Predviđanje srčanog napada u R programskom jeziku

1. Analiza skupa podataka

Skup podataka je vezan za dijagnostičke podatke i analiza krvi i srca. Na osnovu ovih podataka se može izvršiti predikcija ili verovatnoća da će osoba imati srčani napad. Svi atributi su već bili brojčane vrednosti tako da nije bilo potrebe za numeričkom kategorizacijom atributa ili dodeljivanjem nekih numeričkih vrednosti, kao niti bilo kakvog skaliranja vrednosti.

Srčani napad je uzrokovan otkazom dela srčanog mišića usled prestanka dovoda krvi. Dovod krvi je obično prekinut usled stvaranja ugruška u krvi u arteriji koja snabdeva srčani mišić. Ako neki od delova srčanog mišića otkaže, osoba oseća jak bol u grudima i električnu nestabilnost tkiva srčanog mišića.

Atributi:
1) age – **godine starosti,**
2) sex – **pol (1 = muško, 0 = žensko),**
3) cpt = chest pain type (4 vrednosti) – **tip bola u grudima,**
4) restbps = resting blood pressure – **krvni pritisak u mirovanju,**
5) chol = serum cholestoral in mg/dl – **holesteralni serum,**
6) fbs = fasting blood sugar > 120 mg/dl – **visok šećer u krvi,**
7) restecg = resting electrocardiographic results (vrednosti 0,1,2) – **ostalih elektrokardiografskih rezultata,**
8) thalach = maximum heart rate achieved – **maksimalni broj otkucaja srca,**
9) exang = exercise induced angina – **angina izazvana vežbanjem,**
10) oldpeak = ST depression induced by exercise relative to rest – **ST depresija izazvana vežbanjem u odnosu na odmor,**
11) slope = the slope of the peak exercise ST segment – **nagib otkucaja u toku fizičkog napora,**
12) ca = number of major vessels (0-3) colored by flourosopy – **broj glavnih žila obojen flourosopijom**,
13) thal: 0 = **normalno**; 1 = **fiksni defekt**; 2 = **reverzibilni defekt**
14) target: 0= **manja šansa za srčani napad**; 1= **veća šansa za srčani napad**

Ispod je prikaz strukture skupa podataka kao i prvog head skupa.

```
'data.frame':   303 obs. of  14 variables:
 $ age    : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex    : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cpt    : int  3 2 1 1 0 0 1 1 2 2 ...
 $ restbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol   : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs    : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg: int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalach: int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang  : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak: num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope  : int  0 0 2 2 2 1 1 2 2 2 ...
 $ ca     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thal   : int  1 2 2 2 2 1 2 3 3 2 ...
 $ target : int  1 1 1 1 1 1 1 1 1 1 ...
```

	age <int>	sex <int>	cpt <int>	restbps <int>	chol <int>	fbs <int>	restecg <int>	thalach <int>	exang <int>	oldpeak <dbl>	slope <int>	ca <int>	thal <int>	target <int>
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
7	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
8	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
9	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
10	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

10 rows

Na osnovu predstavljenih podataka, može se izvršiti predikcija ili verovatnoća za šansu dobijanja srčanog napada upotrebom stabala odlučivanja (Random Forest) koja će izdvojiti najbitnije atribute prema informativnosti ansamblom stabala koji je jedan od osnovnih načina za binarnu klasifikaciju.

2. Čišćenje podataka

Najpre je bilo potrebno pravilno izmeniti prvu kolonu „age“.

```
head(data, 10)
```

	d.žage <int>	sex <int>	cpt <int>	restbps <int>
1	63	1	3	145
2	37	1	2	130
3	41	0	1	130
4	56	1	1	120
5	57	0	0	120
6	57	1	0	140
7	56	0	1	140
8	44	1	1	120
9	52	1	2	172
10	57	1	2	150

1-10 of 10 rows | 1-10 of 14 columns

```{r}
names(data)[1] <- "age" # Rename the first column.
head(data, 10)
```

	age <int>	sex <int>	cpt <int>	restbps <int>
1	63	1	3	145
2	37	1	2	130
3	41	0	1	130
4	56	1	1	120
5	57	0	0	120
6	57	1	0	140
7	56	0	1	140
8	44	1	1	120
9	52	1	2	172
10	57	1	2	150

Zatim je bilo potrebno proveriti da li postoje nedostajuće vrednosti (NA vrednosti). U koršćenom skupu podataka ih nije bilo od početka, ali je prikazano da bi rešili problem tako što bi za nedostajuće vrednosti uneli srednje vrednosti atributa.

```{r}
colSums(data, na.rm = TRUE, dims = 1) # Sum by columns without NA values.
```

```
   age    sex    cpt restbps   chol    fbs restecg thalach  exang oldpeak  slope     ca   thal target 
 16473    207    293  39882  74618     45    160  45343     99     315    424    221    701    165 
```

```{r}
colSums(is.na(data)) # There are no missig values.
```

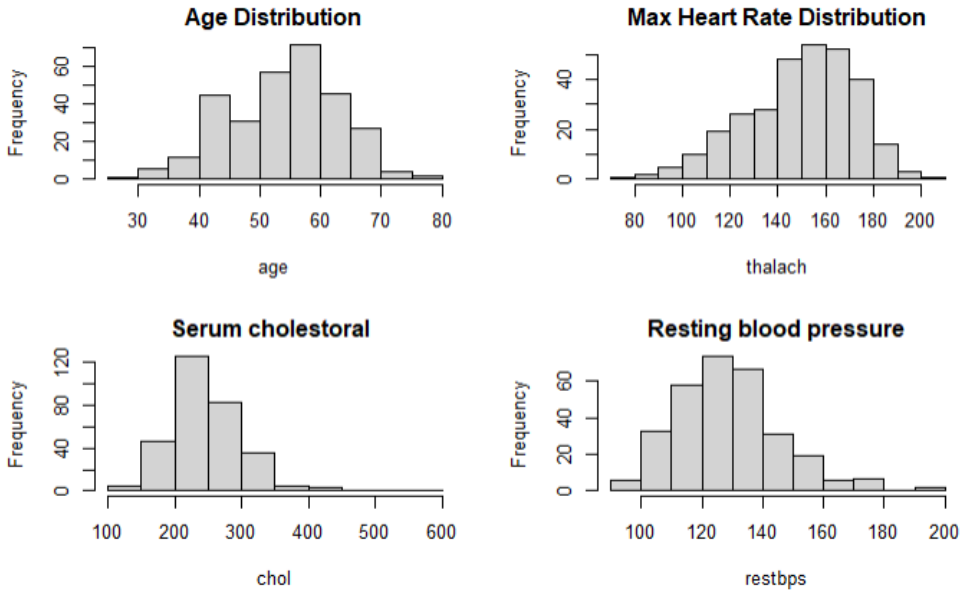
```
   age    sex    cpt restbps   chol    fbs restecg thalach  exang oldpeak  slope     ca   thal target 
     0      0      0       0      0      0       0       0      0       0      0      0      0      0 
```

```{r}
# If there are some missing values, replace it with mean value.
data$thalach[is.na(data$thalach)] <- mean(data$thalach, na.rm = TRUE)
na.omit(data) # Exclude missing values.
```

3. Vizualizacija podataka

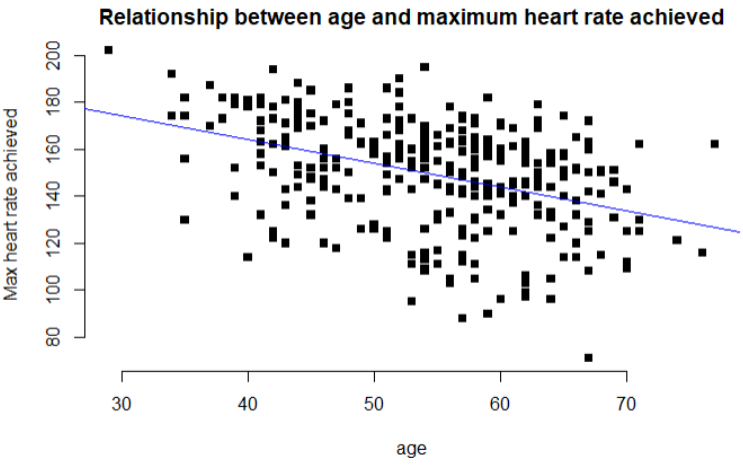
Prikaz histograma sa raspodelama za godine, najveći otkucaj srca, holesteralnog seruma i krvnog pritiska u stanju mirovanja. Možemo primetiti da raspodela godina i maksimalnog pulsa imaju slične raspodele, kao i holerestarlni serum sa krvnim pritiskom u toku mirovanja.

```{r}
par(mfrow = c(2, 2))
hist(data$age, main = "Age Distribution", xlab = "age")
hist(data$thalach, main = "Max Heart Rate Distribution", xlab = "thalach")
hist(data$chol, main = "Serum cholestoral ", xlab = "chol")
hist(data$restbps, main = "Resting blood pressure", xlab = "restbps")
#hist(data$age, main = "Pateints Age Distribution", xlab = "age")
```



Ispod vidimo da je mala zavisnost između atrbita godina i maksimalnog pulsa ali da opet postoji zavisnost, s obzirom da je retka velika zavisnost između 2 originalna atributa od svih ostalih atributa u skupu podataka.

```{r}
x <- data$age
y <- data$thalach
plot(x, y, main = "Relationship between age and maximum heart rate achieved", xlab = "age", ylab = "Max heart rate achieved", pch = 15, frame = FALSE)
abline(lm(y ~ x, data = data), col = "blue") # Correlation is weak.
```



4. Statistički parametri i korelacija

Ispod je naredbom **summary()** prikazana lista parametara za svaki atribut redom:

- Minimalna vrednost,
- Prvi kvartil (25%),
- Drugi Kvartil - Medijana (50%),
- Srednja vrednost,
- Treći kvartil (75%),
- Maksimalna vrednost.

```r
summary(data) # Statistic informations
```

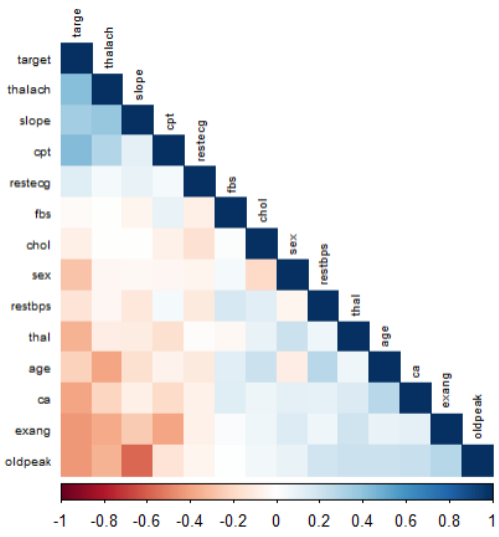
```
      age             sex              cpt            restbps          chol            fbs             restecg          thalach         exang           oldpeak
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0   Median :0.0000   Median :1.0000   Median :153.0   Median :0.0000   Median :0.80
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6   Mean   :0.3267   Mean   :1.04
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20
     slope             ca              thal           target
 Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000
 Median :1.000   Median :0.0000   Median :2.000   Median :1.0000
 Mean   :1.399   Mean   :0.7294   Mean   :2.314   Mean   :0.5446
 3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000
 Max.   :2.000   Max.   :4.0000   Max.   :3.000   Max.   :1.0000
```

Proveravamo da li su tipovi numerički kako bi uspešno kreirali korelacionu matricu koja je prikazana kao Heatmap ispod. Korelacija se meri prosečnom udaljenošću vrednosti od prave koja je podešena tako da je prosečno rastojanje tačaka minimalno podešavajući **b0** i **b1** parametre.

```r
sapply(data, class)
```

```
      age       sex       cpt    restbps      chol       fbs   restecg   thalach     exang   oldpeak     slope        ca      thal    targe
"integer" "integer" "integer" "integer" "integer" "integer" "integer" "numeric" "integer" "numeric" "integer" "integer" "integer" "integer
```

```r
library(corrplot)
corMatrix <- cor(data) #Compute all correlations between each other and form correlation matrix.
corrplot(corMatrix,order = "FPC",method = "color",type = "lower", tl.cex = 0.6, tl.col = "black")
# Principal Component Analysis is not necessary since there are not to much strong correlations between features and target values.
```



5. Priprema podataka za kreiranje modela

Izvršili smo transforamciju skupa podataka tasko što smo kategoričke vrednosti **as.factor()** funkcijom konvertovali u kategoričke vrednosti kako bi se pravilno upotrebile u obučavanju modela kod Random Forest algoritma.

```{r}
data <- transform(
  data,
  age=as.integer(age),
  sex=as.factor(sex),
  cpt=as.factor(cpt),
  restbps=as.integer(restbps),
  chol=as.integer(chol),
  fbs=as.factor(fbs),
  restecg=as.factor(restecg),
  thalach=as.integer(thalach),
  exang=as.factor(exang),
  oldpeak=as.numeric(oldpeak),
  slope=as.factor(slope),
  ca=as.factor(ca),
  thal=as.factor(thal),
  target=as.factor(target)
)
str(data)
```

```
'data.frame':   303 obs. of  14 variables:
 $ age    : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex    : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
 $ cpt    : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
 $ restbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol   : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs    : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
 $ restecg: Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
 $ thalach: int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang  : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
 $ oldpeak: num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope  : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 3 2 2 3 3 3 ...
 $ ca     : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 1 ...
 $ thal   : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
 $ target : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

Skup podatak smo podelili na trening (75%) i validacioni test skup (25%)

```{r}
# Split the data to training and test sets.
smp_size=floor(0.75*nrow(data))
train_ind=sample(seq_len(nrow(data)),size=smp_size)
train_set=subset(data[train_ind,], sample = TRUE)
test_set= subset(data[-train_ind,], sample = FALSE)
dim(train_set)
dim(test_set)
```

```
[1] 227  14
[1] 76 14
```

6. Obučavanje i testiranje modela

Model smo obučili koristeći Random Forest klasifikator za target kolonu i trening podatke. Koristili smo podrazumevani broj stabala = 500. Na kraju vidimo prikaz konfuzione matrice sa tačnim i pogrešnim klasifikacijama i OOB procenom greške od 19.38%.

```{r}
library(randomForest)
# Train model with Random Forest.
rfModel <- randomForest(formula =target ~ ., data=train_set, importance=TRUE)
rfModel
```

```
Call:
 randomForest(formula = target ~ ., data = train_set, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 19.38%
Confusion matrix:
   0   1 class.error
0 78  25   0.2427184
1 19 105   0.1532258
```

Zatim je obučeni model testiran da bi se utvrdila greška klasifikacije nad neviđenim podacima i prikazali smo tačne i pogrešno klasifikovane konfuzionom matricom.

```{r}
# Test the model
predicted = predict(rfModel, newdata=test_set[-14])
confusionMatrix = table(test_set[,14], predicted)
confusionMatrix
```

```
   predicted
     0  1
  0 30  5
  1  3 38
```

Zatim je na kraju tačno prikazana tačnost modela što iznosi približno 90% tačnosti, što je odličan rezultat za Random Forest algoritam, mada pošto je ovo donekle randomiziran algoritam, neće se uvek dobiti ova tačnost ali će varirati između 80-90% u zavinsoti i od broja stabala, maksimalne dubine stabla i ostlaih parametara.

```{r}
# Calculate the Model Accuracy in percentage.
model_accuracy = mean(predicted == test_set$target)
model_accuracy
```

```
[1] 0.8947368
```

7.Zaključak

Za ovakve podatke gde se za ciljni atribut traži klasifikacija je idealno za početak koristiti Random Forest koji je ansambl više stabala odluke gde se biraju favoriti i pronalazi značajnost atributa. Iako je Random Forest ograničen nivoom nasleđivanja kategorijalnih varijabli, ipak je jedan od najboljih modela za klasifikaciju. Lak je i jednostavan za razumevanje i korišćenje stoga što nam govori kako je izvršena neka klasifikacija tako što možemo videti putem od korena (glavnog atributa) do lista (rezultata klasifikacije) i opet nam pruža dobru klasifikaciju u poslovnom scenariju. Takođe se može uporediti Random Forest sa ostalim kompleksnijim modelima poput Logističke Regresije, Metode Nosećih Vektora (SVM) ili Neuronske mreže.