HEART DISEASE

PROJECT - SESSION A1

Student: Brînzaș Sebastian-Vlad

Professor: Lector Dr. Raluca Mureșan

1. Description of the Dataset

Heart disease is broad term used for diseases and conditions affecting the heart and circulatory system. They are also referred as cardiovascular diseases. It is a major cause of disability all around the world. Since heart is amongst the most vital organs of the body, its diseases affect other organs and part of the body as well. There are several different types and forms of heart diseases. The most common ones cause narrowing or blockage of the coronary arteries, malfunctioning in the valves of the heart, enlargement in the size of heart and several others leading to **heart failure** and **heart attack**.

For this report regarding heart diseases i will use **Heart Disease UCI** | Kaggle dataset, which is a dataset with 14 attributes where the goal field refers to the presence of hearth disease in the patient.

Attribute Information:

- age: The person's age in years
- sex: The person's sex (1 = male, 0 = female)
- cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
- chol: The person's cholesterol measurement in mg/dl
- fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach: The person's maximum heart rate achieved
- exang: Exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
- ca: The number of major vessels (0-3)
- thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
- target: Heart disease (0 = no, 1 = yes)

This dataset is a real data including important features of patients. We will visualize the data, describe the dataset through descriptive statistics techniques, perform hypothesis testing on several variables and build the predictable model, then evaluate it’s performance using confusion matrices.

2. Data Preparation

- A copy of the original dataset is done for backup before processing it.
- I made a few changes in data to some of the variables that seem categorical: **sex, cp, fbs, exang, restecg, slope and thal.**
- I have converted some variables to factors that are used to manage the levels and the order of the categorical values.
- With “n_miss()” function of “naniar” library we check for missing values, but none is found.
- With “distinct()” function of “dplyr” library we search in our dataset for duplicates. One row is found and we drop it.
- With “lapply(heart_data, levels)” function we are checking the levels of variables in columns. Through this function we found a few outliners in “ca” column which should have values only between 0-3. Also, the “thal” column had a few rows of value “0” which are outliners. After noticing the 5 outliners variables, i performed operations to drop these rows.

3. Descriptive Statistics Techniques

Visualizing heart disease proportion in the dataset

```
   0    1
0.46 0.54
> |
```

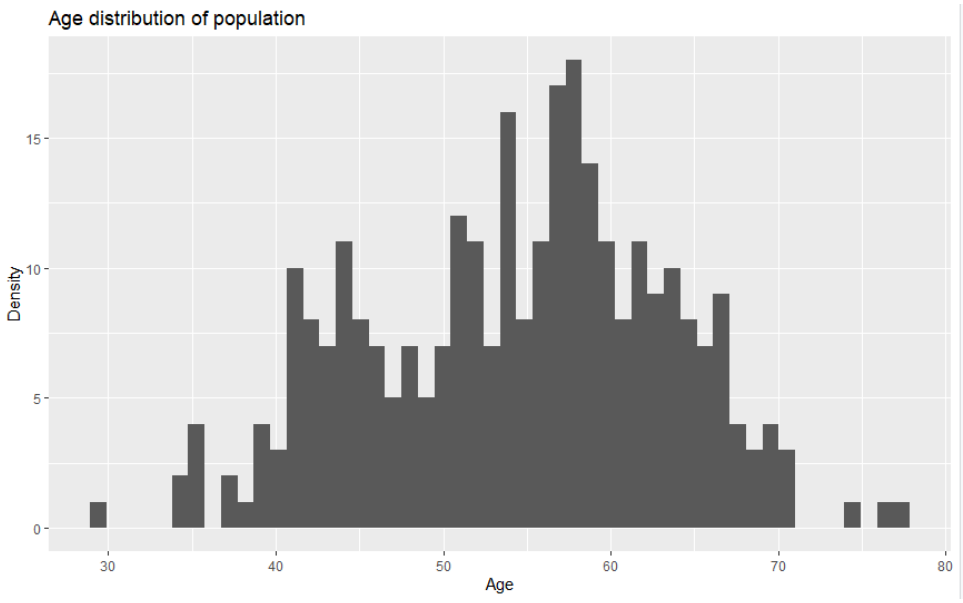
54% of people have heart disease and 46% dont

Sex proportion in the patients

```
female   male
  0.32   0.68
> |
```

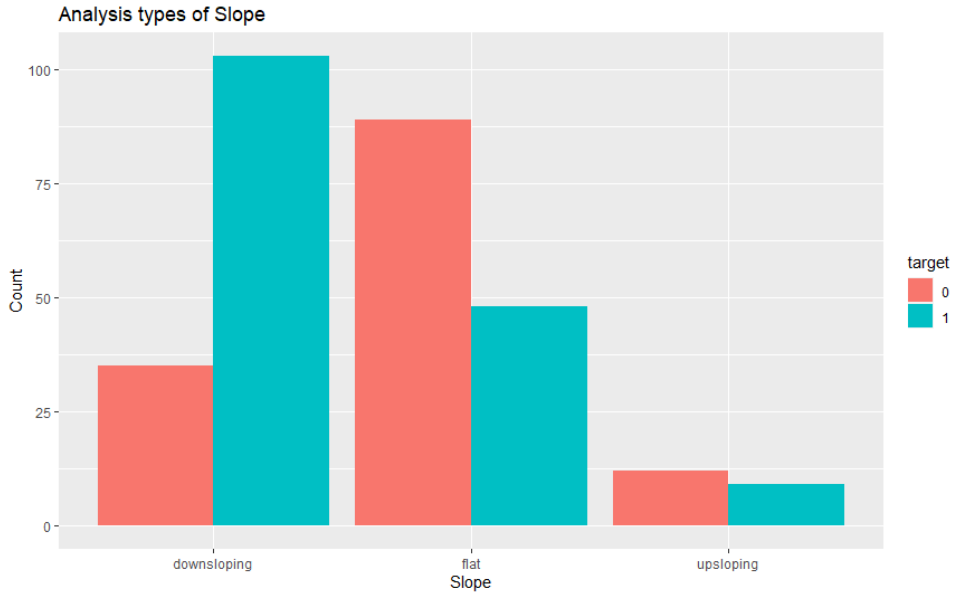
Patiens are: 32% female and 68% male

Age distribution of population



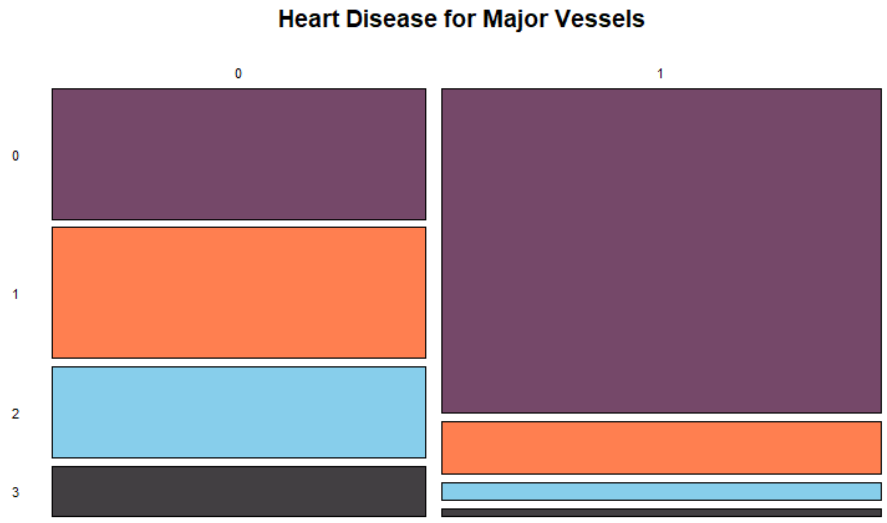
The majority of people lies in the age of 40 to 70 have more chance of heart disease

Slope of the Peak Exercise ST Segment(Slope)



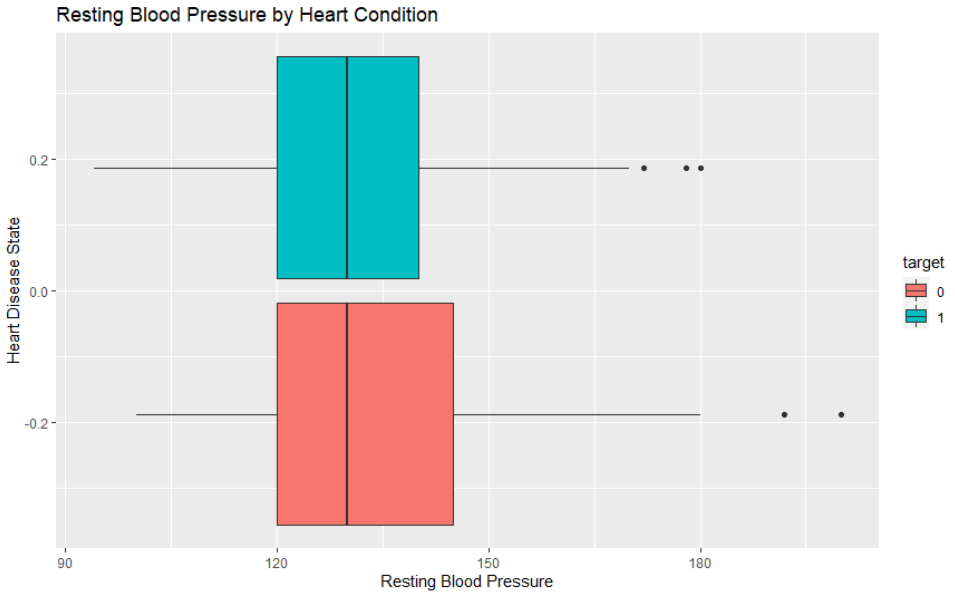
- For "Downsloping" the no heart disease is of 34 than the heart disease patient far above 100.
- For "Flat" the no heart disease is far above 90 than the heart disease patient is of 45.
- For "Upsloping" the no heart disease is of below 20 than the heart disease patient is of below 15.

Number of Major Vessels (ca)



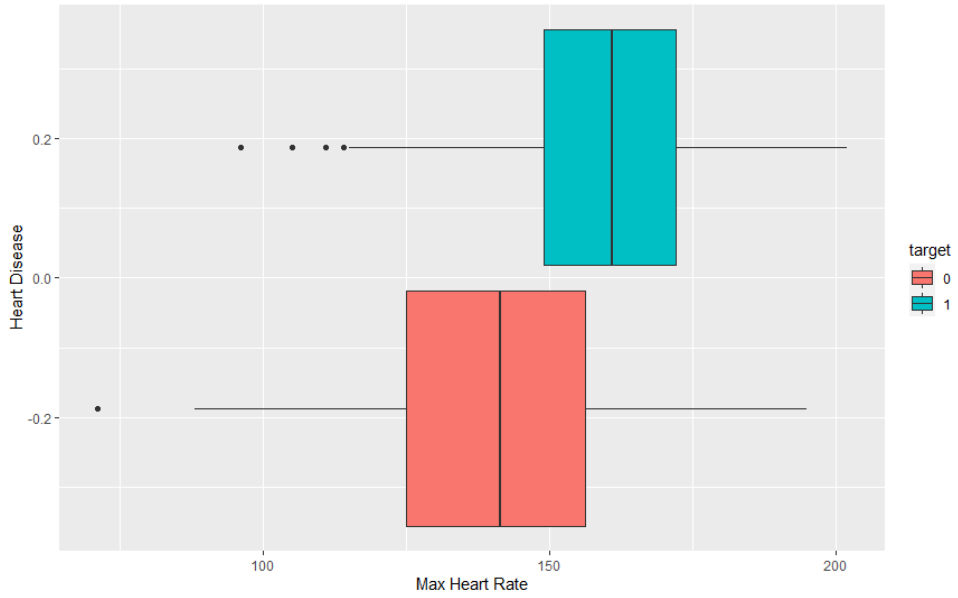
The majority of the people having heart disease have no major vessel.

Resting Blood Sugars separated by heart disease state



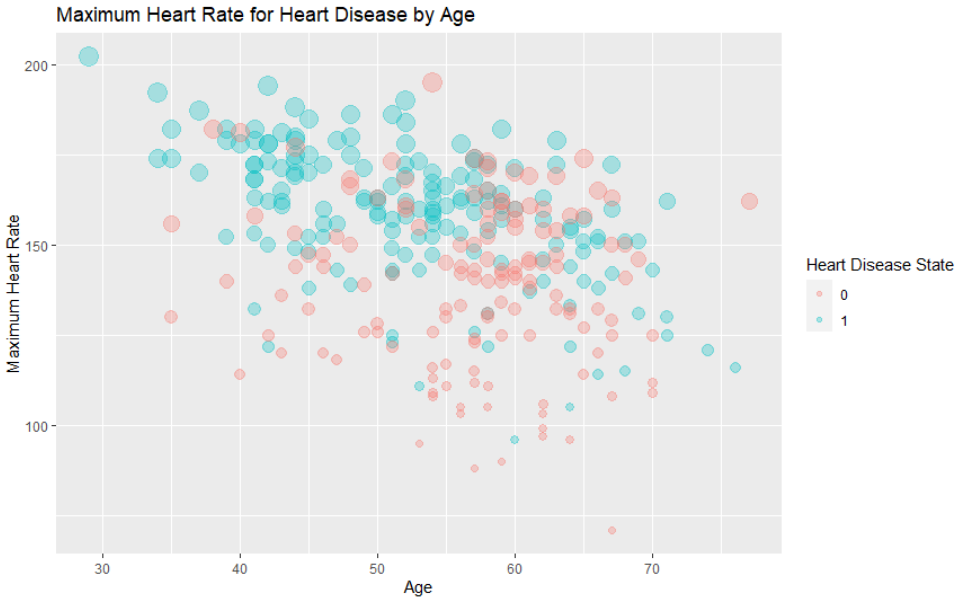
The plot above shows an interquartile range of resting blood sugar is slightly higher for the no heart disease plot. But the medians of both the box plot looks the same.

Heart Rate vs heart condition



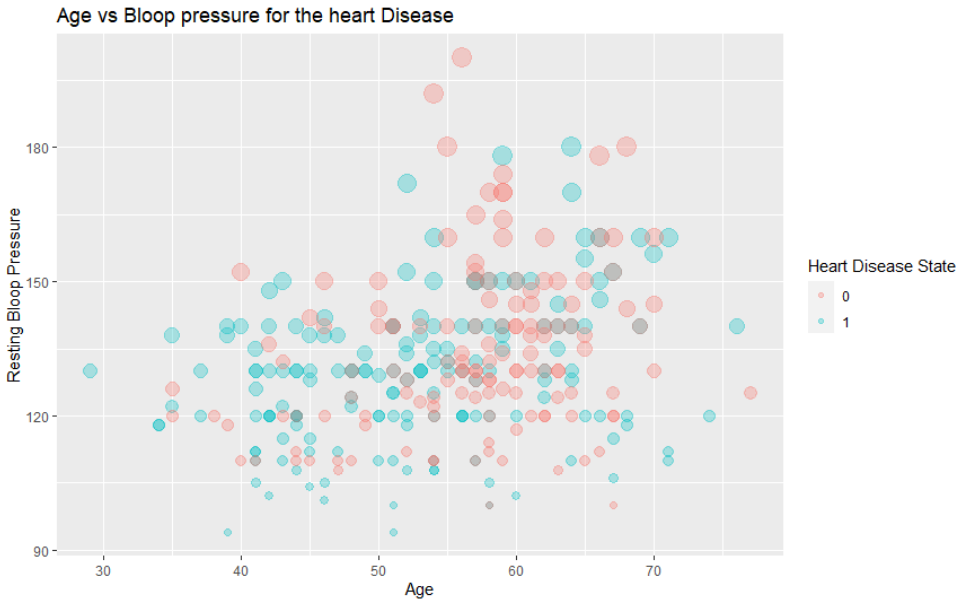
The interquartile range of heart rate shows a higher max heart rate achieved in patients with heart disease.

Age vs Max rate for Heart Disease



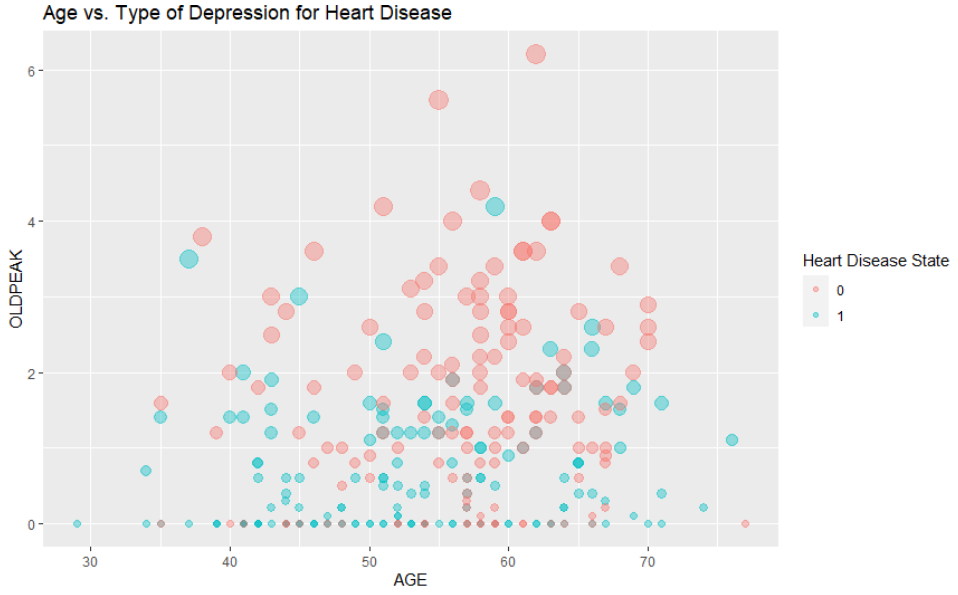
It shows, that when heart rate goes up most of are heart disease patient than no disease.

Age vs. Bloop pressure for the Heart Disease



When the blood pressure/sugar lies between 100-150 than the major number heart diseases patient found. As you in the chart, the more bigger blue dot means no heart disease at 180 blood pressure/sugar. Same time, the bigger red dot means heart disease at 150-180 blood presuure/sugar.

Age vs. type of depression for the Heart Disease



Major number heart disease patient lies in age group between 45 to 60.

4. Hypothesis testing

Mann-Whitney-Wilcoxon test

- Null hypothesis: Distribution functions of trestbps and target are equal

trestbps p-value = 1.913e-06 < .05 => trestbps is not normally distributed

```
        Wilcoxon rank sum test with continuity correction

data:  heart_data$trestbps by heart_data$target
W = 12476, p-value = 0.02918
alternative hypothesis: true location shift is not equal to 0
```

Null hypothesis: Distribution functions of trestbps and target are equal.

p-value = 0.02918 < .05 => nonidentical populations

- Null hypothesis: Distribution functions of age and target are equal

age p-value = 0.006045 < .05 => age is not normally distributed

```
        Wilcoxon rank sum test with continuity correction

data:  heart_data$age by heart_data$target
W = 13880, p-value = 4.314e-05
alternative hypothesis: true location shift is not equal to 0
```

p-value = 4.314e-05 < .05 => nonidentical populations

- Null hypothesis: Distribution functions of chol and target are equal

chol p-value = 8.986e-09 < .05 => chol is not normally distributed

```
        Wilcoxon rank sum test with continuity correction

data:  heart_data$chol by heart_data$target
W = 12285, p-value = 0.05563
alternative hypothesis: true location shift is not equal to 0
```

p-value = 0.05563 > .05 => identical populations

- Null hypothesis: Distribution functions of chol and fasting blood sugar are equal

```
        Wilcoxon rank sum test with continuity correction

data:  heart_data$chol by heart_data$fbs
W = 5270, p-value = 0.7446
alternative hypothesis: true location shift is not equal to 0
```

p-value = 0.7446 > .05 => identical populations

Kruskal-Wallis Test

- Null hypothesis: Distribution functions of trestbps and type of chest pain are equal

```
        Kruskal-Wallis rank sum test

data:  trestbps by cp
Kruskal-Wallis chi-squared = 7.5448, df = 3, p-value = 0.05642
```

p-value = 0.05642 > 0.5 => identical populations

- Null hypothesis: Distribution functions of age and type of chest pain are equal

```
        Kruskal-Wallis rank sum test

data:  age by cp
Kruskal-Wallis chi-squared = 11.864, df = 3, p-value = 0.007865
```

p-value = 0.007865 < .05 => Nonidentical populations

- Null hypothesis: Distribution functions of cholestoral and no. of major vessel colored by flourosopy are equal

```
        Kruskal-Wallis rank sum test

data:  chol by ca
Kruskal-Wallis chi-squared = 6.0807, df = 3, p-value = 0.1077
```

p-value = 0.1077 > .05 => Identical populations

- Null hypothesis: Distribution functions of resting blood pressure and no. of major vessel colored by flourosopy are equal

```
        Kruskal-Wallis rank sum test

data:  trestbps by ca
Kruskal-Wallis chi-squared = 4.8709, df = 3, p-value = 0.1815
```

p-value = 0.1815 > .05 => Identical populations

5. Creating a Predictive Model

Train the model using training data (heart_data_train) using glm function to run a logistic regression. I've split the dataset into two subsets for training and testing.

dependent variable: target

independent variables:

'age','sex','cp','trestbps','chol','fbs','restecg', 'thalach', 'exang','oldpeak', 'slope', 'ca' and 'thal'.

Check model summary:

```
Call:
glm(formula = target ~ age + sex + cp + trestbps + chol + fbs +
    restecg + thalach + exang + oldpeak + slope + ca + thal,
    family = "binomial", data = heart_data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4244  -0.2600   0.1128   0.4409   2.4591

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.896e+00  1.455e+03  -0.001 0.998960
age                         -2.118e-03  3.430e-02  -0.062 0.950763
sexmale                     -2.096e+00  7.933e-01  -2.642 0.008237 **
cpatypical angina           -1.219e+00  9.567e-01  -1.275 0.202400
cpnon-anginal pain          -1.736e-01  8.500e-01  -0.204 0.838153
cptypical angina            -2.230e+00  8.543e-01  -2.611 0.009034 **
trestbps                    -2.425e-02  1.649e-02  -1.470 0.141434
chol                        -8.299e-03  6.421e-03  -1.293 0.196169
fbstrue                      7.769e-01  8.269e-01   0.940 0.347448
restecgNothing to note       9.863e+00  1.455e+03   0.007 0.994593
restecgST-T Wave abnormality 1.067e+01  1.455e+03   0.007 0.994150
thalach                      1.704e-02  1.715e-02   0.993 0.320517
exangyes                    -6.156e-01  6.444e-01  -0.955 0.339374
oldpeak                     -5.743e-01  3.380e-01  -1.699 0.089279 .
slopeflat                   -7.601e-01  6.278e-01  -1.211 0.225961
slopeupsloping              -1.116e+00  1.143e+00  -0.976 0.328956
ca1                         -2.243e+00  6.916e-01  -3.244 0.001180 **
ca2                         -3.884e+00  1.099e+00  -3.534 0.000409 ***
ca3                         -2.486e+00  1.267e+00  -1.961 0.049836 *
thalnormal                  -7.266e-02  1.150e+00  -0.063 0.949637
thalreversible defect       -1.511e+00  5.818e-01  -2.597 0.009407 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 260.82  on 188  degrees of freedom
Residual deviance: 111.95  on 168  degrees of freedom
AIC: 153.95

Number of Fisher Scoring iterations: 14
```

Confusion Matrix on training dataset:

```
             Predicted_target
Actual_target FALSE TRUE
            0    76   11
            1    13   89
```

Model accuracy for training dataset:

```
> round(((confusion_r
[1] 0.873
>
```

Confusion Matrix on testing dataset:

```
             Predicted_target
Actual_target FALSE TRUE
            0    45    4
            1    12   46
```

Model accuracy on testing dataset:

```
> round((conf
[1] 0.85
> 
```