A blue and white logo with a building in the background

Description automatically generated

Prirodno-matematički fakultet  
Informatika

Seminarski rad

Predstavljanje I tumačenje skupa podataka  
“**Adult Census Income**”

**Studenti** **Mentor**

Lara Aksić 33/2018 Branko Arsić

Nikola Vukašinović 27/2019

Contents

[Uvod 4](#_Toc177329142)

[Priprema podataka 4](#_Toc177329143)

[Opis podataka 7](#_Toc177329144)

[Nedostajuće vrednosti 9](#_Toc177329145)

[Raspodela vrednosti po kolonama 21](#_Toc177329146)

[Age 21](#_Toc177329147)

[Fnlwgt 22](#_Toc177329148)

[Education 23](#_Toc177329149)

[Education.num 24](#_Toc177329150)

[Marital.status 25](#_Toc177329151)

[Relationship 26](#_Toc177329152)

[Race 27](#_Toc177329153)

[Sex 28](#_Toc177329154)

[Capital.gain 29](#_Toc177329155)

[Capital.loss 31](#_Toc177329156)

[Hours.per.week 33](#_Toc177329157)

[Outliers 35](#_Toc177329158)

[Age 35](#_Toc177329159)

[Fnlwgt 37](#_Toc177329160)

[Education.num 39](#_Toc177329161)

[Capital.gain 43](#_Toc177329162)

[Capital.loss 45](#_Toc177329163)

[Hours.per.week 47](#_Toc177329164)

[Workclass, education, marital.status, sex, occupation, race, relationship, native.country 49](#_Toc177329165)

[Analiza 49](#_Toc177329166)

[Age 49](#_Toc177329167)

[Workclass 51](#_Toc177329168)

[Education 52](#_Toc177329169)

[Education.num 54](#_Toc177329170)

[Marital.status 55](#_Toc177329171)

[Occupation 57](#_Toc177329172)

[Relationship 58](#_Toc177329173)

[Race 60](#_Toc177329174)

[Sex 63](#_Toc177329175)

[Capital.gain 64](#_Toc177329176)

[Capital.loss 65](#_Toc177329177)

[Hours.per.week 66](#_Toc177329178)

[Native.country 68](#_Toc177329179)

[Analiza između više feature-a 69](#_Toc177329180)

[Age ~ sex 69](#_Toc177329181)

[Age ~ race 69](#_Toc177329182)

[Workclass ~ race 70](#_Toc177329183)

[Occupation ~ relationship.status 71](#_Toc177329184)

[Occupation ~ education 72](#_Toc177329185)

[Hours per week ~ sex 73](#_Toc177329186)

[Occupation ~ sex 74](#_Toc177329187)

[Education ~ sex 75](#_Toc177329188)

[Workclass ~ sex 76](#_Toc177329189)

[Selekcija – Feature selection 77](#_Toc177329190)

[Numericke kolone 77](#_Toc177329191)

[Modeli mašinskog učenja 78](#_Toc177329192)

[Podela na train i test skupove 79](#_Toc177329193)

[Resampling 81](#_Toc177329194)

[Oversampling 82](#_Toc177329195)

[Undersampling 82](#_Toc177329196)

[Over i Under sampling 83](#_Toc177329197)

[Data rose 83](#_Toc177329198)

[F-Regression 85](#_Toc177329199)

[Logistička regresija 85](#_Toc177329200)

[Decision tree 103](#_Toc177329201)

[Random forest 107](#_Toc177329202)

[Zaključak 111](#_Toc177329203)

[Najbolji model 113](#_Toc177329204)

[Literatura 114](#_Toc177329205)

# Uvod

Adult Census Income je skup podataka izvucen iz baze podataka biroa za popis iz 1994. godine. Zadatak predviđanja je da se utvrdi da li osoba zarađuje preko 50 hiljada dolara godišnje. Skup podataka koji smo koristili se nalazi u folderu seminarski i naziva se adult.csv.  
Link koji vodi do sajta odakle je preuzet dataset: [Adult Census Income](https://www.kaggle.com/datasets/uciml/adult-census-income)

# Priprema podataka

Za početak ćemo učitati potrebne biblioteke za rad.

library(readr)

library(tidyverse)

library(dplyr)

library(vcd)

library(ggplot2)

library(pROC)

library(car)

library(ROCR)

library(AUC)

library(caret)

library(randomForest)

library(naniar)

library(rpart)

library(ROSE)

library(nortest)

library(moments)

library(psych)

library(broom)

Zatim ćemo učitati naš skup podataka. Učitavanje csv fajla sa funkcijom read\_csv

dataset = read\_csv("adult.csv")

## Rows: 32561 Columns: 15  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (9): workclass, education, marital.status, occupation, relationship, rac...  
## dbl (6): age, fnlwgt, education.num, capital.gain, capital.loss, hours.per.week  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

view(dataset)  
Funkcija view(dataset) pruža nam tabelaran uvid u set podataka.

dim(dataset)

## [1] 32561 15

Funkcija dim(dataset) daje informacije o dimenzijama okvira podataka.   
Dataset se sastoji iz 15 kolona i 32561. reda.

head(dataset)

## # A tibble: 6 × 15  
## age workclass fnlwgt education education.num marital.status occupation   
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 90 ? 77053 HS-grad 9 Widowed ?   
## 2 82 Private 132870 HS-grad 9 Widowed Exec-manager…  
## 3 66 ? 186061 Some-college 10 Widowed ?   
## 4 54 Private 140359 7th-8th 4 Divorced Machine-op-i…  
## 5 41 Private 264663 Some-college 10 Separated Prof-special…  
## 6 34 Private 216864 HS-grad 9 Divorced Other-service  
## # ℹ 8 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>

Funkcija head(dataset) daje nam uvid u prvih 6 redova našeg skupa podataka.

str(dataset)

## spc\_tbl\_ [32,561 × 15] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ age : num [1:32561] 90 82 66 54 41 34 38 74 68 41 ...  
## $ workclass : chr [1:32561] "?" "Private" "?" "Private" ...  
## $ fnlwgt : num [1:32561] 77053 132870 186061 140359 264663 ...  
## $ education : chr [1:32561] "HS-grad" "HS-grad" "Some-college" "7th-8th" ...  
## $ education.num : num [1:32561] 9 9 10 4 10 9 6 16 9 10 ...  
## $ marital.status: chr [1:32561] "Widowed" "Widowed" "Widowed" "Divorced" ...  
## $ occupation : chr [1:32561] "?" "Exec-managerial" "?" "Machine-op-inspct" ...  
## $ relationship : chr [1:32561] "Not-in-family" "Not-in-family" "Unmarried" "Unmarried" ...  
## $ race : chr [1:32561] "White" "White" "Black" "White" ...  
## $ sex : chr [1:32561] "Female" "Female" "Female" "Female" ...  
## $ capital.gain : num [1:32561] 0 0 0 0 0 0 0 0 0 0 ...  
## $ capital.loss : num [1:32561] 4356 4356 4356 3900 3900 ...  
## $ hours.per.week: num [1:32561] 40 18 40 40 40 45 40 20 40 60 ...  
## $ native.country: chr [1:32561] "United-States" "United-States" "United-States" "United-States" ...  
## $ income : chr [1:32561] "<=50K" "<=50K" "<=50K" "<=50K" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. age = col\_double(),  
## .. workclass = col\_character(),  
## .. fnlwgt = col\_double(),  
## .. education = col\_character(),  
## .. education.num = col\_double(),  
## .. marital.status = col\_character(),  
## .. occupation = col\_character(),  
## .. relationship = col\_character(),  
## .. race = col\_character(),  
## .. sex = col\_character(),  
## .. capital.gain = col\_double(),  
## .. capital.loss = col\_double(),  
## .. hours.per.week = col\_double(),  
## .. native.country = col\_character(),  
## .. income = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

Funkcijom str(dataset) smo saznali da se dataset sastoji iz 9 kategorijskih(chr) kolona i 6 numerickih(dbl) kolona. Predstavićemo dataset i objasniti svaku varijablu.

summary(dataset)

## age workclass fnlwgt education   
## Min. :17.00 Length:32561 Min. : 12285 Length:32561   
## 1st Qu.:28.00 Class :character 1st Qu.: 117827 Class :character   
## Median :37.00 Mode :character Median : 178356 Mode :character   
## Mean :38.58 Mean : 189778   
## 3rd Qu.:48.00 3rd Qu.: 237051   
## Max. :90.00 Max. :1484705   
## education.num marital.status occupation relationship   
## Min. : 1.00 Length:32561 Length:32561 Length:32561   
## 1st Qu.: 9.00 Class :character Class :character Class :character   
## Median :10.00 Mode :character Mode :character Mode :character   
## Mean :10.08   
## 3rd Qu.:12.00   
## Max. :16.00   
## race sex capital.gain capital.loss   
## Length:32561 Length:32561 Min. : 0 Min. : 0.0   
## Class :character Class :character 1st Qu.: 0 1st Qu.: 0.0   
## Mode :character Mode :character Median : 0 Median : 0.0   
## Mean : 1078 Mean : 87.3   
## 3rd Qu.: 0 3rd Qu.: 0.0   
## Max. :99999 Max. :4356.0   
## hours.per.week native.country income   
## Min. : 1.00 Length:32561 Length:32561   
## 1st Qu.:40.00 Class :character Class :character   
## Median :40.00 Mode :character Mode :character   
## Mean :40.44   
## 3rd Qu.:45.00   
## Max. :99.00

Funkcija summary() pruža sažetak podataka u različitim vrstama objekata kao što su vektori, faktori, data frame-ovi i drugi objekti. Takođe dobijamo vrednosti poput: maksimuma, minimuma, medijane, broja nedostajućih vrednosti, vrednosti 1. kvartila, 3. kvartila.

## Opis podataka

Naš skup podataka sadrži sledeće kolone:

1. age - numerička kolona, odnosi se na godine pojedinca  
   (min, max) = (17, 90)
2. workclass - kategorijska kolona, odnosi se na klasifikaciju zaposlenja
3. fnlwgt - numerička kolona, težinski faktori u datotekama Current Population Survey(CPS) kontrolišu se nezavisnim procenama američke civilne, neinstitucionalizovane populacije. Koristi se tri seta kontrola:

Procena populacije od 16 godina i starijih za svaku saveznu državu.

Kontrole za hispanisko poreklo prema uzrastu i polu.

Kontrole prema rasi, uzrastu i polu.

Koristimo sva tri seta kontrola u našem programu za težinsko prilagođavanje i “usklađujemo” ih 6 puta tako da na kraju ponovo uzimamo u obzir sve kontrole koje smo koristili.

(min, max) = (12285, 1484705)

1. education - kategorijska kolona, odnosi se na stepen obrazovanja
2. education.num - numerička kolona, odnosi se na stepen obrazovanja

(min, max) = (1.00, 16.00)

1. marital.status - kategorijska kolona, odnosi se na bračni status
2. occupation - kategorijska kolona, odnosi se na vrstu zanimanja
3. relationship - kategorijska kolona, odnosi se na status veze pojedinca
4. race - kategorijska kolona, odnosi se na rasu pojedinca
5. sex - kategorijska kolona, odnosi se na pol pojedinca
6. capital.gain - numerička kolona, odnosi se na kapitalnu dobit

(min, max) = (0, 99999)

1. capital.loss - numerička kolona, odnosi se na kapitalni gubitak

(min, max) = (0.0, 4356.0)

1. hours.per.week - numerička kolona, odnosi se na broj radnih sati nedeljno

(min, max) = (1.00, 99.00)

1. native.country - kategorijska kolona, odnosi se na zemlju porekla
2. income - kolona koju ćemo prediktovati, odnosi se na to da li pojedinac zaradjuje manje ili više od 50k dolara godišnje.

A screenshot of a computer program

Description automatically generated

A white background with text

Description automatically generated

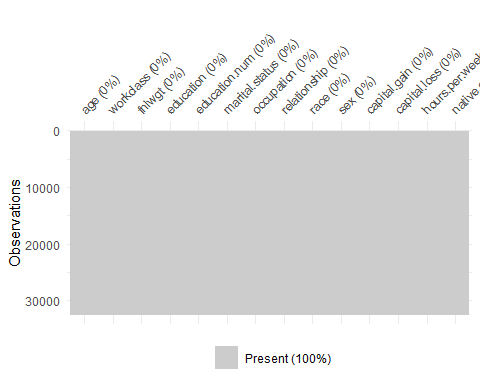
Imamo uvid u sve jedinstvene vrednosti za svaku kategorijsku kolonu.

## Nedostajuće vrednosti

(colMeans(is.na(dataset)))\*100

## age workclass fnlwgt education education.num   
## 0 0 0 0 0   
## marital.status occupation relationship race sex   
## 0 0 0 0 0   
## capital.gain capital.loss hours.per.week native.country income   
## 0 0 0 0 0

vis\_miss(dataset[1:14])



Zaključujemo da dataset ne sadrži NA vrednosti. Ali smo iz prethodnih funkcija uočili da postoje “?” vrednosti za pojedine kolone koje ćemo mi smatrati NA vrednostima. Da su postojale vrednosti ” ” ili “-” i njih bi uzimali kao nedostajuće.

Kolone koje sadrže vrednosti “?” su: 1. workclass 2. occupation 3. native.country

missing\_workclass = sum(dataset$workclass == "?")  
missing\_occupation = sum(dataset$occupation == "?")  
missing\_native.country = sum(dataset$native.country == "?")  
  
print(missing\_workclass)

## [1] 1836

print(((missing\_workclass/nrow(dataset))\*100))

## [1] 5.638647

print(missing\_occupation)

## [1] 1843

print(((missing\_occupation/nrow(dataset))\*100))

## [1] 5.660146

print(missing\_native.country)

## [1] 583

print(((missing\_native.country/nrow(dataset))\*100))

## [1] 1.790486

Funkcijom sum(dataset$column == “?”) dobili smo broj pojavljivanja vrednosti “?” u svakoj od kolona u kojim smo uočili da se pojavljuje.

Za kolonu workclass našli smo da se vrednost “?” pojavljuje 1836 puta.   
Za kolonu occupation našli smo da se vrednost “?” pojavljuje 1843 puta.   
Za kolonu native.country našli smo da se vrednost “?” pojavljuje 583 puta.

Funkcijom ((missing\_column/nrow(dataset))\*100) dobili smo udeo pojavljivanja vrednosti “?” u svakoj od kolona u kojima smo uočili da se pojavljuje.

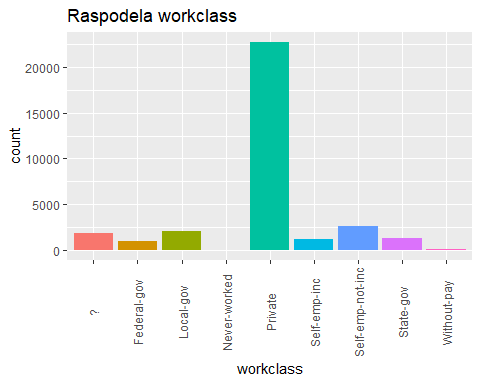
Za kolonu workclass našli smo da se vrednost “?” pojavljuje 5.638647%   
Za kolonu occupation našli smo da se vrednost “?” pojavljuje 5.660146%   
Za kolonu native.country našli smo da se vrednost “?” pojavljuje 1.790486%

Sređivanje “?” vrednosti za kolone koje imaju takve vrednosti:

xtabs(~ workclass, data = dataset)

## workclass  
## ? Federal-gov Local-gov Never-worked   
## 1836 960 2093 7   
## Private Self-emp-inc Self-emp-not-inc State-gov   
## 22696 1116 2541 1298   
## Without-pay   
## 14

ggplot(data = dataset, mapping = aes(workclass, fill=workclass)) + geom\_bar() + labs(title="Raspodela workclass") + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")



Za kolonu workclass dobili smo da ima 22696 podataka sa vrednošću “Private” što čini ~70% ukupnog broja podataka. Tako da ćemo sve vrednosti “?”, kojih ima 1836, zameniti sa vrednošću “Private”.

dataset = dataset %>% mutate(workclass = ifelse(workclass == "?", "Private", workclass))  
xtabs(~ workclass, data = dataset)

## workclass  
## Federal-gov Local-gov Never-worked Private   
## 960 2093 7 24532   
## Self-emp-inc Self-emp-not-inc State-gov Without-pay   
## 1116 2541 1298 14

Posle popunjavanja ”?” vrednosti, iscrtaćemo ponovo graf da vidimo novu raspodelu.

ggplot(data = dataset, mapping = aes(workclass, fill=workclass)) + geom\_bar() + labs(title="Raspodela workclass") + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")

A graph with a green bar and white text

Description automatically generated

Native.country je sledeća kolona koju sređujemo

xtabs(~ native.country, data = dataset)

## native.country  
## ? Cambodia   
## 583 19   
## Canada China   
## 121 75   
## Columbia Cuba   
## 59 95   
## Dominican-Republic Ecuador   
## 70 28   
## El-Salvador England   
## 106 90   
## France Germany   
## 29 137   
## Greece Guatemala   
## 29 64   
## Haiti Holand-Netherlands   
## 44 1   
## Honduras Hong   
## 13 20   
## Hungary India   
## 13 100   
## Iran Ireland   
## 43 24   
## Italy Jamaica   
## 73 81   
## Japan Laos   
## 62 18   
## Mexico Nicaragua   
## 643 34   
## Outlying-US(Guam-USVI-etc) Peru   
## 14 31   
## Philippines Poland   
## 198 60   
## Portugal Puerto-Rico   
## 37 114   
## Scotland South   
## 12 80   
## Taiwan Thailand   
## 51 18   
## Trinadad&Tobago United-States   
## 19 29170   
## Vietnam Yugoslavia   
## 67 16

ggplot(data = dataset, mapping = aes(native.country, fill=native.country)) + geom\_bar() + labs(title="Raspodela native.country") + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")

A graph with text overlay

Description automatically generated

Za kolonu native.country dobili smo da ima 29170 podataka sa vrednošću “United-States” što čini ~90% ukupnog broja podataka. Tako da ćemo sve vrednosti “?”, kojih ima 583, zameniti sa vrednošću “United-States”.

dataset = dataset %>% mutate(native.country = ifelse(native.country == "?", "United-States", native.country))  
xtabs(~ native.country, data = dataset)

## native.country  
## Cambodia Canada   
## 19 121   
## China Columbia   
## 75 59   
## Cuba Dominican-Republic   
## 95 70   
## Ecuador El-Salvador   
## 28 106   
## England France   
## 90 29   
## Germany Greece   
## 137 29   
## Guatemala Haiti   
## 64 44   
## Holand-Netherlands Honduras   
## 1 13   
## Hong Hungary   
## 20 13   
## India Iran   
## 100 43   
## Ireland Italy   
## 24 73   
## Jamaica Japan   
## 81 62   
## Laos Mexico   
## 18 643   
## Nicaragua Outlying-US(Guam-USVI-etc)   
## 34 14   
## Peru Philippines   
## 31 198   
## Poland Portugal   
## 60 37   
## Puerto-Rico Scotland   
## 114 12   
## South Taiwan   
## 80 51   
## Thailand Trinadad&Tobago   
## 18 19   
## United-States Vietnam   
## 29753 67   
## Yugoslavia   
## 16

ggplot(data = dataset, mapping = aes(native.country, fill=native.country)) + geom\_bar() + labs(title="Raspodela native.country") + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")

A graph with text overlay

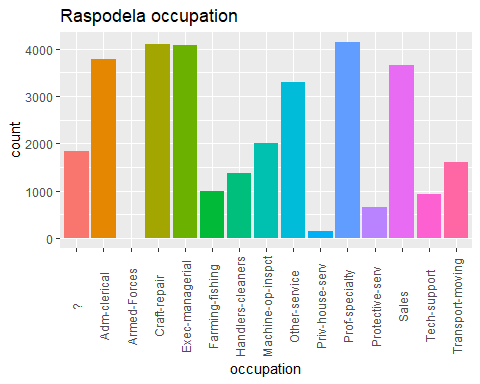
Description automatically generated

Još nam ostaje occupation da sredimo.

xtabs(~ occupation, data = dataset)

## occupation  
## ? Adm-clerical Armed-Forces Craft-repair   
## 1843 3770 9 4099   
## Exec-managerial Farming-fishing Handlers-cleaners Machine-op-inspct   
## 4066 994 1370 2002   
## Other-service Priv-house-serv Prof-specialty Protective-serv   
## 3295 149 4140 649   
## Sales Tech-support Transport-moving   
## 3650 928 1597

ggplot(data = dataset, mapping = aes(occupation, fill=occupation)) + geom\_bar() + labs(title="Raspodela occupation") + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")



Za kolonu occupation ne može se sa lakoćom odrediti kojim ćemo podacima zameniti vrednosti, iz razloga što su podaci ravnomernije raspoređeni.

ggplot(data = dataset, aes(x = education, fill = occupation)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Proportion of Occupations by Education Level",  
 x = "Education",  
 y = "Proportion") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

A graph of occupations by education

Description automatically generated

Vidimo da na višim nivoima obrazovanja, kao što su Bachelors, Masters i Doctorate, najviše su primenjena zanimanja Exec-managerial i Prof-specialty.

Sa nižim obrazovanjima kao što su 1st-4th, 5th-6th, i 7th-8th, imaju veću zastupljenost zanimanja kao što su Handlers-cleaners, Machine-op-inspect, i Other-service.

Zanimanja kao što su Adm-clerical, Sales, i Other-service su zastupljena na više različitih nivoa obrazovanja, što znači da se ljudi sa ovim zanimanjima pojavljuju na raznim obrazovnim nivoima.

Vrednosti za "?" ćemo zameniti odgovarajućom vrednošću iz occupationa na osnovu najčešće pojavljivane vrednosti u education za taj occupation.

xtabs(~ occupation, data = dataset)

## occupation  
## ? Adm-clerical Armed-Forces Craft-repair   
## 1843 3770 9 4099   
## Exec-managerial Farming-fishing Handlers-cleaners Machine-op-inspct   
## 4066 994 1370 2002   
## Other-service Priv-house-serv Prof-specialty Protective-serv   
## 3295 149 4140 649   
## Sales Tech-support Transport-moving   
## 3650 928 1597

#da pronadjemo najcesce zanimanje za svaki nivo obrazovanja  
most\_common\_occupation = dataset %>%  
 filter(occupation != "?") %>%  
 group\_by(education) %>%  
 summarize(most\_common = names(sort(table(occupation), decreasing = TRUE)[1]))  
   
#sada da zamenimo ? vrednosti sa najcescim zanimanjem za odgovarajuci nivo obrazovanja  
dataset = dataset %>%  
 left\_join(most\_common\_occupation, by = "education") %>%  
 mutate(occupation = ifelse(occupation == "?", most\_common, occupation)) %>%  
 select(-most\_common)

ggplot(data = dataset, aes(x = education, fill = occupation)) +  
 geom\_bar(position = "fill") +  
 labs(title = "Proportion of Occupations by Education Level",  
 x = "Education",  
 y = "Proportion") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

A graph of occupations by education

Description automatically generated

xtabs(~ occupation, data = dataset)

## occupation  
## Adm-clerical Armed-Forces Craft-repair Exec-managerial   
## 4333 9 4766 4066   
## Farming-fishing Handlers-cleaners Machine-op-inspct Other-service   
## 994 1370 2002 3654   
## Priv-house-serv Prof-specialty Protective-serv Sales   
## 149 4394 649 3650   
## Tech-support Transport-moving   
## 928 1597

Sada možemo da sačuvamo naš sređen skup podataka kao cist\_dataset

cist\_dataset = dataset  
write.csv(cist\_dataset, "D:/Vezba/3.Godina/Uvod u nauku o podacima/Seminarski/cist\_dataset.csv", row.names = FALSE)  
cist\_dataset = read\_csv("cist\_dataset.csv")

## Rows: 32561 Columns: 15  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (9): workclass, education, marital.status, occupation, relationship, rac...  
## dbl (6): age, fnlwgt, education.num, capital.gain, capital.loss, hours.per.week  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

head(cist\_dataset)

## # A tibble: 6 × 15  
## age workclass fnlwgt education education.num marital.status occupation   
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 90 Private 77053 HS-grad 9 Widowed Craft-repair   
## 2 82 Private 132870 HS-grad 9 Widowed Exec-manager…  
## 3 66 Private 186061 Some-college 10 Widowed Adm-clerical   
## 4 54 Private 140359 7th-8th 4 Divorced Machine-op-i…  
## 5 41 Private 264663 Some-college 10 Separated Prof-special…  
## 6 34 Private 216864 HS-grad 9 Divorced Other-service  
## # ℹ 8 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>

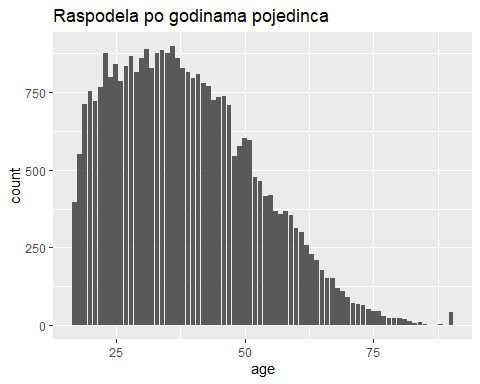
# Raspodela vrednosti po kolonama

## Age

xtabs(~ age, data = cist\_dataset)

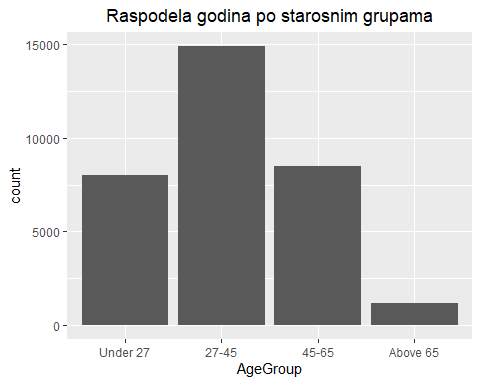
## age  
## 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36   
## 395 550 712 753 720 765 877 798 841 785 835 867 813 861 888 828 875 886 876 898   
## 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56   
## 858 827 816 794 808 780 770 724 734 737 708 543 577 602 595 478 464 415 419 366   
## 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76   
## 358 366 355 312 300 258 230 208 178 150 151 120 108 89 72 67 64 51 45 46   
## 77 78 79 80 81 82 83 84 85 86 87 88 90   
## 29 23 22 22 20 12 6 10 3 1 1 3 43

ggplot(data=cist\_dataset,mapping = aes(x=age)) +   
geom\_bar() +   
labs(title = "Raspodela po godinama pojedinca")



Sa grafika uočavamo da je najveći broj ispitanika mlađi od 60 godina. Kako bi imali bolji uvid možemo ih podeliti na starosne grupe i prikazati tako.

age\_breaks = c(0,27,45,65,Inf)   
age\_labels = c('Under 27','27-45', '45-65', 'Above 65')   
cist\_dataset$AgeGroup = cut(cist\_dataset$age, breaks = age\_breaks, labels=age\_labels)   
  
ggplot(data=cist\_dataset, aes(x=AgeGroup)) +   
geom\_bar(position='dodge') +   
labs(title = "Raspodela godina po starosnim grupama") +   
theme(plot.title = element\_text(hjust = 0.5))



Najviše ispitanika pripada starosnoj grupi od 27 do 45, dok ih najmanje ima u grupi preko 65 godina.

## Fnlwgt

xtabs(~ fnlwgt, data = cist\_dataset)

## fnlwgt  
## 12285 13769 14878 18827 19214 19302 19395 19410 19491 19520   
## 1 1 1 1 1 5 2 1 1 1   
## 19700 19752 19793 19847 19899 19914 20057 20098 20101 20109   
## 1 1 1 2 1 4 2 1 1 1   
## 20179 20296 20308 20323 20333 20438 20469 20507 20511 20534   
## 1 1 1 1 1 1 2 1 1 3   
## 20676 20728 20795 20809 20953 20956 21095 21101 21154 21174   
## 1 1 4 1 1 2 1 1 1 1   
## 21306 21472 21626 21698 21792 21856 21876 21906 22042 22055   
## 1 1 2 3 1 1 1 1 1 1   
## 22154 22155 22186 22201 22211 22245 22313 22328 22418

unique\_fnlwgt = length(unique(cist\_dataset$fnlwgt))  
print(unique\_fnlwgt)

## [1] 21648

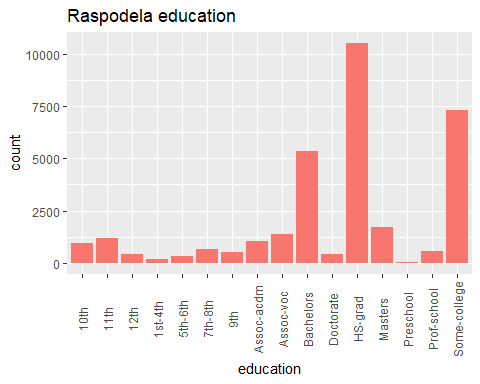
Postoji skoro 22k unique vrednosti, a ostale se ponavljaju manje od 10 puta tako da je raspodela prilično jednaka.

## Education

xtabs(~ education, data = cist\_dataset)

## education  
## 10th 11th 12th 1st-4th 5th-6th 7th-8th   
## 933 1175 433 168 333 646   
## 9th Assoc-acdm Assoc-voc Bachelors Doctorate HS-grad   
## 514 1067 1382 5355 413 10501   
## Masters Preschool Prof-school Some-college   
## 1723 51 576 7291

ggplot(data = cist\_dataset, mapping = aes(education, fill="red")) + geom\_bar() + labs(title="Raspodela education") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")



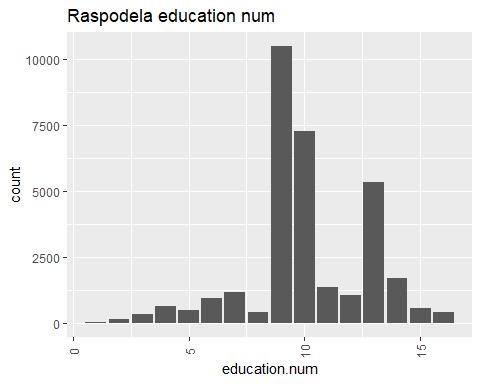
Kod najvećeg broja ispitanika stepen obrazovanja je HS-grad, Some-college i Bachelors.  
Ostale grupe su manje rasprostranjene pri čemu je Preschool u najmanjoj meri.

## Education.num

xtabs(~ education.num, data = cist\_dataset)

## education.num  
## 1 2 3 4 5 6 7 8 9 10 11 12 13   
## 51 168 333 646 514 933 1175 433 10501 7291 1382 1067 5355   
## 14 15 16   
## 1723 576 413

ggplot(data = cist\_dataset, mapping = aes(education.num)) + geom\_bar() + labs(title="Raspodela education num") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")



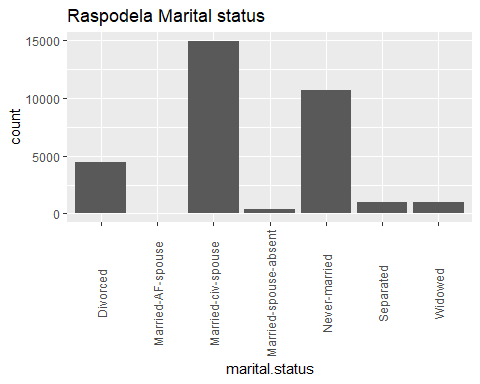
Kod najvećeg broja ispitanika education.num iznosi 9, 10 i 13.  
Ostale grupe su manje rasprostranjene.

## Marital.status

xtabs(~ marital.status, data = cist\_dataset)

## marital.status  
## Divorced Married-AF-spouse Married-civ-spouse   
## 4443 23 14976   
## Married-spouse-absent Never-married Separated   
## 418 10683 1025   
## Widowed   
## 993

ggplot(data = cist\_dataset, mapping = aes(marital.status)) + geom\_bar() + labs(title="Raspodela Marital status") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")



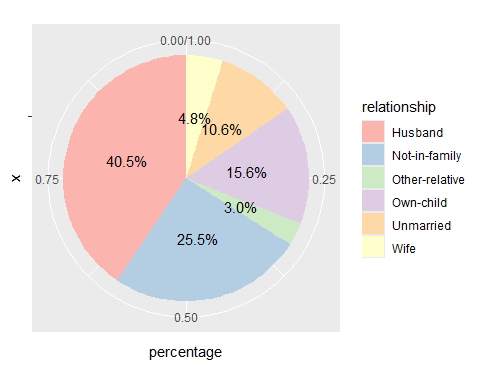
Najviše ima Married-civ-spouse, zatim, Never-married i Divorced. Ostale grupe nisu u tolikoj meri rasprostranjene.

## Relationship

xtabs(~ relationship, data = cist\_dataset)

## relationship  
## Husband Not-in-family Other-relative Own-child Unmarried   
## 13193 8305 981 5068 3446   
## Wife   
## 1568

RelationshipPercent = cist\_dataset %>% group\_by(relationship) %>% count() %>% ungroup() %>% mutate(percentage = `n`/sum(`n`)) %>% arrange(percentage) %>% mutate(p = scales::percent(percentage))  
  
ggplot(RelationshipPercent, aes(x = "", y = percentage, fill = relationship)) + geom\_col() + geom\_text(aes(label = p),position = position\_stack(vjust = 0.5)) + coord\_polar(theta = "y") + scale\_fill\_brewer(palette = "Pastel1")



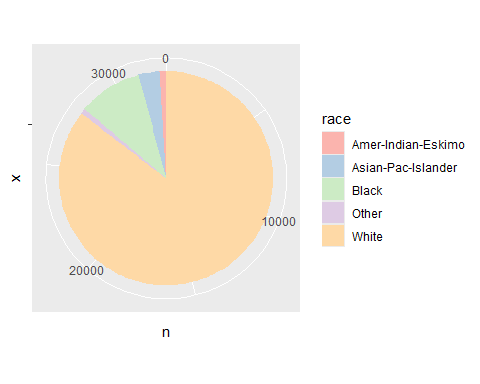
Što se tiče statusa veze u najvećem procentu su tu Husbands sa 40,5%, zatim Not-in-family od 25,5% i Own-child od 15,6%

## Race

xtabs(~ race, data = cist\_dataset)

## race  
## Amer-Indian-Eskimo Asian-Pac-Islander Black Other   
## 311 1039 3124 271   
## White   
## 27816

ggplot(cist\_dataset %>% group\_by(race) %>% count() %>% ungroup(), aes(x = "", y = n, fill = race)) +  
 geom\_col() +  
 coord\_polar(theta = "y") +  
 scale\_fill\_brewer(palette = "Pastel1")



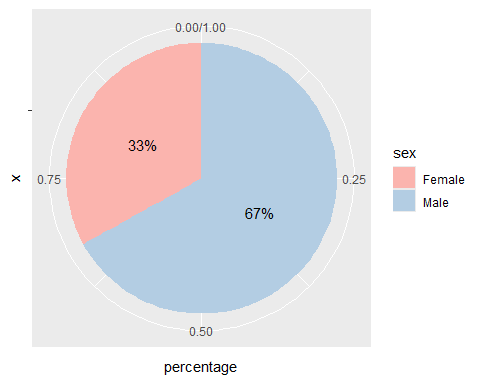
Ubedljivo najviše ima ispitanika koji su bele rase.

## Sex

xtabs(~ sex, data = cist\_dataset)

## sex  
## Female Male   
## 10771 21790

SexPercent = cist\_dataset %>% group\_by(sex) %>% count() %>% ungroup() %>% mutate(percentage = `n`/sum(`n`)) %>% arrange(percentage) %>%  
 mutate(p = scales::percent(percentage))  
  
ggplot(SexPercent, aes(x = "", y = percentage, fill = sex)) + geom\_col() + geom\_text(aes(label = p),position = position\_stack(vjust = 0.5)) + coord\_polar(theta = "y") + scale\_fill\_brewer(palette = "Pastel1")



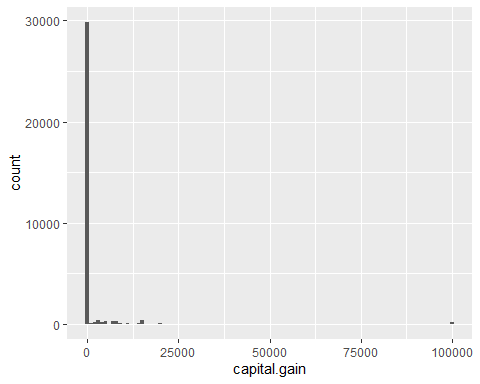
Procentualno je veći broj muškaraca od čak 67%, dok je 33% ženskih ispitanika.

## Capital.gain

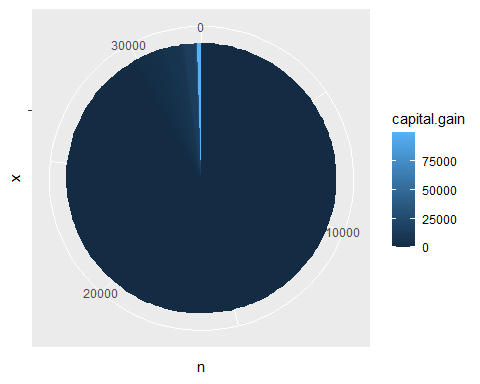
xtabs(~ capital.gain, data = cist\_dataset)

## capital.gain  
## 0 114 401 594 914 991 1055 1086 1111 1151 1173 1409 1424   
## 29849 6 2 34 8 5 25 4 1 8 3 7 3   
## 1455 1471 1506 1639 1797 1831 1848 2009 2036 2050 2062 2105 2174   
## 1 7 15 1 7 7 6 3 4 5 2 9 48   
## 2176 2202 2228 2290 2329 2346 2354 2387 2407 2414 2463 2538 2580   
## 23 16 5 5 6 6 11 1 19 8 11 1 12   
## 2597 2635 2653 2829 2885 2907 2936 2961 2964 2977 2993 3103 3137   
## 20 11 5 31 24 11 3 3 9 8 2 97 37   
## 3273 3325 3411 3418 3432 3456 3464 3471 3674 3781 3818 3887 3908   
## 6 53 24 5 4 2 23 8 14 12 7 6 32   
## 3942 4064 4101 4386 4416 4508 4650 4687 4787 4865 4931 4934 5013   
## 14 42 20 70 12 12 41 3 23 17 1 7 69   
## 5060 5178 5455 5556 5721 6097 6360 6418 6497 6514 6723 6767 6849   
## 1 97 11 5 3 1 3 9 11 5 2 5 27   
## 7298 7430 7443 7688 7896 7978 8614 9386 9562 10520 10566 10605 11678   
## 246 9 5 284 3 1 55 22 4 43 6 12 2   
## 13550 14084 14344 15020 15024 15831 18481 20051 22040 25124 25236 27828 34095   
## 27 41 26 5 347 6 2 37 1 4 11 34 5   
## 41310 99999   
## 2 159

ggplot(cist\_dataset, aes(x = capital.gain)) + geom\_histogram(binwidth=1000)



ggplot(cist\_dataset %>% group\_by(capital.gain) %>% count() %>% ungroup(), aes(x = "", y = n, fill = capital.gain)) +  
 geom\_col() +  
 coord\_polar(theta = "y")



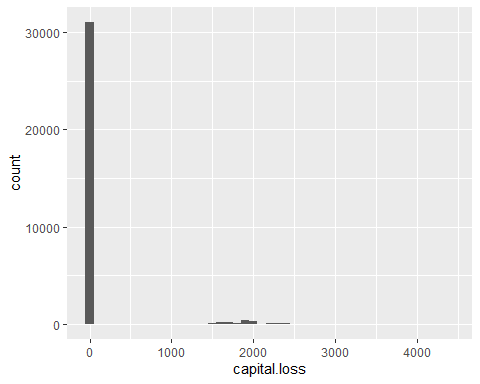
Vrednost 0 se pojavljuje skoro 30k puta što se može videti na oba grafika.

## Capital.loss

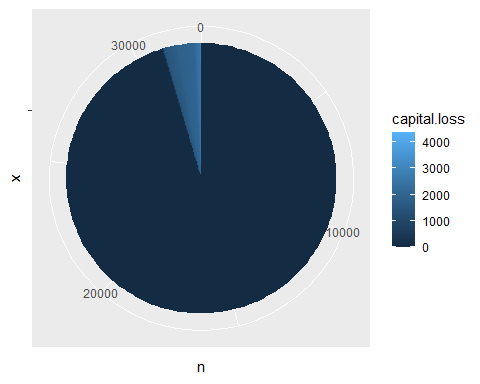
xtabs(~ capital.loss, data = cist\_dataset)

## capital.loss  
## 0 155 213 323 419 625 653 810 880 974 1092 1138 1258   
## 31042 1 4 3 3 12 3 2 6 2 7 2 4   
## 1340 1380 1408 1411 1485 1504 1539 1564 1573 1579 1590 1594 1602   
## 7 7 21 1 51 18 1 25 6 20 40 8 47   
## 1617 1628 1648 1651 1668 1669 1672 1719 1721 1726 1735 1740 1741   
## 9 15 2 9 4 24 34 22 18 4 2 42 24   
## 1755 1762 1816 1825 1844 1848 1876 1887 1902 1944 1974 1977 1980   
## 2 14 2 4 1 51 39 159 202 1 18 168 23   
## 2001 2002 2042 2051 2057 2080 2129 2149 2163 2174 2179 2201 2205   
## 24 21 9 21 6 1 3 2 1 7 15 1 9   
## 2206 2231 2238 2246 2258 2267 2282 2339 2352 2377 2392 2415 2444   
## 6 3 2 6 25 3 1 17 2 20 9 49 12   
## 2457 2467 2472 2489 2547 2559 2603 2754 2824 3004 3683 3770 3900   
## 3 1 1 1 4 12 5 2 10 2 2 2 2   
## 4356   
## 3

ggplot(cist\_dataset, aes(x = capital.loss)) + geom\_histogram(binwidth=100)



ggplot(cist\_dataset %>% group\_by(capital.loss) %>% count() %>% ungroup(), aes(x = "", y = n, fill = capital.loss)) +  
 geom\_col() +  
 coord\_polar(theta = "y")



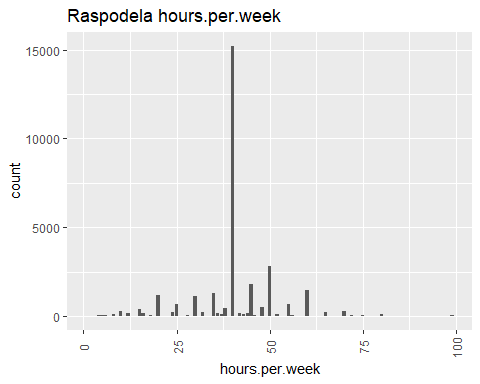
Najviše ima ispitanika sa vrednošću 0 za capital.loss. Takođe postoji mala zastupljenost oko vrednosti 2000.

## Hours.per.week

xtabs(~ hours.per.week, data = cist\_dataset)

## hours.per.week  
## 1 2 3 4 5 6 7 8 9 10 11 12 13   
## 20 32 39 54 60 64 26 145 18 278 11 173 23   
## 14 15 16 17 18 19 20 21 22 23 24 25 26   
## 34 404 205 29 75 14 1224 24 44 21 252 674 30   
## 27 28 29 30 31 32 33 34 35 36 37 38 39   
## 30 86 7 1149 5 266 39 28 1297 220 149 476 38   
## 40 41 42 43 44 45 46 47 48 49 50 51 52   
## 15217 36 219 151 212 1824 82 49 517 29 2819 13 138   
## 53 54 55 56 57 58 59 60 61 62 63 64 65   
## 25 41 694 97 17 28 5 1475 2 18 10 14 244   
## 66 67 68 70 72 73 74 75 76 77 78 80 81   
## 17 4 12 291 71 2 1 66 3 6 8 133 3   
## 82 84 85 86 87 88 89 90 91 92 94 95 96   
## 1 45 13 2 1 2 2 29 3 1 1 2 5   
## 97 98 99   
## 2 11 85

ggplot(data = cist\_dataset, mapping = aes(hours.per.week)) + geom\_bar() + labs(title="Raspodela hours.per.week") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5), legend.position = "none")



Preko 15k ispitanika sa 40 radnih sati nedeljno što je očekivano jer je to standardno radno vreme.

# Outliers

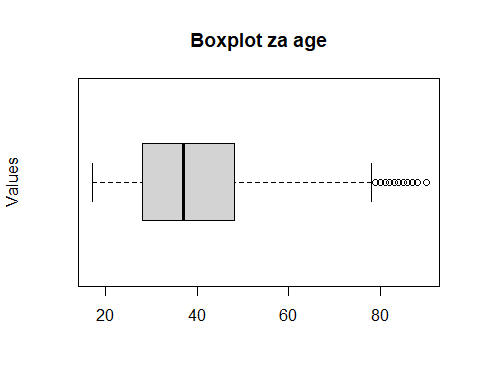
Na decision tree i na random forest outlieri nemaju neki uticaj.

Na logističku regresiju utiču outlieri slično kao za linearnu.

Prvo ćemo da vidimo da li su i u kojoj meri zastupljeni outlieri za numeričke kolone.

## Age

boxplot(cist\_dataset$age, main="Boxplot za age", ylab="Values", horizontal=TRUE)

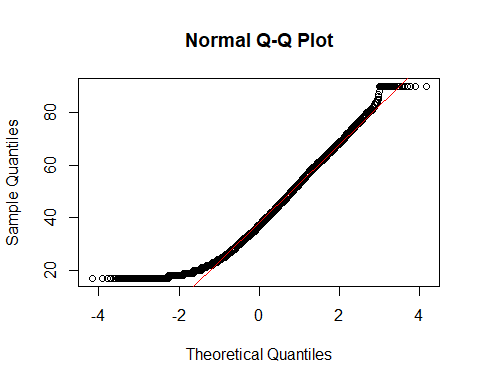


cist\_dataset %>% mutate(z\_score = scale(age)) %>% filter(abs(z\_score) > 3.5)

## # A tibble: 47 × 17  
## age workclass fnlwgt education education.num marital.status occupation  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 90 Private 77053 HS-grad 9 Widowed Craft-rep…  
## 2 90 Private 51744 HS-grad 9 Never-married Other-ser…  
## 3 90 Local-gov 227796 Masters 14 Married-civ-s… Exec-mana…  
## 4 90 Private 87372 Prof-sch… 15 Married-civ-s… Prof-spec…  
## 5 90 Self-emp-not-… 155981 Bachelors 13 Married-civ-s… Prof-spec…  
## 6 90 Private 46786 Bachelors 13 Married-civ-s… Sales   
## 7 90 Private 175491 HS-grad 9 Married-civ-s… Craft-rep…  
## 8 90 Local-gov 153602 HS-grad 9 Married-civ-s… Other-ser…  
## 9 90 Self-emp-not-… 82628 HS-grad 9 Never-married Exec-mana…  
## 10 90 Local-gov 214594 7th-8th 4 Married-civ-s… Protectiv…  
## # ℹ 37 more rows  
## # ℹ 10 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>, AgeGroup <fct>, z\_score <dbl[,1]>

Postoji 47 redova koji imaju z\_score preko 3.5 i to su svi iz AgeGroup-a preko 65, uglavnom 90 godina. Što se vidi i na boxplotu. Realno je da postoje osobe sa tim brojem godina pa ove vrednosti nećemo smatrati kao outliere.

qqnorm(cist\_dataset$age) #grafik normalnosti  
qqline(cist\_dataset$age, col="red") #linija se crta kroz tacke formirane prvim i trećim kvartilom Q1 i Q3

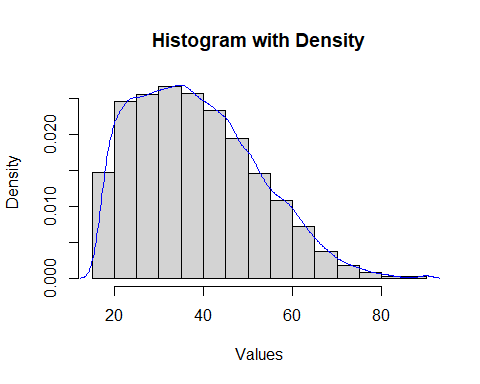


ad\_test = ad.test(cist\_dataset$age)  
print(ad\_test)

##   
## Anderson-Darling normality test  
##   
## data: cist\_dataset$age  
## A = 238.08, p-value < 2.2e-16

Podaci nisu u normalnoj raspodeli, to smo dokazali primenom Anderson-Darling testa normalnosti kao i grafikom normalnosti. Vrednost A ukazuje da je značajno odstupanje od normalnosti.

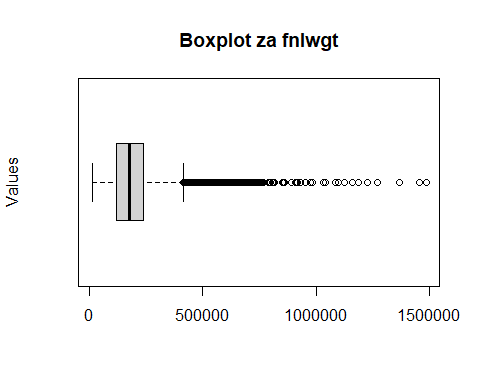
hist(cist\_dataset$age, probability=TRUE, main="Histogram with Density", xlab="Values")  
lines(density(cist\_dataset$age), col="blue") #histogrom sa poligonom frekvencija



Podaci su većinom grupisani između 20 i 50 godina.

## Fnlwgt

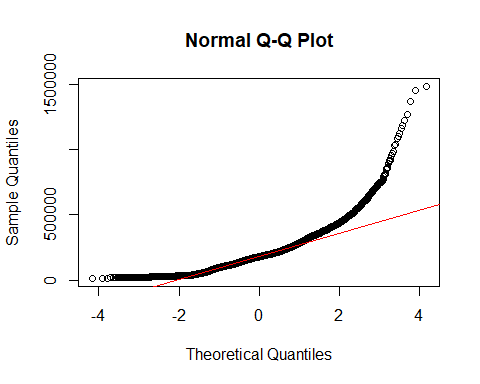
boxplot(cist\_dataset$fnlwgt, main="Boxplot za fnlwgt", ylab="Values", horizontal=TRUE)



cist\_dataset %>% mutate(z\_score = scale(fnlwgt)) %>% filter(abs(z\_score) > 3.5)

## # A tibble: 203 × 17  
## age workclass fnlwgt education education.num marital.status occupation  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 54 Private 816750 HS-grad 9 Married-civ-sp… Craft-rep…  
## 2 53 Private 608184 Bachelors 13 Married-civ-sp… Exec-mana…  
## 3 35 Local-gov 668319 Bachelors 13 Married-civ-sp… Exec-mana…  
## 4 40 Private 566537 Preschool 1 Married-civ-sp… Other-ser…  
## 5 31 Private 651396 HS-grad 9 Never-married Sales   
## 6 51 Private 673764 Masters 14 Never-married Prof-spec…  
## 7 37 Private 588003 Bachelors 13 Married-civ-sp… Exec-mana…  
## 8 36 State-gov 747719 Prof-school 15 Married-civ-sp… Prof-spec…  
## 9 35 Private 589809 Some-college 10 Never-married Exec-mana…  
## 10 31 Private 1033222 Bachelors 13 Never-married Adm-cleri…  
## # ℹ 193 more rows  
## # ℹ 10 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>, AgeGroup <fct>, z\_score <dbl[,1]>

#203 reda sa apsolutnim z\_score-om vecim od 3.5   
  
qqnorm(cist\_dataset$fnlwgt)  
qqline(cist\_dataset$fnlwgt, col="red")



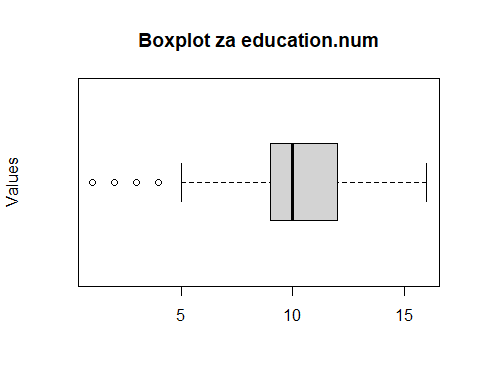
ad\_test = ad.test(cist\_dataset$fnlwgt)  
print(ad\_test)

##   
## Anderson-Darling normality test  
##   
## data: cist\_dataset$fnlwgt  
## A = 386.78, p-value < 2.2e-16

Podaci pokazuju značajno odstupanje od normalnosti, posebno u desnom repu. Ovo sugeriše da skup podataka može da sadrži outliere. Vrednost A ukazuje da je značajno odstupanje od normalnosti.

## Education.num

boxplot(cist\_dataset$education.num, main="Boxplot za education.num", ylab="Values", horizontal=TRUE)



Sa grafika vidimo da su za outliere uzete sve vrednosti ispod 5.

cist\_dataset %>% mutate(z\_score = scale(education.num)) %>% filter(abs(z\_score) > 3.5)

## # A tibble: 51 × 17  
## age workclass fnlwgt education education.num marital.status occupation  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 26 Private 322614 Preschool 1 Married-spouse-abs… Machine-o…  
## 2 40 Private 566537 Preschool 1 Married-civ-spouse Other-ser…  
## 3 27 Private 211032 Preschool 1 Married-civ-spouse Farming-f…  
## 4 32 Private 112137 Preschool 1 Married-civ-spouse Machine-o…  
## 5 53 Local-gov 140359 Preschool 1 Never-married Machine-o…  
## 6 51 Local-gov 241843 Preschool 1 Married-civ-spouse Other-ser…  
## 7 71 Private 235079 Preschool 1 Widowed Craft-rep…  
## 8 31 Private 452405 Preschool 1 Never-married Other-ser…  
## 9 33 Private 239781 Preschool 1 Married-civ-spouse Farming-f…  
## 10 39 Private 362685 Preschool 1 Widowed Other-ser…  
## # ℹ 41 more rows  
## # ℹ 10 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>, AgeGroup <fct>, z\_score <dbl[,1]>

#51 red sa apsolutnim z\_score-om vecim od 3.5 tj svi koji imaju vrednost 1

#Uglavnom ljudi koji nisu isli u skolu posle predskolskog, i to je moguce da od 32561 osoba njih 51 nije zavrsilo nikakvu skolu, ziveli pre toga na selu, dosli iz druge drzave gde im se ne priznaje, nisu hteli, nisu imali para. Po koloni occupation se vidi da su to uglavnom poslovi koji i ne zahtevaju neku skolu. Ali moramo izbaciti one kojima je workclass local-gov I State-gov jer je nemoguce da radi za vladu neko ko ima samo predskolsko obrazovanje, takodje i jedan slucaj iz occupation-a gde je necije zanimanje prof-specialty sto takodje nije realno

ggplot(subset(cist\_dataset, education.num == 1), aes(x = occupation)) +  
 geom\_bar() +  
 labs(title = "Occupation za one kojima je education.num = 1",  
 x = "Occupation",  
 y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

A graph of a graph with text

Description automatically generated with medium confidence

ggplot(subset(cist\_dataset, education.num == 1), aes(x = workclass)) +  
 geom\_bar() +  
 labs(title = "Workclass za one kojima je education.num = 1",  
 x = "Workclass",  
 y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

A graph of a graph with text

Description automatically generated with medium confidence

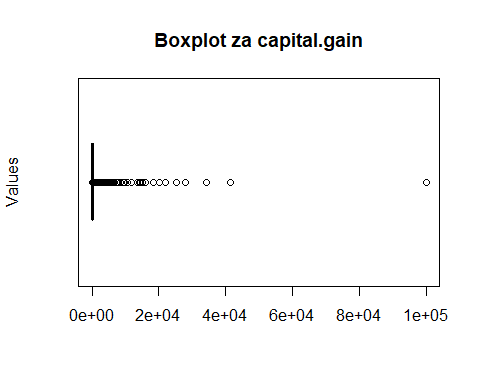
cist\_dataset = cist\_dataset %>% filter(!(education.num == 1 & (workclass == 'Local-gov' | workclass == 'State-gov' | occupation == 'Prof-specialty')))  
#6 ljudi da odstupa od realnih vrednosti  
  
ad\_test = ad.test(cist\_dataset$education.num)  
print(ad\_test)

##   
## Anderson-Darling normality test  
##   
## data: cist\_dataset$education.num  
## A = 1104.6, p-value < 2.2e-16

Podaci nisu u normalnoj raspodeli, dokazano korišćenjem Anderson-Darling testa normalnosti.

## Capital.gain

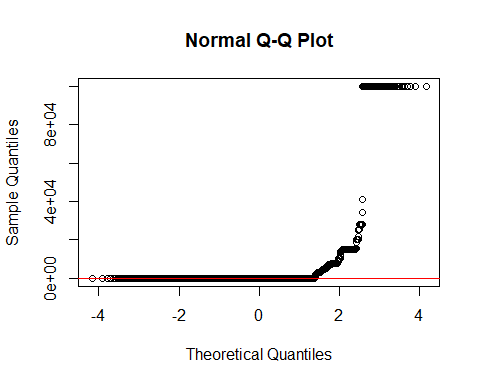
boxplot(cist\_dataset$capital.gain, main="Boxplot za capital.gain", ylab="Values", horizontal=TRUE)



cist\_dataset %>% mutate(z\_score = scale(capital.gain)) %>% filter(abs(z\_score) > 3.5)

## # A tibble: 200 × 17  
## age workclass fnlwgt education education.num marital.status occupation  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 54 Self-emp-inc 166459 Prof-sch… 15 Married-civ-s… Prof-spec…  
## 2 52 Private 152234 HS-grad 9 Married-civ-s… Exec-mana…  
## 3 53 Self-emp-inc 263925 HS-grad 9 Married-civ-s… Sales   
## 4 52 Private 118025 Bachelors 13 Married-civ-s… Exec-mana…  
## 5 46 Private 370119 Prof-sch… 15 Married-civ-s… Prof-spec…  
## 6 43 Private 176270 Bachelors 13 Married-civ-s… Exec-mana…  
## 7 49 Private 159816 Bachelors 13 Married-civ-s… Prof-spec…  
## 8 50 Private 171338 Some-col… 10 Married-civ-s… Exec-mana…  
## 9 22 Self-emp-not-… 202920 HS-grad 9 Never-married Prof-spec…  
## 10 43 Self-emp-inc 172826 Some-col… 10 Married-civ-s… Sales   
## # ℹ 190 more rows  
## # ℹ 10 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>, AgeGroup <fct>, z\_score <dbl[,1]>

#200 redova sa apsolutnim z\_score-om vecim od 3.5 realno je da osoba ima veliku kapitalnu dobit  
  
qqnorm(cist\_dataset$capital.gain)  
qqline(cist\_dataset$capital.gain, col="red")



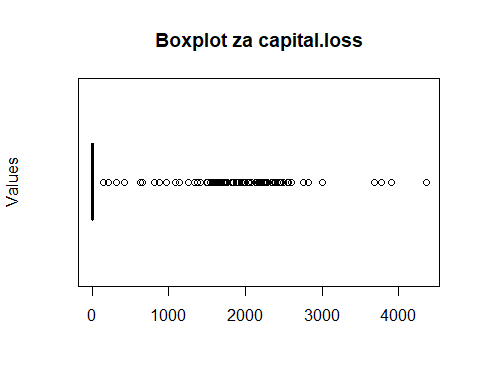
ad\_test = ad.test(cist\_dataset$capital.gain)  
print(ad\_test)

##   
## Anderson-Darling normality test  
##   
## data: cist\_dataset$capital.gain  
## A = 10460, p-value < 2.2e-16

Podaci nisu u normalnoj raspodeli.

## Capital.loss

boxplot(cist\_dataset$capital.loss, main="Boxplot za capital.loss", ylab="Values", horizontal=TRUE)

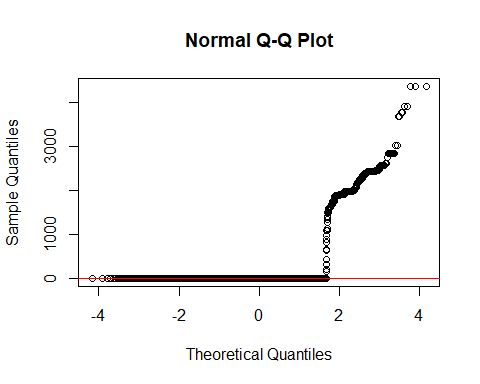


Podaci su u velikoj meri koncetrisani sa vrednošću 0. Pa svi podaci koji imaju vrednost veću od 0 su na grafiku prikazani kao outlieri.

cist\_dataset %>% mutate(z\_score = scale(capital.loss)) %>% filter(abs(z\_score) > 3.5)

## # A tibble: 1,383 × 17  
## age workclass fnlwgt education education.num marital.status occupation  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 90 Private 77053 HS-grad 9 Widowed Craft-rep…  
## 2 82 Private 132870 HS-grad 9 Widowed Exec-mana…  
## 3 66 Private 186061 Some-college 10 Widowed Adm-cleri…  
## 4 54 Private 140359 7th-8th 4 Divorced Machine-o…  
## 5 41 Private 264663 Some-college 10 Separated Prof-spec…  
## 6 34 Private 216864 HS-grad 9 Divorced Other-ser…  
## 7 38 Private 150601 10th 6 Separated Adm-cleri…  
## 8 74 State-gov 88638 Doctorate 16 Never-married Prof-spec…  
## 9 68 Federal-gov 422013 HS-grad 9 Divorced Prof-spec…  
## 10 41 Private 70037 Some-college 10 Never-married Craft-rep…  
## # ℹ 1,373 more rows  
## # ℹ 10 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>, AgeGroup <fct>, z\_score <dbl[,1]>

#1383 redova sa apsolutnim z\_score-om vecim od 3.5, realno je da ljudi imaju kapitalni gubitak  
  
qqnorm(cist\_dataset$capital.loss)  
qqline(cist\_dataset$capital.loss, col="red")



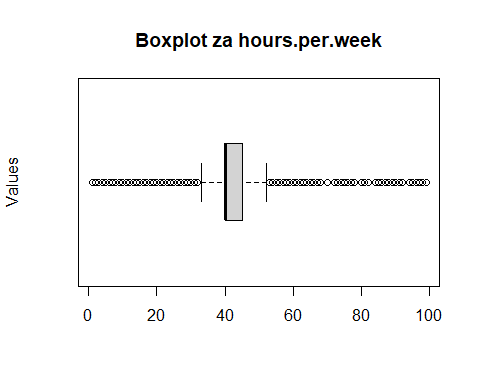
ad\_test = ad.test(cist\_dataset$capital.loss)  
print(ad\_test)

##   
## Anderson-Darling normality test  
##   
## data: cist\_dataset$capital.loss  
## A = 11654, p-value < 2.2e-16

Većina podataka je koncentrisana oko 0. Ali podaci nisu u normalnoj raspodeli.

## Hours.per.week

boxplot(cist\_dataset$hours.per.week, main="Boxplot za hours.per.week", ylab="Values", horizontal=TRUE)

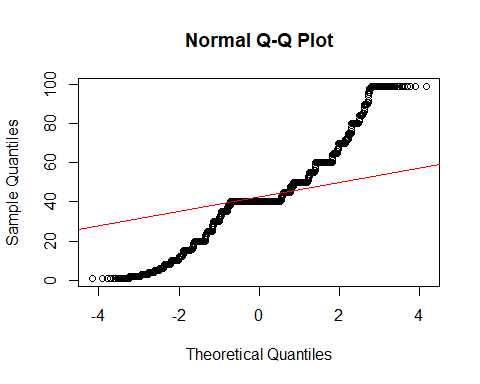


Podaci su u velikoj meri koncentrisani sa vrednošću 40.

cist\_dataset %>% mutate(z\_score = scale(hours.per.week)) %>% filter(abs(z\_score) > 3.5)

## # A tibble: 204 × 17  
## age workclass fnlwgt education education.num marital.status occupation  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 38 Private 175360 10th 6 Never-married Prof-spec…  
## 2 39 Private 348521 Some-colle… 10 Married-civ-s… Farming-f…  
## 3 39 Private 237713 Prof-school 15 Married-civ-s… Sales   
## 4 38 Private 187870 Prof-school 15 Married-civ-s… Prof-spec…  
## 5 33 Private 288825 HS-grad 9 Divorced Craft-rep…  
## 6 41 Self-emp-inc 139916 Assoc-voc 11 Married-civ-s… Sales   
## 7 38 Private 111499 HS-grad 9 Married-civ-s… Sales   
## 8 62 Private 71751 Some-colle… 10 Married-civ-s… Exec-mana…  
## 9 31 Private 147284 Doctorate 16 Married-civ-s… Prof-spec…  
## 10 43 Private 266324 HS-grad 9 Married-civ-s… Craft-rep…  
## # ℹ 194 more rows  
## # ℹ 10 more variables: relationship <chr>, race <chr>, sex <chr>,  
## # capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,  
## # native.country <chr>, income <chr>, AgeGroup <fct>, z\_score <dbl[,1]>

#204 reda sa apsolutnim z\_score-om vecim od 3.5 uglavnom svi koji imaju preko 84 sata rada. U Americi se sve preko 40 sati racuna kao prekovremeno zato ovi podaci deluju kao outlieri, ali je skroz realno da ljudi rade prekovremeno pa cak u nekim slucajevima i po 99 sati nedeljno.  
  
  
qqnorm(cist\_dataset$hours.per.week)  
qqline(cist\_dataset$hours.per.week, col="red")



ad\_test = ad.test(cist\_dataset$hours.per.week)  
print(ad\_test)

##   
## Anderson-Darling normality test  
##   
## data: cist\_dataset$hours.per.week  
## A = 1764.4, p-value < 2.2e-16

Podaci nisu u normalnoj raspodeli. Sa značajnim odstupanjem od normalnosti.

## Workclass, education, marital.status, sex, occupation, race, relationship, native.country

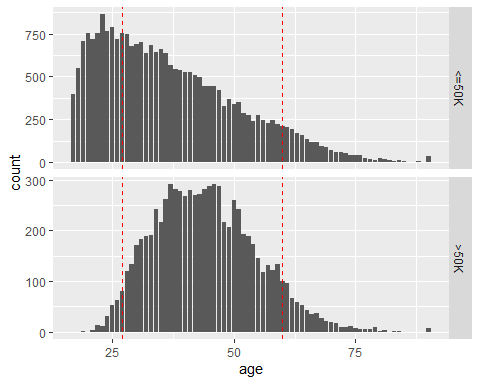
Outlieri su uglavnom povezani sa numeričkim podacima i označavaju neobično visoke ili niske vrednosti pa ih nećemo primenjivati na kategorijske kolone.

# Analiza

U nastavku ćemo prikazati vrednost income-a u odnosu na parametare koji opisuju jednog pojedinca. Income - da li pojedinac zaradjuje 0(<=50k) ili 1(>50k).

## Age

cist\_dataset %>%   
 ggplot() + aes(x = age) + geom\_bar() + geom\_vline(xintercept = c(27, 60), col = "red", linetype = "dashed") +  
 facet\_grid(income ~ ., scales = "free\_y")



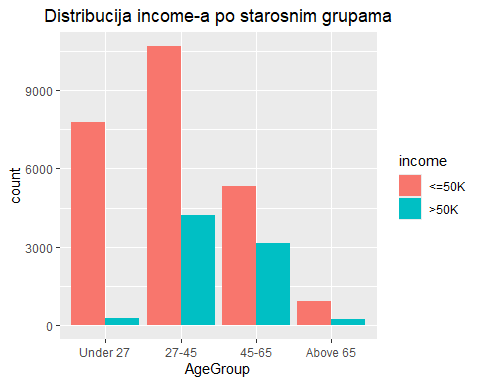
Sa grafika vidimo da >50k imaju u najvećem broju pojedinci između 27 i 60 godina. Dok za <=50k, pravilo je da linearno opada broj ljudi sa povećanjem godina počevši od 30.

ggplot(data=cist\_dataset,mapping = aes(x=age, fill=income )) +   
geom\_bar() +   
labs(title = "Distribucija income-a po godinama pojedinca") +   
theme(plot.title = element\_text(hjust = 0.5))

A graph of a number of people

Description automatically generated

ggplot(data=cist\_dataset, aes(x=AgeGroup,fill=income)) +   
geom\_bar(position='dodge') +   
labs(title = "Distribucija income-a po starosnim grupama") +   
theme(plot.title = element\_text(hjust = 0.5))



## Workclass

gg\_workclass = ggplot(cist\_dataset, aes(x = workclass, fill = factor(income))) +   
 geom\_bar(position = "dodge") +   
 xlab("Workclass") +   
 ylab("Count") +  
 scale\_fill\_discrete(name = "Yearly income over 50k", labels = c("No", "Yes")) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
gg\_workclass

A graph with different colored bars

Description automatically generated

Najviše ima pojedinaca kojima je workclass Private, shodno tome, najviše je njih i među onima koji zarađuju više od 50k i među onima koji ne zarađuju. Procentualno, najviše onih koji zarađuju preko 50k je iz grupe Self-emp-inc.

## Education

gg\_education = ggplot(cist\_dataset, aes(x = education, fill = factor(income))) +   
 geom\_bar() +   
 xlab("Education") +   
 ylab("Count") +  
 scale\_fill\_discrete(name = "Yearly income over 50k", labels = c("No", "Yes")) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
gg\_education

A graph of a number of people

Description automatically generated with medium confidence

Procentualno, najviše ima onih koji zarađuju preko 50k sa nivoom edukacije Doctorate, Prof-school i Masters. Brojčano ih najviše ima iz grupe Bachelors.

education.income = xtabs(~ education + income, data = cist\_dataset)  
education.income

## income  
## education <=50K >50K  
## 10th 871 62  
## 11th 1115 60  
## 12th 400 33  
## 1st-4th 162 6  
## 5th-6th 317 16  
## 7th-8th 606 40  
## 9th 487 27  
## Assoc-acdm 802 265  
## Assoc-voc 1021 361  
## Bachelors 3134 2221  
## Doctorate 107 306  
## HS-grad 8826 1675  
## Masters 764 959  
## Preschool 51 0  
## Prof-school 153 423  
## Some-college 5904 1387

education.income.prop = prop.table(education.income, margin = 1)  
education.income.prop

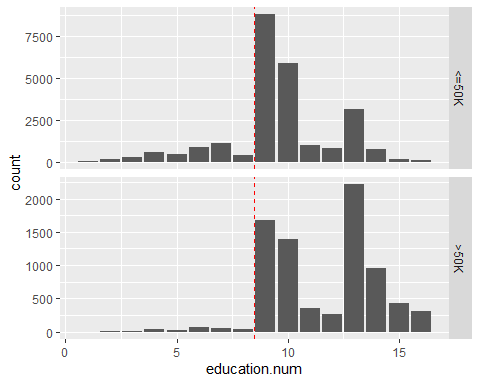
## income  
## education <=50K >50K  
## 10th 0.93354770 0.06645230  
## 11th 0.94893617 0.05106383  
## 12th 0.92378753 0.07621247  
## 1st-4th 0.96428571 0.03571429  
## 5th-6th 0.95195195 0.04804805  
## 7th-8th 0.93808050 0.06191950  
## 9th 0.94747082 0.05252918  
## Assoc-acdm 0.75164011 0.24835989  
## Assoc-voc 0.73878437 0.26121563  
## Bachelors 0.58524743 0.41475257  
## Doctorate 0.25907990 0.74092010  
## HS-grad 0.84049138 0.15950862  
## Masters 0.44341265 0.55658735  
## Preschool 1.00000000 0.00000000  
## Prof-school 0.26562500 0.73437500  
## Some-college 0.80976546 0.19023454

Najveće grupe za edukaciju koje zarađuju preko 50k procentualno su Doctorate(74%), Prof-school(73%) i Masters(56%), praćeni sa Bachelors(41%).

Sve ispod HS-grad jedva da ima neke brojke sa takvim pojedincima, uglavnom od 3-7%

## Education.num

cist\_dataset %>%   
 ggplot() + aes(x = education.num) + geom\_bar() + geom\_vline(xintercept = c(8.5), col = "red", linetype = "dashed") +  
 facet\_grid(income ~ ., scales = "free\_y")

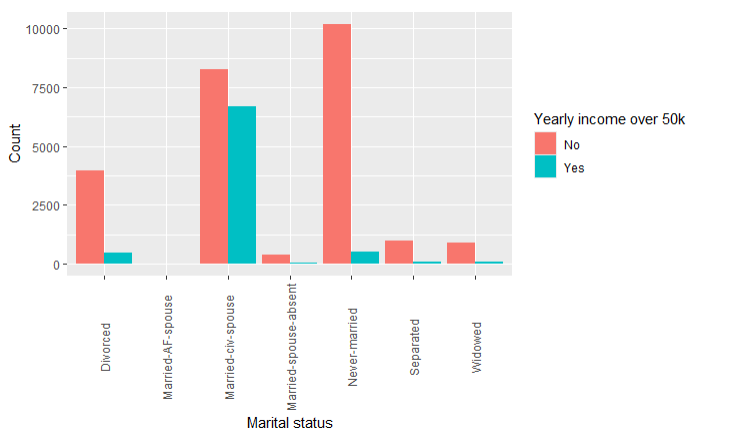


Malo je onih koji imaju education.num manji od 9 i oni u većini zarađuju manje od 50k.

Sa povećanjem education.num povećava se procentualno broj onih koji zarađuju više od 50k.

## Marital.status

gg\_marital.status = ggplot(cist\_dataset, aes(x = marital.status, fill = factor(income))) +   
 geom\_bar(position = "dodge") +   
 xlab("Marital status") +   
 ylab("Count") +  
 scale\_fill\_discrete(name = "Yearly income over 50k", labels = c("No", "Yes")) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
gg\_marital.status

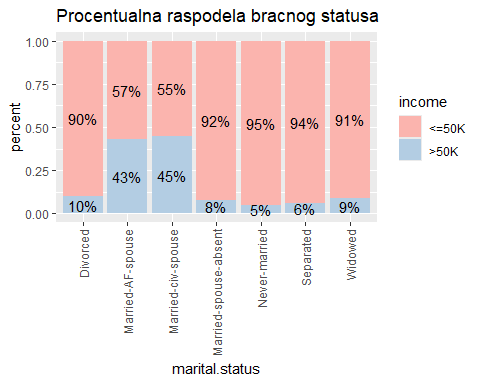


Married-civ-spouse i procentualno i brojčano najviše njih zarađuje preko 50k.  
Never-married u velikoj meri ne zarađuju 50k.

MarriedPercent = cist\_dataset %>% group\_by(marital.status,income) %>% summarise(n = n()) %>% mutate(percent = round(prop.table(n),2)) %>% arrange(percent) %>% mutate(p = scales::percent(percent))

## `summarise()` has grouped output by 'marital.status'. You can override using  
## the `.groups` argument.

ggplot(MarriedPercent, aes(x = marital.status, y = percent, fill = income)) +  
 geom\_col() +labs(title="Procentualna raspodela bracnog statusa") +  
 geom\_text(aes(label = p), position = position\_stack(vjust = 0.5)) +  
 scale\_fill\_brewer(palette = "Pastel1") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))



Married-AF-spuse i Married-civ-spouse u ove dve grupe income je skoro ravnomerno raspoređen na one koji zarađuju preko 50k i one koji zarađuju manje od 50k. Dok kod svih ostalih grupa preko 90% je onih koji zarađuju manje od 50k.

## Occupation

gg\_occupation = ggplot(cist\_dataset, aes(x = occupation, fill = factor(income))) +   
 geom\_bar(position = "dodge") +   
 xlab("Occupation") +   
 ylab("Count") +  
 scale\_fill\_discrete(name = "Yearly income over 50k", labels = c("No", "Yes")) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
gg\_occupation

A graph of different colored bars

Description automatically generated

Pojedinci iz grupa Exec-managerial i Prof-specialty imaju najveći procenat onih koji zarađuju više od 50k. S druge strane, Priv-house-serv, Armed-Forces, Adm-clerical i Other-service imaju procentualno najmanje onih koji zarađuju više od 50k.

## Relationship

gg\_relationship = ggplot(cist\_dataset, aes(x = relationship, fill = factor(income))) +   
 geom\_bar(position = "dodge") +   
 xlab("Relationship") +   
 ylab("Count") +  
 scale\_fill\_discrete(name = "Yearly income over 50k", labels = c("No", "Yes")) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
gg\_relationship

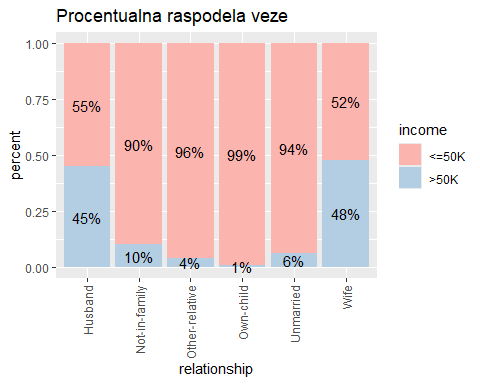
A graph of different colored bars

Description automatically generated

RelationshipPercent2 = cist\_dataset %>% group\_by(relationship,income) %>% summarise(n = n()) %>% mutate(percent = round(prop.table(n),2)) %>% arrange(percent) %>% mutate(p = scales::percent(percent))

## `summarise()` has grouped output by 'relationship'. You can override using the  
## `.groups` argument.

ggplot(RelationshipPercent2, aes(x = relationship, y = percent, fill = income)) +  
 geom\_col() + labs(title="Procentualna raspodela veze") +  
 geom\_text(aes(label = p), position = position\_stack(vjust = 0.5)) +  
 scale\_fill\_brewer(palette = "Pastel1") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))



Oni pojedinci koji su Husband i Wife maltene u podjednakoj meri pripadaju onima koji zarađuju više od 50k i onima koji zarađuju manje od 50k. S tim da je dosta više pojedinaca iz grupe Husband. Dok sa druge strane, osobe koje su Not-in-family, Other-service, Own-child i Unmarried preko 90% pripadaju onima koji zarađuju manje od 50k.

## Race

gg\_race = ggplot(cist\_dataset, aes(x = race, fill = factor(income))) +   
 geom\_bar(position = "dodge") +   
 xlab("Race") +   
 ylab("Count") +  
 scale\_fill\_discrete(name = "Yearly income over 50k", labels = c("No", "Yes")) +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))  
  
gg\_race

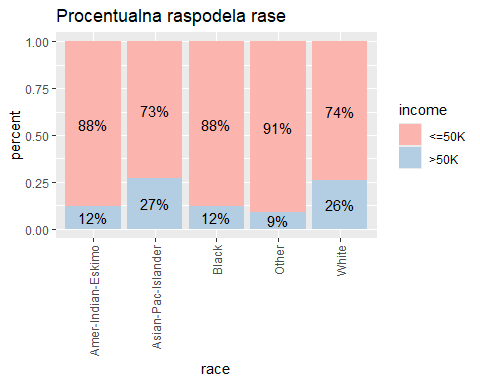
A graph with different colored squares

Description automatically generated

RacePercent1 = cist\_dataset %>% group\_by(race,income) %>% summarise(n = n()) %>% mutate(percent = round(prop.table(n),2)) %>% arrange(percent) %>% mutate(p = scales::percent(percent))

## `summarise()` has grouped output by 'race'. You can override using the  
## `.groups` argument.

ggplot(RacePercent1, aes(x = race, y = percent, fill = income)) +  
 geom\_col() +labs(title="Procentualna raspodela rase") +  
 geom\_text(aes(label = p), position = position\_stack(vjust = 0.5)) +  
 scale\_fill\_brewer(palette = "Pastel1") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))



race.income = xtabs(~ race + income, data = cist\_dataset)  
race.income

## income  
## race <=50K >50K  
## Amer-Indian-Eskimo 275 36  
## Asian-Pac-Islander 763 276  
## Black 2737 387  
## Other 246 25  
## White 20699 7117

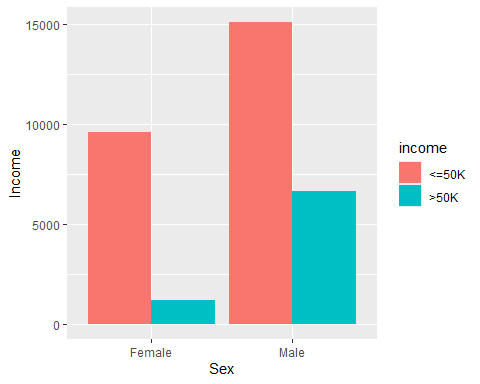
race.income.prop = prop.table(race.income, margin = 1)  
race.income.prop

## income  
## race <=50K >50K  
## Amer-Indian-Eskimo 0.88424437 0.11575563  
## Asian-Pac-Islander 0.73435996 0.26564004  
## Black 0.87612036 0.12387964  
## Other 0.90774908 0.09225092  
## White 0.74414006 0.25585994

Najviše pojedinaca po Race-u pripada grupi White. Oni uz grupu Asian-Pac-Islander imaju najveći procenat onih koji zarađuju više od 50k.

## Sex

gg\_sex <- ggplot(cist\_dataset, aes(x = factor(sex), fill = income)) +  
 geom\_bar(position = "dodge") +  
 ylab("Income") +  
 xlab("Sex")   
gg\_sex



Zaključujemo da su za >50k uglavnom zastupljeni muškarci, dok za <=50k, procentualno, više su zastupljene žene.

sex.income = xtabs(~ sex + income, data = cist\_dataset)  
sex.income

## income  
## sex <=50K >50K  
## Female 9592 1179  
## Male 15128 6662

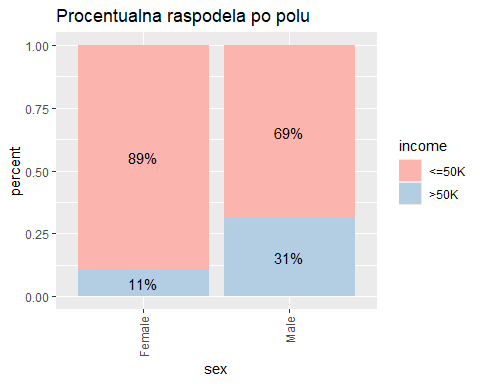
sex.income.prop = prop.table(sex.income, margin = 1)  
sex.income.prop

## income  
## sex <=50K >50K  
## Female 0.8905394 0.1094606  
## Male 0.6942634 0.3057366

SexPercent1 = cist\_dataset %>% group\_by(sex,income) %>% summarise(n = n()) %>% mutate(percent = round(prop.table(n),2)) %>% arrange(percent) %>% mutate(p = scales::percent(percent))

## `summarise()` has grouped output by 'sex'. You can override using the `.groups`  
## argument.

ggplot(SexPercent1, aes(x = sex, y = percent, fill = income)) +  
 geom\_col() +labs(title="Procentualna raspodela po polu") +  
 geom\_text(aes(label = p), position = position\_stack(vjust = 0.5)) +  
 scale\_fill\_brewer(palette = "Pastel1") +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))



## Capital.gain

gg\_capital.gain = ggplot(cist\_dataset, aes(x = capital.gain, fill=income )) +   
 geom\_histogram(binwidth=3000)   
  
gg\_capital.gain

A graph with numbers and lines

Description automatically generated

Pojedinci koji imaju capital.gain veći od 5000 uglavnom zarađuju više od 50k. Dok među onima kojima je capital.gain 0 procentualno su zastupljeniji oni koji zarađuju manje od 50k.

## Capital.loss

gg\_capital.loss = ggplot(cist\_dataset, aes(x = capital.loss, fill=income)) +   
 geom\_histogram(binwidth=500)   
  
gg\_capital.loss

A graph with numbers and lines

Description automatically generated

Pojedinci koji imaju capital.loss veći od 2000 uglavnom zarađuju više od 50k. Dok među onima kojima je capital. loss 0 procentualno su zastupljeniji oni koji zarađuju manje od 50k.

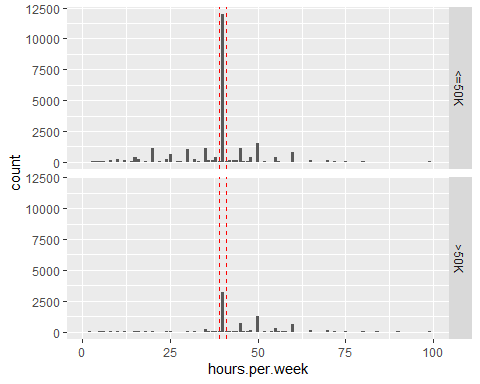
## Hours.per.week

ggplot(data=cist\_dataset,mapping = aes(x=hours.per.week, fill=income )) +   
geom\_bar() +   
labs(title = "Distribucija income-a po nedeljnim satima rada") +   
theme(plot.title = element\_text(hjust = 0.5))

A graph with red and blue lines

Description automatically generated

cist\_dataset %>%   
 ggplot() + aes(x = hours.per.week) + geom\_bar() + geom\_vline(xintercept = c(39, 41), col = "red", linetype = "dashed") +  
 facet\_grid(income ~ ., scales = "free\_x")



Očekivano, najveća grupa ljudi je ona koja radi 40h nedeljno pa je među njima i najviše ljudi koji zarađuju preko 50k tako i onih koji ne zaradjuju preko 50k.

## Native.country

gg\_native.country = ggplot(cist\_dataset, aes(x = native.country, fill = factor(income))) +   
 geom\_bar() +   
 xlab("native.country") +   
 ylab("Count") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
  
gg\_native.country

A graph with different colored bars

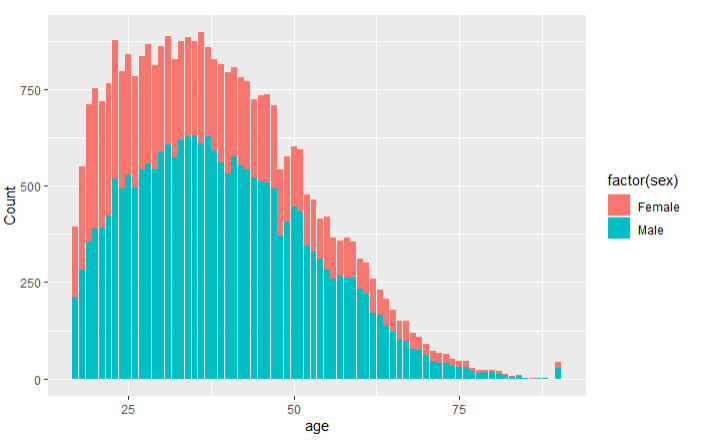
Description automatically generated

Očekivano, najviše ljudi je iz United-States pa je među njima i najviše ljudi koji zarađuju preko 50k tako i onih koji ne zaradjuju preko 50k.

# Analiza između više feature-a

## Age ~ sex

gg\_age\_sex = ggplot(cist\_dataset, aes(x = age, fill = factor(sex))) +   
 geom\_bar() +   
 xlab("age") +   
 ylab("Count")   
  
gg\_age\_sex



Sa povećanjem godina procentualno raste broj muškaraca u odnosu na broj žena. Sa manjim brojem godina ujednačenija je raspodela.

## Age ~ race

gg\_age\_race = ggplot(cist\_dataset, aes(x = age, fill = factor(race))) +   
 geom\_bar() +   
 xlab("age") +   
 ylab("Count")  
  
gg\_age\_race

A graph of a number of people

Description automatically generated

U svakoj od godina dominiraju ljudi bele rase, praćeni ljudima crne rase.

## Workclass ~ race

ggplot(cist\_dataset, aes(x = workclass, fill = factor(race))) +   
 geom\_bar() +  
 facet\_wrap(~ race) +  
 xlab("workclass") +  
 ylab("Count") +  
 theme\_minimal()+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

A graph with different colored bars

Description automatically generated with medium confidence

Za svaku rasu dominira “Private” sektor.

## Occupation ~ relationship.status

ggplot(cist\_dataset, aes(x = occupation, fill = factor(relationship))) +   
 geom\_bar() +  
 facet\_wrap(~ relationship) +  
 xlab("occupation") +  
 ylab("Count") +  
 theme\_minimal()+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

A graph of different colored bars

Description automatically generated

Brojčano je najviše Husband, a najmanje Other relative i Wife. Husband najviše rade u Craft-reapairu i exec-managerial. Wife, Own-Child i Unmarried najviše rade u Adm-clerical.

## Occupation ~ education

heatmap\_data = cist\_dataset %>%  
 count(occupation, education) %>%  
 na.omit()   
  
 ggplot(heatmap\_data, aes(x = education, y = occupation, fill = n)) +   
 geom\_tile() +  
 scale\_fill\_gradient(low = "beige", high = "red") +   
 xlab("education") +  
 ylab("Occupation") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

A screen shot of a graph

Description automatically generated

Najviše je onih koji imaju obrazovanje HS-grad i rade kao Craft-repair. Takođe, možemo da vidimo da oni koji imaju obrazovanje Bachleros se bave Prof-specialy i Exec-managerial. Dok oni koji završe Some-college rade kao Adm-clerical.

## Hours per week ~ sex

ggplot(cist\_dataset, aes(x = hours.per.week, fill = factor(sex))) +   
 geom\_bar() +   
 xlab("hours.per.week") +   
 ylab("Count")

A graph with red and blue lines

Description automatically generated

Sa povećanjem radnih sati uvećava se i zastupljenost muškaraca u odnosu na žene.

## Occupation ~ sex

ggplot(cist\_dataset, aes(x = occupation, fill = factor(sex))) +   
 geom\_bar() +   
 xlab("occupation") +   
 ylab("Count") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

A graph with different colored bars

Description automatically generated

Procentualno, najviše muškaraca se bavi Craft-repair, Farming-fishing, Handlers-cleaners, Protective-serv i Transport-moving. Dok se žene bave Adm-clerical, Priv-house-serv i Other-Service.

## Education ~ sex

ggplot(cist\_dataset, aes(x = education, fill = factor(sex))) +   
 geom\_bar() +   
 xlab("education") +   
 ylab("Count") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

A graph of a graph with different colored bars

Description automatically generated with medium confidence

U svakoj od edukacijskih grupa muškarci preovladavaju. Žena ima najviše u Some-college i HS-grad u odnosu na druge grupe.

## Workclass ~ sex

ggplot(cist\_dataset, aes(x = workclass, fill = factor(sex))) +   
 geom\_bar() +   
 xlab("workclass") +   
 ylab("Count") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

A graph of a bar chart

Description automatically generated

U svim kategorijama zaposlenja dominiraju muškarci, vredna pomena je kategorija “Private” u kojoj ima i dosta žena.

# Selekcija – Feature selection

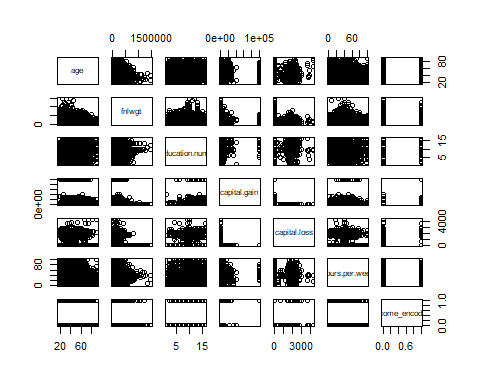
## Numericke kolone

Napravićemo novu kolonu income\_encoded koja će imati enkodirate vrednosti iz kolone income gde će <=50k uzimati vrednost 0, dok će >50k uzimati vrednost 1.

cist\_dataset$income\_encoded = ifelse(cist\_dataset$income == "<=50K", 0, 1)  
  
numeric\_columns = cist\_dataset %>% select(where(is.numeric))  
  
cor(numeric\_columns)

## age fnlwgt education.num capital.gain  
## age 1.00000000 -0.0766458679 0.03652719 0.0776744982  
## fnlwgt -0.07664587 1.0000000000 -0.04319463 0.0004318858  
## education.num 0.03652719 -0.0431946327 1.00000000 0.1226301147  
## capital.gain 0.07767450 0.0004318858 0.12263011 1.0000000000  
## capital.loss 0.05777454 -0.0102517117 0.07992296 -0.0316150630  
## hours.per.week 0.06875571 -0.0187684906 0.14812273 0.0784086154  
## income\_encoded 0.23403710 -0.0094625572 0.33515395 0.2233288182  
## capital.loss hours.per.week income\_encoded  
## age 0.05777454 0.06875571 0.234037103  
## fnlwgt -0.01025171 -0.01876849 -0.009462557  
## education.num 0.07992296 0.14812273 0.335153953  
## capital.gain -0.03161506 0.07840862 0.223328818  
## capital.loss 1.00000000 0.05425636 0.150526312  
## hours.per.week 0.05425636 1.00000000 0.229689066  
## income\_encoded 0.15052631 0.22968907 1.000000000

pairs(numeric\_columns) #grafici rasejanja za svaki moguci par varijabli

  
Pošto nam je prediktovana kolona binarna tj <=50k(0) ili >50k(1) linearna regresija nije najbolja opcija.

Korišćenjem logističkog regresionog modela trebalo bi da dobijemo bolje rezultate predikcije s obzirom da prediktujemo binarnu vrednost.

Na prvi pogled nema kolinearnosti što je veoma bitno za kreiranje modela logističke regresije jer bilo koje dve kolone nisu u linearnoj zavisnosti jedna od druge, pa ih ne moramo spajati ili brisati. Svakako ćemo u nastavku detaljnije proveriti multikolinearnost korišćenjem VIF-a.

# Modeli mašinskog učenja

Podelićemo ceo skup na test i train skupove u odnosu 80/20 gde ćemo 80% koristiti za treniranje, ostalih 20% ćemo koristiti za testiranje.

Izbrisaćemo kolonu income, ostavićemo enkodiranu kolonu gde će "<=50k" biti predstavljeno sa 0 dok će ">50k" biti predstavljeni brojem 1.

## Podela na train i test skupove

cist\_dataset = cist\_dataset %>% select(-income)  
  
set.seed(123)  
  
#stratifikacija na osnovu kolone native.country jer tu imamo najvise nivoa za faktore.  
train\_index = createDataPartition(cist\_dataset$native.country, p = 0.80, list = FALSE)

## Warning in createDataPartition(cist\_dataset$native.country, p = 0.8, list =  
## FALSE): Some classes have a single record ( Holand-Netherlands ) and these will  
## be selected for the sample

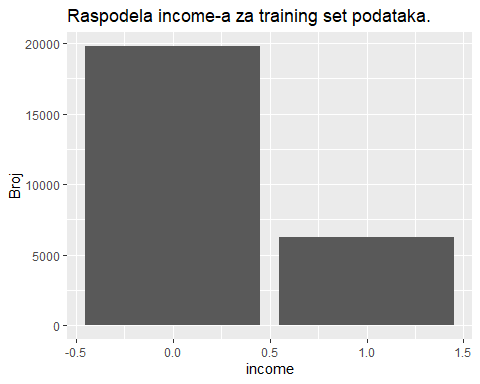
train = cist\_dataset[train\_index,]  
test = cist\_dataset[-train\_index,]

train = train[train$native.country != 'Holand-Netherlands', ]

Brišemo red sa vrednošću Holand-Netherlands jer postoji samo jedna vrednost i to u train skupu dok u test skupu ne postoji, a zbog različitih levela za svaku kolonu to nam pravi problem.

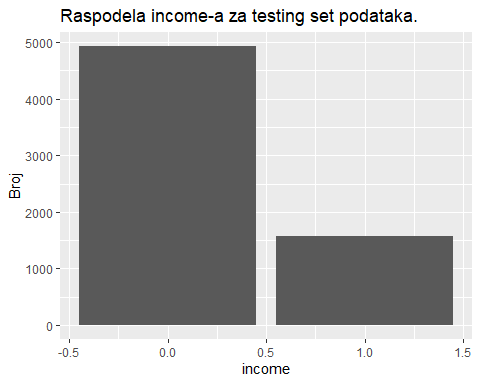
Raspodela income-a na train skup podataka.

gg\_train = ggplot(train, aes(x=income\_encoded)) +  
 geom\_bar() +  
 labs(title="Raspodela income-a za training set podataka.",  
 x="income",  
 y="Broj")  
  
print(gg\_train)



Raspodela income-a za test skup podataka.

gg\_train = ggplot(test, aes(x=income\_encoded)) +  
 geom\_bar() +  
 labs(title="Raspodela income-a za testing set podataka.",  
 x="income",  
 y="Broj")  
  
print(gg\_train)



Procentualna raspodela vrednosti za kolonu income\_encoded u train skupu podataka:

prop.table(table(train$income\_encoded))

##   
## 0 1   
## 0.7568397 0.2431603

Dobili smo da otprilike 76% zarađuje manje od 50k dok oko 24% zarađuje više. Iz toga znamo da nisu izbalansirani podaci.

## Resampling

#Random Over-Sampling Examples sluzi za generisanje sintetickih balansiranih skupova podataka iz nasih podataka  
  
proba = rpart(income\_encoded ~ ., data = train)  
  
proba\_prediction = predict(proba, newdata = test)  
  
accuracy.meas(test$income\_encoded,proba\_prediction[])

##   
## Call:   
## accuracy.meas(response = test$income\_encoded, predicted = proba\_prediction[])  
##   
## Examples are labelled as positive when predicted is greater than 0.5   
##   
## precision: 0.758  
## recall: 0.511  
## F: 0.305

Prag za vrednosti je 0.5 (Treshold)   
Precision nam je 0.758 što znači da su oko 75.8% pozitivnih predikcija našeg modela tačne. Pošto je vrednost između 60% i 80% ovo je ok vrednost.

Recall nam je 0.511 što znači da je oko 51% pozitivnih i manje-više isto toliko negativnih instanci. Pošto je vrednost manja od 60% ovo se uzima kao loša vredost.

F score = 0.305 što je relativno nisko pošto je manja od 0.5.

roc.curve(test$income\_encoded, proba\_prediction[], plotit = F)

## Area under the curve (AUC): 0.853

AUC vrednost od 0.853 označava da naš model nije loš, ali je svakako potrebno da se prvo izbalansira.

Raspodela vrednosti income\_encoded za train skup:

xtabs(~income\_encoded, data = train)

## income\_encoded  
## 0 1   
## 19724 6337

### Oversampling

set.seed(123)  
data\_oversampling = ovun.sample(income\_encoded ~ ., data = train, method = "over",N = 39448)$data  
  
table(data\_oversampling$income\_encoded)

##   
## 0 1   
## 19724 19724

### Undersampling

set.seed(123)  
data\_undersampling = ovun.sample(income\_encoded ~ ., data = train, method = "under", N = 12674, seed = 1)$data  
  
table(data\_undersampling$income\_encoded)

##   
## 0 1   
## 6337 6337

Podaci su sad balansirani. Ostaje nam da iskombinujemo ove dve vrednosti.

dim(train)

## [1] 26061 16

### Over i Under sampling

set.seed(123)  
data\_both = ovun.sample(income\_encoded ~ ., data = train, method  
= "both", p=0.5, N=26061, seed = 1)$data  
  
table(data\_both$income\_encoded)

##   
## 0 1   
## 13141 12920

### Data rose

data\_rose = ROSE(income\_encoded ~ ., data = train, seed = 1)$data  
  
table(data\_rose$income\_encoded)

##   
## 0 1   
## 13141 12920

Provera:

tree.rose = rpart(income\_encoded ~ ., data = data\_rose)  
tree.over = rpart(income\_encoded ~ ., data = data\_oversampling)  
tree.under = rpart(income\_encoded ~ ., data = data\_undersampling)  
tree.both = rpart(income\_encoded ~ ., data = data\_both)  
  
predict\_rose = predict(tree.rose, newdata = test)  
predict\_over = predict(tree.over, newdata = test)  
predict\_under = predict(tree.under, newdata = test)  
predict\_both = predict(tree.both, newdata = test)  
  
roc\_rose = roc(test$income\_encoded, predict\_rose)

roc\_over = roc(test$income\_encoded, predict\_over)

roc\_under = roc(test$income\_encoded, predict\_under)

roc\_both = roc(test$income\_encoded, predict\_both)

plot(roc\_rose, col = "blue", main = "ROC Curve Comparison")

lines(roc\_over, col = "red")

lines(roc\_under, col = "green")

lines(roc\_both, col = "purple")

legend("bottomright", legend = c("ROSE", "Oversampling", "Undersampling", "  
Both"),  
 col = c("blue", "red", "green", "purple"), lty = 1)

A graph of a number of different colored lines

Description automatically generated with medium confidence  
Ne vidimo Oversampling liniju zato sto je skoro ista kao i Undersampling linija.

print(roc\_rose)

print(roc\_over)

print(roc\_under)

print(roc\_both)

rose (AUC): 0.7968  
Oversampling (AUC): 0.8738   
Undersampling (AUC): 0.8597   
Both (AUC): 0.8726

Najbolji rezultat dobijamo oversampling metodom.

X\_train = train[, -ncol(train)]  
Y\_train = train[, ncol(train)]  
  
data\_oversampled = ROSE(income\_encoded ~ ., data = train, seed = 1)$data  
  
X\_oversampled = data\_oversampled[, -ncol(data\_oversampled)]  
Y\_oversampled = data\_oversampled[, ncol(data\_oversampled)]

### F-Regression

X\_train = X\_oversampled  
Y\_train = Y\_oversampled  
  
fr\_model = lm(Y\_train ~ ., data = X\_train)  
  
f\_regression = summary(fr\_model)$fstatistic  
  
p\_values = pf(f\_regression[1], f\_regression[2], f\_regression[3], lower.tail = FALSE)  
  
significant\_features = names(cist\_dataset)[p\_values < 0.05]  
significant\_features

## [1] "age" "workclass" "fnlwgt" "education"   
## [5] "education.num" "marital.status" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hours.per.week" "native.country" "AgeGroup" "income\_encoded"

Iako se f-regression inače koristi uz linearnu regresiju, hteli smo da proverimo značaj feature-a ovog modela. Na osnovu rezultata zaključujemo da sve kolone možemo uzeti u obzir.

## Logistička regresija

Logistička regresija se koristi za modelovanje verovatnoće da se dogodi određeni događaj koji ima binarni izlaz (kod nas income\_encoded ima izlaz 0 ili 1).   
Binomijalna raspodela se koristi za modeliranje slučajeva gde se događaji mogu podeliti u dve kategorije (uspeh i neuspeh).

model3 = glm(formula = Y\_train ~ ., data.frame(X\_train, Y\_train), family = "binomial")  
  
summary(model3)

##   
## Call:  
## glm(formula = Y\_train ~ ., family = "binomial", data = data.frame(X\_train,   
## Y\_train))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -4.858e+00 8.852e-01 -5.488 4.07e-08  
## age 1.429e-02 2.634e-03 5.425 5.81e-08  
## workclassLocal-gov -5.419e-01 1.167e-01 -4.642 3.44e-06  
## workclassNever-worked -1.342e+01 3.184e+02 -0.042 0.966386  
## workclassPrivate -4.546e-01 9.699e-02 -4.687 2.77e-06  
## workclassSelf-emp-inc -2.280e-01 1.298e-01 -1.757 0.078914  
## workclassSelf-emp-not-inc -8.022e-01 1.122e-01 -7.147 8.84e-13  
## workclassState-gov -8.939e-01 1.302e-01 -6.867 6.55e-12  
## workclassWithout-pay -1.389e+01 1.955e+02 -0.071 0.943350  
## fnlwgt 3.158e-07 1.632e-07 1.935 0.053044  
## education11th 2.617e-01 1.910e-01 1.370 0.170698  
## education12th 5.609e-01 2.403e-01 2.334 0.019598  
## education1st-4th -8.207e-01 4.106e-01 -1.999 0.045651  
## education5th-6th -9.690e-02 2.945e-01 -0.329 0.742109  
## education7th-8th -4.937e-01 2.048e-01 -2.411 0.015907  
## education9th 2.043e-01 2.134e-01 0.958 0.338225  
## educationAssoc-acdm 1.231e+00 2.011e-01 6.124 9.12e-10  
## educationAssoc-voc 1.382e+00 1.838e-01 7.521 5.45e-14  
## educationBachelors 1.748e+00 1.981e-01 8.825 < 2e-16  
## educationDoctorate 3.012e+00 2.984e-01 10.092 < 2e-16  
## educationHS-grad 7.614e-01 1.488e-01 5.116 3.12e-07  
## educationMasters 2.019e+00 2.218e-01 9.103 < 2e-16  
## educationPreschool -1.146e+01 1.599e+02 -0.072 0.942860  
## educationProf-school 2.645e+00 2.688e-01 9.840 < 2e-16  
## educationSome-college 1.050e+00 1.603e-01 6.550 5.77e-11  
## education.num 1.254e-02 1.968e-02 0.637 0.523988  
## marital.statusMarried-AF-spouse 3.916e+00 7.351e-01 5.327 9.98e-08  
## marital.statusMarried-civ-spouse 2.275e+00 2.374e-01 9.582 < 2e-16  
## marital.statusMarried-spouse-absent 3.310e-02 1.894e-01 0.175 0.861277  
## marital.statusNever-married -3.607e-01 7.556e-02 -4.774 1.80e-06  
## marital.statusSeparated -3.680e-01 1.424e-01 -2.584 0.009760  
## marital.statusWidowed 1.688e-01 1.348e-01 1.252 0.210524  
## occupationArmed-Forces -1.234e+01 2.985e+02 -0.041 0.967033  
## occupationCraft-repair -2.108e-01 7.427e-02 -2.838 0.004539  
## occupationExec-managerial 7.290e-01 7.518e-02 9.697 < 2e-16  
## occupationFarming-fishing -8.999e-01 1.253e-01 -7.180 6.99e-13  
## occupationHandlers-cleaners -8.234e-01 1.338e-01 -6.154 7.55e-10  
## occupationMachine-op-inspct -4.616e-01 9.566e-02 -4.825 1.40e-06  
## occupationOther-service -1.023e+00 1.033e-01 -9.902 < 2e-16  
## occupationPriv-house-serv -2.190e+00 8.200e-01 -2.671 0.007572  
## occupationProf-specialty 4.962e-01 7.854e-02 6.318 2.64e-10  
## occupationProtective-serv 2.634e-01 1.260e-01 2.090 0.036602  
## occupationSales 2.255e-01 7.892e-02 2.858 0.004262  
## occupationTech-support 4.432e-01 1.140e-01 3.887 0.000101  
## occupationTransport-moving -1.868e-01 9.345e-02 -1.998 0.045667  
## relationshipNot-in-family 7.802e-01 2.372e-01 3.289 0.001006  
## relationshipOther-relative -4.310e-01 2.278e-01 -1.892 0.058511  
## relationshipOwn-child -3.931e-01 2.339e-01 -1.681 0.092834  
## relationshipUnmarried 5.058e-01 2.504e-01 2.020 0.043377  
## relationshipWife 1.445e+00 9.752e-02 14.814 < 2e-16  
## raceAsian-Pac-Islander 4.948e-01 2.371e-01 2.087 0.036883  
## raceBlack 3.891e-01 1.930e-01 2.016 0.043779  
## raceOther 6.025e-02 3.340e-01 0.180 0.856872  
## raceWhite 4.638e-01 1.817e-01 2.553 0.010693  
## sexMale 1.056e+00 6.811e-02 15.500 < 2e-16  
## capital.gain 7.681e-05 3.370e-06 22.791 < 2e-16  
## capital.loss 4.453e-04 3.671e-05 12.131 < 2e-16  
## hours.per.week 2.529e-02 1.571e-03 16.099 < 2e-16  
## native.countryCanada -1.988e+00 8.228e-01 -2.416 0.015689  
## native.countryChina -2.632e+00 8.431e-01 -3.121 0.001802  
## native.countryColumbia -4.191e+00 1.112e+00 -3.770 0.000164  
## native.countryCuba -1.813e+00 8.370e-01 -2.166 0.030340  
## native.countryDominican-Republic -3.881e+00 1.022e+00 -3.796 0.000147  
## native.countryEcuador -1.525e+00 1.040e+00 -1.467 0.142479  
## native.countryEl-Salvador -3.207e+00 8.924e-01 -3.594 0.000326  
## native.countryEngland -1.725e+00 8.479e-01 -2.035 0.041854  
## native.countryFrance -6.531e-01 1.079e+00 -0.605 0.545104  
## native.countryGermany -1.777e+00 8.288e-01 -2.145 0.031987  
## native.countryGreece -2.002e+00 9.744e-01 -2.055 0.039920  
## native.countryGuatemala -1.604e+00 1.045e+00 -1.535 0.124707  
## native.countryHaiti -1.720e+00 9.264e-01 -1.857 0.063364  
## native.countryHonduras -4.013e+00 1.970e+00 -2.037 0.041685  
## native.countryHong -1.854e+00 1.011e+00 -1.833 0.066756  
## native.countryHungary -2.619e+00 1.112e+00 -2.355 0.018516  
## native.countryIndia -3.138e+00 8.313e-01 -3.775 0.000160  
## native.countryIran -2.007e+00 9.647e-01 -2.081 0.037454  
## native.countryIreland -1.315e+00 1.008e+00 -1.305 0.192042  
## native.countryItaly -1.020e+00 8.703e-01 -1.171 0.241399  
## native.countryJamaica -2.875e+00 8.920e-01 -3.224 0.001266  
## native.countryJapan -1.700e+00 8.694e-01 -1.955 0.050545  
## native.countryLaos -1.306e+00 1.107e+00 -1.179 0.238314  
## native.countryMexico -2.710e+00 8.065e-01 -3.361 0.000777  
## native.countryNicaragua -2.624e+00 1.121e+00 -2.340 0.019282  
## native.countryOutlying-US(Guam-USVI-etc) -1.622e+01 3.190e+02 -0.051 0.959451  
## native.countryPeru -2.867e+00 1.063e+00 -2.696 0.007022  
## native.countryPhilippines -1.348e+00 8.011e-01 -1.682 0.092498  
## native.countryPoland -2.013e+00 8.813e-01 -2.284 0.022383  
## native.countryPortugal -2.409e+00 1.121e+00 -2.150 0.031549  
## native.countryPuerto-Rico -1.653e+00 8.669e-01 -1.907 0.056487  
## native.countryScotland -3.209e+00 1.219e+00 -2.632 0.008476  
## native.countrySouth -2.799e+00 8.568e-01 -3.266 0.001089  
## native.countryTaiwan -3.075e+00 8.809e-01 -3.490 0.000482  
## native.countryThailand -3.532e+00 1.319e+00 -2.678 0.007416  
## native.countryTrinadad&Tobago -3.473e+00 1.156e+00 -3.005 0.002656  
## native.countryUnited-States -2.010e+00 7.833e-01 -2.566 0.010301  
## native.countryVietnam -3.018e+00 9.116e-01 -3.311 0.000929  
## native.countryYugoslavia -7.155e-01 9.663e-01 -0.740 0.459046  
## AgeGroup27-45 1.286e+00 7.755e-02 16.582 < 2e-16  
## AgeGroup45-65 1.481e+00 1.041e-01 14.226 < 2e-16  
## AgeGroupAbove 65 6.167e-01 1.682e-01 3.668 0.000245  
##   
## (Intercept) \*\*\*  
## age \*\*\*  
## workclassLocal-gov \*\*\*  
## workclassNever-worked   
## workclassPrivate \*\*\*  
## workclassSelf-emp-inc .   
## workclassSelf-emp-not-inc \*\*\*  
## workclassState-gov \*\*\*  
## workclassWithout-pay   
## fnlwgt .   
## education11th   
## education12th \*   
## education1st-4th \*   
## education5th-6th   
## education7th-8th \*   
## education9th   
## educationAssoc-acdm \*\*\*  
## educationAssoc-voc \*\*\*  
## educationBachelors \*\*\*  
## educationDoctorate \*\*\*  
## educationHS-grad \*\*\*  
## educationMasters \*\*\*  
## educationPreschool   
## educationProf-school \*\*\*  
## educationSome-college \*\*\*  
## education.num   
## marital.statusMarried-AF-spouse \*\*\*  
## marital.statusMarried-civ-spouse \*\*\*  
## marital.statusMarried-spouse-absent   
## marital.statusNever-married \*\*\*  
## marital.statusSeparated \*\*   
## marital.statusWidowed   
## occupationArmed-Forces   
## occupationCraft-repair \*\*   
## occupationExec-managerial \*\*\*  
## occupationFarming-fishing \*\*\*  
## occupationHandlers-cleaners \*\*\*  
## occupationMachine-op-inspct \*\*\*  
## occupationOther-service \*\*\*  
## occupationPriv-house-serv \*\*   
## occupationProf-specialty \*\*\*  
## occupationProtective-serv \*   
## occupationSales \*\*   
## occupationTech-support \*\*\*  
## occupationTransport-moving \*   
## relationshipNot-in-family \*\*   
## relationshipOther-relative .   
## relationshipOwn-child .   
## relationshipUnmarried \*   
## relationshipWife \*\*\*  
## raceAsian-Pac-Islander \*   
## raceBlack \*   
## raceOther   
## raceWhite \*   
## sexMale \*\*\*  
## capital.gain \*\*\*  
## capital.loss \*\*\*  
## hours.per.week \*\*\*  
## native.countryCanada \*   
## native.countryChina \*\*   
## native.countryColumbia \*\*\*  
## native.countryCuba \*   
## native.countryDominican-Republic \*\*\*  
## native.countryEcuador   
## native.countryEl-Salvador \*\*\*  
## native.countryEngland \*   
## native.countryFrance   
## native.countryGermany \*   
## native.countryGreece \*   
## native.countryGuatemala   
## native.countryHaiti .   
## native.countryHonduras \*   
## native.countryHong .   
## native.countryHungary \*   
## native.countryIndia \*\*\*  
## native.countryIran \*   
## native.countryIreland   
## native.countryItaly   
## native.countryJamaica \*\*   
## native.countryJapan .   
## native.countryLaos   
## native.countryMexico \*\*\*  
## native.countryNicaragua \*   
## native.countryOutlying-US(Guam-USVI-etc)   
## native.countryPeru \*\*   
## native.countryPhilippines .   
## native.countryPoland \*   
## native.countryPortugal \*   
## native.countryPuerto-Rico .   
## native.countryScotland \*\*   
## native.countrySouth \*\*   
## native.countryTaiwan \*\*\*  
## native.countryThailand \*\*   
## native.countryTrinadad&Tobago \*\*   
## native.countryUnited-States \*   
## native.countryVietnam \*\*\*  
## native.countryYugoslavia   
## AgeGroup27-45 \*\*\*  
## AgeGroup45-65 \*\*\*  
## AgeGroupAbove 65 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 36126 on 26060 degrees of freedom  
## Residual deviance: 20394 on 25961 degrees of freedom  
## AIC: 20594  
##   
## Number of Fisher Scoring iterations: 13

Anova(model3, type = "III")

## Analysis of Deviance Table (Type III tests)  
##   
## Response: Y\_train  
## LR Chisq Df Pr(>Chisq)   
## age 29.51 1 5.565e-08 \*\*\*  
## workclass 98.53 7 < 2.2e-16 \*\*\*  
## fnlwgt 3.74 1 0.05301 .   
## education 169.36 15 < 2.2e-16 \*\*\*  
## education.num 0.41 1 0.52398   
## marital.status 149.40 6 < 2.2e-16 \*\*\*  
## occupation 621.48 13 < 2.2e-16 \*\*\*  
## relationship 382.14 5 < 2.2e-16 \*\*\*  
## race 9.73 4 0.04523 \*   
## sex 248.80 1 < 2.2e-16 \*\*\*  
## capital.gain 736.46 1 < 2.2e-16 \*\*\*  
## capital.loss 157.05 1 < 2.2e-16 \*\*\*  
## hours.per.week 267.96 1 < 2.2e-16 \*\*\*  
## native.country 134.40 39 2.069e-12 \*\*\*  
## AgeGroup 413.35 3 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Model koji uključuje sve kolone ima AIC vrednost 20594. Za kolonu education.num, fnlwgt i race smo dobili da je p-vrednost veća od 0.05 pa njih verovatno nećemo da uzimamo u obzir.

prediction\_model3 = predict(model3, test, type="response")   
  
roc\_curve3 = roc(test$income\_encoded, prediction\_model3)

plot(roc\_curve3)

auc(roc\_curve3)

optimal\_treshold3 = coords(roc\_curve3, "best",best.method="youden")$threshold

print(optimal\_treshold3)

conf\_matrix\_model3 = confusionMatrix(table(ifelse(prediction\_model3 > 0.71, 1, 0), test$income\_encoded))  
conf\_matrix\_model3

A graph with a curve

Description automatically generated

Area under the curve: 0.8988

[1] 0.4983544

## Confusion Matrix and Statistics  
##   
##   
## 0 1  
## 0 4493 545  
## 1 496 959  
##   
## Accuracy : 0.8397   
## 95% CI : (0.8305, 0.8485)  
## No Information Rate : 0.7684   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.5444   
##   
## Mcnemar's Test P-Value : 0.1368   
##   
## Sensitivity : 0.9006   
## Specificity : 0.6376   
## Pos Pred Value : 0.8918   
## Neg Pred Value : 0.6591   
## Prevalence : 0.7684   
## Detection Rate : 0.6920   
## Detection Prevalence : 0.7759   
## Balanced Accuracy : 0.7691   
##   
## 'Positive' Class : 0   
##

print(conf\_matrix\_model3$overall["Accuracy"])

## Accuracy   
## 0.8396735

print(conf\_matrix\_model3$byClass["Pos Pred Value"])

## Pos Pred Value   
## 0.8918222

print(paste(round(conf\_matrix\_model3$byClass["Sensitivity"], 2)))

## [1] "0.9"

print(paste(round(conf\_matrix\_model3$byClass["F1"], 2)))

## [1] "0.9"

Odds ratio nam pomaže da protumačimo važnost svakog feature-a u modelu.

odds\_ratios = exp(coef(model3))  
  
# konvertujemo odds ratios u data frame  
odds\_ratios\_df = data.frame(  
 term = names(odds\_ratios),  
 odds\_ratio = odds\_ratios  
)  
  
# definisemo listu kategorickih i numerickih varijabli  
categorical\_prefixes = c("native.country", "marital.status", "workclass", "education", "occupation", "relationship", "race", "sex", "AgeGroup")  
numerical\_vars = c("age", "fnlwgt", "capital.gain", "capital.loss", "hours.per.week", "education.num")  
  
# funkcija da proverimo da li je varijabla kategoricka  
is\_categorical = function(term, prefixes) {  
 any(sapply(prefixes, function(prefix) grepl(paste0("^", prefix), term)))  
}  
  
# odvajamo kategoricki i numericki odds ratio  
categorical\_odds\_ratios = odds\_ratios\_df[sapply(odds\_ratios\_df$term, function(term) is\_categorical(term, categorical\_prefixes)), ]  
numerical\_odds\_ratios = odds\_ratios\_df[odds\_ratios\_df$term %in% numerical\_vars, ]  
  
# funkcija da grupise odds ratio za svaku kategoricku varijablu  
aggregate\_odds\_ratios = function(df, variable\_name) {  
 # izdvajamo redove koji se poklapaju sa variable\_name prefix  
 category\_rows = df[grepl(paste0("^", variable\_name), df$term), ]  
   
 # racunamo mean odds ratio  
 avg\_odds\_ratio = mean(category\_rows$odds\_ratio)  
   
 return(data.frame(term = variable\_name, odds\_ratio = avg\_odds\_ratio))  
}  
  
# grupisemo odds ratios za svaku kategoricku varijablu  
aggregated\_odds\_ratios = do.call(rbind, lapply(categorical\_prefixes, function(prefix) {  
 aggregate\_odds\_ratios(categorical\_odds\_ratios, prefix)  
}))  
  
# kombinujemo grupisane kategoricke varijable ratios i numericke varijable  
combined\_odds\_ratios = rbind(aggregated\_odds\_ratios, numerical\_odds\_ratios)  
combined\_odds\_ratios

## term odds\_ratio  
## 1 native.country 0.1375253  
## 2 marital.status 10.5894251  
## 3 workclass 0.4099830  
## 4 education 4.2082558  
## 5 occupation 0.8780396  
## 6 relationship 1.8810862  
## 7 race 1.4420275  
## 8 sex 2.8739590  
## 9 AgeGroup 3.2887953  
## age age 1.0143886  
## fnlwgt fnlwgt 1.0000003  
## education.num education.num 1.0126216  
## capital.gain capital.gain 1.0000768  
## capital.loss capital.loss 1.0004454  
## hours.per.week hours.per.week 1.0256155

# iscrtaj iskombinovane odds ratios  
ggplot(combined\_odds\_ratios, aes(x = reorder(term, odds\_ratio), y = odds\_ratio)) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 labs(title = "Odds Ratios za svaku varijablu",  
 x = "Varijabla",  
 y = "Odds Ratio") +  
 theme\_minimal()

A graph with a bar graph

Description automatically generated

# Provera multikolinearnosti sa Variance Inflation Factor(VIF)

print(vif(model3))

## GVIF Df GVIF^(1/(2\*Df))  
## age 3.510472 1 1.873625  
## workclass 1.621663 7 1.035135  
## fnlwgt 1.053391 1 1.026348  
## education 13.807369 15 1.091450  
## education.num 7.968238 1 2.822807  
## marital.status 50.942742 6 1.387576  
## occupation 3.250794 13 1.046386  
## relationship 93.132717 5 1.573658  
## race 2.820451 4 1.138387  
## sex 2.479679 1 1.574700  
## capital.gain 1.043809 1 1.021670  
## capital.loss 1.020987 1 1.010439  
## hours.per.week 1.162009 1 1.077965  
## native.country 4.026406 39 1.018018  
## AgeGroup 3.822798 3 1.250442

Koeficijenti su jedna od najbitnijih stvari kod logisticke regresije, ovde od log-odds pravimo odds ratios, takođe proveravamo i multikolinearnost između varijabli i zaključujemo da nema značajnijih problema jer su VIF vrednosti u opsegu od 1 do 5 i dosta manje od 10.

model21 = glm (Y\_train ~ marital.status + education + AgeGroup + sex + native.country, data.frame(X\_train, Y\_train), family = "binomial")  
  
summary(model21)

Anova(model21, type = "III")

##   
## Call:  
## glm(formula = Y\_train ~ marital.status + education + AgeGroup +   
## sex + native.country, family = "binomial", data = data.frame(X\_train,   
## Y\_train))  
##   
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 36126 on 26060 degrees of freedom  
## Residual deviance: 22868 on 25996 degrees of freedom  
## AIC: 22998  
##   
## Number of Fisher Scoring iterations: 12

## Analysis of Deviance Table (Type III tests)  
##   
## Response: Y\_train  
## LR Chisq Df Pr(>Chisq)   
## marital.status 3545.2 6 < 2.2e-16 \*\*\*  
## education 3182.8 15 < 2.2e-16 \*\*\*  
## AgeGroup 973.8 3 < 2.2e-16 \*\*\*  
## sex 130.2 1 < 2.2e-16 \*\*\*  
## native.country 151.9 39 3.001e-15 \*\*\*  
## ---

Pošto AIC vrednost treba da bude što manja, bolji nam je model3 model jer je ovde AIC vrednost 22998.

print(vif(model21))

## GVIF Df GVIF^(1/(2\*Df))  
## marital.status 1.512992 6 1.035110  
## education 1.400421 15 1.011289  
## AgeGroup 1.205995 3 1.031710  
## sex 1.212125 1 1.100965  
## native.country 1.280956 39 1.003179

Nema multikolinearnosti između varijabli.

Uradićemo predikciju i nacrtaćemo roc krivu.

prediction\_model21 = predict(model21, test, type="response")   
  
roc\_curve21 = roc(test$income\_encoded, prediction\_model21)

plot(roc\_curve21)

auc(roc\_curve21)

optimal\_treshold21 = coords(roc\_curve21, "best", best.method="youden")$threshold

print(optimal\_treshold21)

conf\_matrix\_model21 = confusionMatrix(table(ifelse(prediction\_model21 > 0.71, 1, 0), test$income\_encoded))  
conf\_matrix\_model21

Area under the curve: 0.8689

[1] 0.4580813

## Confusion Matrix and Statistics  
##   
##   
## 0 1  
## 0 4558 707  
## 1 431 797  
##   
## Accuracy : 0.8247   
## 95% CI : (0.8153, 0.8339)  
## No Information Rate : 0.7684   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.4739   
##   
## Mcnemar's Test P-Value : 3.581e-16   
##   
## Sensitivity : 0.9136   
## Specificity : 0.5299   
## Pos Pred Value : 0.8657   
## Neg Pred Value : 0.6490   
## Prevalence : 0.7684   
## Detection Rate : 0.7020   
## Detection Prevalence : 0.8109   
## Balanced Accuracy : 0.7218   
##   
## 'Positive' Class : 0   
##

A graph with a curve

Description automatically generated  
print(conf\_matrix\_model21$overall["Accuracy"])

## Accuracy   
## 0.8247343

print(conf\_matrix\_model21$byClass["Pos Pred Value"])

## Pos Pred Value   
## 0.865717

print(paste(round(conf\_matrix\_model21$byClass["Sensitivity"], 2)))

## [1] "0.91"

print(paste(round(conf\_matrix\_model21$byClass["F1"], 2)))

## [1] "0.89"

AUC iznosi 0.8689.

model2 = glm (Y\_train ~ marital.status + education + AgeGroup + sex, data.frame(X\_train, Y\_train), family = "binomial")  
  
summary(model2)

##   
## Call:  
## glm(formula = Y\_train ~ marital.status + education + AgeGroup +   
## sex, family = "binomial", data = data.frame(X\_train, Y\_train))  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.554830 0.149203 -30.528 < 2e-16 \*\*\*  
## marital.statusMarried-AF-spouse 3.205060 0.620569 5.165 2.41e-07 \*\*\*  
## marital.statusMarried-civ-spouse 1.910993 0.053714 35.577 < 2e-16 \*\*\*  
## marital.statusMarried-spouse-absent -0.178078 0.168778 -1.055 0.29138   
## marital.statusNever-married -0.469361 0.065965 -7.115 1.12e-12 \*\*\*  
## marital.statusSeparated -0.364131 0.127688 -2.852 0.00435 \*\*   
## marital.statusWidowed -0.000715 0.121592 -0.006 0.99531   
## education11th 0.142323 0.178035 0.799 0.42405   
## education12th 0.533656 0.219887 2.427 0.01523 \*   
## education1st-4th -1.199598 0.370205 -3.240 0.00119 \*\*   
## education5th-6th -0.627513 0.261969 -2.395 0.01660 \*   
## education7th-8th -0.603809 0.187275 -3.224 0.00126 \*\*   
## education9th -0.051643 0.198130 -0.261 0.79436   
## educationAssoc-acdm 1.665415 0.149865 11.113 < 2e-16 \*\*\*  
## educationAssoc-voc 1.690345 0.143172 11.806 < 2e-16 \*\*\*  
## educationBachelors 2.466150 0.129766 19.005 < 2e-16 \*\*\*  
## educationDoctorate 3.867475 0.204393 18.922 < 2e-16 \*\*\*  
## educationHS-grad 0.854826 0.126825 6.740 1.58e-11 \*\*\*  
## educationMasters 2.878773 0.140557 20.481 < 2e-16 \*\*\*  
## educationPreschool -11.615692 99.254968 -0.117 0.90684   
## educationProf-school 3.637592 0.181683 20.022 < 2e-16 \*\*\*  
## educationSome-college 1.355923 0.128542 10.549 < 2e-16 \*\*\*  
## AgeGroup27-45 1.637383 0.065760 24.900 < 2e-16 \*\*\*  
## AgeGroup45-65 2.033167 0.069679 29.179 < 2e-16 \*\*\*  
## AgeGroupAbove 65 1.291242 0.109055 11.840 < 2e-16 \*\*\*  
## sexMale 0.481525 0.041970 11.473 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 36126 on 26060 degrees of freedom  
## Residual deviance: 23020 on 26035 degrees of freedom  
## AIC: 23072  
##   
## Number of Fisher Scoring iterations: 12

Anova(model2, type = "III")

## Analysis of Deviance Table (Type III tests)  
##   
## Response: Y\_train  
## LR Chisq Df Pr(>Chisq)   
## marital.status 3573.9 6 < 2.2e-16 \*\*\*  
## education 3337.2 15 < 2.2e-16 \*\*\*  
## AgeGroup 1011.9 3 < 2.2e-16 \*\*\*  
## sex 130.8 1 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

prediction\_model2 = predict(model2, test, type="response") # na testnom skupu  
  
roc\_curve2 = roc(test$income\_encoded, prediction\_model2)

plot(roc\_curve2)

auc(roc\_curve2)

optimal\_treshold2 = coords(roc\_curve2, "best", best.method="youden")$threshold

print(optimal\_treshold2)

conf\_matrix\_model2 = confusionMatrix(table(ifelse(prediction\_model2 > 0.71, 1, 0), test$income\_encoded))  
conf\_matrix\_model2

## Confusion Matrix and Statistics  
##   
##   
## 0 1  
## 0 4566 712  
## 1 423 792  
##   
## Accuracy : 0.8252   
## 95% CI : (0.8157, 0.8344)  
## No Information Rate : 0.7684   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.4736   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9152   
## Specificity : 0.5266   
## Pos Pred Value : 0.8651   
## Neg Pred Value : 0.6519   
## Prevalence : 0.7684   
## Detection Rate : 0.7032   
## Detection Prevalence : 0.8129   
## Balanced Accuracy : 0.7209   
##   
## 'Positive' Class : 0   
##

print(conf\_matrix\_model2$overall["Accuracy"])

## Accuracy   
## 0.8251964

print(conf\_matrix\_model2$byClass["Pos Pred Value"])

## Pos Pred Value   
## 0.8651004

print(paste(round(conf\_matrix\_model2$byClass["Sensitivity"], 2)))

## [1] "0.92"

print(paste(round(conf\_matrix\_model2$byClass["F1"], 2)))

## [1] "0.89"

Provera da li je model overfitovan.

train\_predictions2 = predict(model2, train, type="response") # na train skupu  
  
predicted\_factors = factor(ifelse(train\_predictions2 > 0.71, 1, 0), levels = c(0, 1))  
actual\_factors = factor(train$income\_encoded, levels = c(0, 1))  
train\_accuracy = confusionMatrix(predicted\_factors, actual\_factors)

A screenshot of a computer

Description automatically generated

Nismo uočili značajan overfiting ni underfiting pošto su slične vrednosti kako za accuracy tako i za ostale PPV(pozitivne prediktovane vrednosti).   
Generalno je model dobar i nisu drastične razlike u tačnosti kada se koristi testni skup iako ga model do sada nije video.   
Logistička regresija je generalno manje sklona overfitingu zbog svoje jednostavnosti.

## Decision tree

modelDT2 = rpart (Y\_train ~ ., data.frame(X\_train, Y\_train), method = "class")  
  
prediction\_modelDT2 = predict(modelDT2, test, type="prob")  
  
prediction\_modelDT2 = prediction\_modelDT2[, 2]  
  
roc\_curvedt2 = roc(test$income\_encoded, prediction\_modelDT2)

plot(roc\_curvedt2)

auc(roc\_curvedt2)

prediction\_modelDT2 = ifelse(prediction\_modelDT2 >= 0.5, 1, 0)

conf\_matrix\_modelDT2 = confusionMatrix(table(prediction\_modelDT2, test$income\_encoded))  
conf\_matrix\_modelDT2

A graph with a line

Description automatically generated

Area under the curve: 0.5849  
## Confusion Matrix and Statistics  
##   
##   
## prediction\_modelDT2 0 1  
## 0 4771 1183  
## 1 218 321  
##   
## Accuracy : 0.7842   
## 95% CI : (0.774, 0.7942)  
## No Information Rate : 0.7684   
## P-Value [Acc > NIR] : 0.001194   
##   
## Kappa : 0.2188   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9563   
## Specificity : 0.2134   
## Pos Pred Value : 0.8013   
## Neg Pred Value : 0.5955   
## Prevalence : 0.7684   
## Detection Rate : 0.7348   
## Detection Prevalence : 0.9170   
## Balanced Accuracy : 0.5849   
##   
## 'Positive' Class : 0   
##

print(conf\_matrix\_modelDT2$overall["Accuracy"])

## Accuracy   
## 0.7842292

print(conf\_matrix\_modelDT2$byClass["Pos Pred Value"])

## Pos Pred Value   
## 0.80131

print(conf\_matrix\_modelDT2$byClass["Sensitivity"])

## Sensitivity   
## 0.9563039

print(conf\_matrix\_modelDT2$byClass["F1"])

## F1   
## 0.871973

# Koji featuri su bitni  
print(modelDT2$variable.importance)

## capital.gain capital.loss education.num education fnlwgt   
## 9385.247137 441.658255 194.530765 189.011161 7.814938   
## native.country   
## 5.737277

#Napravicemo model sa capital.gain, capital.loss, education.num, education  
  
modelDT = rpart (Y\_train ~ capital.gain + capital.loss + education.num + education, data.frame(X\_train, Y\_train), method = "class")  
  
  
prediction\_modelDT = predict(modelDT, test, type="prob")  
  
prediction\_modelDT = prediction\_modelDT[, 2]  
  
roc\_curvedt = roc(test$income\_encoded, prediction\_modelDT)

plot(roc\_curvedt)

auc(roc\_curvedt)

prediction\_modelDT = ifelse(prediction\_modelDT >= 0.5, 1, 0)  
table(prediction\_modelDT, test$income\_encoded)

##   
## prediction\_modelDT 0 1  
## 0 4771 1183  
## 1 218 321

#AUC je samo 0.5849

conf\_matrix\_modelDT = confusionMatrix(table(prediction\_modelDT, test$income\_encoded))  
conf\_matrix\_modelDT

## Confusion Matrix and Statistics  
##   
##   
## prediction\_modelDT 0 1  
## 0 4771 1183  
## 1 218 321  
##   
## Accuracy : 0.7842   
## 95% CI : (0.774, 0.7942)  
## No Information Rate : 0.7684   
## P-Value [Acc > NIR] : 0.001194   
##   
## Kappa : 0.2188   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9563   
## Specificity : 0.2134   
## Pos Pred Value : 0.8013   
## Neg Pred Value : 0.5955   
## Prevalence : 0.7684   
## Detection Rate : 0.7348   
## Detection Prevalence : 0.9170   
## Balanced Accuracy : 0.5849   
##   
## 'Positive' Class : 0   
##

print(conf\_matrix\_modelDT$overall["Accuracy"])

## Accuracy   
## 0.7842292

print(conf\_matrix\_modelDT$byClass["Pos Pred Value"])

## Pos Pred Value   
## 0.80131

print(conf\_matrix\_modelDT$byClass["Sensitivity"])

## Sensitivity   
## 0.9563039

print(conf\_matrix\_modelDT$byClass["F1"])

## F1   
## 0.871973

Poređenjem ova 2 modela, dobili smo da nema nikakve razlike između njih iako smo menjali broj feature-a, AUC vrednost je 0.5849

## Random forest

ModelRF uključuje sve feature.

library(randomForest)  
  
Y\_train = factor(Y\_train)  
modelRF = randomForest(Y\_train ~ ., data = data.frame(X\_train, Y\_train), ntree = 100)  
  
prediction\_modelRF1 = predict(modelRF, newdata = test)  
  
conf\_matrix\_modelRF1 = confusionMatrix(table(prediction\_modelRF1, test$income\_encoded))  
conf\_matrix\_modelRF1

## Confusion Matrix and Statistics  
##   
##   
## prediction\_modelRF1 0 1  
## 0 4829 1007  
## 1 160 497  
##   
## Accuracy : 0.8203   
## 95% CI : (0.8107, 0.8295)  
## No Information Rate : 0.7684   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.3714   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9679   
## Specificity : 0.3305   
## Pos Pred Value : 0.8275   
## Neg Pred Value : 0.7565   
## Prevalence : 0.7684   
## Detection Rate : 0.7437   
## Detection Prevalence : 0.8988   
## Balanced Accuracy : 0.6492   
##   
## 'Positive' Class : 0   
##

print(conf\_matrix\_modelRF1$overall["Accuracy"])

## Accuracy   
## 0.820268

print(conf\_matrix\_modelRF1$byClass["Pos Pred Value"])

## Pos Pred Value   
## 0.8274503

print(conf\_matrix\_modelRF1$byClass["Sensitivity"])

## Sensitivity   
## 0.9679294

print(conf\_matrix\_modelRF1$byClass["F1"])

## F1   
## 0.892194

#Nivo znacajnosti na osnovu random forest algoritma  
  
importance(modelRF)

## MeanDecreaseGini  
## age 623.41632  
## workclass 155.61624  
## fnlwgt 311.88965  
## education 590.94901  
## education.num 636.24407  
## marital.status 1194.16359  
## occupation 721.49123  
## relationship 1184.09877  
## race 54.48696  
## sex 110.36299  
## capital.gain 5264.52318  
## capital.loss 1005.18835  
## hours.per.week 433.80146  
## native.country 125.21537  
## AgeGroup 535.30696

varImpPlot(modelRF)

A graph with text on it

Description automatically generated

feature\_weight = data.frame( Feature = row.names(importance(modelRF)), MeanDecreaseGini = importance(modelRF))  
  
gg\_feature\_weight = ggplot(feature\_weight, aes(x = reorder(Feature, MeanDecreaseGini), y = MeanDecreaseGini)) +  
geom\_bar(stat = "identity", fill = "green") +  
theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
labs(title = "Nivo značajnosti prediktora", x = "Prediktor", y = "Nivo značajnosti")  
  
gg\_feature\_weight

A graph with green and white squares

Description automatically generated

modelRF2 = randomForest(Y\_train ~ marital.status + relationship + capital.gain + capital.loss, data = data.frame(X\_train, Y\_train), ntree = 100)  
  
prediction\_modelRF2 = predict(modelRF2, newdata = test)  
  
conf\_matrix\_modelRF2 = confusionMatrix(table(prediction\_modelRF2, test$income\_encoded))  
conf\_matrix\_modelRF2

## Confusion Matrix and Statistics  
##   
##   
## prediction\_modelRF2 0 1  
## 0 4838 1179  
## 1 151 325  
##   
## Accuracy : 0.7952   
## 95% CI : (0.7851, 0.8049)  
## No Information Rate : 0.7684   
## P-Value [Acc > NIR] : 1.147e-07   
##   
## Kappa : 0.2441   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9697   
## Specificity : 0.2161   
## Pos Pred Value : 0.8041   
## Neg Pred Value : 0.6828   
## Prevalence : 0.7684   
## Detection Rate : 0.7451   
## Detection Prevalence : 0.9267   
## Balanced Accuracy : 0.5929   
##   
## 'Positive' Class : 0   
##

print(conf\_matrix\_modelRF2$overall["Accuracy"])

## Accuracy   
## 0.795164

print(conf\_matrix\_modelRF2$byClass["Pos Pred Value"])

## Pos Pred Value   
## 0.8040552

print(conf\_matrix\_modelRF2$byClass["Sensitivity"])

## Sensitivity   
## 0.9697334

print(conf\_matrix\_modelRF2$byClass["F1"])

## F1   
## 0.8791568

# Zaključak

Rezultati za model3(logistička regresija sa svim feature-ima) dobijeni na osnovu konfuzione matrice conf\_matrix\_model3 i korišćenjem praga od 0.71 su:

AUC 0.87

Accuracy 0.825

Precision 0.865

Recall 0.92

F1-score 0.89

----------------

Rezultati za model21 (logistička regresija gde smo za feature koristili: marital.status, education , AgeGroup, sex, native.country) dobijeni na osnovu konfuzione matrice conf\_matrix\_model21 i korišćenjem praga od 0.7:

AUC 0.87

Accuracy 0.825

Precision 0.87

Recall 0.91

F1-score 0.89

----------------

Rezultati za model2 (logistička regresija gde smo za feature koristili: marital.status, education , AgeGroup, sex) dobijeni na osnovu konfuzione matrice conf\_matrix\_model2 i korišćenjem praga od 0.7:

AUC 0.87

Accuracy 0.825

Precision 0.865

Recall 0.92

F1-score 0.89

----------------

Rezultati za modelDT2 (decision tree sa svim feature-ima) dobijeni na osnovu konfuzione matrice conf\_matrix\_modelDT2:

AUC 0.585

Accuracy 0.78

Precision 0.80

Recall 0.96

F1-score 0.87

----------------

Rezultati za modelDT (decision tree sa capital.gain, capital.loss, education.num, education) dobijeni na osnovu konfuzione matrice conf\_matrix\_modelDT:

AUC 0.585

Accuracy 0.78

Precision 0.80

Recall 0.96

F1-score 0.87

----------------

Rezultati za modelRF (random forest sa svim feature-ima) dobijeni na osnovu konfuzione matrice conf\_matrix\_modelRF1:

Accuracy 0.82

Precision 0.83

Recall 0.97

F1-score 0.89

----------------

Rezultati za modelRF2 (random forest sa marital.status, relationship, capital.gain, capital.loss) dobijeni na osnovu konfuzione matrice conf\_matrix\_modelRF2:

Accuracy 0.80

Precision 0.83

Recall 0.97

F1-score 0.88

----------------

## Najbolji model

**model2** u kojem koristimo 0.7 za treshold jer predvidjamo 0 i 1

ima najbolji AUC - 0.87

ima najbolji Accuracy - 0.825

ima dobar precision - 0.865

ima dobar recall - 0.92

ima najbolji f1-score - 0.89

# Literatura

[1] <https://topepo.github.io/caret/data-splitting.html>  
[2] <https://www.geeksforgeeks.org/data-visualization-in-r/?ref=outind>

[3] <https://www.geeksforgeeks.org/data-analysis-using-r/?ref=outind>

[4] <https://www.geo.fu-berlin.de/en/v/soga-r/Basics-of-statistics/Logistic-Regression/Logistic-Regression-in-R---An-Example/index.html>

[5] <https://www.simplilearn.com/tutorials/data-science-tutorial/random-forest-in-r>

[6]<https://www.geeksforgeeks.org/decision-tree-for-regression-in-r-programming/>

[8] <https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-f-ffclassification-problems/>

[7] Softverski alati 2 – Uvod u programski jezik R (dr Miloš Ivanović, Tatjana Bošković)

[8] Literatura sa predavanja (Predavanja, Vežbe)