Heart Project

Gabriele

12/7/2021

Contents

Introduction

The goal of this project is to predict whether certain patients will have a heart disease, performing data analysis on vitals measured by a cardiologist.

We will evaluate the correlations among the data along with the accuracy of various AI machine learning algorithms that learn & improve from experience. In the end we will highlight the most important features for the prediction and we will provide an example of a fictional patient.

——> EXPLORATORY ANALYSIS

```
## The Dataset has  303  rows and  14  columns
```

The variables in the data set are:

```
##  [1] "age"      "sex"      "cp"       "trestbps" "chol"     "fbs"
##  [7] "restecg"  "thalach"  "exang"    "oldpeak"  "slope"    "ca"
## [13] "thal"     "target"
```

1. ***age (#)***
2. ***sex*** : 1= Male, 0= Female (Binary)
3. ***(cp)*** chest pain type (4 values -Ordinal):Value 1: typical angina ,Value 2: atypical angina, Value 3: non-anginal pain , Value 4: asymptomatic
4. ***(trestbps)*** resting blood pressure (#)
5. ***(chol)*** serum cholesterol in mg/dl (#)
6. ***(fbs)*** fasting blood sugar > 120 mg/dl(Binary)(1 = true; 0 = false)
7. ***(restecg)*** resting electrocardiography results(values 0,1,2)
8. ***(thalach)*** maximum heart rate achieved (#)
9. ***(exang)*** exercise induced angina (binary) (1 = yes; 0 = no)
10. ***(oldpeak)*** = ST depression induced by exercise relative to rest (#)
11. ***(slope)*** of the peak exercise ST segment (Ordinal) (Value 1: up sloping , Value 2: flat , Value 3: down sloping)
12. ***(ca)*** number of major vessels (0-3, Ordinal) colored by fluoroscopy
13. ***(thal)*** maximum heart rate achieved - (Ordinal): 3 = normal; 6 = fixed defect; 7 = reversible defect

The variables are of different types:

```
## '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 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trestbps: 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 ...
```

Continuous: which is quantitative data that can be measured Ordinal Data: Categorical data that has a order to it (0,1,2,3, etc) Binary Data: data whose unit can take on only two possible states (0 &1)

The dataset has no missing value

```
##      age      sex       cp trestbps     chol      fbs  restecg  thalach
## "No N/A" "No N/A" "No N/A" "No N/A" "No N/A" "No N/A" "No N/A" "No N/A"
##    exang  oldpeak    slope       ca     thal   target
## "No N/A" "No N/A" "No N/A" "No N/A" "No N/A" "No N/A"
```

As we can see from the descriptive statistics provided below we have a big dispersion of data in the trestbps (resting blood pressure) and in the chol (cholesterol) feature. Thereafter cholesterol has a leptokurtic distribution and it is right skewed as we will show in the next graphs.

```
##          vars   n   mean    sd median trimmed   mad min   max range  skew
## age         1 303  54.37  9.08   55.0   54.54 10.38  29  77.0  48.0 -0.20
## sex         2 303   0.68  0.47    1.0    0.73  0.00   0   1.0   1.0 -0.78
## cp          3 303   0.97  1.03    1.0    0.86  1.48   0   3.0   3.0  0.48
## trestbps    4 303 131.62 17.54  130.0  130.44 14.83  94 200.0 106.0  0.71
## chol        5 303 246.26 51.83  240.0  243.49 47.44 126 564.0 438.0  1.13
## fbs         6 303   0.15  0.36    0.0    0.06  0.00   0   1.0   1.0  1.97
## restecg     7 303   0.53  0.53    1.0    0.52  0.00   0   2.0   2.0  0.16
## thalach     8 303 149.65 22.91  153.0  150.98 22.24  71 202.0 131.0 -0.53
## exang       9 303   0.33  0.47    0.0    0.28  0.00   0   1.0   1.0  0.74
## oldpeak    10 303   1.04  1.16    0.8    0.86  1.19   0   6.2   6.2  1.26
## slope      11 303   1.40  0.62    1.0    1.46  1.48   0   2.0   2.0 -0.50
## ca         12 303   0.73  1.02    0.0    0.54  0.00   0   4.0   4.0  1.30
## thal       13 303   2.31  0.61    2.0    2.36  0.00   0   3.0   3.0 -0.47
## target     14 303   0.54  0.50    1.0    0.56  0.00   0   1.0   1.0 -0.18
##          kurtosis   se
## age         -0.57 0.52
## sex         -1.39 0.03
## cp          -1.21 0.06
## trestbps     0.87 1.01
## chol         4.36 2.98
## fbs          1.88 0.02
## restecg     -1.37 0.03
## thalach     -0.10 1.32
## exang       -1.46 0.03
## oldpeak      1.50 0.07
## slope       -0.65 0.04
## ca           0.78 0.06
## thal         0.25 0.04
## target      -1.97 0.03
```

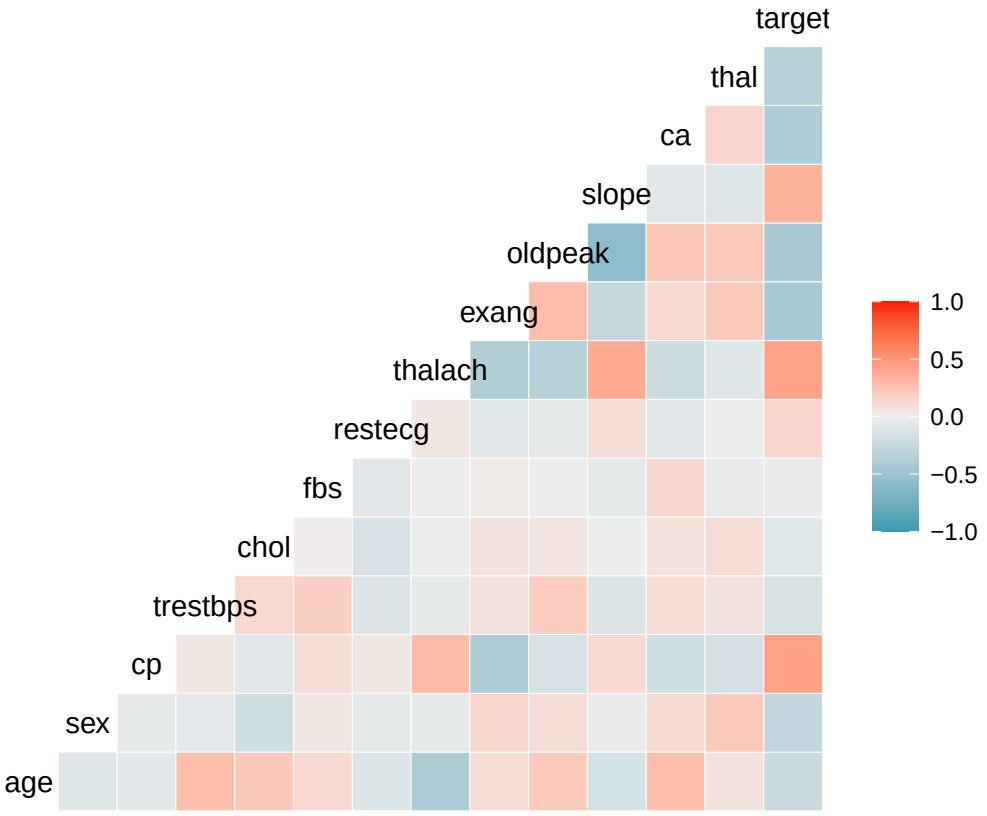
Before investigating the correlations among the variables, let's check if there is a right proportion in the predictor variable between 0 (No heart disease) and 1 (Heart disease)

```
##
##   0   1
## 138 165
```

```
##   Var1 Freq Percentuale
## 1    0  138    45.54455
## 2    1  165    54.45545
```

Let's show now the correlations among the target and the other features

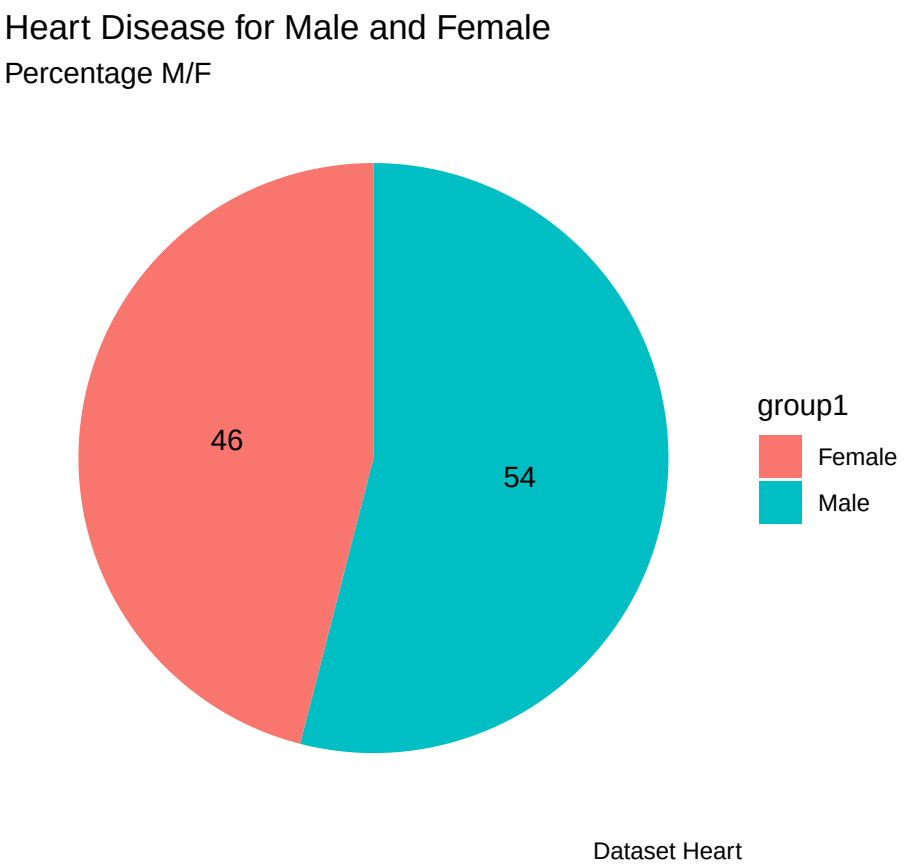
```
##                Target
## target     1.00000000
## cp         0.43379826
## thalach    0.42174093
## slope      0.34587708
## restecg    0.13722950
## fbs       -0.02804576
## chol      -0.08523911
## trestbps  -0.14493113
## age       -0.22543872
## sex       -0.28093658
## thal      -0.34402927
## ca        -0.39172399
## oldpeak   -0.43069600
## exang     -0.43675708
```



Let's get deeper into the correlation among the main variables.

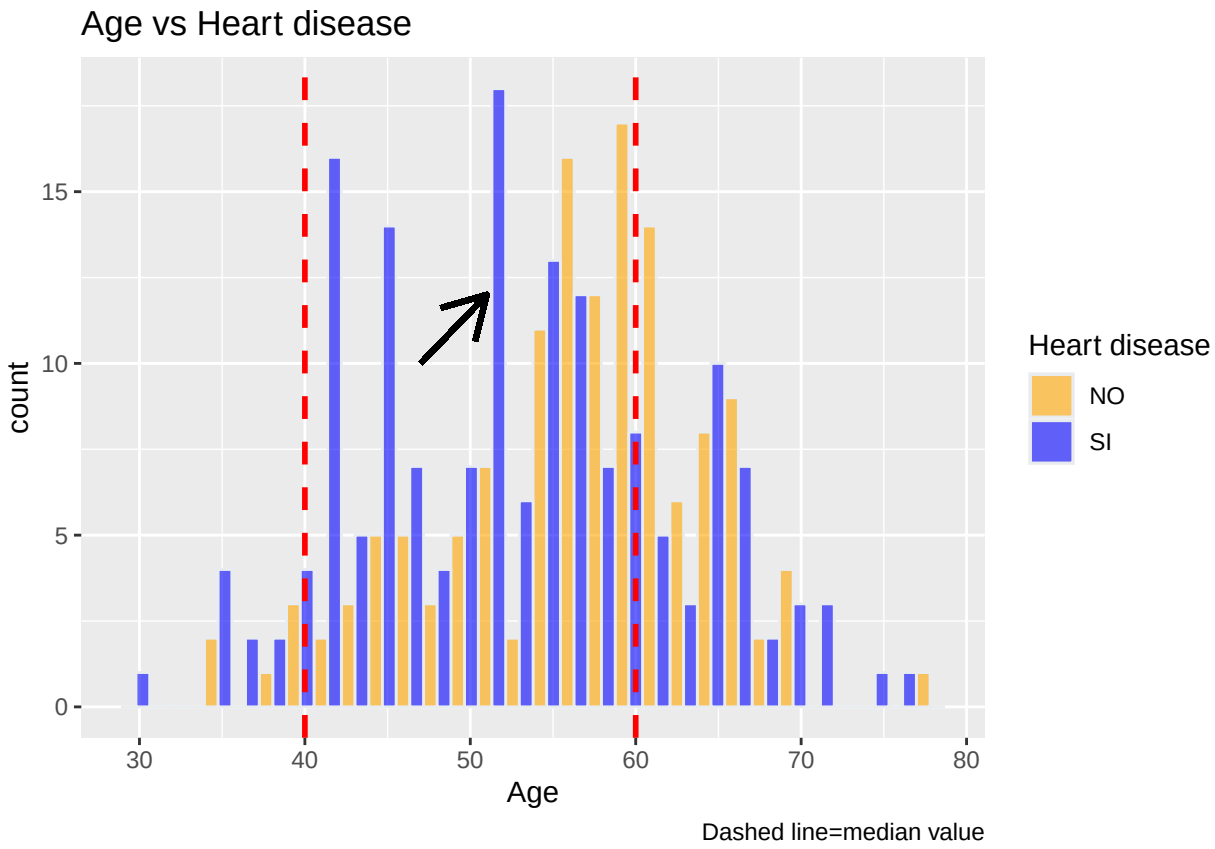
SEX

Male have more probability to have a heart disease



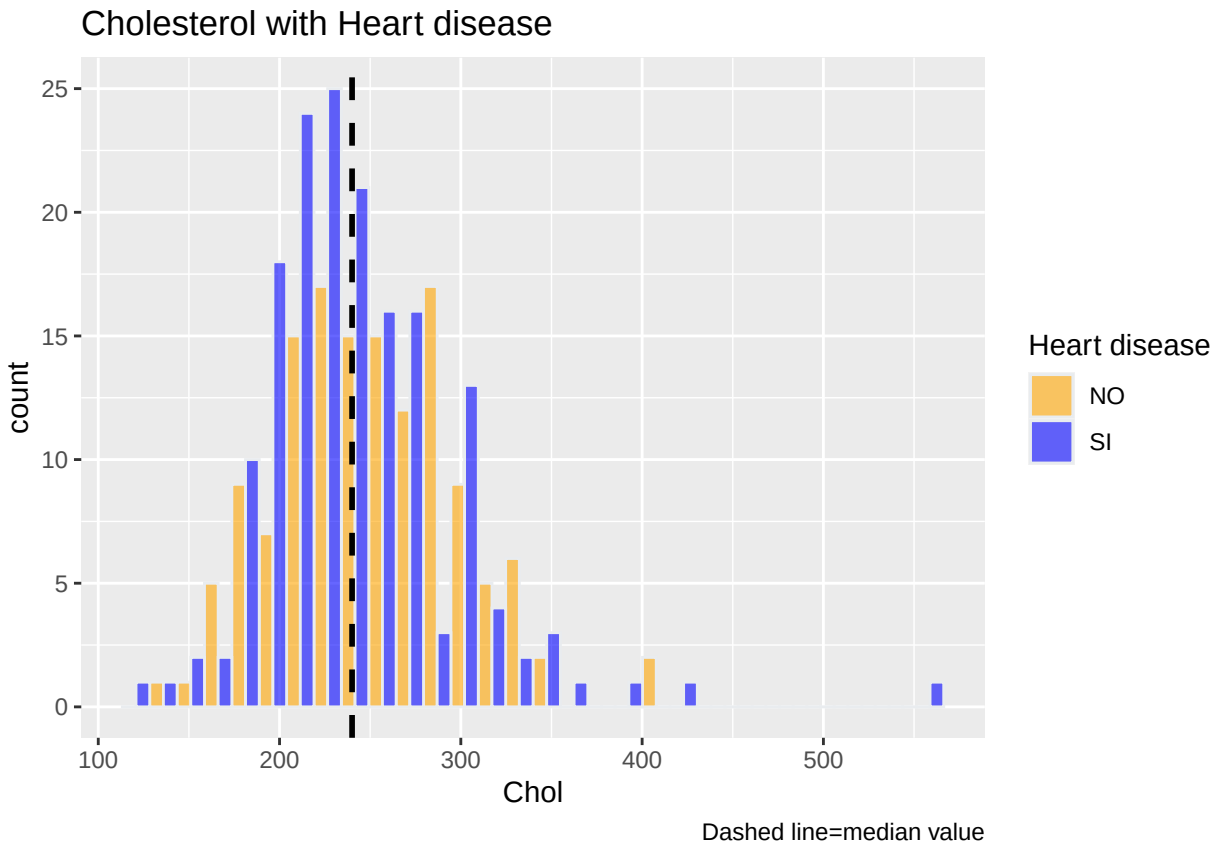
AGE

The distribution of Age vs heart disease is approximately normal, simmetric and mesokurti, as shown by the previous descriptive statistics table (skewness=-0.20, Kurtosis=-0.50). As litterature reports, the graph shows that the maximum incidence is over 50.

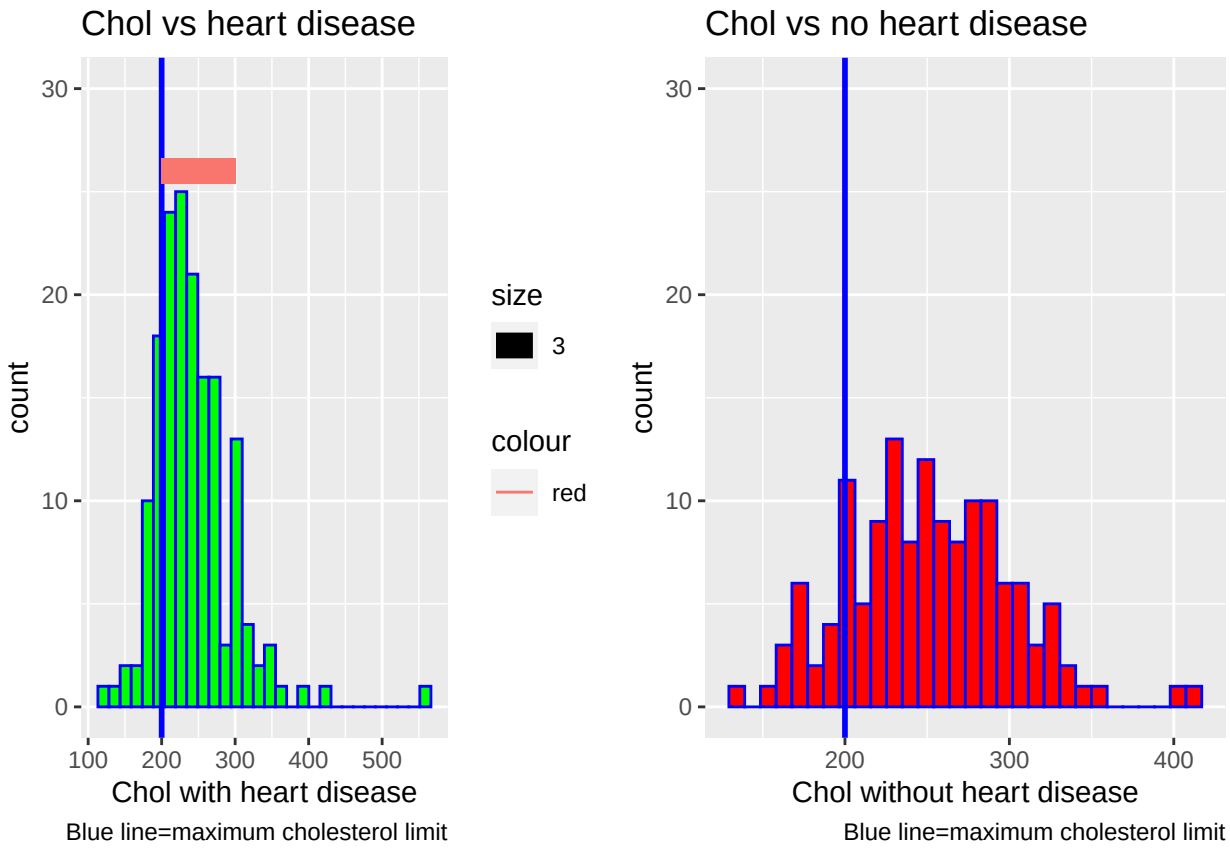


CHOLESTEROL

**Cholesterol is another important feratute for the development of heart disease. As shown in the graph below the median value among the patients with heart disease is above th normal upper limit range of 200 mg/dl

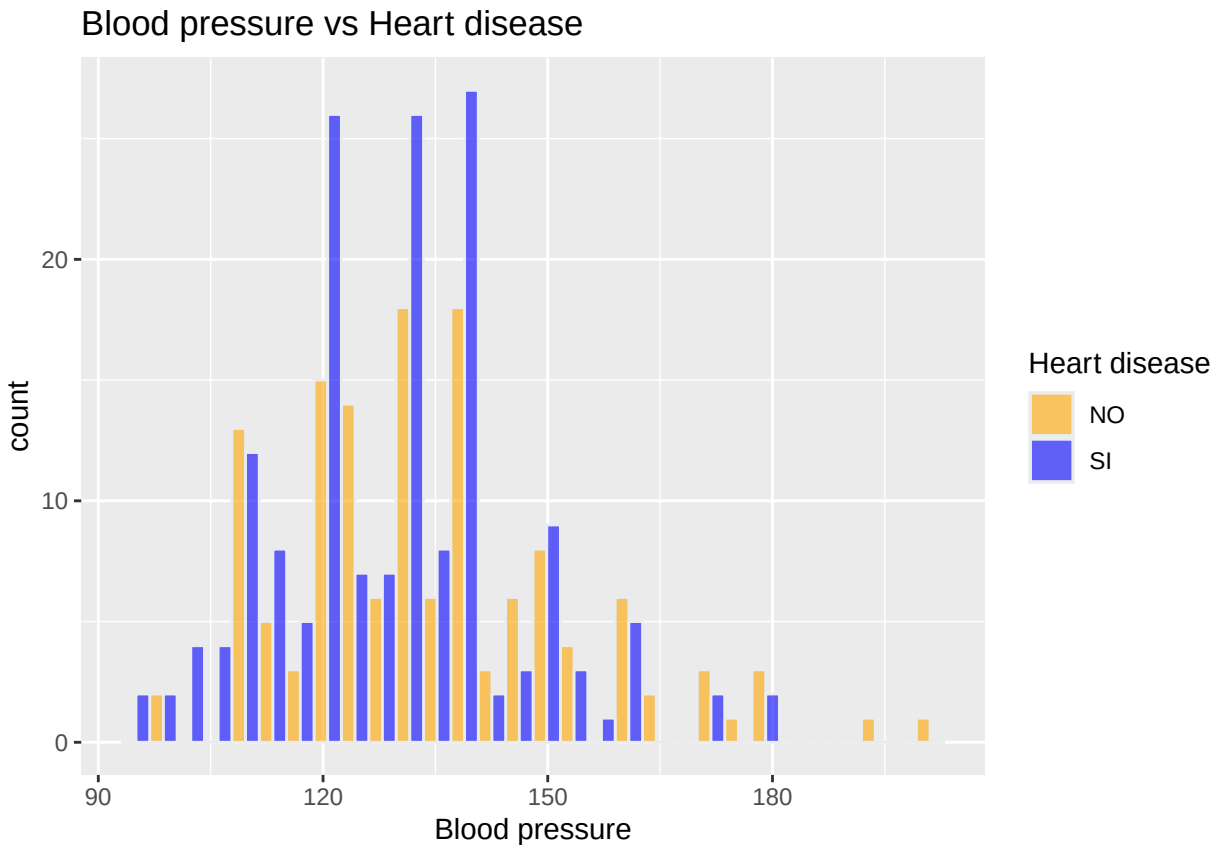


We can see it better in the graph below, where the number of patients with heart disease is far beyond those without heart disease

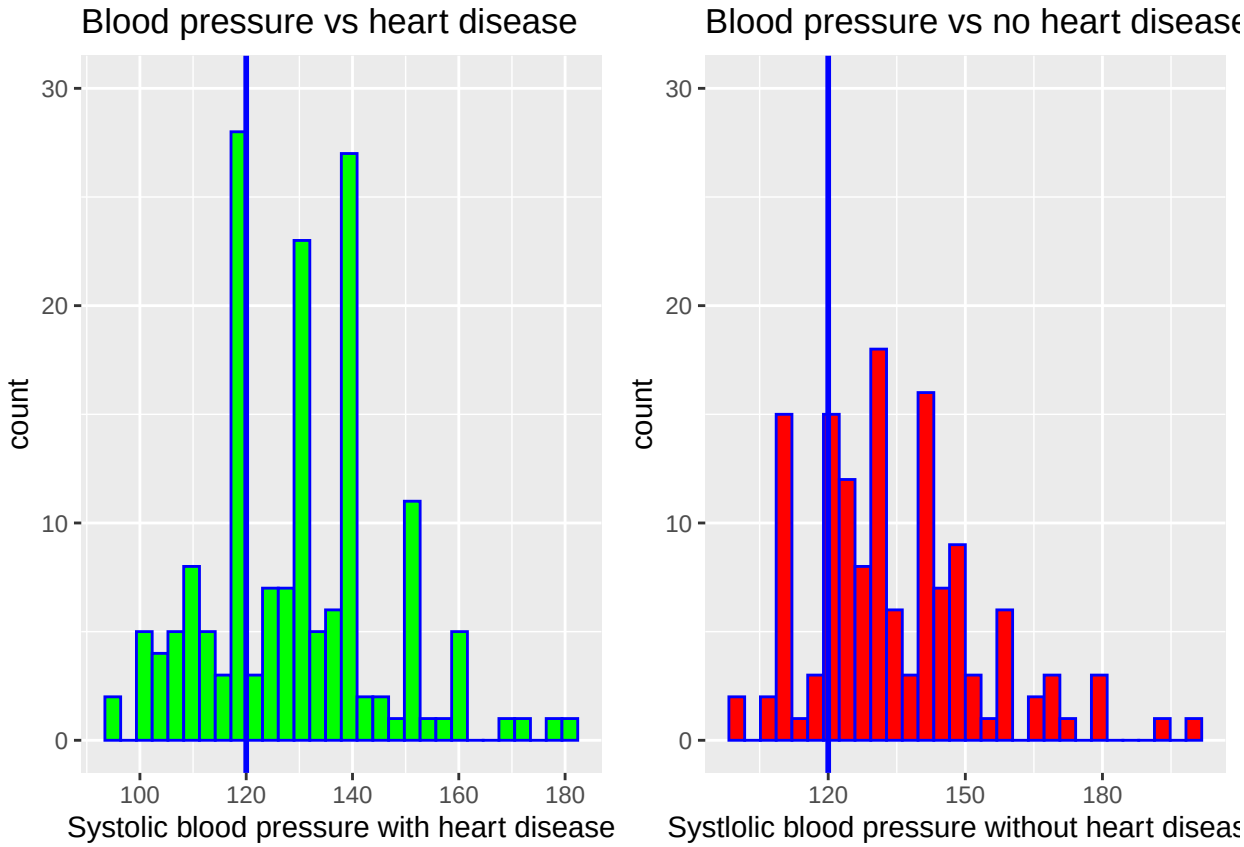


RESTING BLOOD PRESSURE

Many people who have high blood pressure have no idea they have it because it has no signs or symptoms. That's why we call high blood pressure the "silent killer."The American Heart Association warns of many possible consequences of high blood pressure, such as: angina (chest pain), damage to the heart and coronary arteries, peripheral artery disease, stroke.

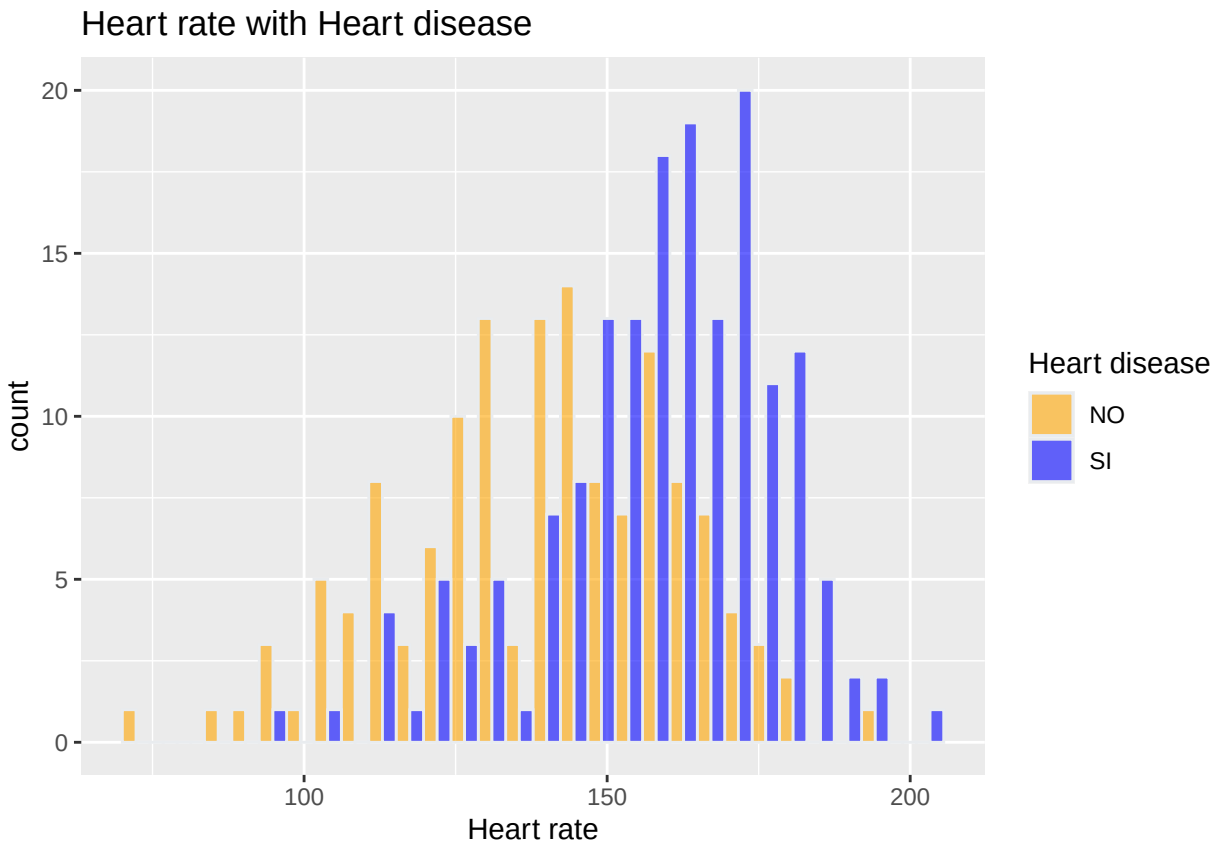


The American Heart Association (AHA) advise that blood pressure numbers below 120/80 millimeters of mercury (mm Hg) are normal. When readings range from 120–129 mm Hg systolic and less than 80 mm Hg diastolic, the person has elevated blood pressure.The graph shows how the persons with heart disease tend to have a higher pressure in relation to those without heart disease.

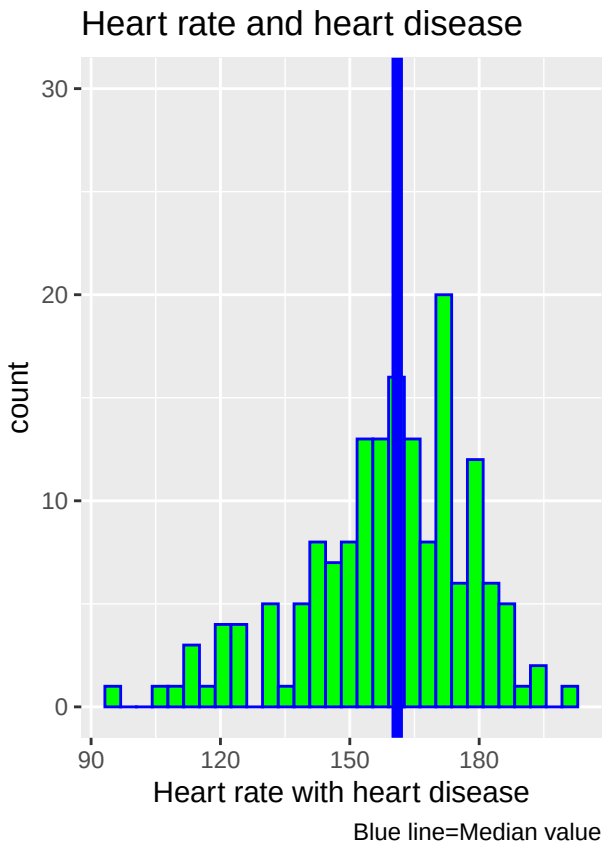


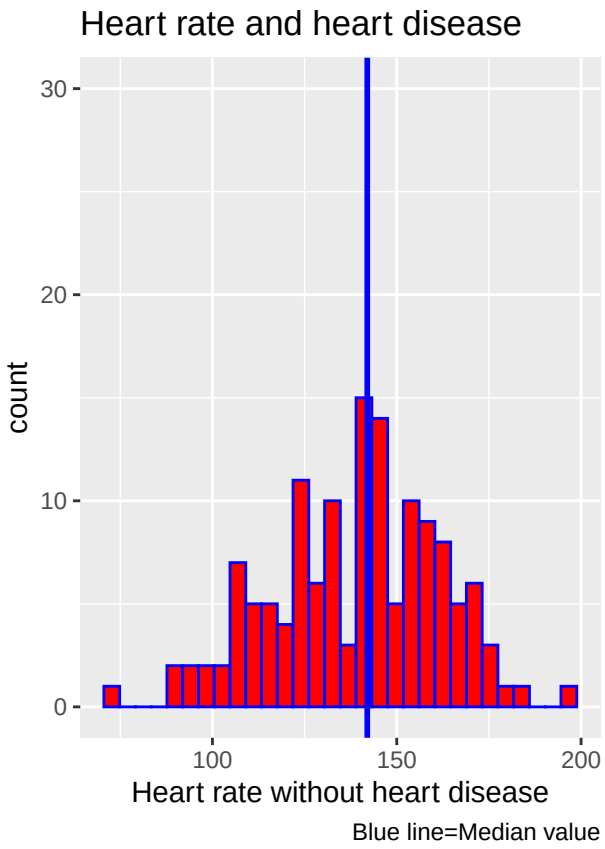
MAXIMUM HEART RATE

Your resting heart rate is the number of times your heart beats per minute when you're at rest.The rate can be affected by factors like stress, anxiety, hormones, medication, and how physically active you are. An athlete or more active person may have a resting heart rate as low as 40 beats per minute. Almost all the epidemiological studies that aimed to answer the question of the relationship between heart rate and all-cause or cardiovascular morbidity and mortality reported that a high heart rate was associated with a higher risk of all-cause mortality and cardiovascular events. We can see that persons with heart disease tend to have a a higher heart rate.

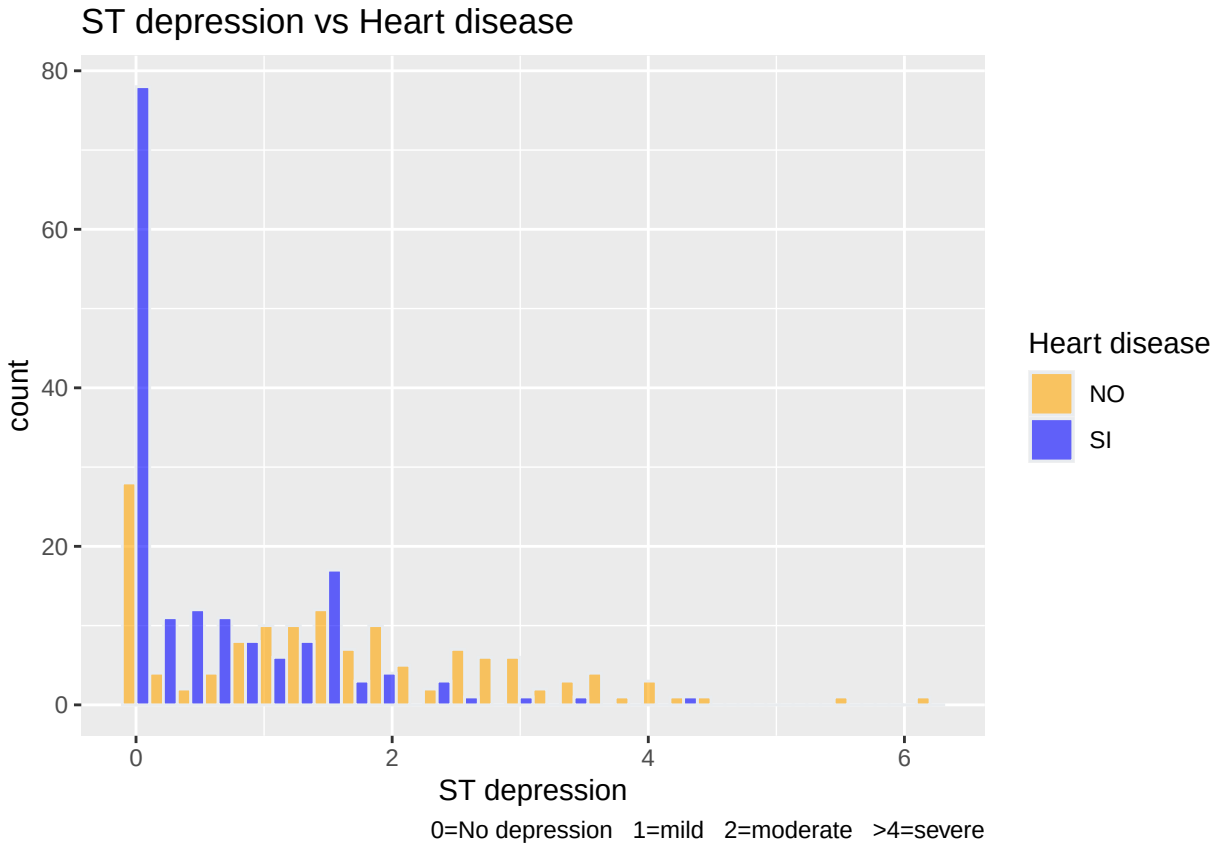


More specifically patients with heart disease have in our dataset a median heart rate over 150 while patients without heart disase below 150.





##OLDPEAK OR ST DEPRESSION ST depression refers to a finding on an electrocardiogram, wherein the trace in the ST segment is abnormally low below the baseline. ST-segment depression has been identified as a marker for adverse cardiac events in patients and it is often a sign of myocardial ischemia. In our dataset the majority of patients with heart disease had no/mild ST depression.



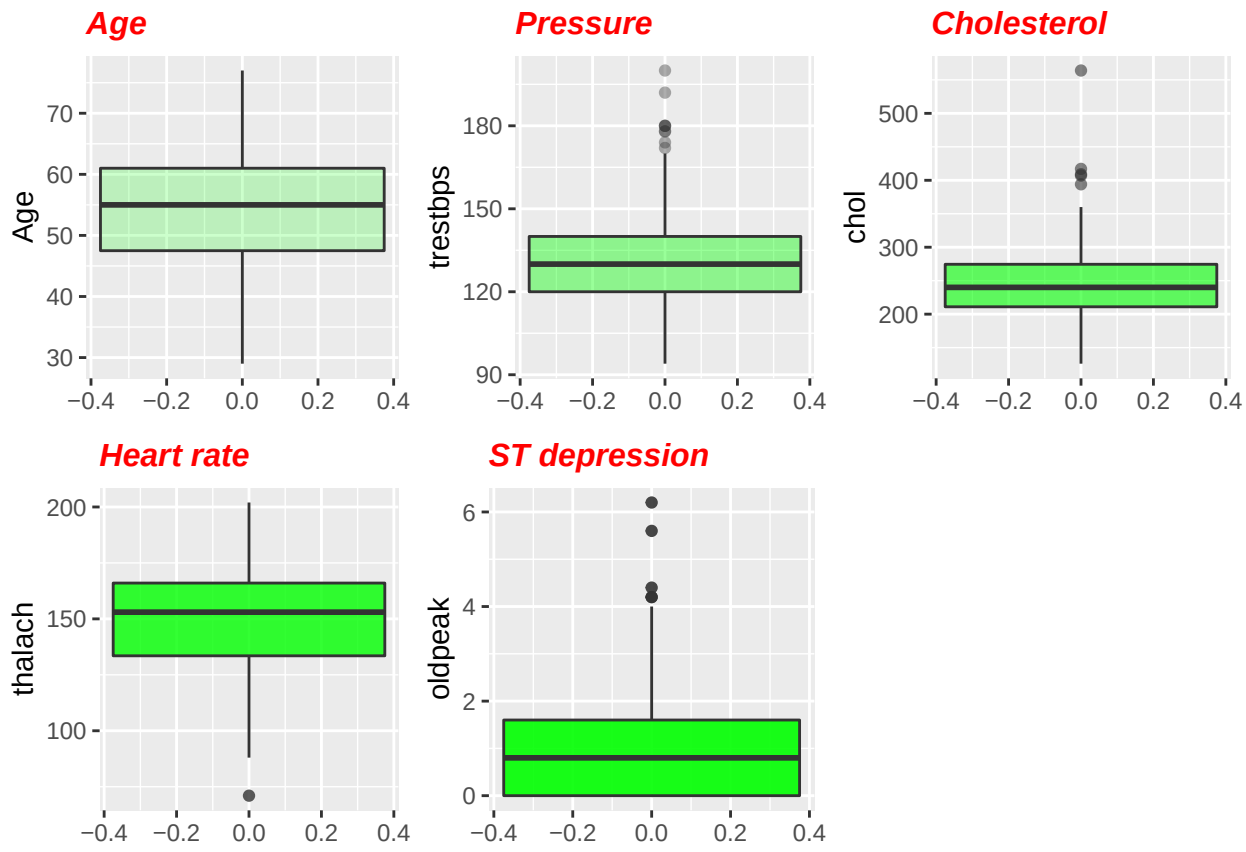
——> DATA CLEANING

Before buiding a train and a test set, let's find if we have some ouliers in the main features like age, rest blood pressure (trest), clolesterol (chol), maximum hear rate (thalac) and ST depression induced by exercise relative to rest (oldpeak), that could distort our final results

```
##          vars   n   mean    sd median trimmed   mad min   max range  skew
## age         1 303  54.37  9.08   55.0   54.54 10.38  29  77.0  48.0 -0.20
## trestbps    2 303 131.62 17.54  130.0  130.44 14.83  94 200.0 106.0  0.71
## chol        3 303 246.26 51.83  240.0  243.49 47.44 126 564.0 438.0  1.13
## thalach     4 303 149.65 22.91  153.0  150.98 22.24  71 202.0 131.0 -0.53
## oldpeak     5 303   1.04  1.16    0.8    0.86  1.19   0   6.2   6.2  1.26
##          kurtosis   se
## age         -0.57 0.52
## trestbps     0.87 1.01
## chol         4.36 2.98
## thalach     -0.10 1.32
## oldpeak      1.50 0.07
```

```
##       age           trestbps          chol          thalach         oldpeak
##  Min.   :29.00   Min.   : 94.0   Min.   :126.0   Min.   : 71.0   Min.   :0.00
##  1st Qu.:47.50   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.5   1st Qu.:0.00
##  Median :55.00   Median :130.0   Median :240.0   Median :153.0   Median :0.80
##  Mean   :54.37   Mean   :131.6   Mean   :246.3   Mean   :149.6   Mean   :1.04
```

```
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:166.0   3rd Qu.:1.60
##  Max.   :77.00   Max.   :200.0   Max.   :564.0   Max.   :202.0   Max.   :6.20
```



#Let's erase the values that exceed 3 standard deviation for rest blood pressure, cholesterol and ST depression

---

——> TRAIN AND TEST SET BUILDING

```
## Trainset dimension is 236 14
```

```
## Testset dimension is  58 14
```

—-> MODELING

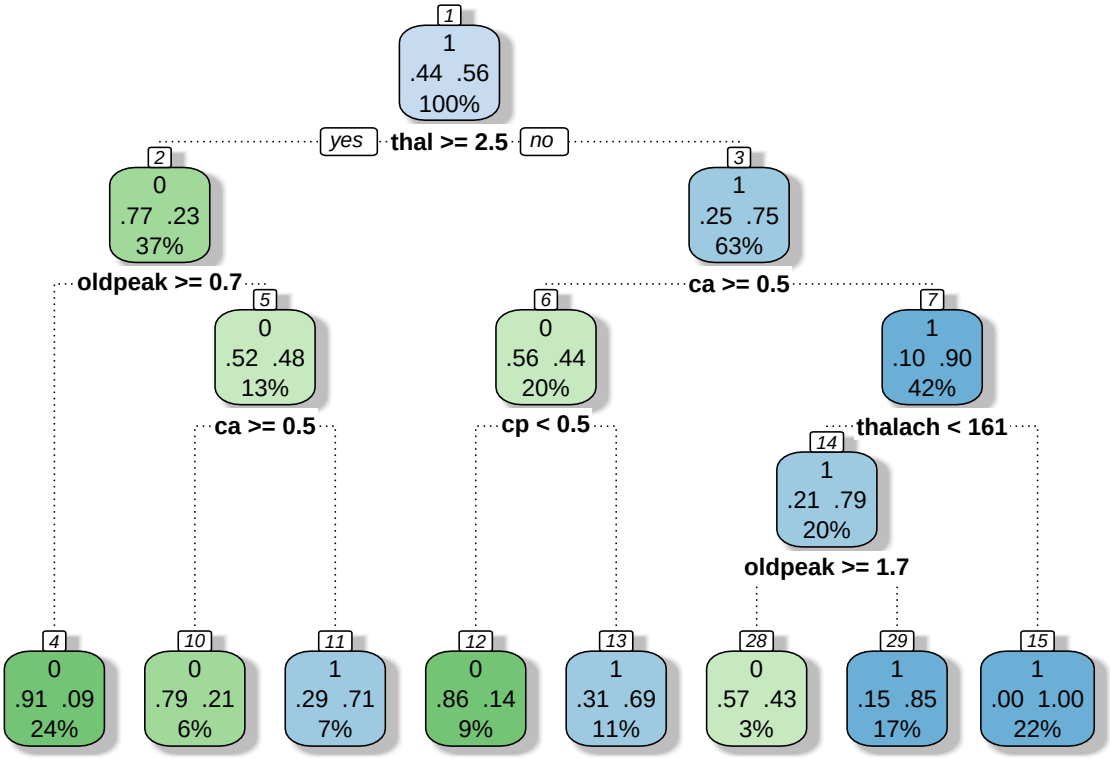
We want to predict if a pazient will have a heart disease (1) or will not. For this classification purpose We will take into account 6 machine learning algorithm:

- Decision Tree

- Random Forest
- Logistin regression
- Knn Algorithm
- Extreme Gradient Boosting Algorithm
- Naive Bayes

Each of them will be trained to find the best tune parameters. All the models will be finally compared with the accuracy.

1. ———> Decision Tree



Let's tune the Decision Tree with the best parameters and let's see the Confusion Matrix and the Accuracy of the model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 20  7
##          1  6 25
##
##                Accuracy : 0.7759
##                  95% CI : (0.6473, 0.8749)
##     No Information Rate : 0.5517
##     P-Value [Acc > NIR] : 0.0003327
```

```
##
##                   Kappa : 0.5485
##
##  Mcnemar's Test P-Value : 1.0000000
##
##             Sensitivity : 0.7692
##             Specificity : 0.7812
##          Pos Pred Value : 0.7407
##          Neg Pred Value : 0.8065
##              Prevalence : 0.4483
##          Detection Rate : 0.3448
##    Detection Prevalence : 0.4655
##       Balanced Accuracy : 0.7752
##
##        'Positive' Class : 0
##
```

```
## Accuracy for the Decision Tree Model is  0.7758621
```

2. ———> Random Forest

Let's see the Confusion Matrix and the Accuracy of the model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 19  4
##          1  7 28
##
##                Accuracy : 0.8103
##                  95% CI : (0.6859, 0.9013)
##     No Information Rate : 0.5517
##     P-Value [Acc > NIR] : 3.332e-05
##
##                   Kappa : 0.6124
##
##  Mcnemar's Test P-Value : 0.5465
##
##             Sensitivity : 0.7308
##             Specificity : 0.8750
##          Pos Pred Value : 0.8261
##          Neg Pred Value : 0.8000
##              Prevalence : 0.4483
##          Detection Rate : 0.3276
##    Detection Prevalence : 0.3966
##       Balanced Accuracy : 0.8029
##
##        'Positive' Class : 0
##
```

```
## Accuracy for the Random Forest Model is  0.8103448
```

3. ———> Logistic Regression Algorithm

Let's see the Confusion Matrix and the Accuracy of the model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 19  9
##          1  7 23
##
##                Accuracy : 0.7241
##                  95% CI : (0.591, 0.8334)
##     No Information Rate : 0.5517
##     P-Value [Acc > NIR] : 0.005315
##
##                   Kappa : 0.4463
##
##  Mcnemar's Test P-Value : 0.802587
##
##             Sensitivity : 0.7308
##             Specificity : 0.7188
##          Pos Pred Value : 0.6786
##          Neg Pred Value : 0.7667
##              Prevalence : 0.4483
##          Detection Rate : 0.3276
##    Detection Prevalence : 0.4828
##       Balanced Accuracy : 0.7248
##
##        'Positive' Class : 0
##
```

```
## Accuracy for the Logistic Regression model is  0.7241379
```

4. ———> Knn Algorithm

Let's see the Confusion Matrix and the Accuracy of the model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 16  9
##          1 10 23
##
##                Accuracy : 0.6724
##                  95% CI : (0.5366, 0.7899)
##     No Information Rate : 0.5517
##     P-Value [Acc > NIR] : 0.04178
##
##                   Kappa : 0.3353
##
```

```
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.6154
##             Specificity : 0.7188
##          Pos Pred Value : 0.6400
##          Neg Pred Value : 0.6970
##              Prevalence : 0.4483
##          Detection Rate : 0.2759
##    Detection Prevalence : 0.4310
##       Balanced Accuracy : 0.6671
##
##        'Positive' Class : 0
##
```

```
##  Accuracy
## 0.6724138
```

```
## Accuracy for the Knn model is  0.6724138
```

5. ———> Extreme Gradient Boosting Algorithm

Let's see the Confusion Matrix and the Accuracy of the model:

```
##   nrounds  eta max_depth gamma colsample_bytree min_child_weight subsample
## 1     100 0.01         2     0              0.4                1       0.5
```

```
## eXtreme Gradient Boosting
##
## 236 samples
##  13 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (3 fold)
## Summary of sample sizes: 157, 157, 158
## Resampling results:
##
##   Accuracy   Kappa
##   0.8602726  0.7131138
##
## Tuning parameter 'nrounds' was held constant at a value of 100
## Tuning
##  held constant at a value of 1
## Tuning parameter 'subsample' was held
##  constant at a value of 0.5
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 19  4
##          1  7 28
```

```
##
##                Accuracy : 0.8103
##                  95% CI : (0.6859, 0.9013)
##     No Information Rate : 0.5517
##     P-Value [Acc > NIR] : 3.332e-05
##
##                   Kappa : 0.6124
##
##  Mcnemar's Test P-Value : 0.5465
##
##             Sensitivity : 0.7308
##             Specificity : 0.8750
##          Pos Pred Value : 0.8261
##          Neg Pred Value : 0.8000
##              Prevalence : 0.4483
##          Detection Rate : 0.3276
##    Detection Prevalence : 0.3966
##       Balanced Accuracy : 0.8029
##
##        'Positive' Class : 0
##
```

```
##  Accuracy
## 0.8103448
```

```
## Accuracy for the Extreme Boosting Gradient model is  0.8103448
```

6. ———> Naive Bayes Algorithm

Let's see the Confusion Matrix and the Accuracy of the model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 18  4
##          1  8 28
##
##                Accuracy : 0.7931
##                  95% CI : (0.6665, 0.8883)
##     No Information Rate : 0.5517
##     P-Value [Acc > NIR] : 0.0001106
##
##                   Kappa : 0.5756
##
##  Mcnemar's Test P-Value : 0.3864762
##
##             Sensitivity : 0.6923
##             Specificity : 0.8750
##          Pos Pred Value : 0.8182
##          Neg Pred Value : 0.7778
##              Prevalence : 0.4483
```

```
##          Detection Rate : 0.3103
##    Detection Prevalence : 0.3793
##       Balanced Accuracy : 0.7837
##
##        'Positive' Class : 0
##
```

```
## Accuracy for the Naive Bayes model is  0.7931034
```

———> FINAL COMPARISON AMONG MODELS

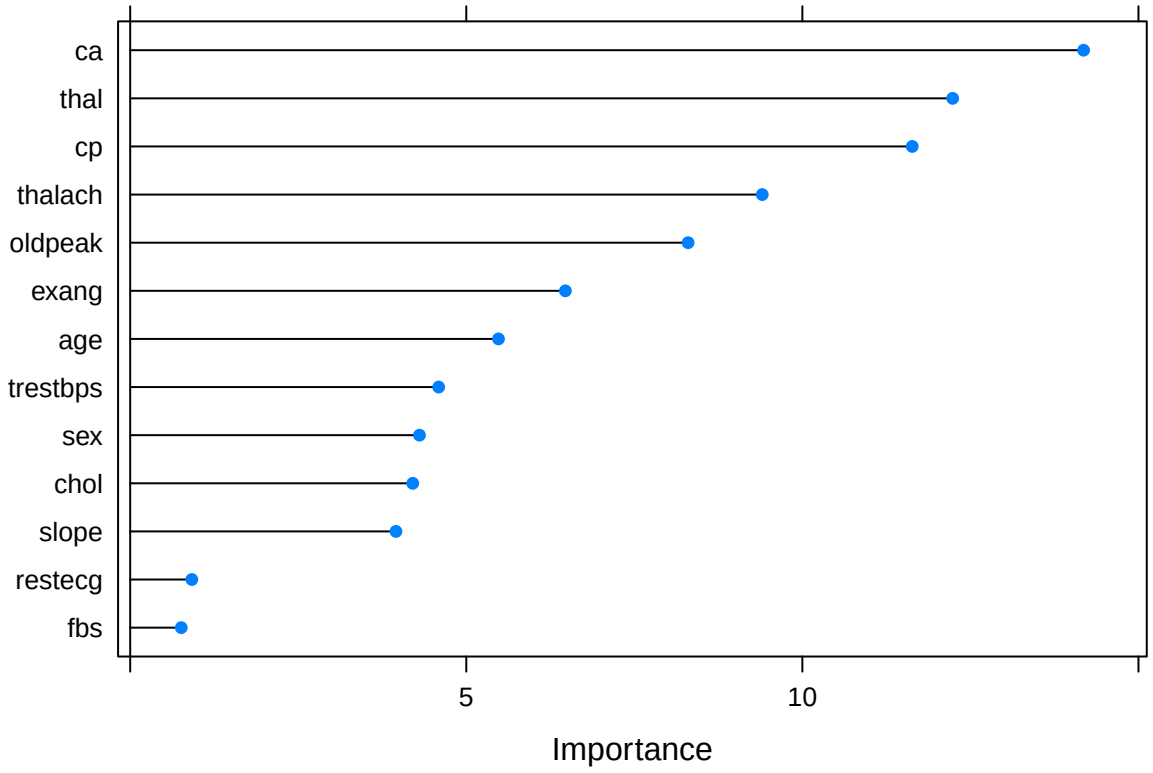
```
## Accuracy per varie tecniche:
```

	knn	logisticR	NaiveB	RandomF	ExtremeB	DecisionT
Accuracy	0.6724138	0.7241379	0.7931034	0.8103448	0.8103448	0.7758621

The Best Algorithms are the Random Forest and the Extreme Boost Gradient. They obtain the same result, but the processing time is longer with the Extreme Boost Gradient so the best algorithm, taking into account the effectiveness and the processing time, is the Random Forest. It gets a good result in terms of Accuracy and it isn't time consuming.

Importance Variables

In the end let's see the contribution of each variable to the Random Forest Algorithm.



—-> CONCLUSION

Let's imagine to have data provided by a cardiologist who has measured vitals on patients that have reported various cardiac symptoms. The goal of this project is to predict whether these patients have a risk to develop a cardiac disease.In this project we have trained 5 different machine learning models in order to achieve the goal to classify the patient at risk (0=no cardiac disease, 1 cardiac disease). The 5 machine learning models used as classification algorithms have been: **Decision Tree, Random Forest, Logistic regression, Knn Algorithm,Extreme Gradient Boosting Algorithm,Naive Bayes.** The algorithm that provided the best results, evaluated in terms of Accuracy and more efficient in terms of processing time, has been the Random Forest Algorithm.