HEART FAILURE PREDICTION ANALYSIS

Greta Angolani

March 2024

Supervised Learning Project DSE

1 Abstract

The aim of this paper is to explore the relationship between some common symptoms and other physiological characteristics and the presence of Heart Disease. I did this using four different supervised binary classification approaches like logistic regression, linear discriminant analysis, decision tree and random forest.

2 Introduction

The dataset Heart Failure Prediction Dataset was found on Kaggle and contains 918 observations with 12 attributes.

The 12 attributes are:

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS ¿ 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of ¿ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

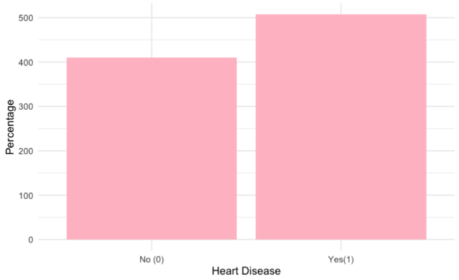
    As we can see also from the summary of the dataset, some variables are numerical and others are categorical, and they will be treated in different ways during the data preparation analysis.

    Research questions:

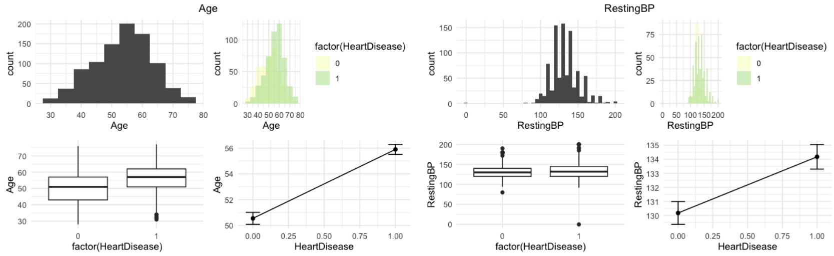
    The research has been made in order to predict, starting from some characteristics and features in the dataset, the presence of Heart Disease or not. So the Heart Disease feature in the dataset represents our target variable that we need to predict.

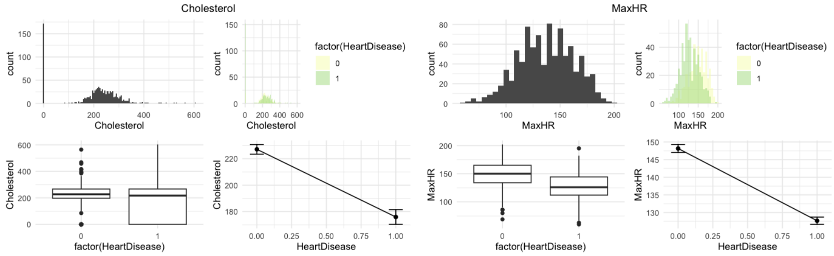
3 Data Pre-Processing

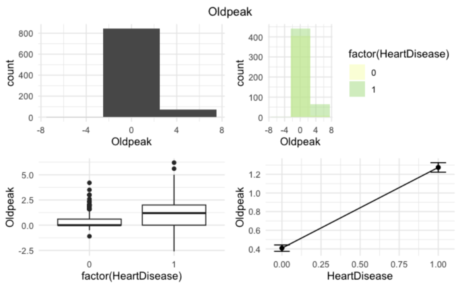
The first step is to check for any missing or duplicates values, which this dataset doesn't have. Also is needed to check if the target variable to predict, which is HeartDisease, has balanced values; to do this I have plotted the frequency which are 44.66% for negative values and 55.34% for positive values, and we can considered them to be balanced. I have also plotted the correspondent graph:

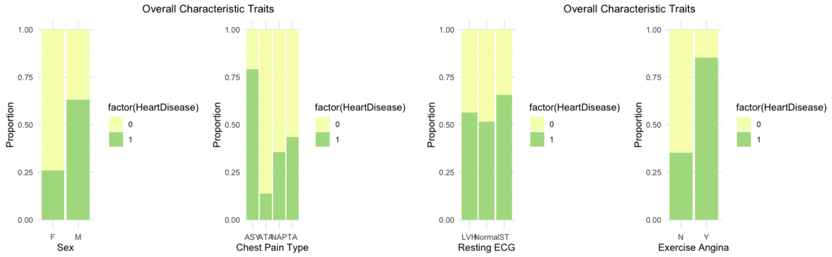


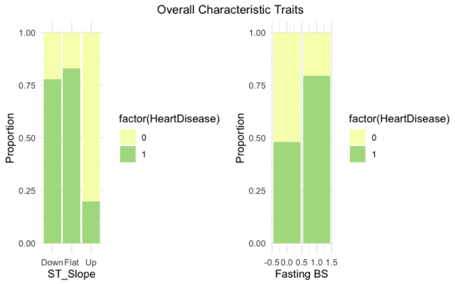
In this section I have plotted the graphs to inspect the distributions of the numerical and of the categorical features; as we can see there are some outliers in the dataset, which at the end I decided to keep since removing them were not changing the results.











After that another important things to do is to scale the numerical variables and to encode the categorical ones. This was done thanks to the scale function and one hot - enconding, keeping out from this process only one variable ”FastingBS” cause this was already encoded in the correct form. I have then checked for the correlation between variables in order to avoid multicollinearity problems in the analysis. As we can see in the figure the most highly correlated variables are:

(a) Sex M with Sex F = -1 (negative correaltion)

(b) ExerciseAnginaY and ExerciseAnginaN = -1 (negative correaltion)

(c) ST_SlopeUP with ST_SlopeFlat = -0.86 (negative correaltion)

(d) RestingECGNormal with RestingECGLVH = -0.62 (negative correaltion)

(e) RestingECGNormal with RestingECGST = -0.6 (negative correaltion)

(f) ChestpainTypeASY with ChestpainTypeATA and ChestpainTypeNAP = -0.52, -0.58 respectively (negative correlation)

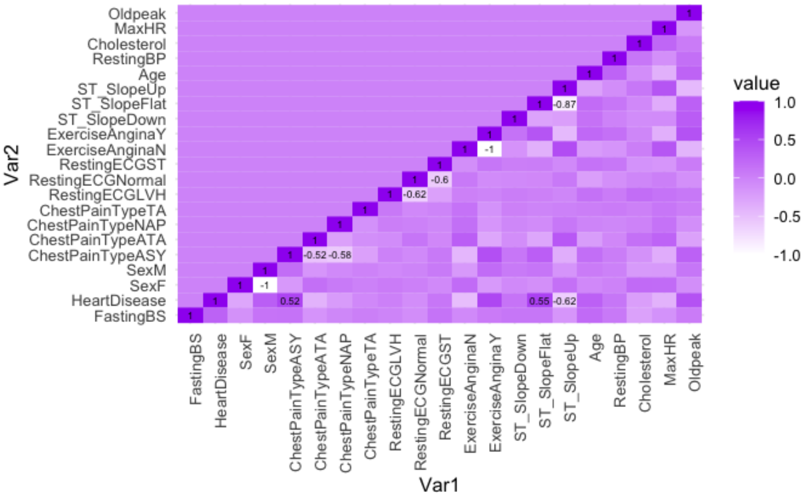


Figure 4: Correlation Matrix

I have also checked the correlation between all the variables with the target variable HeartDisease:

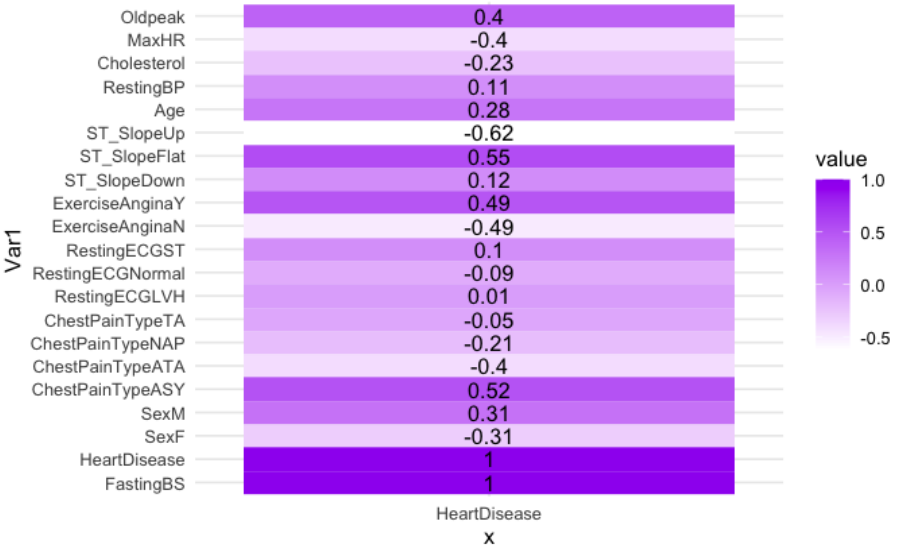


Figure 5: Correlation Matrix

Considering the correlation between all the variables with HeartDisease we can see that ST_Slope_Flat, ChestPainType_ASY, ExerciseAngina_Y, Oldpeak, Sex_M, Age, FastingBS are the variables with higher values of correaltion, and I can suppose that these are the variables that better can predict the presence of HeartDisease.

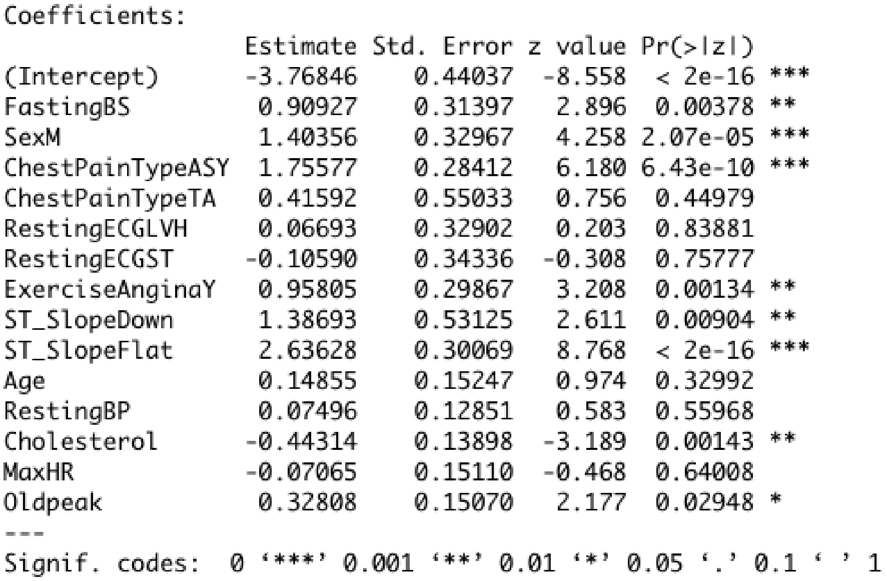
The variables that I had decided to exclude from the analysis are: "Sex F", "ChestPainTypeATA", "ChestPainTypeNAP", RestingECGNNormal", "ExerciseAnginaN", "ST_SlopeUP".

Before applying the supervised approaches, the dataset, which now consists has 6 variables less, needs to be splitted in train and test.

Numeric Feature

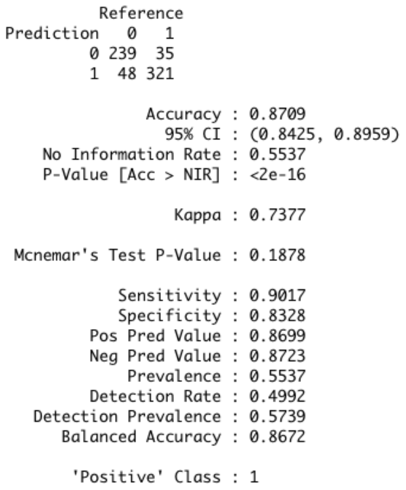
4 Logistic Regression

Logistic regression works taking as input multiple features and giving as an output values between 0 and 1. A default threshold was used at the beginning, corresponding to 0.5. Values above the threshold are assigned to the positive class (1 in this case indicating the presence of Heart Disease) and values below the threshold are assigned to the negative class (0 in this case). Below the logistic regression fit

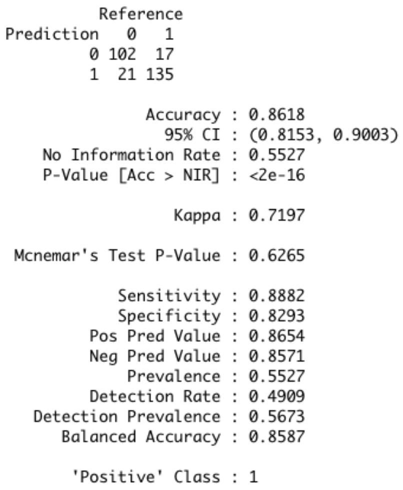


```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -3.76846    0.44037  -8.558  < 2e-16 ***
FastingBS          0.90927    0.31397   2.896  0.00378 ** 
SexM               1.40356    0.32967   4.258 2.07e-05 ***
ChestPainTypeASY   1.75577    0.28412   6.180 6.43e-10 ***
ChestPainTypeTA    0.41592    0.55033   0.756  0.44979    
RestingECGLVH      0.06693    0.32902   0.203  0.83881    
RestingECGST      -0.10590    0.34336  -0.308  0.75777    
ExerciseAnginaY    0.95805    0.29867   3.208  0.00134 ** 
ST_SlopeDown       1.38693    0.53125   2.611  0.00904 ** 
ST_SlopeFlat       2.63628    0.30069   8.768  < 2e-16 ***
Age                0.14855    0.15247   0.974  0.32992    
RestingBP          0.07496    0.12851   0.583  0.55968    
Cholesterol       -0.44314    0.13898  -3.189  0.00143 ** 
MaxHR             -0.07065    0.15110  -0.468  0.64008    
Oldpeak            0.32808    0.15070   2.177  0.02948 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The confusion matrix indicates a good performance by the logistic regression.



```
          Reference
Prediction   0   1
         0 239  35
         1  48 321

               Accuracy : 0.8709
                 95% CI : (0.8425, 0.8959)
    No Information Rate : 0.5537
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7377

 Mcnemar's Test P-Value : 0.1878

            Sensitivity : 0.9017
            Specificity : 0.8328
         Pos Pred Value : 0.8699
         Neg Pred Value : 0.8723
             Prevalence : 0.5537
         Detection Rate : 0.4992
   Detection Prevalence : 0.5739
      Balanced Accuracy : 0.8672

       'Positive' Class : 1
```



```
          Reference
Prediction   0   1
         0 102  17
         1  21 135

               Accuracy : 0.8618
                 95% CI : (0.8153, 0.9003)
    No Information Rate : 0.5527
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7197

 Mcnemar's Test P-Value : 0.6265

            Sensitivity : 0.8882
            Specificity : 0.8293
         Pos Pred Value : 0.8654
         Neg Pred Value : 0.8571
             Prevalence : 0.5527
         Detection Rate : 0.4909
   Detection Prevalence : 0.5673
      Balanced Accuracy : 0.8587

       'Positive' Class : 1
```

The ROC curve indicates a good performance, with 0.926 area under the cruve (AUC)

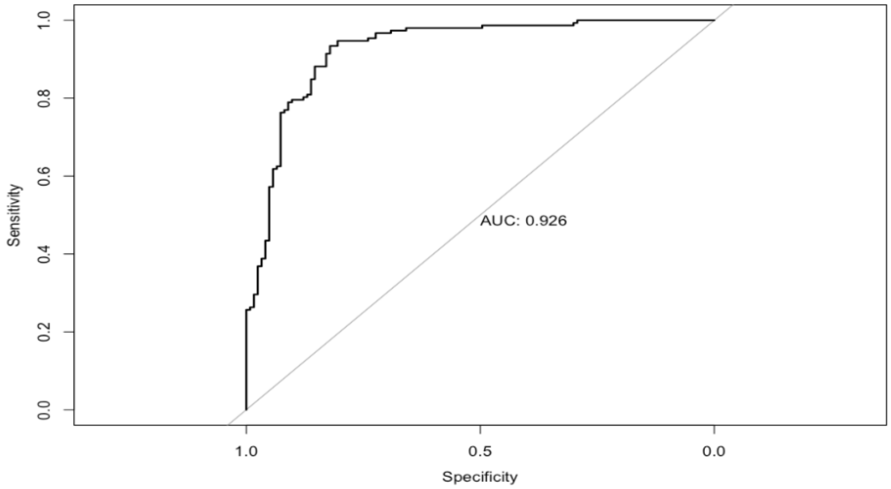


Figure 7: ROC curve

In this model, the accuracy and the sensitivity are the metrics to optimize. The default threshold chosen of 0.5 is able to achieve an 86.18% test accuracy. However other thresholds between 0 and 1 were explored at 0.01 increment. Below the correspondent threshold interval graphs:

As we can see, a threshold of 0.5 is very good but not the best one in

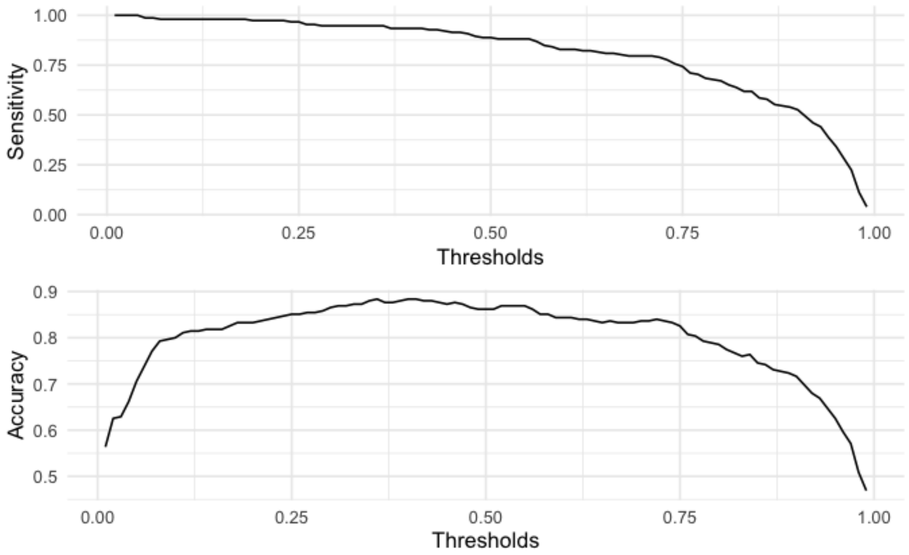


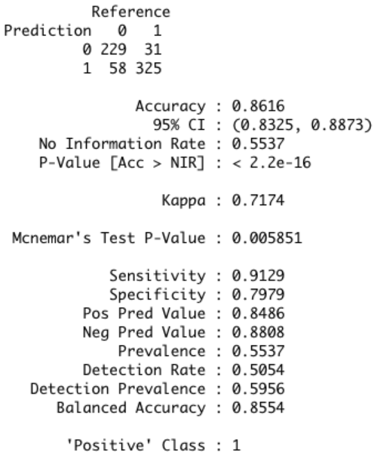
Figure 8: Sensitivity and Accuracy

terms of optimization of sensitivity and accuracy, that can be higher with a lower threshold. The option explored are:

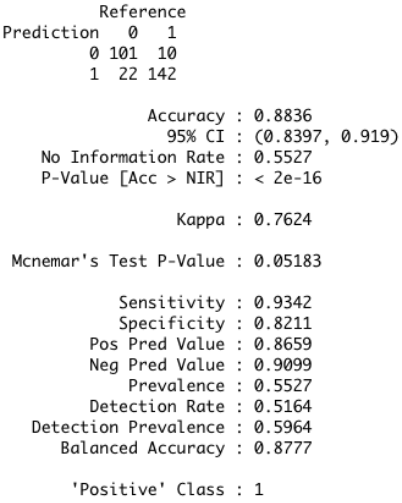
(a) Increasing Sensitivity: a threshold between 0.30 ¡= t ¡= 0.40 is able to increase both, accuracy and sensitivity

(b) Maximizing Accuracy: a threshold between 0.36 ¡= t ¡= 0.41 also increase both, and for t= 0.4 we have that accuracy is maximized and at the same time also sensitivity has improved

I have chosen option 2, Maximizing accuracy, and below are the new values with a threshold t = 0.4



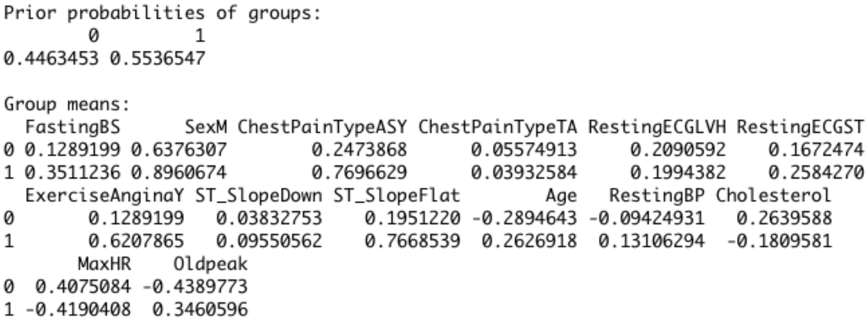
```
          Reference
Prediction   0   1
         0 229  31
         1  58 325

               Accuracy : 0.8616
                 95% CI : (0.8325, 0.8873)
    No Information Rate : 0.5537
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7174

 Mcnemar's Test P-Value : 0.005851

            Sensitivity : 0.9129
            Specificity : 0.7979
         Pos Pred Value : 0.8486
         Neg Pred Value : 0.8808
             Prevalence : 0.5537
         Detection Rate : 0.5054
   Detection Prevalence : 0.5956
      Balanced Accuracy : 0.8554

       'Positive' Class : 1
```



```
          Reference
Prediction   0   1
         0 101  10
         1  22 142

               Accuracy : 0.8836
                 95% CI : (0.8397, 0.919)
    No Information Rate : 0.5527
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7624

 Mcnemar's Test P-Value : 0.05183

            Sensitivity : 0.9342
            Specificity : 0.8211
         Pos Pred Value : 0.8659
         Neg Pred Value : 0.9099
             Prevalence : 0.5527
         Detection Rate : 0.5164
   Detection Prevalence : 0.5964
      Balanced Accuracy : 0.8777

       'Positive' Class : 1
```

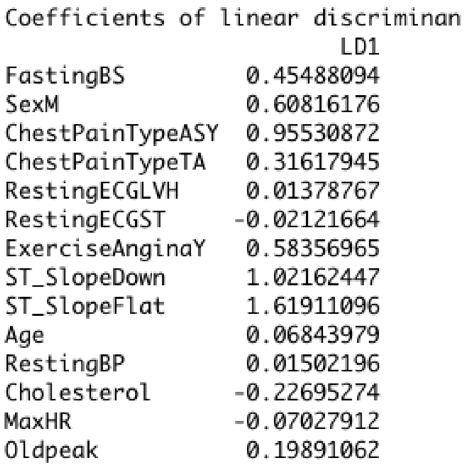
5 Linear Discriminant Analysis (LDA)

Using the Linear Discriminant analysis on the dataset yields the following results:

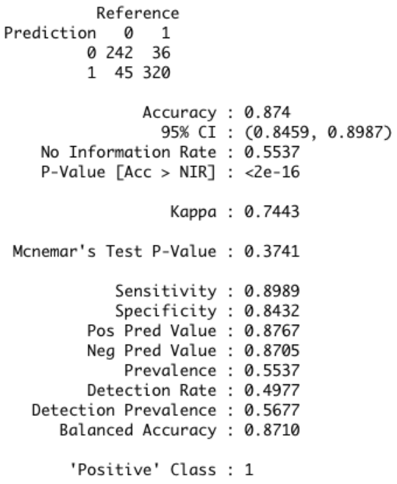


```
Prior probabilities of groups:
        0         1 
0.4463453 0.5536547 

Group means:
  FastingBS      SexM ChestPainTypeASY ChestPainTypeTA RestingECGLVH RestingECGST
0 0.1289199 0.6376307        0.2473868      0.05574913     0.2090592    0.1672474
1 0.3511236 0.8960674        0.7696629      0.03932584     0.1994382    0.2584270
  ExerciseAnginaY ST_SlopeDown ST_SlopeFlat        Age   RestingBP Cholesterol
0       0.1289199   0.03832753    0.1951220 -0.2894643 -0.09424931   0.2639588
1       0.6207865   0.09550562    0.7668539  0.2626918  0.13106294  -0.1809581
       MaxHR    Oldpeak
0  0.4075084 -0.4389773
1 -0.4190408  0.3460596
```

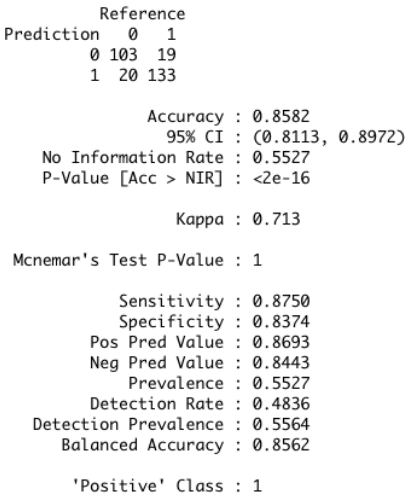
As we can see below the performance is comparable to the Logistic Regression, in terms of accuracy and sensitivity (the values are almost the same)



```
Coefficients of linear discriminan
                          LD1
FastingBS          0.45488094
SexM               0.60816176
ChestPainTypeASY   0.95530872
ChestPainTypeTA    0.31617945
RestingECGLVH      0.01378767
RestingECGST      -0.02121664
ExerciseAnginaY    0.58356965
ST_SlopeDown       1.02162447
ST_SlopeFlat       1.61911096
Age                0.06843979
RestingBP          0.01502196
Cholesterol       -0.22695274
MaxHR             -0.07027912
Oldpeak            0.19891062
```

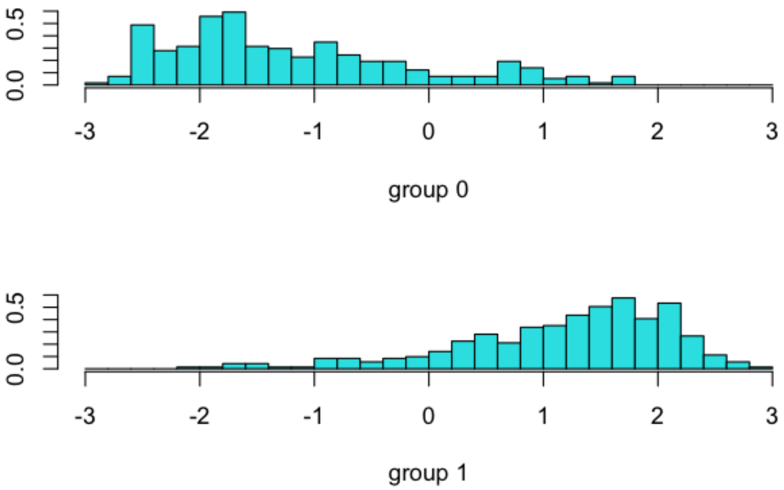


```
          Reference
Prediction   0   1
         0 242  36
         1  45 320

               Accuracy : 0.874
                 95% CI : (0.8459, 0.8987)
    No Information Rate : 0.5537
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7443

 Mcnemar's Test P-Value : 0.3741

            Sensitivity : 0.8989
            Specificity : 0.8432
         Pos Pred Value : 0.8767
         Neg Pred Value : 0.8705
             Prevalence : 0.5537
         Detection Rate : 0.4977
   Detection Prevalence : 0.5677
      Balanced Accuracy : 0.8710

       'Positive' Class : 1
```



```
          Reference
Prediction   0   1
         0 103  19
         1  20 133

               Accuracy : 0.8582
                 95% CI : (0.8113, 0.8972)
    No Information Rate : 0.5527
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.713

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8750
            Specificity : 0.8374
         Pos Pred Value : 0.8693
         Neg Pred Value : 0.8443
             Prevalence : 0.5527
         Detection Rate : 0.4836
   Detection Prevalence : 0.5564
      Balanced Accuracy : 0.8562

       'Positive' Class : 1
```

The plot below shows the spread of the linear combination of the two most dominant lags in the LDA. The two response classes have different centers and spreads, indicating that they can be distinguished well by the LDA model.



6 Decision Tree

The fitted tree model is below, and it can be easily read by starting at the root and moving along the paths of the data point, until a leaf is reached.

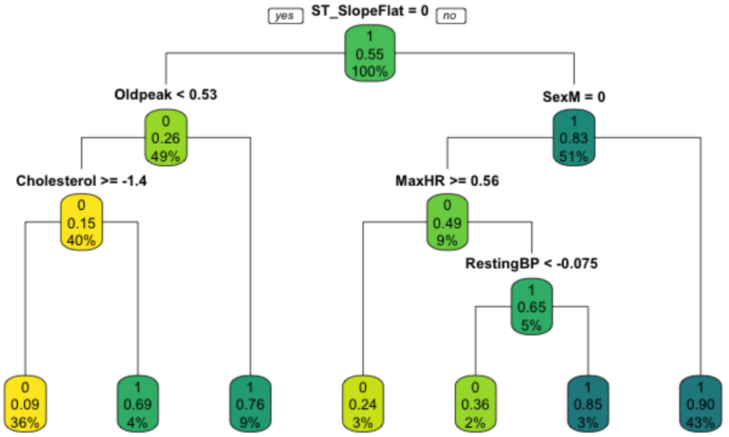
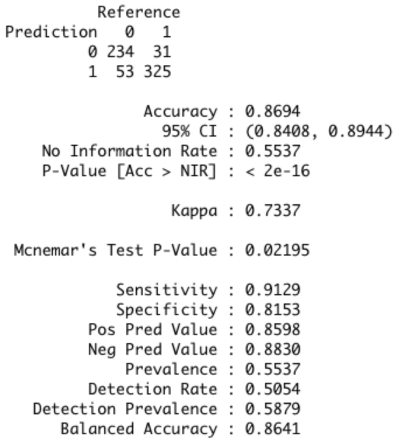
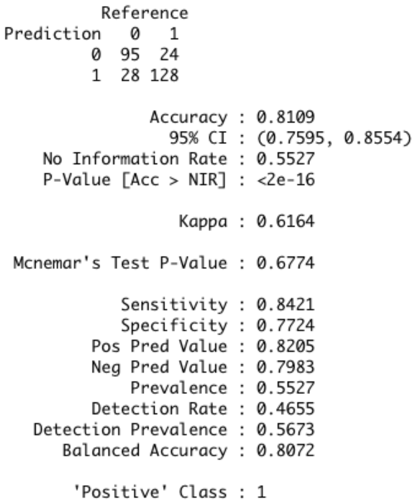


Figure 11: DECISION TREE



```
          Reference
Prediction   0   1
         0 234  31
         1  53 325

               Accuracy : 0.8694
                 95% CI : (0.8408, 0.8944)
    No Information Rate : 0.5537
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7337

 Mcnemar's Test P-Value : 0.02195

            Sensitivity : 0.9129
            Specificity : 0.8153
         Pos Pred Value : 0.8598
         Neg Pred Value : 0.8830
             Prevalence : 0.5537
         Detection Rate : 0.5054
   Detection Prevalence : 0.5879
      Balanced Accuracy : 0.8641
```



```
          Reference
Prediction   0   1
         0  95  24
         1  28 128

               Accuracy : 0.8109
                 95% CI : (0.7595, 0.8554)
    No Information Rate : 0.5527
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6164

 Mcnemar's Test P-Value : 0.6774

            Sensitivity : 0.8421
            Specificity : 0.7724
         Pos Pred Value : 0.8205
         Neg Pred Value : 0.7983
             Prevalence : 0.5527
         Detection Rate : 0.4655
   Detection Prevalence : 0.5673
      Balanced Accuracy : 0.8072

       'Positive' Class : 1
```

The values obtained for accuracy and sensitivity are quite good, but it could have been worked better.

6.1 Pruned Decision Tree

As we saw the training accuracy for the previous decision tree is 86.94%, while the testing accuracy is 81.09%. Since the testing accuracy is less than the training accuracy, this can lead us to overfitting problem. I have tried to avoid the problem pruning the tree, but unfortunately this could only slightly improved the testing accuracy, which now is 81.82%.

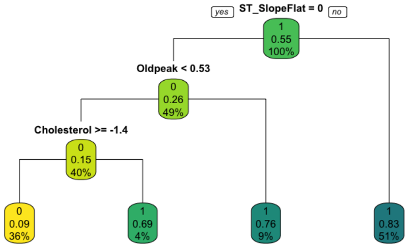
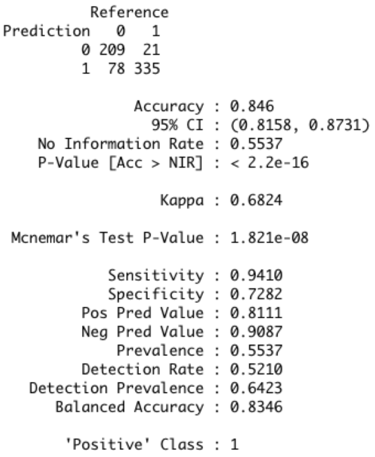
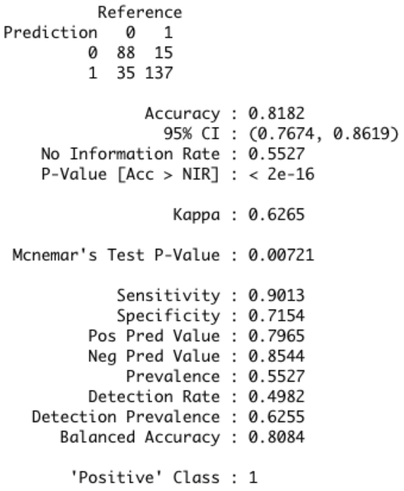


Figure 13: PRUNED DECISION TREE

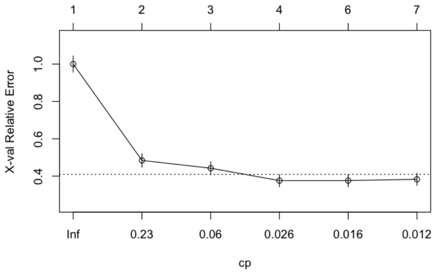


```
          Reference
Prediction   0   1
         0 209  21
         1  78 335

               Accuracy : 0.846
                 95% CI : (0.8158, 0.8731)
    No Information Rate : 0.5537
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6824

 Mcnemar's Test P-Value : 1.821e-08

            Sensitivity : 0.9410
            Specificity : 0.7282
         Pos Pred Value : 0.8111
         Neg Pred Value : 0.9087
             Prevalence : 0.5537
         Detection Rate : 0.5210
   Detection Prevalence : 0.6423
      Balanced Accuracy : 0.8346

       'Positive' Class : 1
```



```
          Reference
Prediction   0   1
         0  88  15
         1  35 137

               Accuracy : 0.8182
                 95% CI : (0.7674, 0.8619)
    No Information Rate : 0.5527
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6265

 Mcnemar's Test P-Value : 0.00721

            Sensitivity : 0.9013
            Specificity : 0.7154
         Pos Pred Value : 0.7965
         Neg Pred Value : 0.8544
             Prevalence : 0.5527
         Detection Rate : 0.4982
   Detection Prevalence : 0.6255
      Balanced Accuracy : 0.8084

       'Positive' Class : 1
```

The complexity parameter (CP) chosen as an optimal threshold is 0.02, based on the following plot:

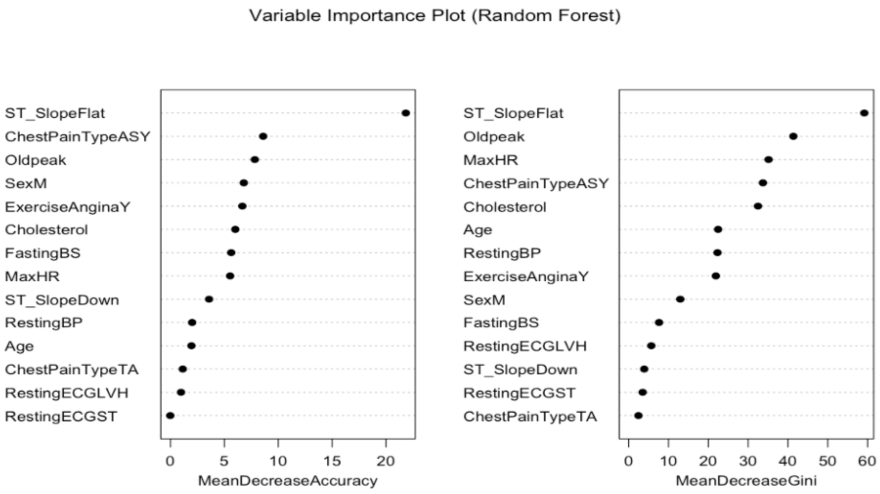


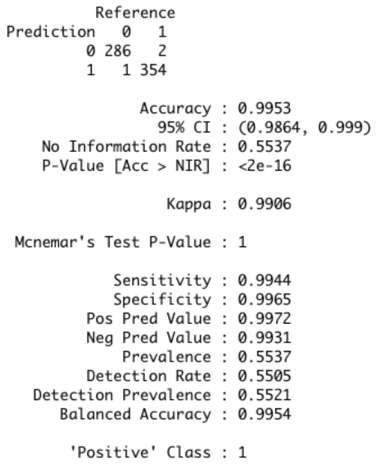
7 Random Forest

Random Forest improve on bagged tree by de-correlating the trees and reducing the variance. This is achieved by randomizing the selection features available to the model at each tree split.

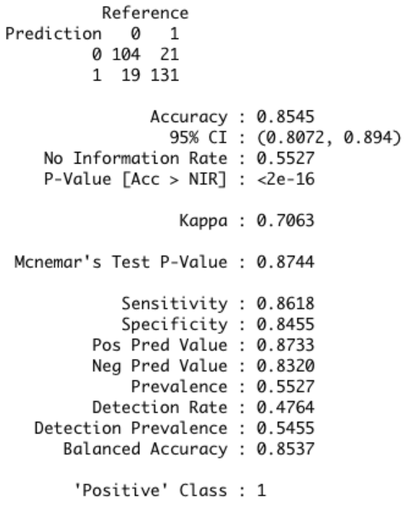
The figure below computes the most important variables for the model, and is clear how the ST_SlopeFlat, Oldpeak and ChestPainType ASY are the 3 most important values to consider and watch out for predicting HeartDisease.

As we can see below the random forest has improved both, accuracy and sensitivity over the prior trees.





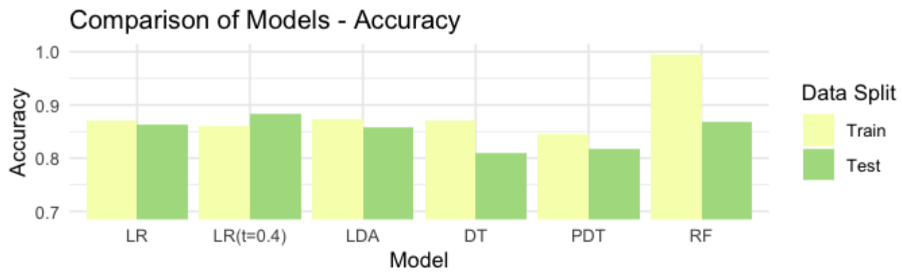
```
          Reference
Prediction   0   1
         0 286   2
         1   1 354
                                           
               Accuracy : 0.9953           
                 95% CI : (0.9864, 0.999)  
    No Information Rate : 0.5537           
    P-Value [Acc > NIR] : <2e-16           
                                           
                  Kappa : 0.9906           
                                           
 Mcnemar's Test P-Value : 1                
                                           
            Sensitivity : 0.9944           
            Specificity : 0.9965           
         Pos Pred Value : 0.9972           
         Neg Pred Value : 0.9931           
             Prevalence : 0.5537           
         Detection Rate : 0.5505           
   Detection Prevalence : 0.5521           
      Balanced Accuracy : 0.9954           
                                           
       'Positive' Class : 1                
```

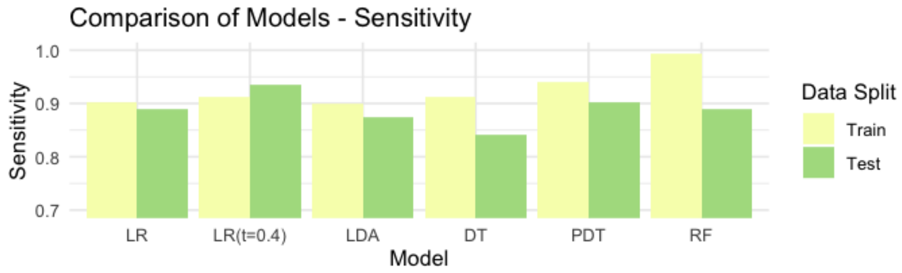


```
          Reference
Prediction   0   1
         0 104  21
         1  19 131
                                           
               Accuracy : 0.8545           
                 95% CI : (0.8072, 0.894)  
    No Information Rate : 0.5527           
    P-Value [Acc > NIR] : <2e-16           
                                           
                  Kappa : 0.7063           
                                           
 Mcnemar's Test P-Value : 0.8744           
                                           
            Sensitivity : 0.8618           
            Specificity : 0.8455           
         Pos Pred Value : 0.8733           
         Neg Pred Value : 0.8320           
             Prevalence : 0.5527           
         Detection Rate : 0.4764           
   Detection Prevalence : 0.5455           
      Balanced Accuracy : 0.8537           
                                           
       'Positive' Class : 1                
```

8 Model Comparison

All the models achieved high accuracy and sensitivity results, confirming the presence of Heart Disease if some symptoms/ characteristics are present. The Random Forest is the model that had performed better than all the others, while the one that had performed worst is the decision tree. The other models had produced very similar perfomances i terms of accuracy and sensibility. Below the model comparison plot:





9 Conclusion

In conclusion of the work, we can say that we have a good overview of which are the main symptoms that cause HeartDisease, and we understood to which one we need to pay more attention than others. Of course, the higher the number of physiological symptoms present in the dataset, the higher is the presence of HeartDisease. Moreover we can see that ”Age”, the only variables which is not a symptom, has a quite high influence in the presence of HeartDisease.