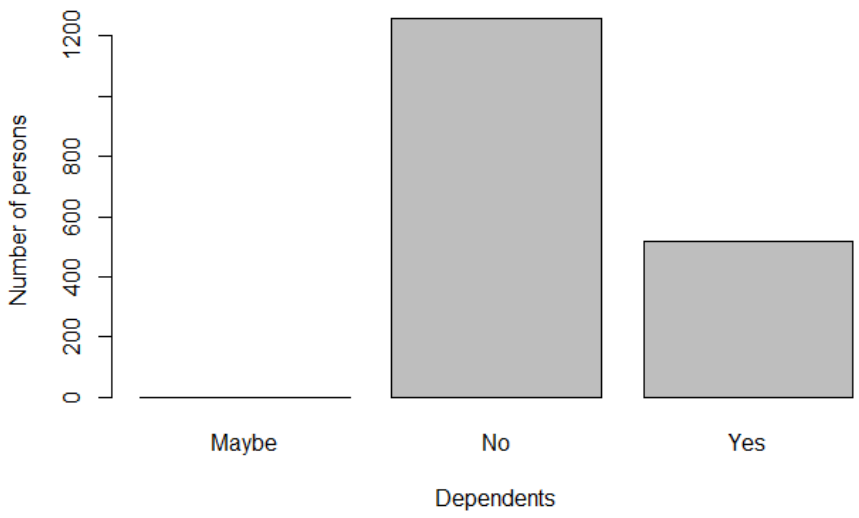


Abbildung 2 - Atribut Dependents

Atribut *PhoneService* ima dvije moguće vrijednosti: No i Yes. Frekventnost ovih varijabli je prekazana barplotom ispod.

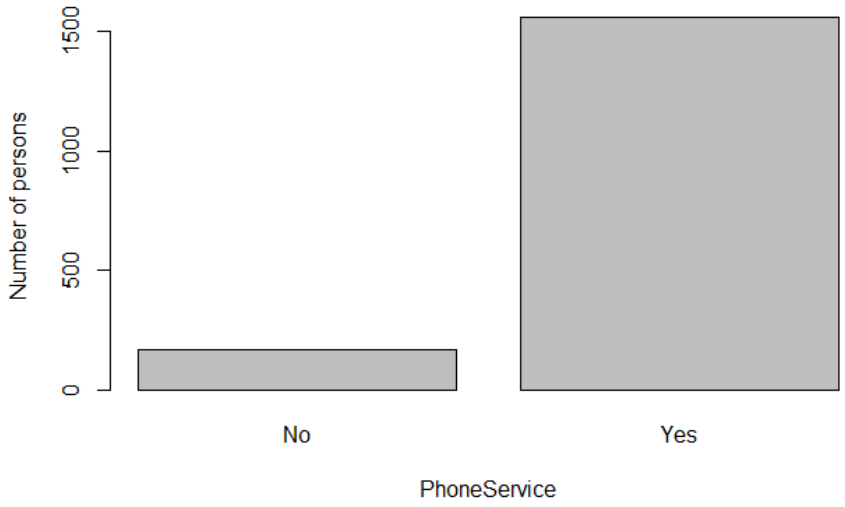


Abbildung 3 - Atribut PhoneService

Atribut *MultipleLines* ima tri moguće vrijednosti: No, No Phone Service i Yes. Frekventnost ovih varijabli je prekazana barplotom ispod.

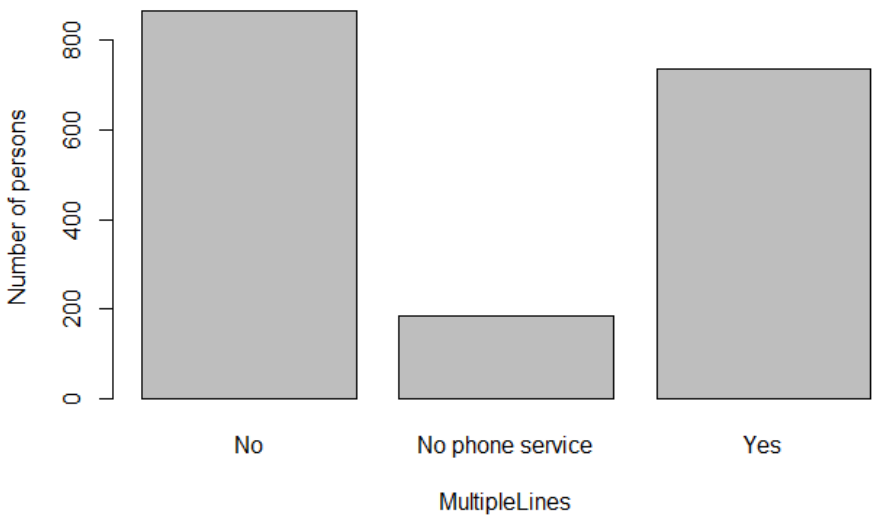


Abbildung 4 - Atribut MultipleLines

Atribut *InternetService* ima tri moguće vrijednosti: DSL, Fiber Optic i No. Frekventnost ovih varijabli je prekazana barplotom ispod.

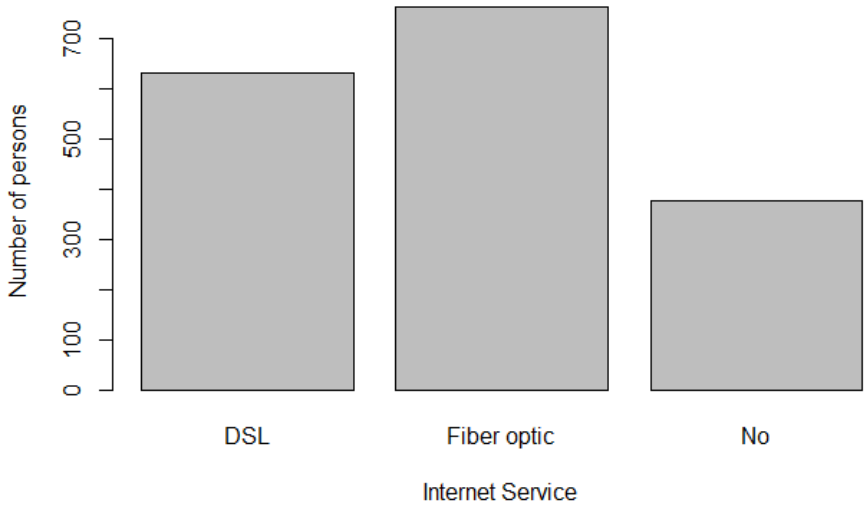


Abbildung 5 - Atribut InternetService

Atribut *StreamingTV* ima tri moguće vrijednosti: No, No Internet Service i Yes. Frekventnost ovih varijabli je prekazana barplotom ispod.

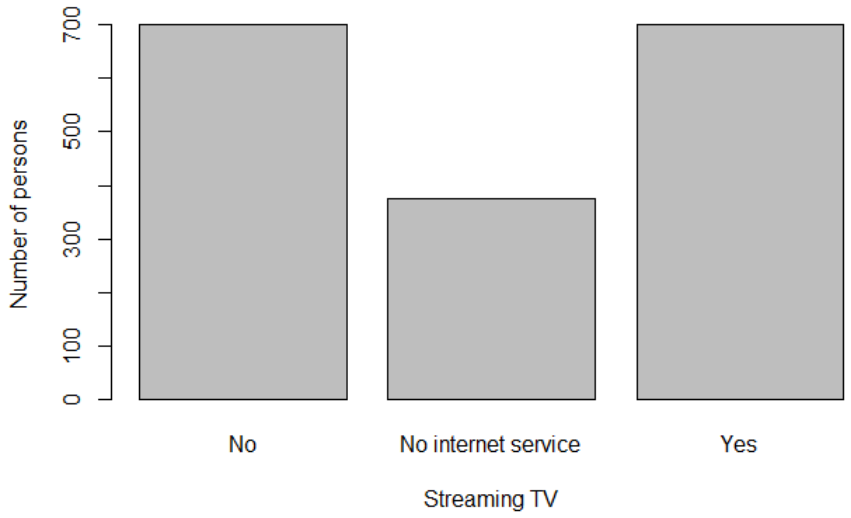


Abbildung 6 - Atribut StreamingTV

Atribut *StreamingMovies* ima tri moguće vrijednosti: No, No Internet Service i Yes. Frekventnost ovih varijabli je prekazana barplotom ispod.

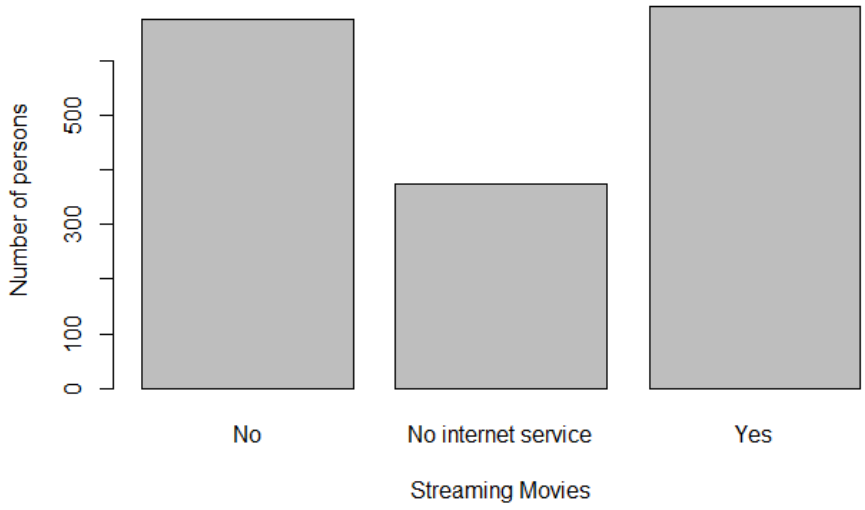


Abbildung 7 - Atribut StreamingMovies

Atribut *Contract* ima tri moguće vrijednosti: Month-to-Month, One year i Two year. Frekventnost ovih varijabli je prekazana barplotom ispod.

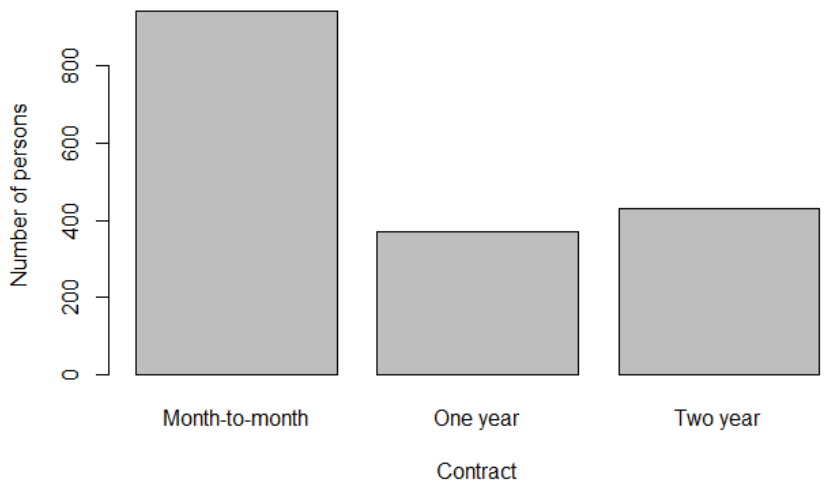


Abbildung 8 - Atribut Contract

Atribut *PaymentMethod* ima pet mogućih vrijednosti: abcd, Bank transfer (automatic), Credit card (automatic), Electronic check i Mailed check. Frekventnost ovih varijabli je prekazana barplotom ispod.

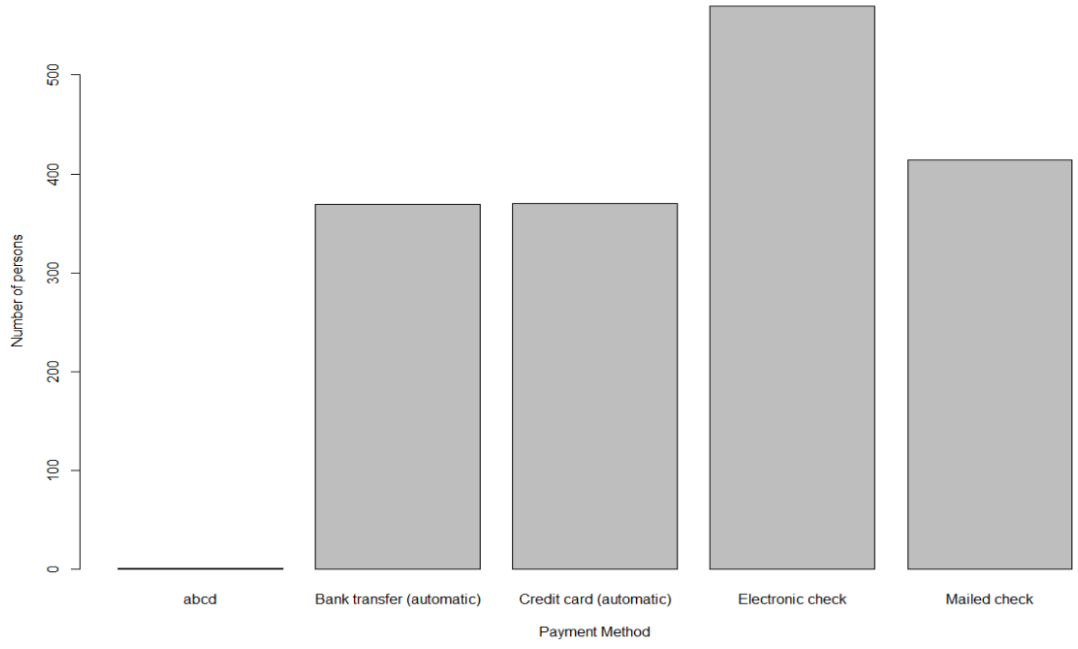


Abbildung 9 - Atribut PaymentMethod

1.1.2. PROCJENA LOKACIJE I VARIJABILNOSTI

Procjene lokacije i varijabilnosti su primjenjene nad numeričkim atributima.

Procjena lokacije uključuje osnovni pregled svih numeričkih atributa: minimalna vrijednost, vrijednost prvog kvartala, median vrijednost, mean vrijednost, vrijednost trećeg kvartala, maksimalna vrijednost i ukupni broj nedostajućih (NA) vrijednosti. Također je i ručno implementirana funkcija za dobivanje trimmed mean vrijednosti za vrijednost paramenra p=444.

```
Pregled osnovnih podataka o numeričkim varijablama

     tenure      MonthlyCharges    TotalCharges      DailyCharges
 Min.   : 0.00   Min.   : -1.22   Min.   :   19.1   Min.   :-0.244
 1st Qu.: 9.00   1st Qu.: 35.00   1st Qu.:  433.4   1st Qu.: 6.940
 Median :29.00   Median : 70.30   Median : 1415.4   Median :14.050
 Mean   :32.52   Mean   : 64.41   Mean   : 2280.6   Mean   :12.855
 3rd Qu.:56.00   3rd Qu.: 89.40   3rd Qu.: 3751.7   3rd Qu.:17.880
 Max.   :72.00   Max.   :118.65   Max.   :10000.0   Max.   :23.730
 NA's   :247     NA's   :256      NA's   :276       NA's   :255

Trimmed mean vrijednost za varijablu tenure (p=4):  30.30636
Trimmed mean vrijednost za varijablu MonthlyCharges (p=4):  67.70152
Trimmed mean vrijednost za varijablu TotalCharges (p=4):  1646.781
Trimmed mean vrijednost za varijablu DailyCharges (p=4):  13.50939
```

Abbildung 10 - Procjena lokacije za numeričke attribute

Da bismo imali ljepši uvid u sve dobivene vrijednosti korištena je funkcija *boxplot* koja iscrtava odgovarajući dijagram za metode procjene lokacije.

Na slici ispod nalazi se boxplot za atribut *Tenure*.

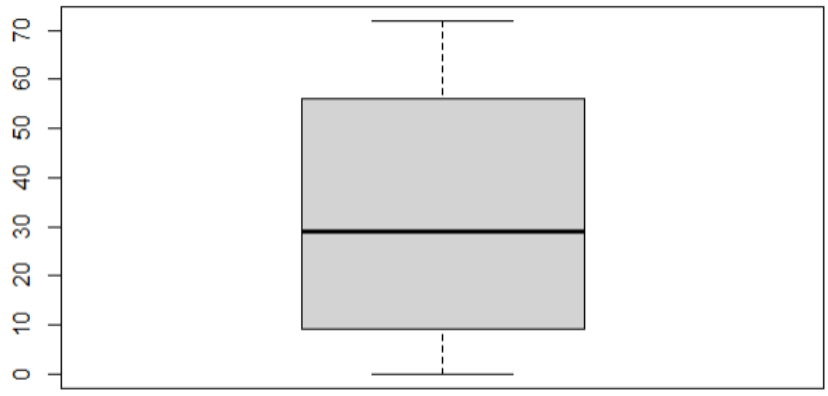


Abbildung 11 - Boxplot za Tenure

Obzirom da je vrijednost median jednaka 29 vidimo da se srednja vrijednost ne nalazi tačno na sredini opsega, već je pomjerena prema dolje. Navedeno važi i za cijeli interkvartalni opseg. Činjenica da se veći broj instanci nalazi između medijana i trećeg kvartala nam govori da postoji veći broj vrijednosti iznad prosjeka.

Situaciju za atribut *MonthlyCharges* možemo vidjeti na sljedećoj slici.

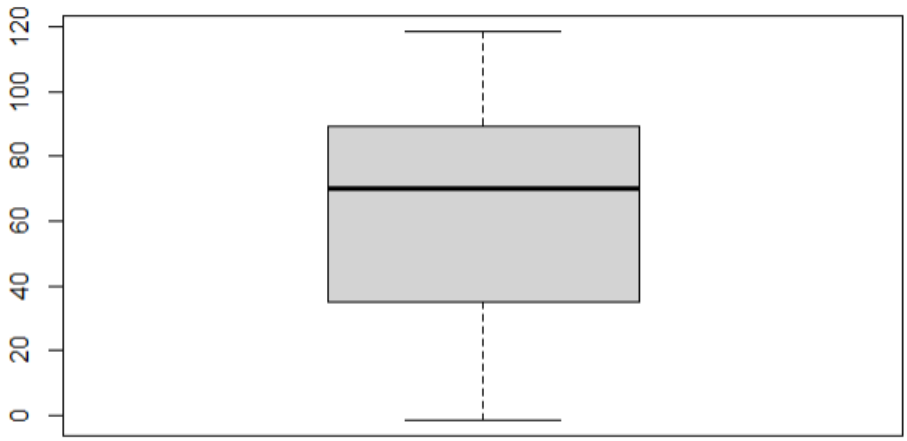


Abbildung 12 - Boxplot za MonthlyCharges

Srednje vrijednost koja je jednaka 70 nalazi se u gornjem dijelu opsega, dok je interkvartalni opseg otprilike na sredini (blago pomjeren ka gore). Za razliku od prethodnog grafika, veći broj instanci se nalazi ispod prosjeka (s obzirom na to da je veliki broj instanci između prvog kvartala i medijana), što znači da su mjesečne naknade za većinu razmatranih osoba ispod 70.

Kada je riječ o atributu *TotalCharges* vidimo da je medijan, ali i cijeli interkvartalni opseg na samom dnu grafika, što znači da postoji izrazita težnja ka manjim opsezima i da su ukupne naknade uglavnom manje od 4000 bez obzira na to što je maksimalna vrijednost totalne naknade 10 000. Dalje, vidimo da su iznosi totalne naknade uglavnom veći od srednje vrijednosti koja iznosi 1415 jer postoji više instanci između medijane i trećeg kvartala u odnosu na broj instanci između medijane i prvog kvartala.

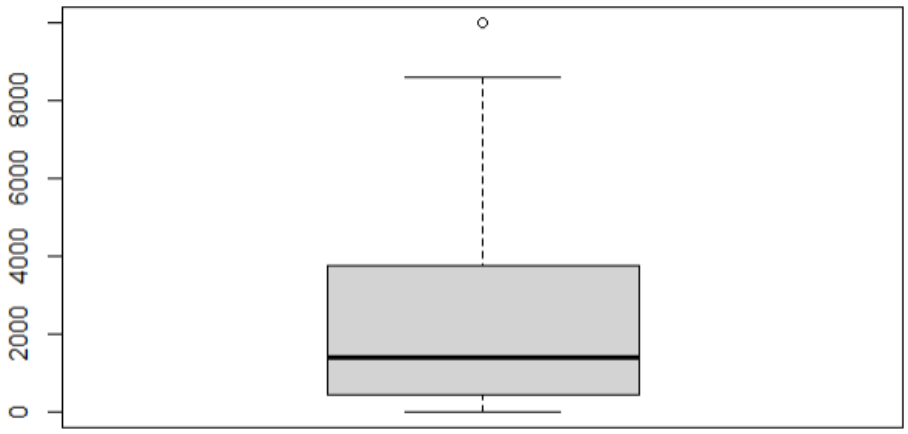


Abbildung 13 - Boxplot za TotalCharges

Posljednji boxplot grafik odnosi se na atribut *DailyCharges* i prikazan je u nastavku.

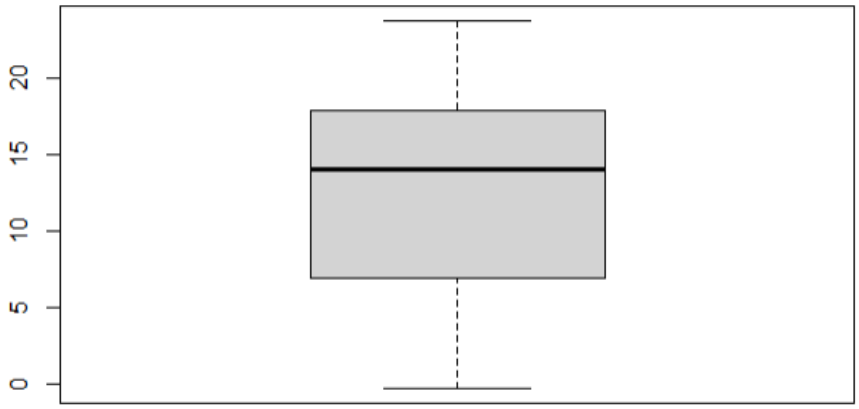


Abbildung 14 - Boxplot za DailyCharges

Obzirom da ovaj grafik izgleda identično kao i boxplot za MonthlyCharges detaljnija analiza nije potrebna.

Nakon što je obrađena procjena lokacije, izvršena je i procjena varijabilnosti. Procjena varijabilnosti se također odnosi isključivo na numeričke varijable i obuhvata analizu vrijednosti poput srednje apsolutne devijacije, standardne devijacije i varijanse. Sve navedene metrike za sva četiri numerička atributa su prikazana ispod.

```
Srednja apsolutna devijacija

       tenure MonthlyCharges   TotalCharges   DailyCharges
    34.099800      35.730660    1824.191040       7.160958

Standardna devijacija

       tenure MonthlyCharges   TotalCharges   DailyCharges
    24.625128      29.977981    2230.346050       6.017192

Varijansa

       tenure MonthlyCharges   TotalCharges   DailyCharges
    606.3969       898.6794   4974443.5010       36.2066
```

Abbildung 15 - Procjene varijabilnosti za numeričke attribute

Za potrebe vizualizacije procjene varijabilnosti korištena je kombinacija histograma i prave koja pokazuje zastupljenost podataka na grafiku gustoće.

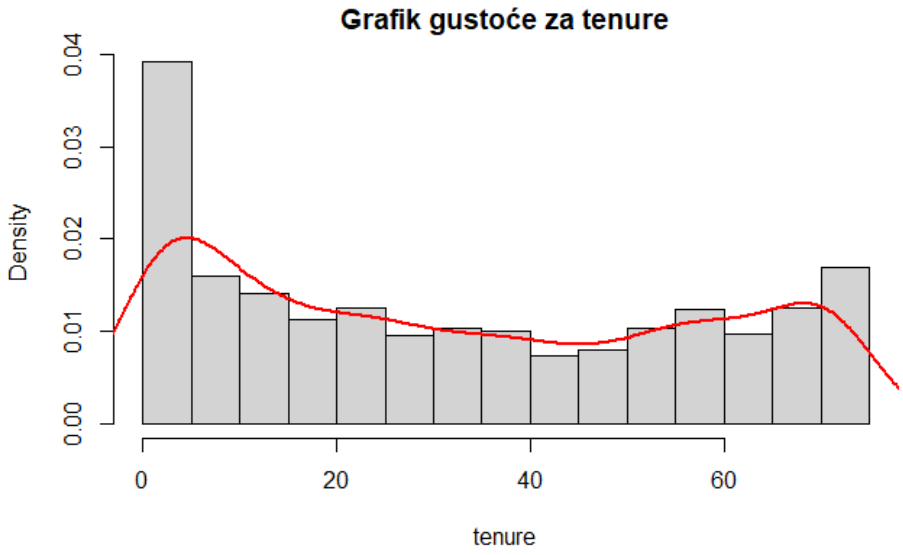


Abbildung 16 - Histogram za Tenure

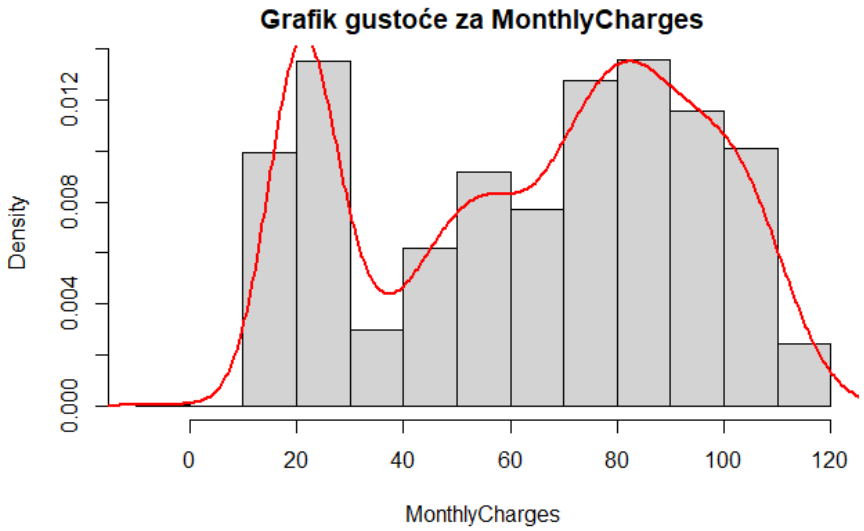


Abbildung 17 - Histogram za MonthlyCharges

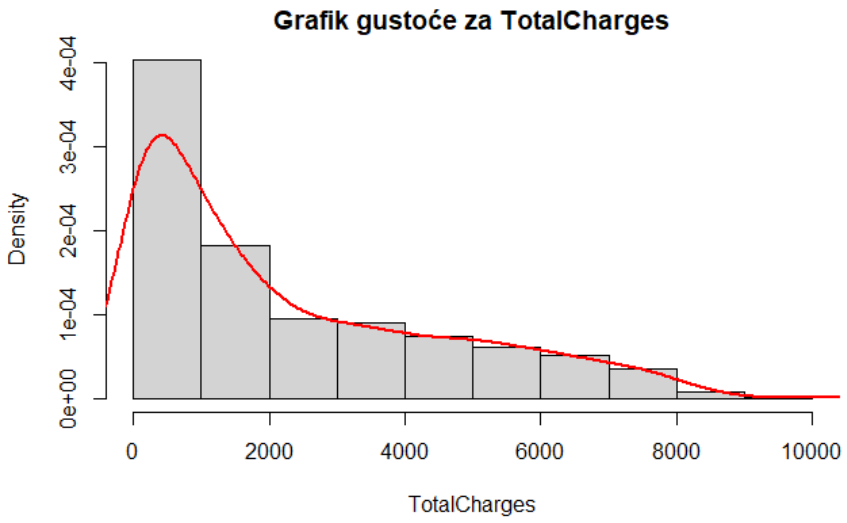


Abbildung 18 - Histogram za TotalCharges

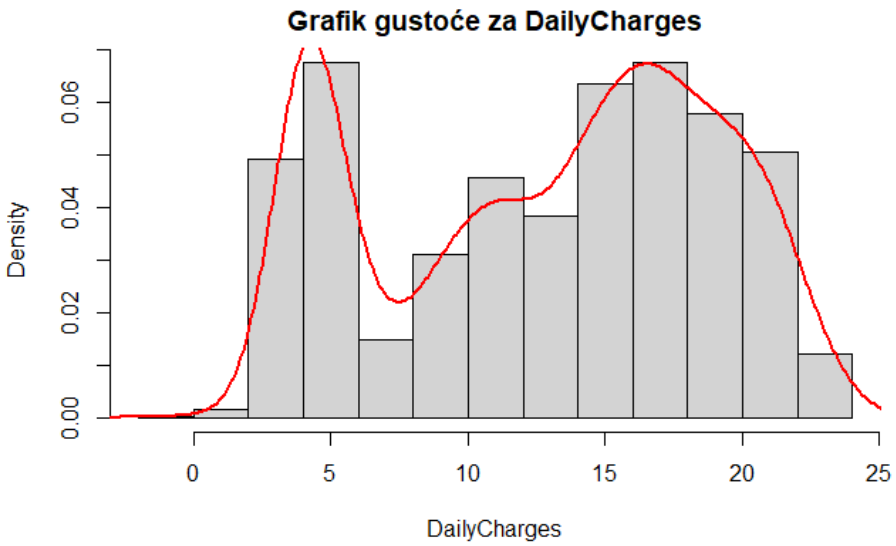


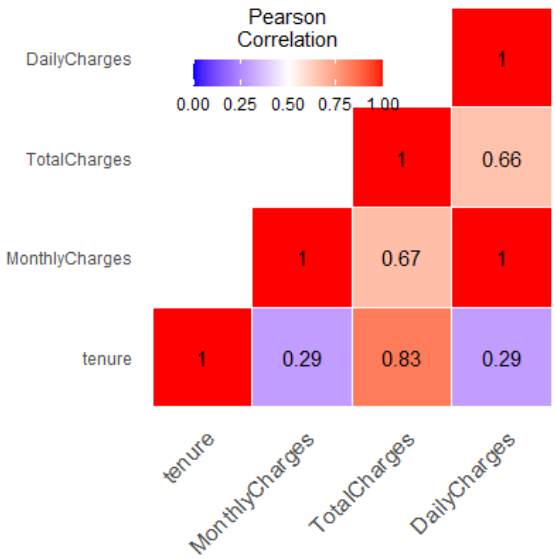
Abbildung 19 - Histogram da DailyCharges

1.1.3. PROCJENA KORELACIJE IZMEĐU VARIJABLI

Za potrebe ove zadaće urađene su sljedeće korelacije između varijabli:

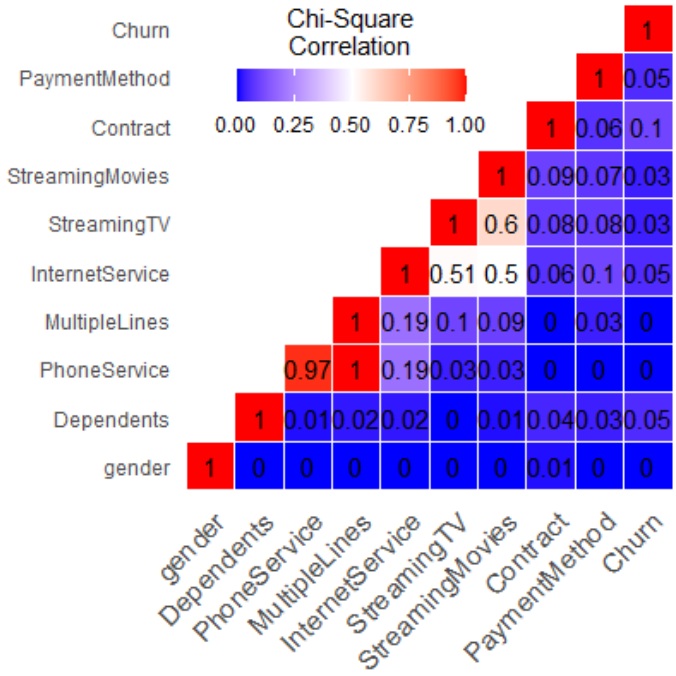
- Metoda Pearsonovog koeficijenta
- Metoda Chi-square koeficijenta

Pearsonov koeficijent je primijenjen za procjenu korelacije između numeričkih varijabli. Sljedeća slika prikazuje rezultat ove analize.



Vidimo da su varijable *DailyCharges* i *MonthlyCharges* u sto postotnoj korelaciji. Također, visok stepen korelacije imaju i varijable *TotalCharges* i *Tenure* (83%).

Što se tiče kategoričkih varijabli, za njih je korištena chi-square matrica iz koje se može analizirati korelacija između svih varijabli.



1.2. PREPROCESIRANJE PODATAKA

Procesiranje podataka je nezaobilazan korak prilikom kreiranja predikcijskih modela. Za potrebe našeg rada izvršene su sljedeće metode koje su specifične za ovu fazu:

- Popunjavanje nedostajućih vrijednosti
- Odbacivanje outliera
- Odbacivanje atributa s visokim stepenom korelacije
- Transformacija podataka

1.2.1. POPUNJAVANJE NEDOSTAJUĆIH VRIJEDNOSTI

Broj NA vrijednosti u bilo kojem redu našeg data seta iznosi preko 400, a kako je ranije naglašeno da skup podataka sadrži ukupno 2000 instanci, bilo je neprihvatljivo ignorisati i izbrisati ovakve vrijednosti jer bi takav pristup doveo do gubitka značajnog broja podataka. Umjesto toga, nedostajuće vrijednosti su popunjene na sljedeći način:

- Nedostajuće vrijednosti kategoričkih varijabli su popunjene najfrekventnijim vrijednostima iz određenog atributa
- Nedostajuće vrijednosti numeričkih varijabli su popunjene median vrijednošću određenog atributa

Na ovaj način smo izvršili prečišćavanje podataka.

1.2.2. ODBACIVANJE OUTLIERA

Kada je riječ o nepodobnim vrijednostima, odnosno outlierima, izbačena su ukupno 3 reda, što ne predstavlja veliki gubitak podataka:

- Izbačen je red u kojem je iznos za *DailyCharges* bio negativan
- Izbačen je red u kojem je vrijednost za *Dependents* bila „Maybe“
- Izbačen je red u kojem je vrijednost za *PaymentMethod* bila „abcd“

Navedene vrijednosti predstavljaju outliere koji su se ponovili samo jednom te su iz tog razloga izbačeni iz data seta.

1.2.3. ODBACIVANJE ATRIBUTA S VISOKIM STEPENOM KORELACIJE

U ovom koraku su izbačeni svi atributi, odnosno kolone koje su nepotrebne, tj. koje nemaju nikakvog uticaja na rezultatnu kolonu.

Na prvom mjestu je to kolona MonthlyCharges za koju je prethodno rečeno da je u sto procentnoj korelaciji sa kolonom DailyCharges. Također, bez obzira na matricu korelacije, izbačeni su i atributi Gender, MultipleLines, PhoneService, StreamingTV i StreamingMovies. Intuitivno se dolazi do zaključka da ovi atributi ne mogu imati nikakve veze sa otkazivanjem pretplate.

1.2.4. TRANSFORMACIJA PODATAKA

Što se tiče transformacije podataka, ona je izvršena nad preostalim numeričkim atributima: *DailyCharges*, *Tenure* i *TotalCharges*.

Za *DailyCharges* i *Tenure* korištena je min-max normalizacija, dok je za *TotalCharges* korištena decimal-normalizacija obzirom da vrijednosti kolena *TotalCharges* imaju najveći opseg i da decimal-normalizacija rješava problem „stisnutosti podataka“ koju uzrokuje min-max normalizacija.

Način normalizacija pomenutih varijabli prikazan je u nastavku.

```r
#min max normalizacija za daily charges
library(dplyr)
max <- max(podaci$DailyCharges)
min <- min(podaci$DailyCharges)
podaci <- mutate(podaci, DailyCharges= (DailyCharges- min) / (max-min))
```

```r
#min max normalizacija za tenure

max <- max(podaci$tenure)
min <- min(podaci$tenure)
podaci <- mutate(podaci, tenureMinMax= (tenure- min) / (max-min))
```

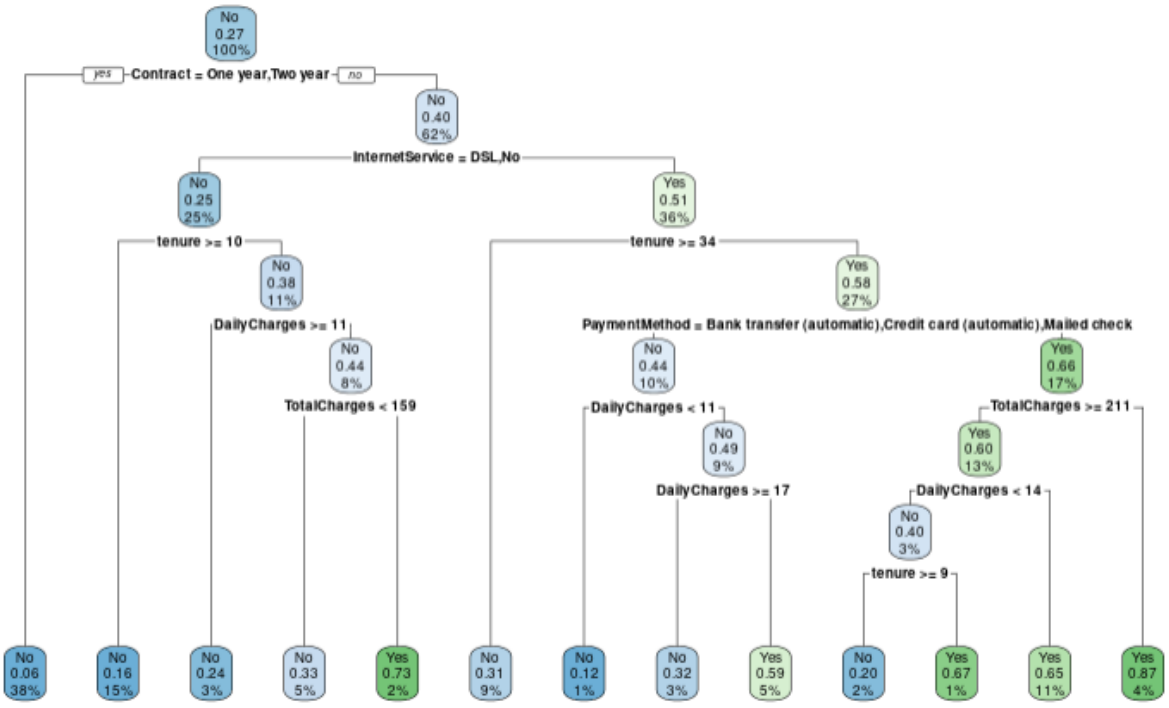
```r
#decimal normalizacija za total charges


max <- max(podaci$TotalCharges)
j <- 0
while (10 ** j < max)
 j = j + 1
podaci <- mutate(podaci, TotalCharges = TotalCharges/ 10 ** j)
```

2. IZGRADNJA MODELA KLASIFIKACIJE

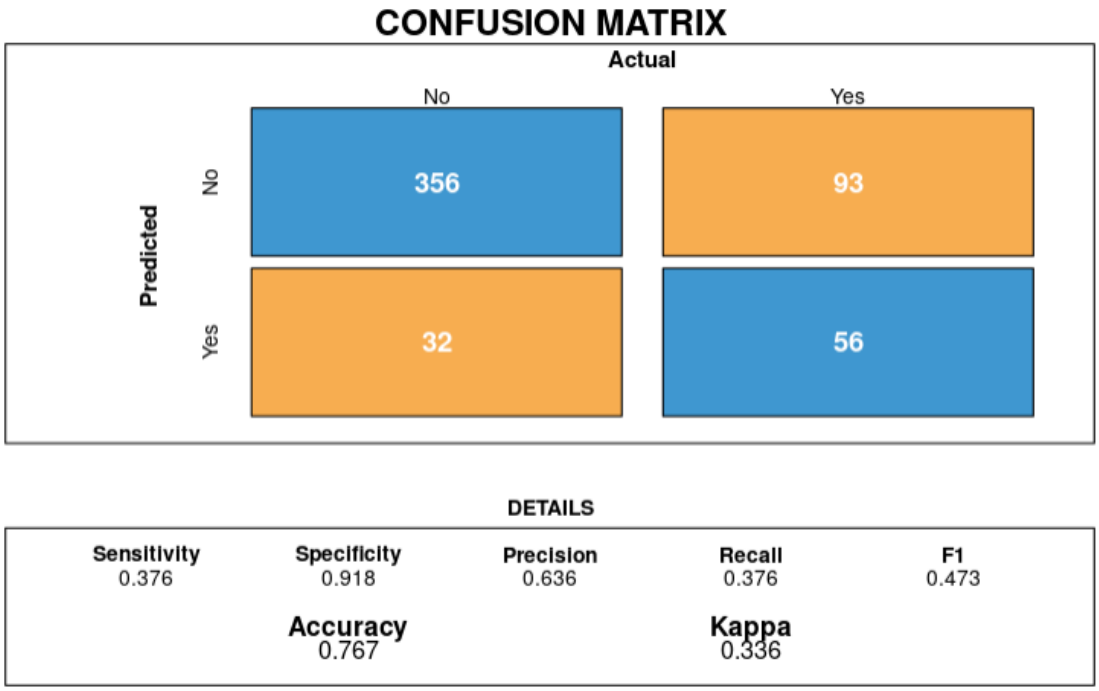
2.1. IZGRADITI MODELA KLASIFIKACIJE I NJIHOVO EVALUIRANJE

2.1.1 DRVO ODLUČIVANJA KOJE KAO MJERU ATRIBUTA SELEKCIJE KORISTI INFORMACIJSKU DOBIT

Najprije smo kreirali drvo odlučivanja za informacijsku dobit, te smo podijelili naš data set na dva dijela: trening set i testni set, pri čemu je trening set sadržavao 70 % instanci originalnog data seta, dok je testni set sadržavao 30% originalnog seta. Prikaz drveta i pravila odlučivanja vidimo na sljedećoj slici.



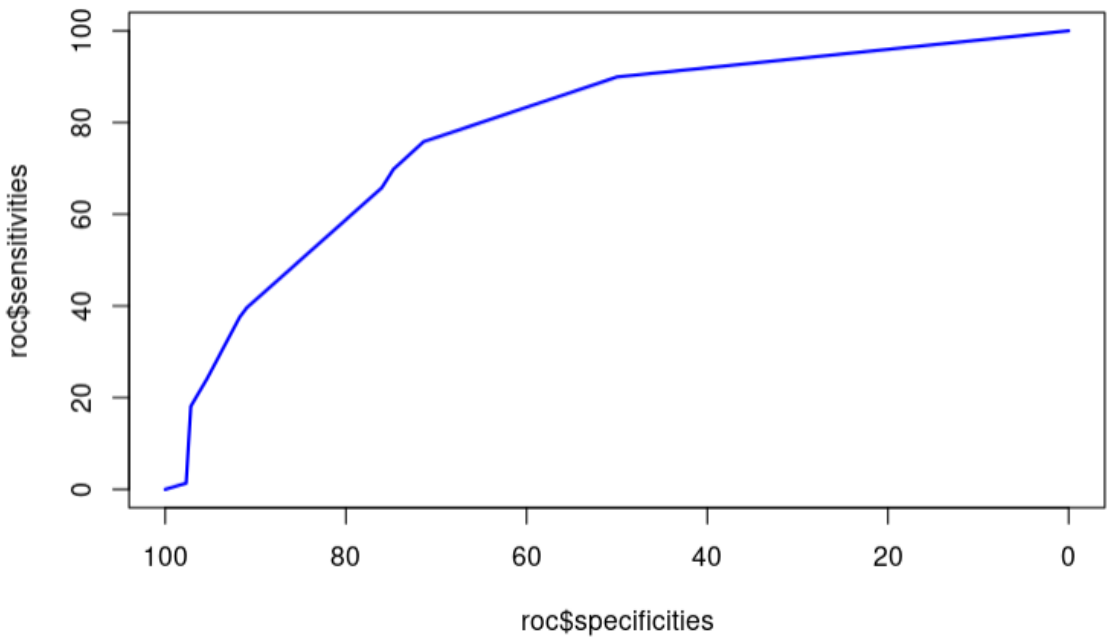
Nakon toga izvršili smo treniranje drveta nad trening setom, i njegovo testiranje nad testnim setom. Na sljedećoj slici je prikazana konfuzijska matrica kao rezultat nad testnim skupom podataka.



Iz konfuzijske matrice možemo očitati mjere procjene za evaluaciju ovog kasifikatora.

Senzitivnost koja ima vrijednost 0.376, predstavlja da je 62.4% instanci pogrešno označeno kao negativno. Ovaj broj u konfuzijskoj matrici je smješten u gornjem desnom uglu, te nam govori da je u našem slučaju 93 instance imalo FN vrijednost. U našem slučaju ova vrijednost je zabrinjujuće velika, te doprinosi lošoj evaluaciji ovog klasifikatora.
Specifičnost ima vrijednost 0.918, što predstavlja da je 8.2% instance označeno kao netačno pozitivno. Ovaj broj je u konfuzijskoj matrici predstavljen u donjem desnom uglu, ten am govori da je u našem slučaju 32 instance imalo FP vrijednost.
Tačnost predstavlja broj tačno klasificiranih instanci, te u našem slučaju ima vrijednost 0.767, što znači da je klasificirao 76.7% insanci kao tačne. Ovaj broj je u konfuzijskoj matrici smješten u plavim pravougaonicima, u donjem desnom uglu, ten am govori da je za naš testni set, čak 56 instanci tačno klasificirano kao tačne (TP).
Kappa statistika je u našem slučaju 0.3336 .
F1 ima vrijednost 0.851.

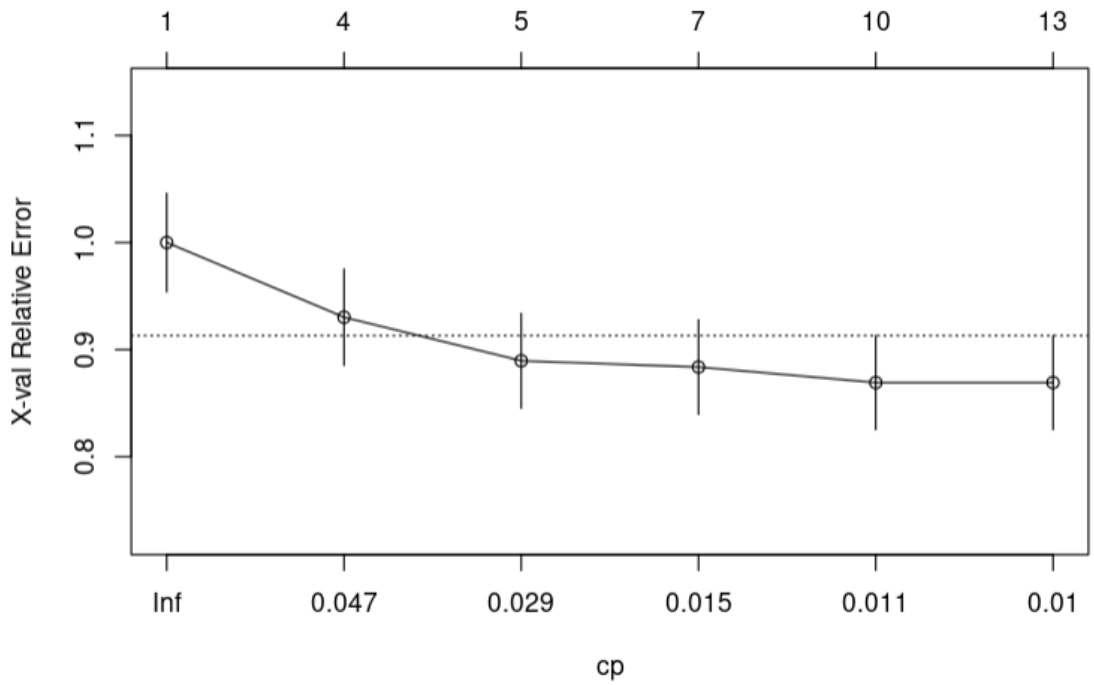
ROC krivu vidimo na sljedećoj slici.



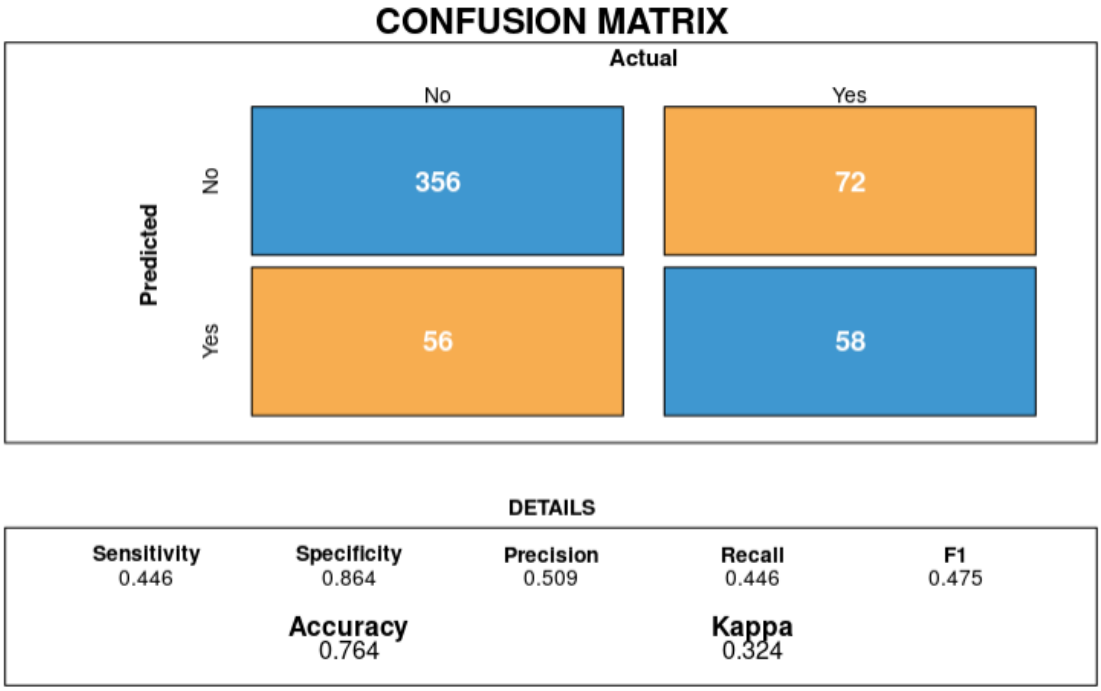
Sa slike možemo primijetiti da je ROC kriva blizu dijagonale, odnosno odaljena od gorenjeg lijevog ugla, gdje i specifičnost i senzitivnost imaju vrijednost 100, te zbog toga predstavlja da nam je klasifikator loš.

Drvo odlučivanja za informacijsku dobit nakon prečišćavanja:

Najprije smo prikazali grafik kompleksnosti drveta odlučivanja kako bismo odredili najmanju cross-validacijsku grešku. U našem slučaju najmanja cross-validacijska greška se dobiva kada koristimo 10 listova. Vrijednost najmanje cross-validacijske greške iznosi 0.011, što vidimo na sljedećoj slici.



Nakon prečišćavanja drveta koristeći funckiju prune sa parametrom cp = 0.011, dobivamo sljedeće rezultate i konfuzijsku matricu.

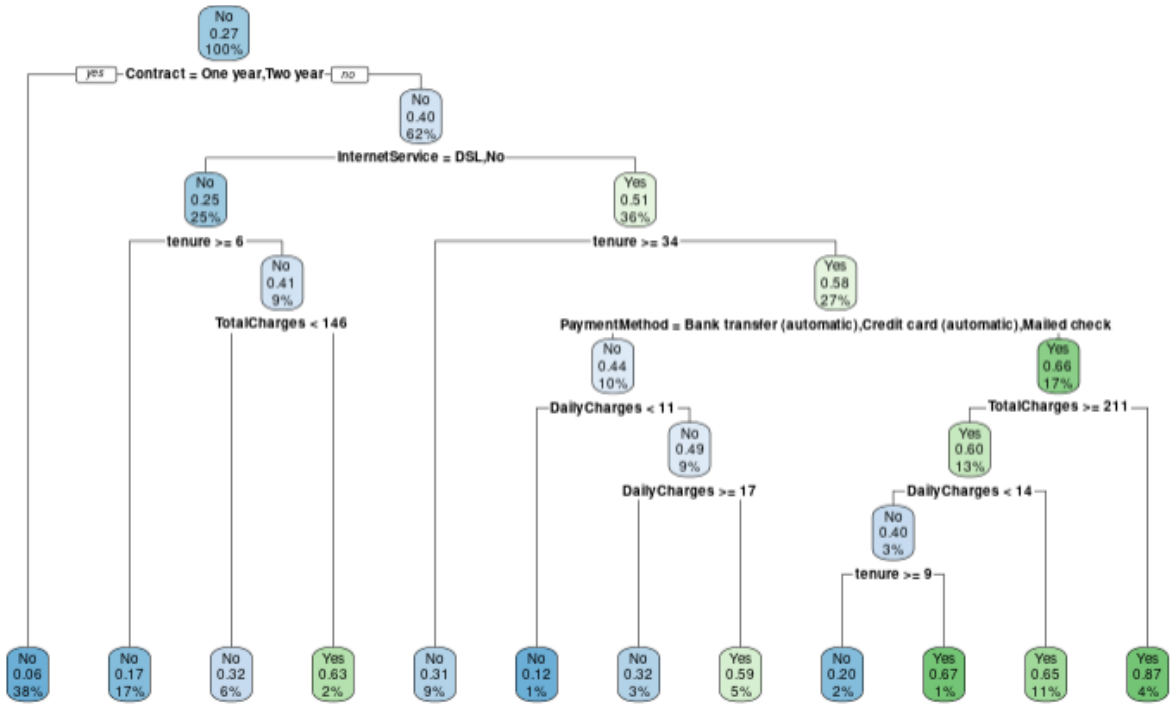


Vidimo da se nakon prečišćavanja drveta senszitivnost povećala, dok su se tačnost, specifičnost i kappa vrijednost neznačajno smanjile.

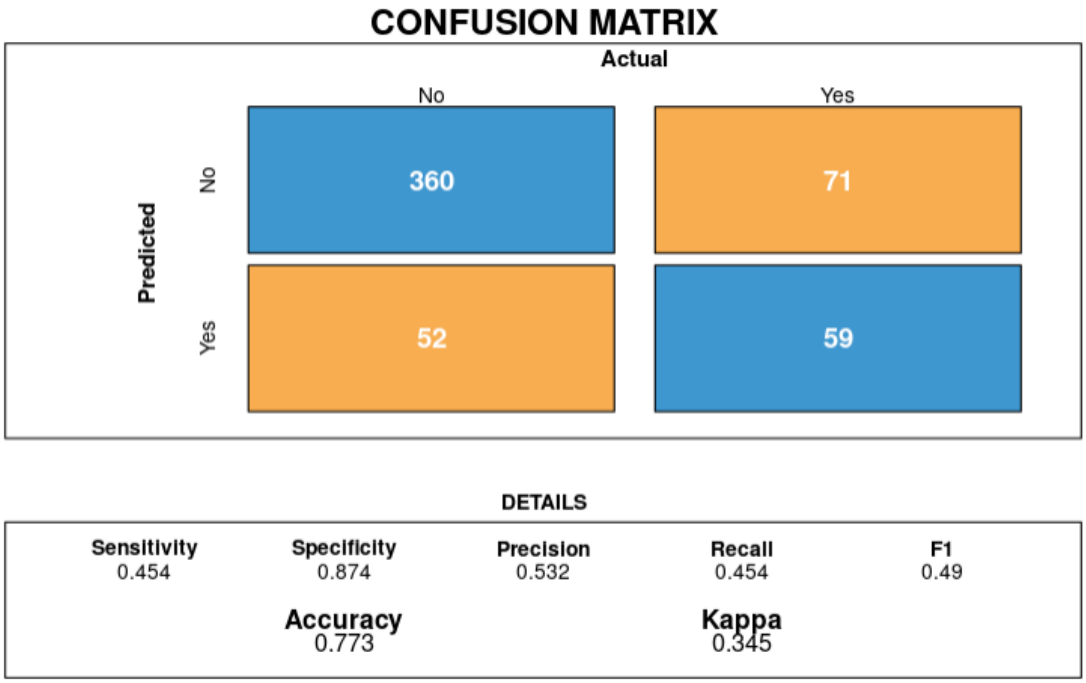
2.1.2 DRVO ODLUČIVANJA KOJE KAO MJERU ATRIBUTA SELEKCIJE KORISTI GINI INDEKS

Kao i u prethodnom primjeru drva odlulčivanja sa infromacijskom dobiti, tako smo i ovdje, postupak kreiranja drveta ponovili identično. Nakon kreiranja, smo podijelili podatke iz seta na isti omjer tesnih i trening podataka, te testirali drvo.

Prikaz drveta nakon treniranja je na sljedećoj slici.



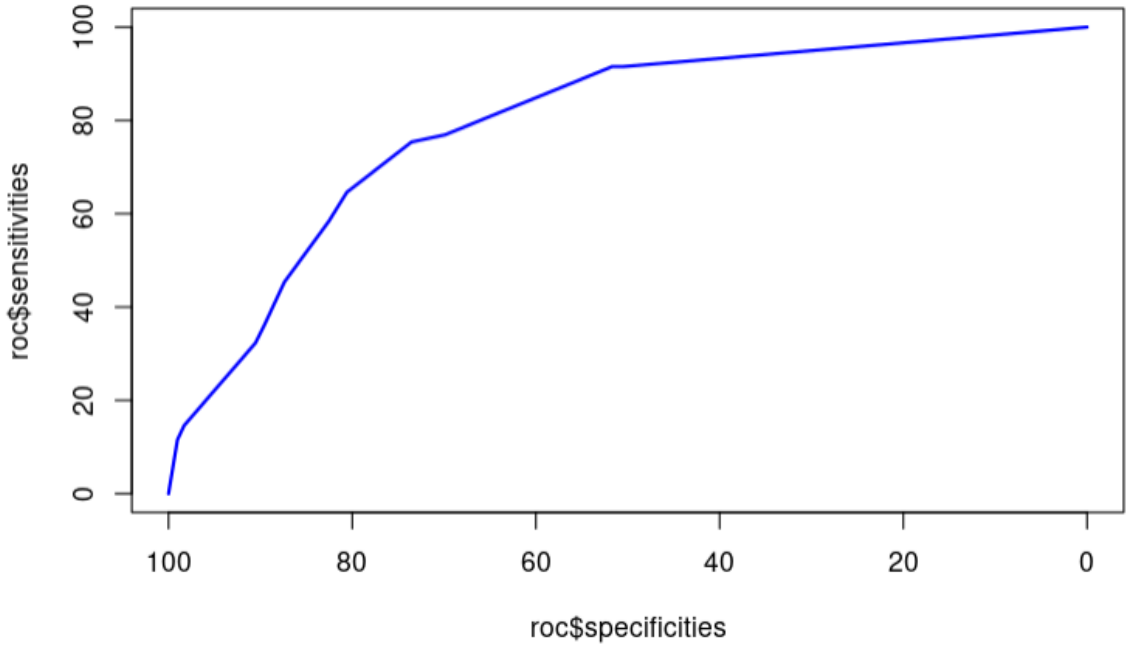
Ispod se nalazi slika koja prikazuje konfuzijsku matricu ovog drva odlučivanja.



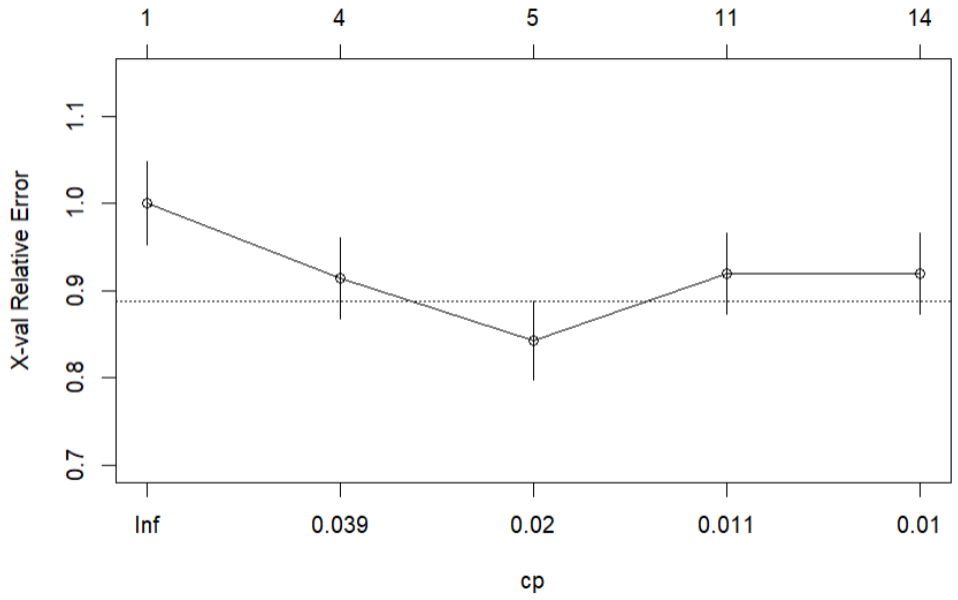
Iz konfuzijske matrice možemo očitati mjere procjene za evaluaciju ovog kasifikatora.

Senzitivnost koja ima vrijednost 0.454, predstavlja da je 54.6% instanci pogrešno označeno kao negativno. Ovaj broj u konfuzijskoj matrici je smješten u gornjem desnom uglu, te nam govori da je u našem slučaju 71 instanci imalo FN vrijednost. U našem slučaju ova vrijednost je zabrinjujuće velika, te doprinosi lošoj evaluaciji ovog klasifikatora. Međutim primjetimo drastično poboljšanje u odnosu na prethodno drvo odlučivanja, s obzirom da je za prethodno drvo 93 instance bilo pogrešno klasificirano, dakle poboljšanje je za ukupno 22 instance iz testnog seta.
Specifičnost ima vrijednost 0.874, što predstavlja da je 12.6% instanci označeno kao netačno pozitivno. Ovaj broj je u konfuzijskoj matrici predstavljen u donjem desnom uglu, te nam govori da je u našem slučaju 52 instance imalo FP vrijednost. Za specifičnost u ovom drvu primjećujemo pogršanje, što se odražava negativno na naše drvo u odnosu na prethodno. Vidimo da je u ovom slučaju 20 instanci više bilo pogrešno klasificirano kao netačno.
Tačnost predstavlja broj tačno klasificiranih instanci, te u našem slučaju ima vrijednost 0.773, što znači da je klasificirao 77.3% insanci kao tačne. Vidimo da je tačnost u ovom drvetu minimalno poboljšana u odnosu na prošlo.
Kappa statistika je u našem slučaju 0.345 , došlo je do blagog poboljšanja.
F1 ima vrijednost 0.49, te vidimo da je poboljšana u odnosu na prethodni primjer.

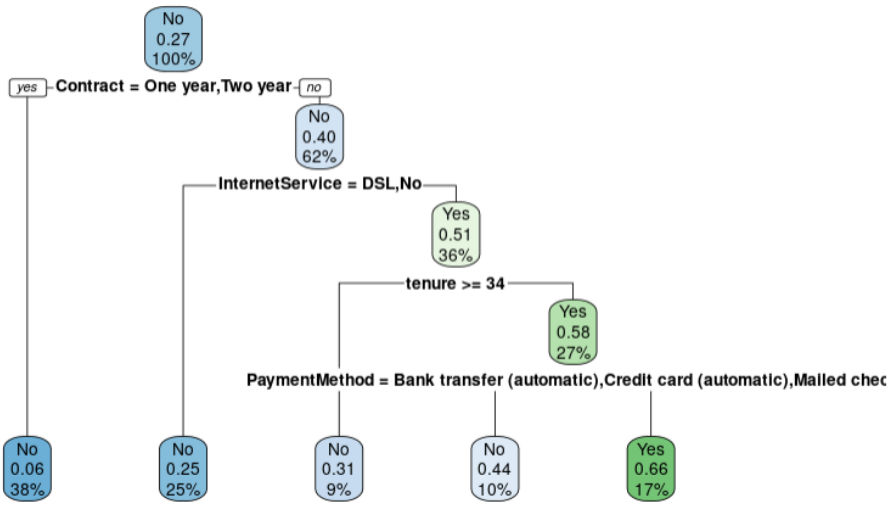
ROC krivu vidimo na sljedećoj slici. Način kreiranja ROC krive za ovo drvo uradili smo identično kao i za drvo odlučivanja sa informacijskom dobiti.



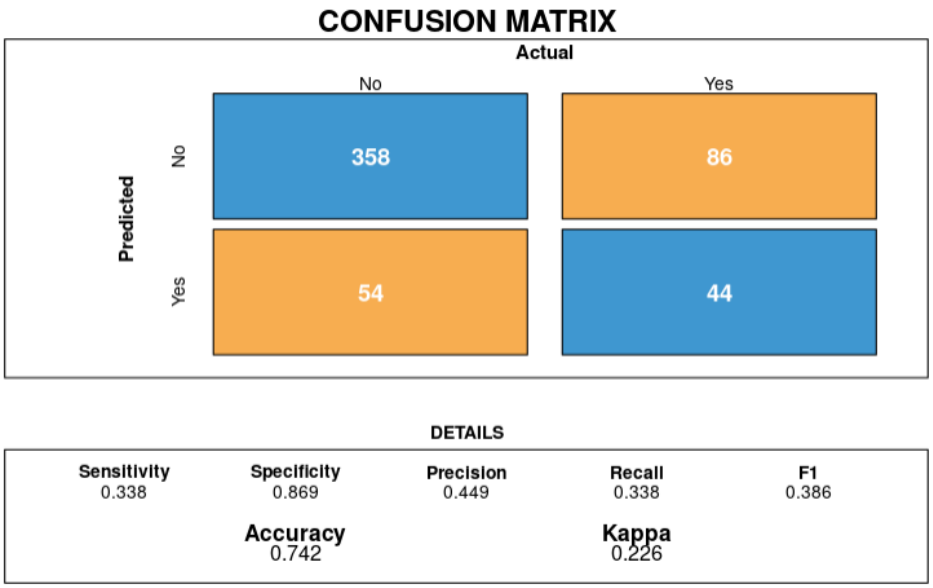
Nakon što smo predstavili grafik kompleksnosti drveta odlučivanja očitavamo da je minimalna vrijednost cross-validacijske greške 0.02, te se doibiva kada je broj listova 5.



Nakon što smo odradili prečišćavanje drveta odlučivanja sa gini indeksom koristeći parametar cp = 0.02, dobivamo izgled drveta kao na sljedećoj slici i vidimo drastičnu redukciju čvorova i pravila odlučivanja.



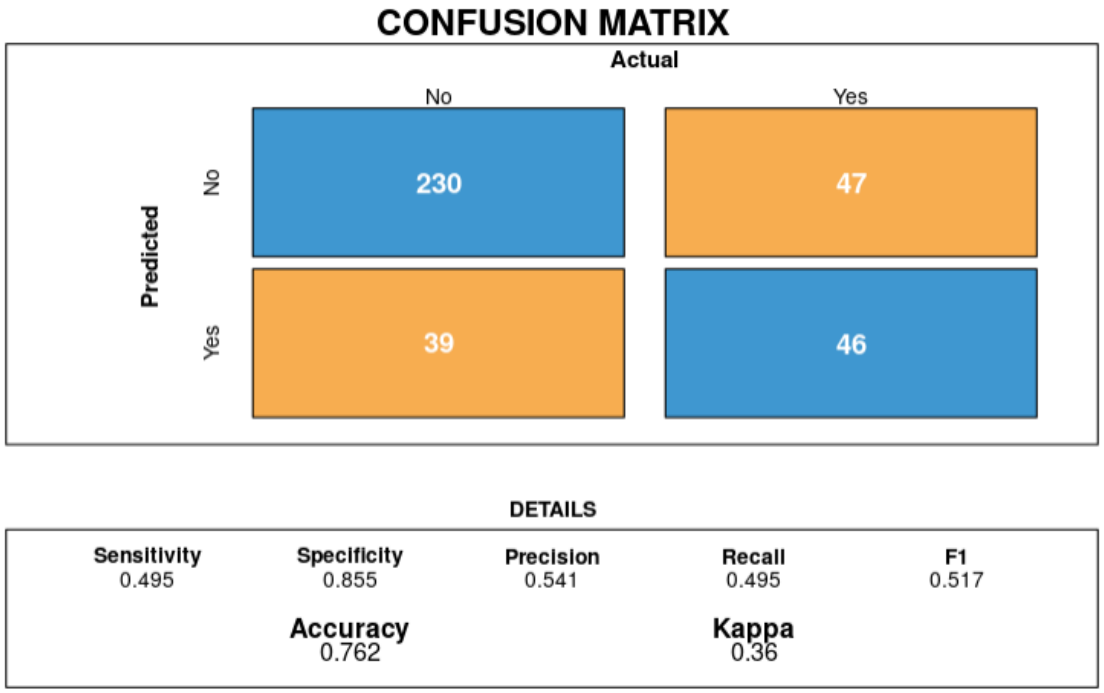
Konfuzijsku matricu vidimo na sljedećoj slici.



Vidimo da su se nakon prečišćavanja drveta sve vrijednosti smanjile.

2.1.3 C5.0 MODEL KLASIFIKACIJE

Za kreiranje ovo modela klasifikacije je korišten isti postupak kao i u prethodna dva, međutim za njega ne možemo prikazati grafički izgled čvorova i pravila odlučivanja.
Nakon treniranja i testiranja ovog modela klasifikacije na isti način kao što smo trenirali i testirali prethodna dva modela, dobivamo sljedeću konfuzijsku matricu.



Iz konfuzijske matrice možemo očitati mjere procjene za evaluaciju ovog kasifikatora.

Senzitivnost koja ima vrijednost 0.495, predstavlja da je 50.5% instanci pogrešno označeno kao negativno. Primjetimo poboljšanje.
Specifičnost ima vrijednost 0.855, što predstavlja da je 14.5% instanci označeno kao netačno pozitivno. Primjetimo pogoršanje.
Tačnost predstavlja broj tačno klasificiranih instanci, te u našem slučaju ima vrijednost 0.762, što znači da je klasificirao 76.2% insanci kao tačne. Primjetimo pogoršanje.
Kappa statistika je u našem slučaju 0.36. Primjetimo poboljšanje.
F1 ima vrijednost 0.517. Primjetimo poboljšanje.
ROC kriva ne postoji za ovaj model klasifikacije.

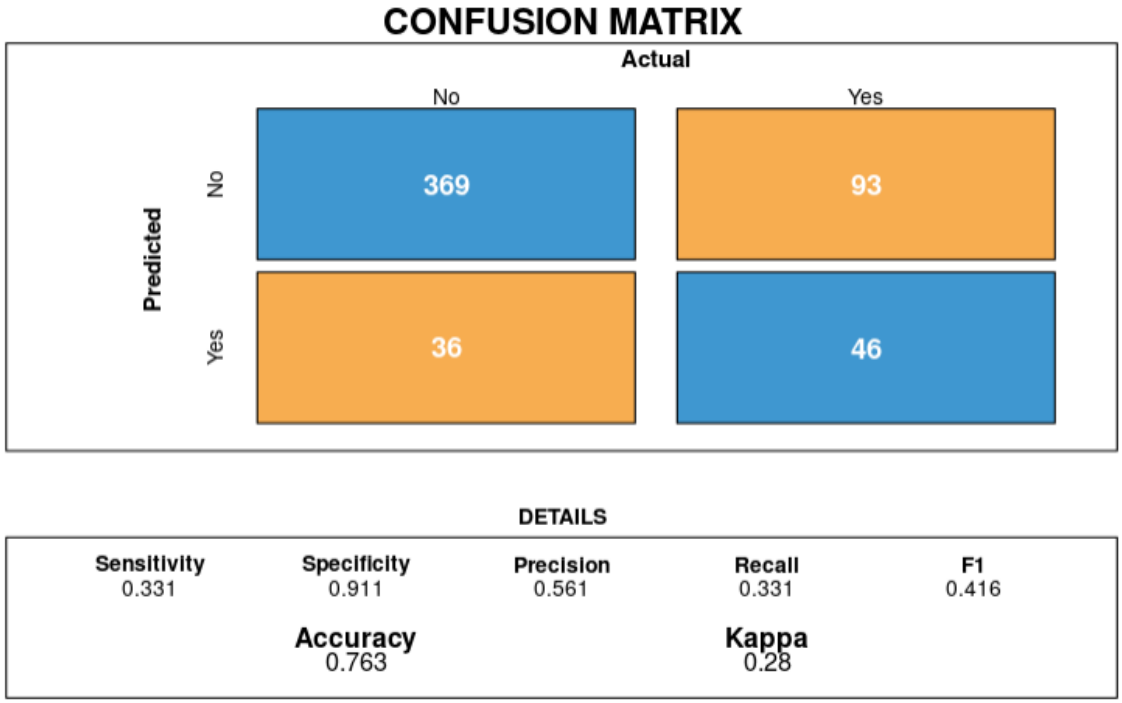
Također za ovaj model klasifikacije nismo radili ni prečišćavanje stabla.

2.2. MPLEMENTIRATI PREDIKCIJSKE MODELE SA METODAMA HOLDOUT, K-FOLD I K-FOLD-BOOTSRAPING

2.2.1 DRVO ODLUČIVANJA SA INFORMACIJSKOM DOBITI

2.2.1.1 HOLDOUT METODA

Na slici ispod su prikazani rezultati klasifikacije za ovako definisan model klasifikacije koristeći nasumičnu raspodjelu podataka na trening i testni podskup podataka. Vidimo da se senzitivnost smanjila, specifičnost neznatno smanjila, tačnost neznatno povećala i kappa vrijednost povećala u odnosu na vrijednosti ovih metrika inicijalnog modela bez uvođenja nasumičnosti u raspodjelu podataka po podskupovima.



2.2.1.2 K-FOLD METODA

Kao što se vidi na slici ispod , u slučaju k=25 dobija se srednja tačnost 76.78% i Kappa statistika 33.33%. Obje ove srednje tačnosti više su nego u slučaju holdout klasifikatora bez korištenja k-fold validacije. Ovo nam pokazuje da korištenje samo jednog trening i testnog skupa podataka ne daje pouzdane rezultate s obzirom da bi se prilikom svakog sljedećeg pokretanja dobio drugačiji rezultat.

```
25-fold validacija
Najveća tačnost: 0.8352941 , fold: 6, najveća kappa: 0.4894942 , fold: 12
Najmanja tačnost: 0.691358 , fold: 7, najmanja kappa: 0.155194 , fold: 15
Srednja tačnost: 0.7677799, srednja kappa: 0.3332586
```

2.2.1.3 K-FOLD BOOTSTRAPING METODA

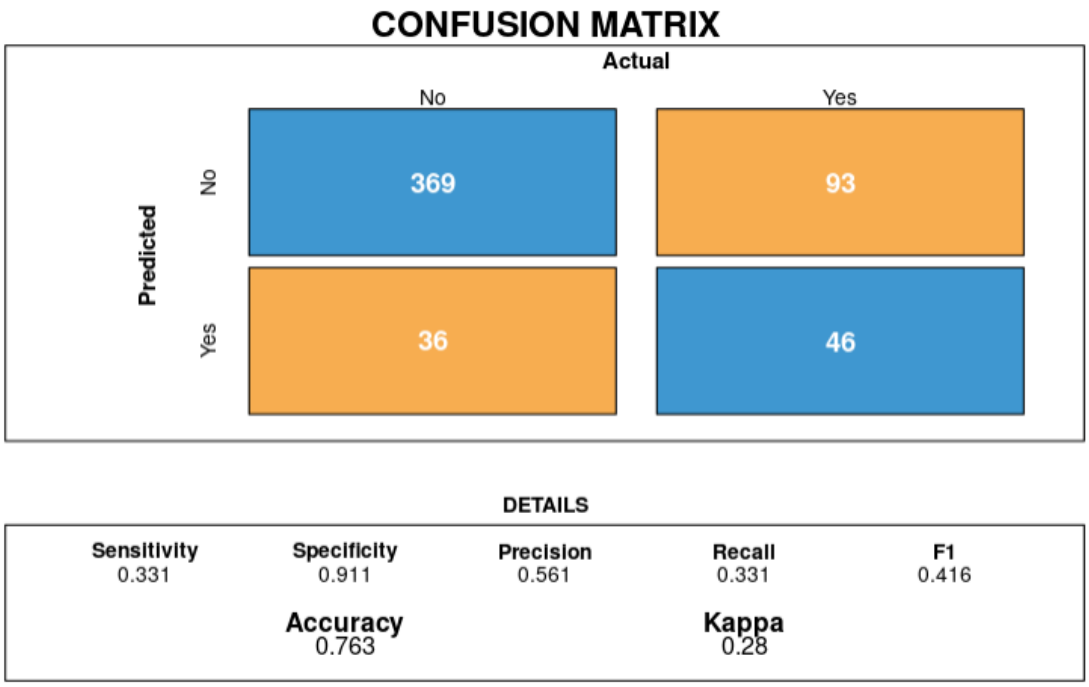
Kao što se vidi na slici ispod , u slučaju k=25 dobija se srednja tačnost 76.94% i Kappa statistika 30.74%. Rezultati su slični kao i za k-fold validaciju.

```
25-fold validacija
Najveća tačnost: 0.8902439 , fold: 11, najveća kappa: 0.6586494 , fold: 11
Najmanja tačnost: 0.6904762 , fold: 7, najmanja kappa: 0.08990011 , fold: 19
Srednja tačnost: 0.7694901, srednja kappa: 0.3074286
```

2.2.2 DRVO ODLUČIVANJA SA GINI INDEKSOM

2.2.2.1 HOLDOUT METODA

Na slici ispod su prikazani rezultati klasifikacije za ovako definisan model klasifikacije koristeći nasumičnu raspodjelu podataka na trening i testni podskup podataka. Vidimo da se senzitivnost i kappa smanjila, tačnost neznatno smanjila a specifičnost neznatno povećala u odnosu na vrijednosti ovih metrika inicijalnog modela bez uvođenja nasumičnosti u raspodjelu podataka po podskupovima.



2.2.2.2 K-FOLD METODA

Kao što se vidi na slici ispod , u slučaju k=25 dobija se srednja tačnost 76.82% i Kappa statistika 33.52%. Obje ove srednje tačnosti više su nego u slučaju holdout klasifikatora bez korištenja k-fold validacije. Ovo nam pokazuje da korištenje samo jednog trening i testnog skupa podataka ne daje pouzdane rezultate s obzirom da bi se prilikom svakog sljedećeg pokretanja dobio drugačiji rezultat.

```
25-fold validacija
Najveća tačnost: 0.8352941 , fold: 6, najveća kappa: 0.4894942 , fold: 12
Najmanja tačnost: 0.691358 , fold: 7, najmanja kappa: 0.155194 , fold: 15
Srednja tačnost: 0.7682808, srednja kappa: 0.335232
```

2.2.2.3 K-FOLD BOOTSTRAPING METODA

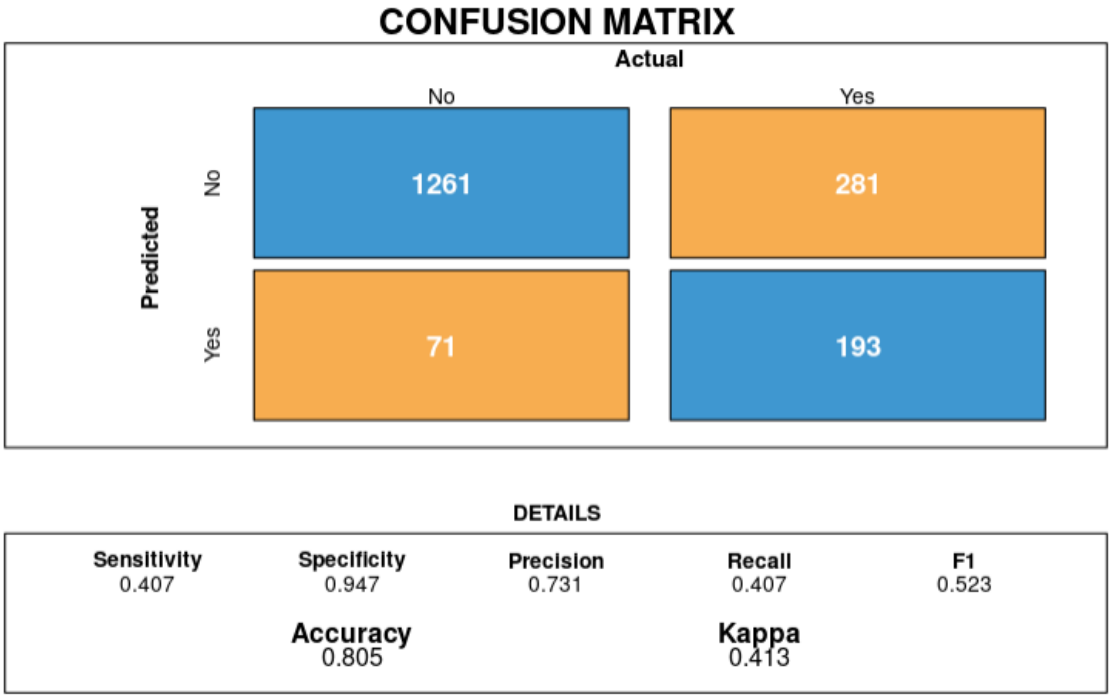
Kao što se vidi na slici ispod , u slučaju k=25 dobija se srednja tačnost 76.18% i Kappa statistika 30.56%. Rezultati su slični kao i za k-fold validaciju.

```
25-fold validacija
Najveća tačnost: 0.8902439 , fold: 11, najveća kappa: 0.6586494 , fold: 11
Najmanja tačnost: 0.6666667 , fold: 17, najmanja kappa: -0.006134969 , fold: 9
Srednja tačnost: 0.7618187, srednja kappa: 0.3056118
```

2.2.3 C5.0 MODEL KLASIFIKACIJE

2.2.3.1 HOLDOUT METODA

Na slici ispod su prikazani rezultati klasifikacije za ovako definisan model klasifikacije koristeći nasumičnu raspodjelu podataka na trening i testni podskup podataka. Vidimo da se senzitivnost smanjila, specifičnost, tačnost I kappa vrijednsost povećale u odnosu na vrijednosti ovih metrika inicijalnog modela bez uvođenja nasumičnosti u raspodjelu podataka po podskupovima.



2.2.3.2 K-FOLD METODA

Kao što se vidi na slici ispod , u slučaju k=25 dobija se srednja tačnost 76.28% i Kappa statistika 35.03%. Obje ove srednje tačnosti manje su nego u slučaju holdout

klasifikatora bez korištenja k-fold validacije. Ovo nam pokazuje da korištenje samo jednog trening i testnog skupa podataka ne daje pouzdane rezultate s obzirom da bi se prilikom svakog sljedećeg pokretanja dobio drugačiji rezultat.

```
25-fold validacija
Najveća tačnost: 0.8135593 , fold: 4, najveća kappa: 0.4817674 , fold: 4
Najmanja tačnost: 0.7087912 , fold: 9, najmanja kappa: 0.1975684 , fold: 3
Srednja tačnost: 0.7628554, srednja kappa: 0.3503675
```

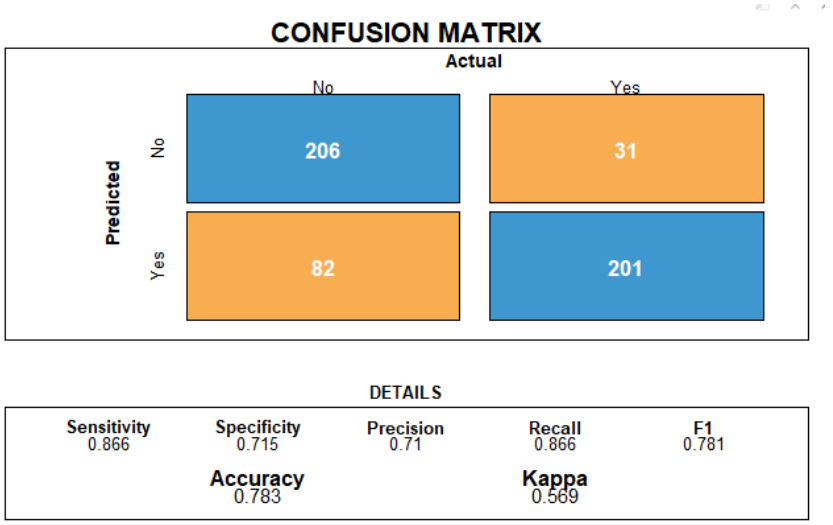
2.2.3.3 K-FOLD BOOTSTRAPING METODA

Kao što se vidi na slici ispod , u slučaju k=25 dobija se srednja tačnost 75.99% i Kappa statistika 35.10%. Rezultati su slični kao i za k-fold validaciju.

```
25-fold validacija
Najveća tačnost: 0.7966102 , fold: 10, najveća kappa: 0.4948387 , fold: 6
Najmanja tačnost: 0.7150838 , fold: 2, najmanja kappa: 0.2221567 , fold: 9
Srednja tačnost: 0.7599327, srednja kappa: 0.3510669
```

2.3. ANALIZA BALANSIRANOST PODATAKA

Balansiranje podataka je urađeno za C5.0 model klasifikacije. Urađen je oversampling zbog zastupljenosti vrijednosti “No” u rezultantnoj koloni. Ukupan broj instanci je postavljan na 2600. Nakon balansiranja podataka konfuzijska matrica za C5.0 je izgledala kaon a slici ispod.



2.4. KORIŠTENJE ANSAMBL TEHNIKA ZA UNAPRJEĐENJE TAČNOSTI KLASIFIKACIJE

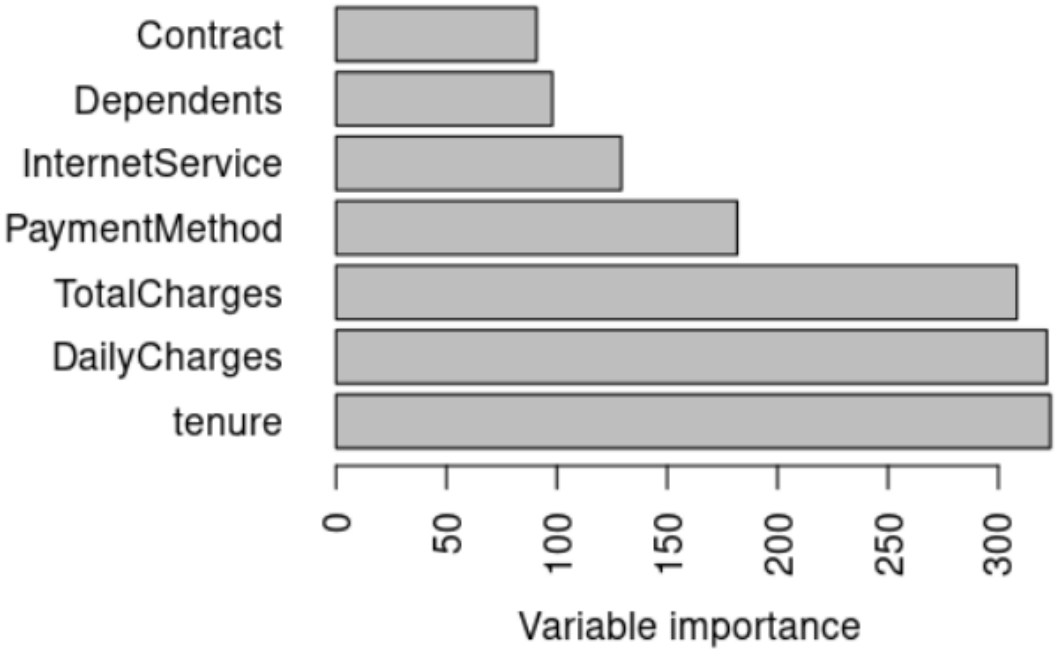
2.4.1. BAGGING MODEL

Nakon treniranja bagging modela dobivamo sljedeće podatke kao na slici ispod.

Tačnost iznosi 78.79%
Senzitivnost iznosi 44.3%
Specifičnost iznosi 88.38%
Kappa vrijednost iznosi 34.41%.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  251  44
       Yes  33  35

               Accuracy : 0.7879
                 95% CI : (0.7422, 0.8288)
    No Information Rate : 0.7824
    P-Value [Acc > NIR] : 0.4289

                  Kappa : 0.3441

 Mcnemar's Test P-Value : 0.2545

            Sensitivity : 0.44304
            Specificity : 0.88380
         Pos Pred Value : 0.51471
         Neg Pred Value : 0.85085
             Prevalence : 0.21763
         Detection Rate : 0.09642
   Detection Prevalence : 0.18733
      Balanced Accuracy : 0.66342

       'Positive' Class : Yes
```

Nakon treniranja bagging modela, prikazujemo ukupni stepen značaja svih pojedinačnih atributa. Na sljedećoj slici vidimo da je najbitnija varijabla za naš data set je tenure, a u stopu je prate DailyCharges i TotalCharges.



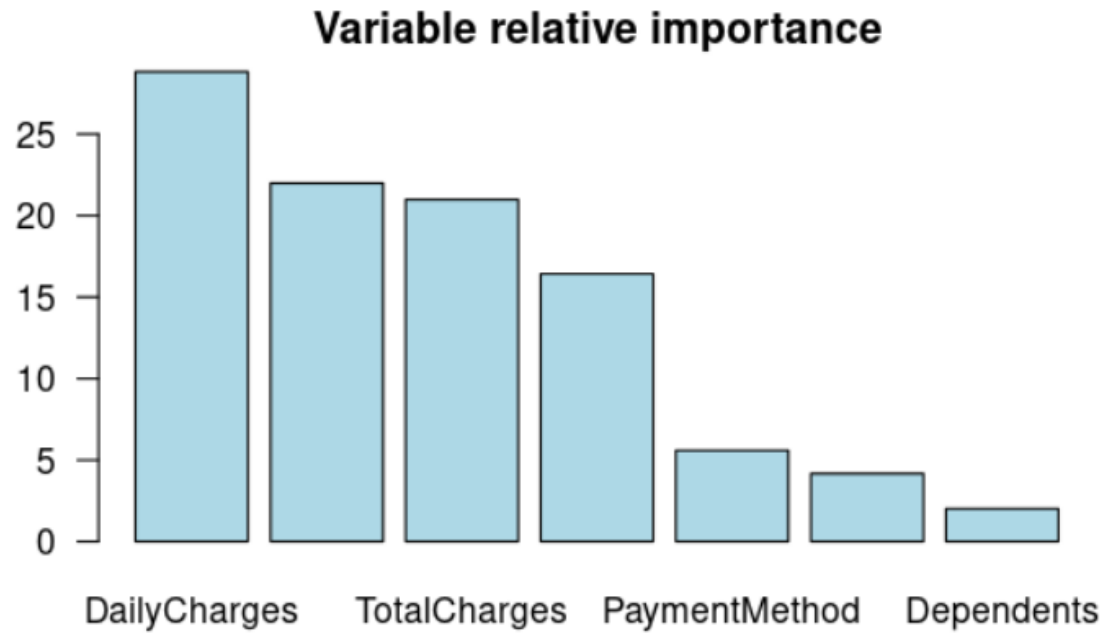
2.4.2 BOOSTING MODEL KORIŠTENJEM ADABOOST

Nakon treniranja bagging modela dobivamo sljedeće podatke kao na slici ispod.

Tačnost iznosi 78.18%
Senzitivnost iznosi 47.96%
Specifičnost iznosi 89.39%
Kappa vrijednost iznosi 40.33%.

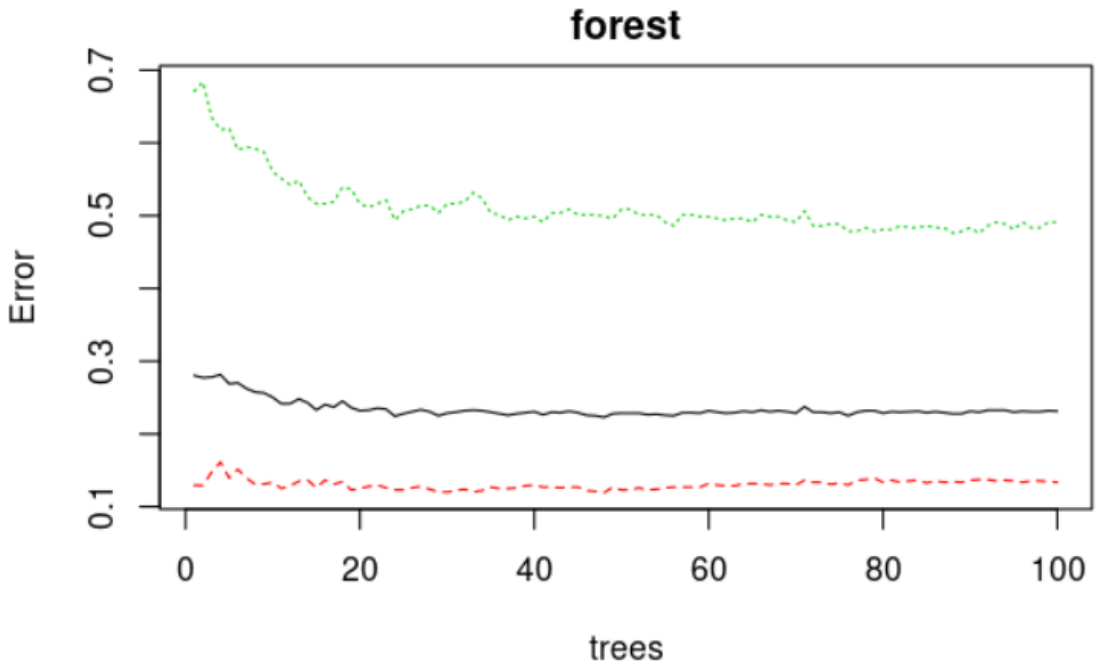
```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  236  51
       Yes  28  47

               Accuracy : 0.7818
                 95% CI : (0.7356, 0.8232)
    No Information Rate : 0.7293
    P-Value [Acc > NIR] : 0.01298

                  Kappa : 0.4033

 Mcnemar's Test P-Value : 0.01332

            Sensitivity : 0.4796
            Specificity : 0.8939
         Pos Pred Value : 0.6267
         Neg Pred Value : 0.8223
             Prevalence : 0.2707
         Detection Rate : 0.1298
   Detection Prevalence : 0.2072
      Balanced Accuracy : 0.6868

       'Positive' Class : Yes
```

Nakon treniranja boosting modela, prikazujemo ukupni stepen značaja svih pojedinačnih atributa, i vidimo da je najznačajnija varijabla DailyCharges.



2.4.3 RANDOM FOREST MODEL

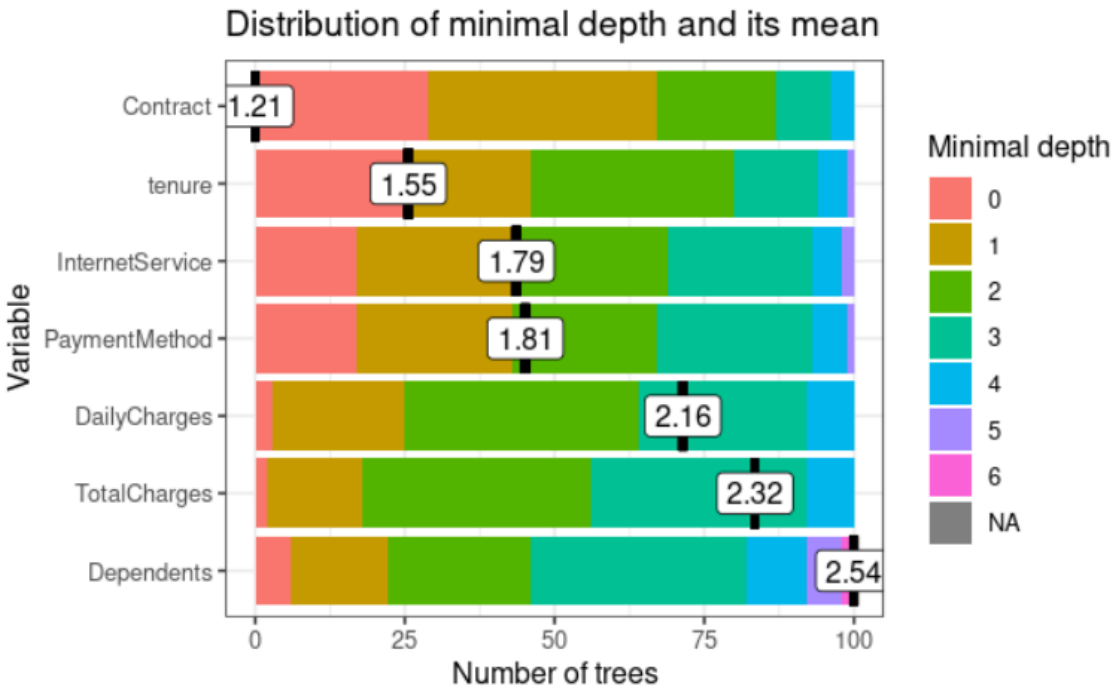
Nakon kreiranja drveta kreira se sljedeći prikaz odakle je vidljiva minimalna, maksimalna i srednja greška klasifikacije pri inkrementalnom dodavanju modela drveta odlučivanja.



Predikcije smo izvršili na identičan način kao i sve do sada. Sa sljedeće slike je vidljivo da tačnost ima vrijednost 77.62% a Kappa statistika 34.12% .

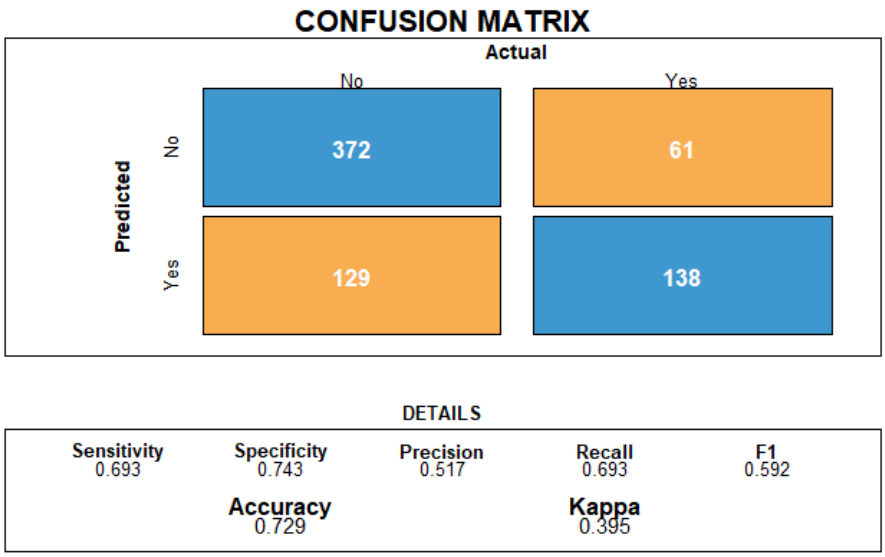
```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  243  41
       Yes  40  38

               Accuracy : 0.7762
                 95% CI : (0.7298, 0.8181)
    No Information Rate : 0.7818
    P-Value [Acc > NIR] : 0.6288

                  Kappa : 0.3412

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.8587
            Specificity : 0.4810
         Pos Pred Value : 0.8556
         Neg Pred Value : 0.4872
             Prevalence : 0.7818
         Detection Rate : 0.6713
   Detection Prevalence : 0.7845
      Balanced Accuracy : 0.6698

       'Positive' Class : No
```

Nakon treniranja random forest modela, moguće je prikazati ukupni stepen značaja svih pojedinačnih atributa u svim modelima. Vidimo da su za naš primjer kao najvažniji označeni contract i tenure.



3. TESTIRANJE NAJBOLJEG MODELA

Od svih kreiranih modela za testiranje je izabran C5.0, a rezultati su prikazani u nastavku.



4. PRAVILO-BAZIRANA KLASIFIKACIJA