Recognition of Heart Disease

STATISTICAL TOOLS

Elianne Mora | Statistical Learning | December 8, 2019

Introduction

In this project we analyze the 'Heart Disease' dataset from the clinical and noninvasive test results of 303 patients at the Cleveland Clinic in Cleveland, Ohio. While the original data collected from the 303 patients contained 75 variables, the final subset that was made available contained only 14. We have focused on the subset data for the application of classification tools; however, only 13 variables were kept. The variable *thal* was eliminated due to inconsistent reporting of the values and their respective meanings. The variables are defined as follows:

ID	*Description*
Age	Patient's age
Sex	Patient's sex (0=female and 1=male)
Cp	The chest pain experienced (Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic)
Trestbps	Patient's resting blood pressure (mm Hg on admission to the hospital)
Chol	Patient's cholesterol levels in mg/dl
Fbs	Person's fasting blood sugar (if > 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' crite
Thalach	The person's maximum heart rate achieved during Stress Test (exercise)
Exang	Exercise induced angina (1=yes, 0=no)
Oldpeak	Stress test depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
Slope	the slope of the peak exercise ST segment (Value 0: upsloping, Value 1: flat, Value 2: downsloping)
Ca	The number of major vessels colored by fluoroscopy
Target	Diagnosis of heart disease (1=yes) or no heart disease (0=no).

In this project we are interested in the predictive capabilities of different classification models for the variable *Target*, given all other records. We wish to find a model that will give an accurate diagnosis with a certain level of significance. In this scenario, wrongly diagnosing a patient with heart disease in the absence of it, is as bad as misdiagnosing a person who is in fact ill. In the first case, the patient will likely be admitted to a hospital, further incur medical expenses, intake unnecessary drugs, be submitted to further tests and procedures, etc. The latter will be sent home at risk of having a stroke, heart attack, further complications affecting daily life, and even death.

It is not clear how the data was collected for this heart disease study. It is unknown whether the patients were scheduled for a visit, selected for a study given their health history, or simply attended the hospital for an "urgent" care visit (no prior appointments). Given the location, the administered exams recorded, research focus and popularity of the Cleveland Clinic, it is highly unlikely that the project data is a simple random sample representative of the population (residents of Ohio), nor that it was collected in an "Emergency Room"; however, we will explore the applications of classification models on this data set, such that it may be used in such a setting.

DATA PROCESSING AND VISUALIZATIONS

The data collected was complete, no missing values nor zeros were found in the continuous variables. The first step taken was the transformation of applicable variables into factors and deletion of the variable *thal*.

```
Heart                    303 obs. of 14 variables
 ï..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 ...
```

```
Heart                    303 obs. of 13 variables
 ï..age : int 63 37 41 56 57 57 56 44 52 57 ...
 sex : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 1 2 2 2 ...
 cp : Factor w/ 4 levels "Typical Angina",..: 4 3 2 2 1 1 2 2 3 3 ...
 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 : Factor w/ 2 levels "Below120","Above120": 2 1 1 1 1 1 1 1 2 1 ...
 restecg : Factor w/ 3 levels "Normal","Abnormal",..: 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 "No","Yes": 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 "Upslopping","Flat",..: 1 1 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 ...
 target : Factor w/ 2 levels "No Heart Disease",..: 2 2 2 2 2 2 2 2 2 2 ...
```

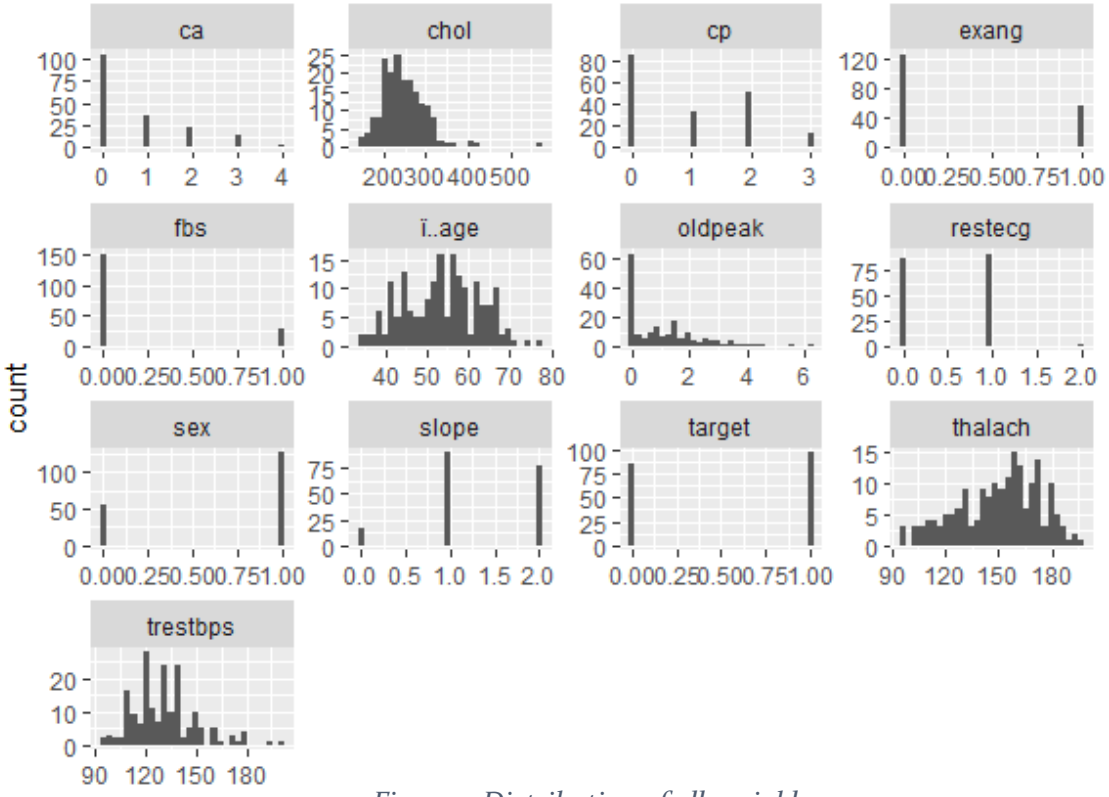
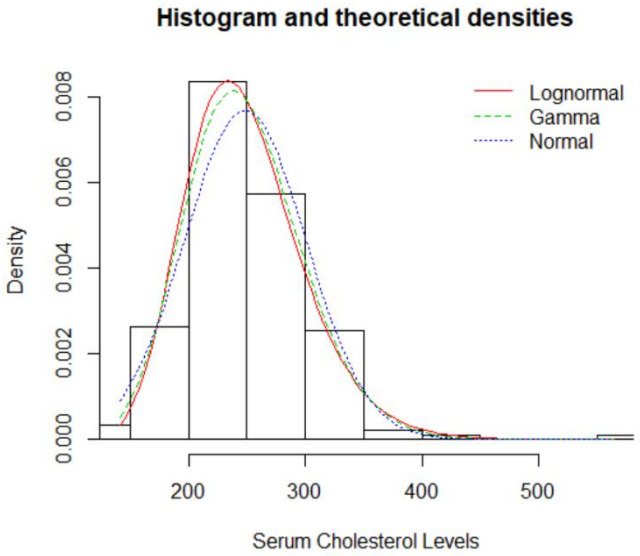
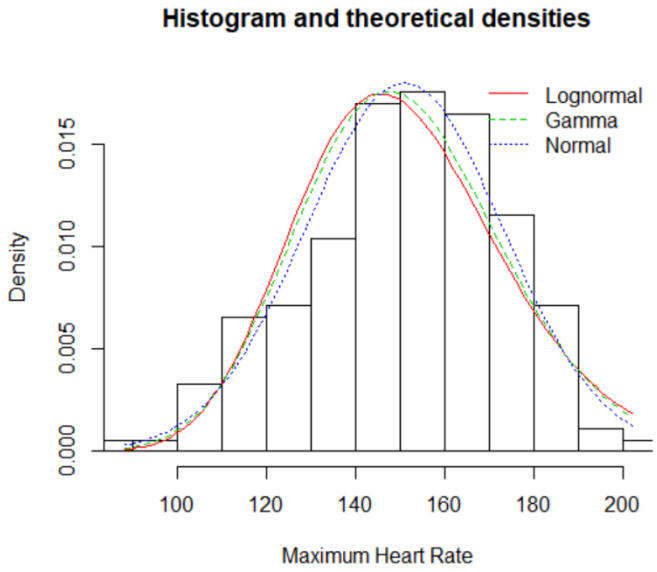
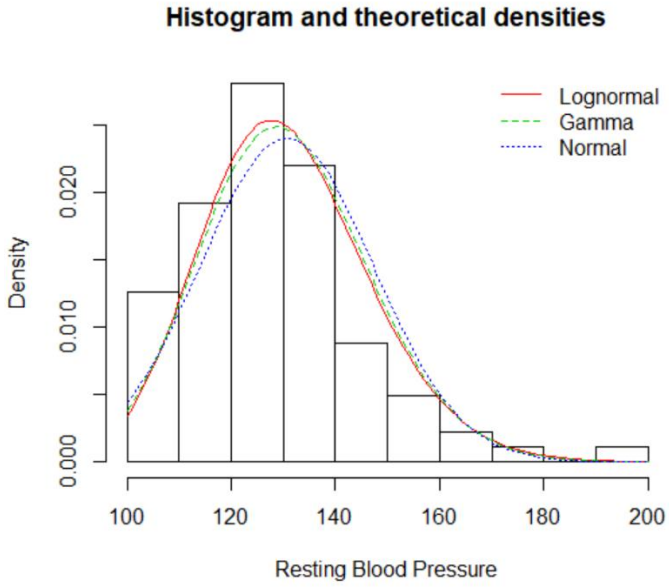


Figure 1. Distribution of all variables.

The dataset was split into training (60%) and testing (40%) sets. It can be observed that the variables recording cholesterol, maximum heart rate and resting blood pressure seem to follow a lognormal distribution.







The full dataset contained 96 females of which 72 were diagnosed with heart disease (75%), while out of 207 males, 93 were diagnosed with heart disease (45%). The *training* dataset contains 64 females and 118 males, and roughly 75% and 43% women and men were positively diagnosed with the disease. The classifications are as follows:

	No Heart Disease	Heart Disease
Female	16	48
Male	67	51

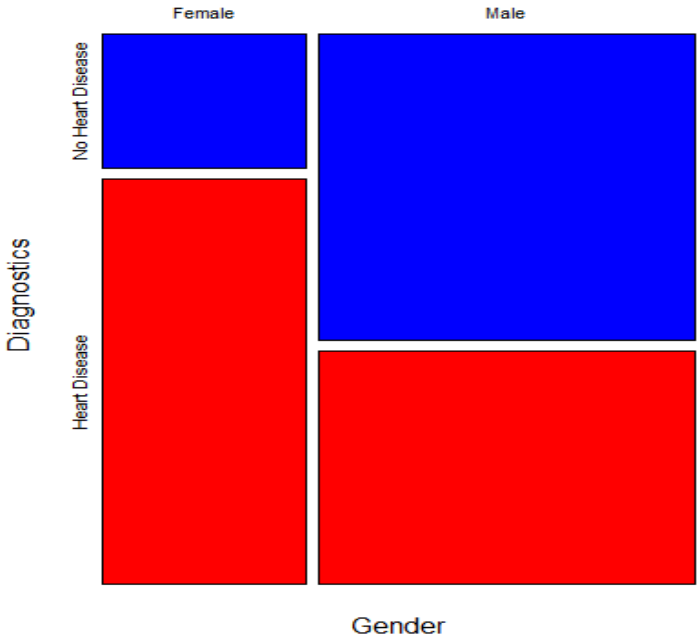


Figure 2. The table shows the diagnosis distribution based on gender for the Training set.

The literature shows that men and women are equally at risk of heart disease. While genetics and family history are significant factors, other characteristics that increase the risk of heart disease are age, unhealthy diets and sedentarism leading to high blood pressure and high cholesterol levels, among many others. Our data, however, shows that most of the women positively diagnosed have cholesterol levels equal to or below those of the women with negative diagnosis.

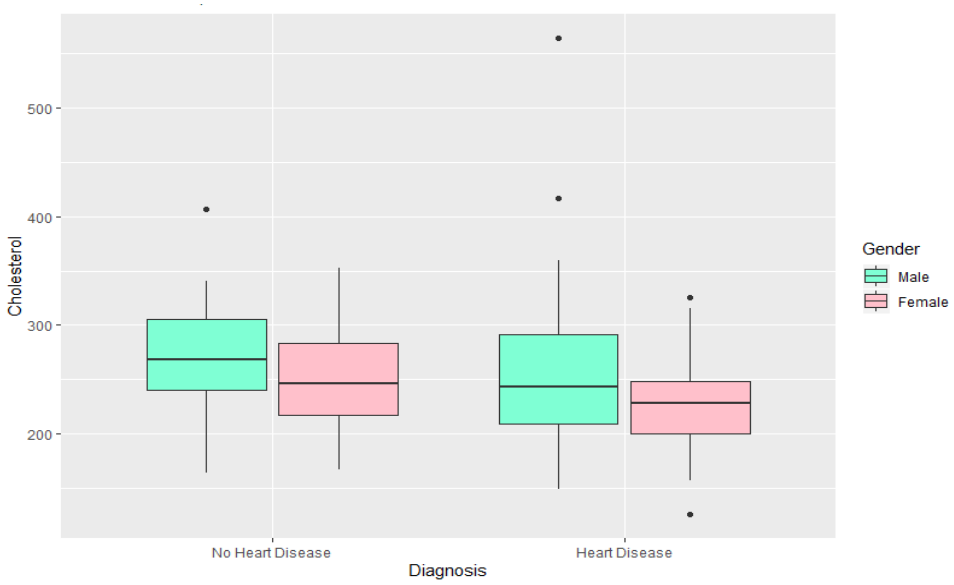


Figure 3. The box plot shows the distribution of diagnosis given cholesterol levels and gender. All outliers were kept for this project.

There is a significantly high outlier for a 67 year-old male, whose cholesterol levels was 564 mg/dl, reported non-anginal chest pain, normal systolic blood pressure of 115 and a normal electrocardiographic measurement, whose diagnosis is positive. There's also a significantly "low" outlier for a 57 year-old woman, with 126 mg/dl cholesterol, non-anginal chest pain, elevated systolic blood pressure at 150, an abnormal electrocardiographic measurement, whose diagnosis is also positive. We have chosen to include all outliers for our study, as they may provide significant components to our classification problem.

CLASSIFICATION MODELING

We first approach this classification problem with Logistic Regression. The model has been built to estimate the binary response (*Heart Disease vs. No Heart Disease)*, given all other variables in the dataset. The fitted model is as follows:

```
Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5096  -0.2921   0.1248   0.5192   2.9223

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.536e+00  3.516e+00   0.437 0.662224
ï..age              3.318e-02  3.340e-02   0.993 0.320522
sexMale            -2.073e+00  6.689e-01  -3.099 0.001939 **
cpAtypical Angina   1.213e+00  6.829e-01   1.776 0.075787 .
cpNon-anginal Pain  2.016e+00  6.445e-01   3.128 0.001760 **
cpAsymptomatic      3.207e+00  9.519e-01   3.369 0.000755 ***
trestbps           -1.522e-02  1.365e-02  -1.115 0.265000
chol               -8.208e-03  5.134e-03  -1.599 0.109919
fbsAbove120         4.527e-01  6.945e-01   0.652 0.514474
restecgAbnormal    -2.211e-01  4.932e-01  -0.448 0.653912
restecgHypertrophy -1.490e+01  2.756e+03  -0.005 0.995687
thalach             2.134e-02  1.493e-02   1.429 0.152998
exangYes           -4.489e-01  5.734e-01  -0.783 0.433660
oldpeak            -6.031e-01  3.269e-01  -1.845 0.065029 .
slopeFlat          -1.008e+00  9.333e-01  -1.080 0.280243
slopeDownslopping   6.258e-01  1.024e+00   0.611 0.541239
ca1                -2.178e+00  6.428e-01  -3.388 0.000705 ***
ca2                -4.154e+00  1.035e+00  -4.015 5.95e-05 ***
ca3                -2.457e+00  1.273e+00  -1.929 0.053709 .
ca4                 1.587e+01  2.049e+03   0.008 0.993819
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 250.90  on 181  degrees of freedom
Residual deviance: 119.99  on 162  degrees of freedom
AIC: 159.99

Number of Fisher Scoring iterations: 16
```

Figure 4. Logistic Regression results with all regressors from the Train dataset.

The coefficients can be interpreted such that for a one-unit increase in x_i the expected change in log odds is β_i. In order to validate the predicted probability of our model, the ROC curve and Confusion Matrix were computed using the *Test* dataset. The area under the ROC curve is 0.93 which is considered to be very good; the larger the AUC, the better the ability of the model to discriminate between

individuals that present heart disease and those who do not. From this plot we can also observe that the optimal cutoff is 0.544 and overall the model is very good in terms of predicting power.

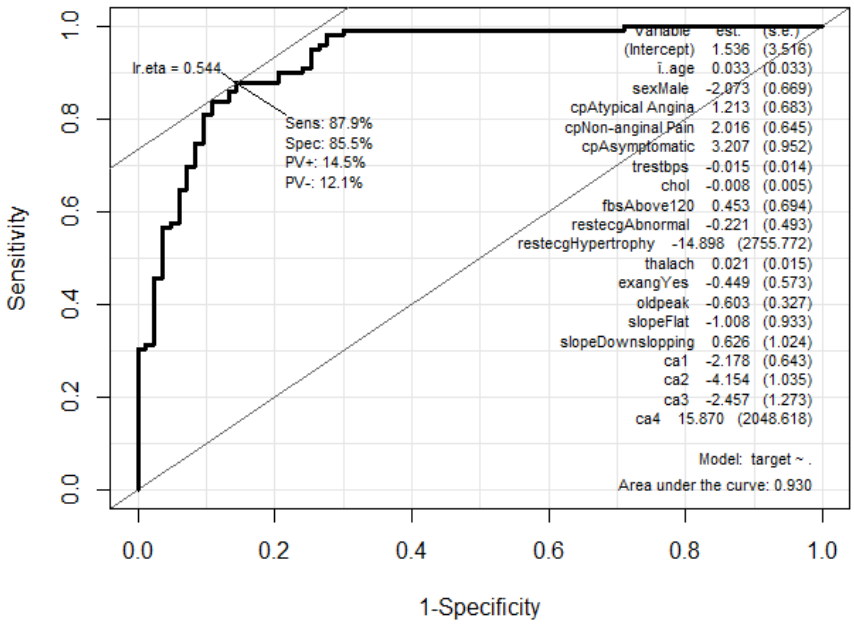


Figure 5. ROC curve for full model

The results from the Confusion Matrix are very similar to the ones found in the ROC curve. The accuracy, sensitivity, and specificity are all above 80%. If we were to always predict/classify a patient's diagnosis at a rate of 54.55% (no information rate) using a dummy model that always predicts "No Heart Disease", our logistic regression model would be much more accurate.

```
Confusion Matrix and Statistics

                Reference
Prediction         No Heart Disease Heart Disease
  No Heart Disease               47            11
  Heart Disease                   8            55

               Accuracy : 0.843
                 95% CI : (0.7657, 0.9027)
    No Information Rate : 0.5455
    P-Value [Acc > NIR] : 3.891e-12

                  Kappa : 0.6848

 Mcnemar's Test P-Value : 0.6464

            Sensitivity : 0.8545
            Specificity : 0.8333
         Pos Pred Value : 0.8103
         Neg Pred Value : 0.8730
             Prevalence : 0.4545
         Detection Rate : 0.3884
   Detection Prevalence : 0.4793
      Balanced Accuracy : 0.8439

       'Positive' Class : No Heart Disease
```

Figure 6. Confusion Matrix for the Logistic Model with λ=0.5

Let us use the Linear Discriminant Analysis. Both logistic regression and LDA give linear logit results, yet the estimates are computed differently. For this approach, we have set the prior probabilities to 0.6 and 0.4, and we have initially left the threshold as 0.5. The sensitivity and specificity have remained close and above 0.8, the accuracy also increased, and the errors have decreased. The AUC remains the same at 0.93.

```
Confusion Matrix and Statistics

                Reference
Prediction         No Heart Disease Heart Disease
  No Heart Disease               48             9
  Heart Disease                   7            57

               Accuracy : 0.8678
                 95% CI : (0.7942, 0.9225)
    No Information Rate : 0.5455
    P-Value [Acc > NIR] : 3.31e-14

                  Kappa : 0.7341

 Mcnemar's Test P-Value : 0.8026

            Sensitivity : 0.8727
            Specificity : 0.8636
         Pos Pred Value : 0.8421
         Neg Pred Value : 0.8906
             Prevalence : 0.4545
         Detection Rate : 0.3967
   Detection Prevalence : 0.4711
      Balanced Accuracy : 0.8682

       'Positive' Class : No Heart Disease
```

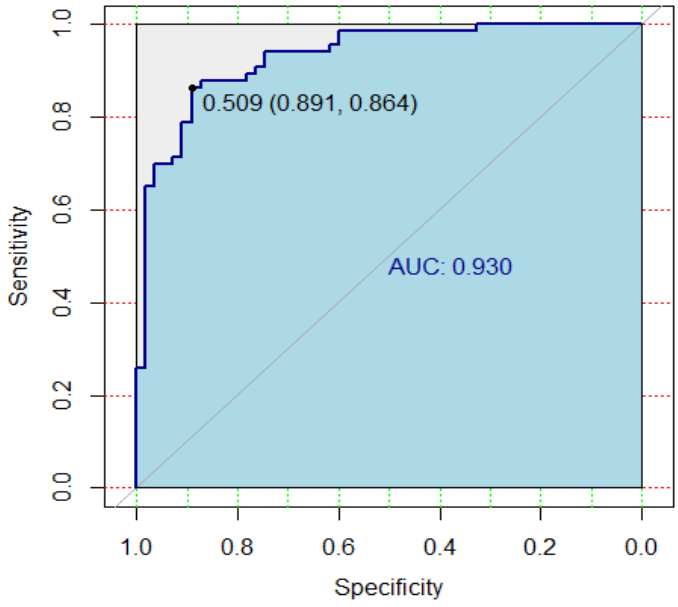


Figure 7. Confusion matrix and ROC curve for the LDA model, with prior probabilities (.6,.4) and λ=0.5

Lastly, we use Random Forest and we independently make a prediction to get the confusion matrix of the model on the *Test* dataset. Calculating the sensitivity from this matrix we get 0.7931, specificity rate of 0.8571, and accuracy of 0.8264. The false positive rate is 0.1429 and the false negative rate is 0.2069.

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

        OOB estimate of  error rate: 15.93%
Confusion matrix:
                No Heart Disease Heart Disease class.error
No Heart Disease               66            17   0.2048193
Heart Disease                  12            87   0.1212121
```

```
predrf<-predict(hrf, newdata = test[-13])
confusionmat=table(test[,13],predrf)
confusionmat
                predrf
                 No Heart Disease Heart Disease
No Heart Disease               46             9
Heart Disease                  12            54
```

CONCLUSIONS

Out of the three models that have been ran thus far, given the metrics provided, we would choose the LDA model for classification, as it provides the highest sensitivity, specificity, and accuracy while giving the lowest error. Given the AUC remained the same at 0.93, the predictive power of this model would also be good. For our data, we did not consider penalizing models, as it was previously stated that it bares equal weight to misdiagnose someone regardless of the incorrect diagnosis.

While the model may be useful in the medical field, the setting of the implementation must be examined. The records in the data set show that the patients were submitted to stress tests and fluoroscopy; therefore, an Emergency Room may not wish to use this LDA model and select a different approach where less variables are considered and the implementation is clearer, say a decision tree.

CITACIONS

Heart Disease Data Set, Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.