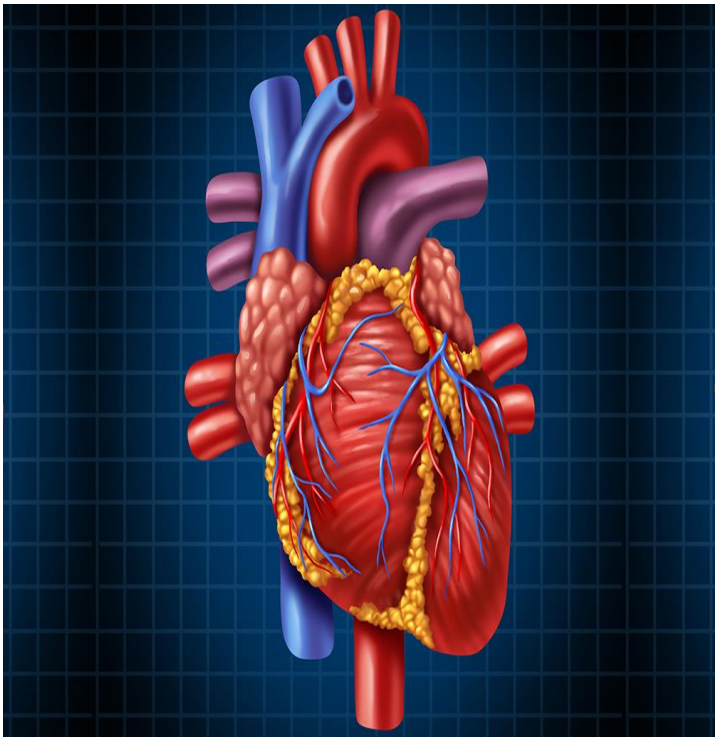
IDENTIFYING THE RISK FACTORS OF HEART DISEASE



FINAL PROJECT
EXST 7142 STATISTICAL DATA MINING
FALL 2021

BY CHRISTOPHER KUETSINYA

MOTIVATION OF THE STUDY

Globally, cardiovascular diseases (CVDs) are the number one cause of death, taking an approximately 17.9 million lives each year, which accounts for 31% of all deaths (WHO, 2020). Four out of five CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age (WHO, 2020). Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels including, coronary heart disease (heart attacks), cerebrovascular diseases (strokes) and heart failure (WHO, 2020). In particular, heart failure occurs when the heart is unable to pump enough blood to the body, and it is usually caused by diabetes, high blood pressure, or other heart conditions or diseases (WHO, 2020).

With fatalities figures on the rise in recent times and the importance of a vital organ such as the heart, identifying the risk factors of heart disease has become a priority for medical doctors and physicians.

In this project, some statistical models will be used to identify top risk factors of heart disease and assess the effect of these factors on a patient's heart condition. This is needed to help in the early detection and management of heart disease or heart failure.

DATA DESCRIPTION & SUMMARY

The dataset for this study was created by combining different independent datasets already available. This dataset entails five heart datasets are combined over eleven common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are 200 observations from V.A. Medical Center-Long Beach, 303

observations from Cleveland Clinic Foundation, 294 observations from Hungarian Institute of Cardiology-Budapest, 123 observations from University Hospital-Zurich-Switzerland, and 270 observations from University Hospital-Basel-Switzerland. These combined datasets produced a total of 1190 observations. However, there were duplicates of 272 observations. After removing these duplicates, a final dataset with 918 observations and 12 variables was obtained. All the five heart datasets can be found under the Index of heart disease datasets from the UCI Machine Learning Repository.

The variables in the dataset have mixed features. 6 of the explanatory/predictor variables are categorical and 5 are numeric. '**HeartDisease'** is the target variable in this study. It has a binary response that indicates the heart condition of a patient. Below are the names and descriptions of each variable in the dataset.

- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- 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]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]

- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- Oldpeak: oldpeak [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: up-sloping, Flat: flat, Down: down-sloping]
- HeartDisease: output [1: heart disease, 0: Normal]

DATA CLEANING AND PROCESSING

The data named Heartsdata.csv was read in R; its dimension and structure were checked. The categorical variables in the dataset were factored to make the data look appropriate for further analysis.

The dataset had no missing values therefore no further cleaning was implemented. The structure of the data was examined again to be sure all the variables had the right features.

EXPLORATORY DATA ANALYSIS

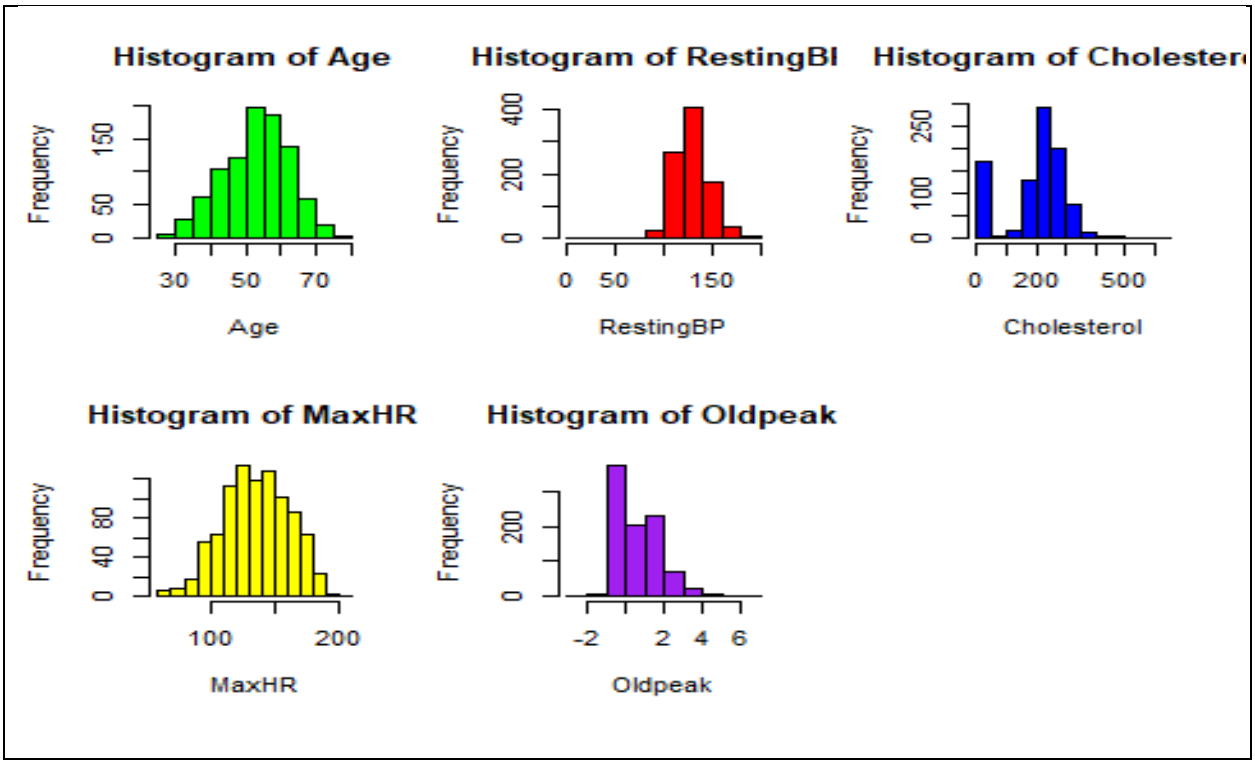
Summary Statistics

In the Hearts dataset, the target variable which indicates the heart condition of a patient has responses ‘0’ and ‘1’. Where ‘0’ indicates normal condition and ‘1’, heart disease. There are 410 patients with normal heart conditions and 508 patients with heart disease according to the data. The ages of patients in this study ranged from 28 years to 77 years, with a mean age of 53.51 years. More males were involved in this study than females. There were 193 females and 725 males.

EDA of Numeric Variables

A pairwise correlation test was conducted for all the numeric variables. The correlation values showed that there is no strong correlation among the variables.

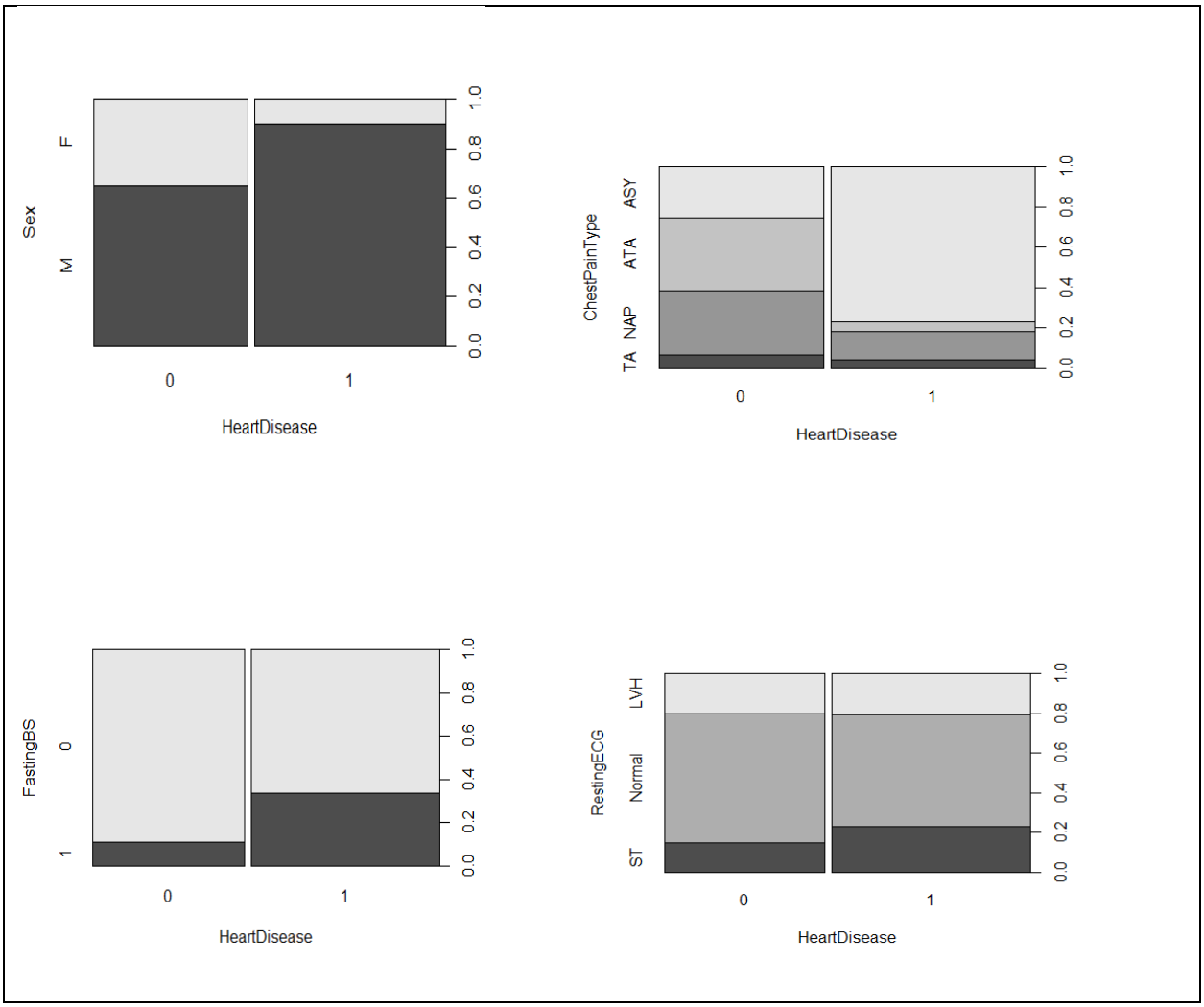
To check the distributions of the numerical variables to ascertain the skewness or normality of the variables, histograms were plotted and the Shapiro-Wilk test for normality was conducted for each variable.

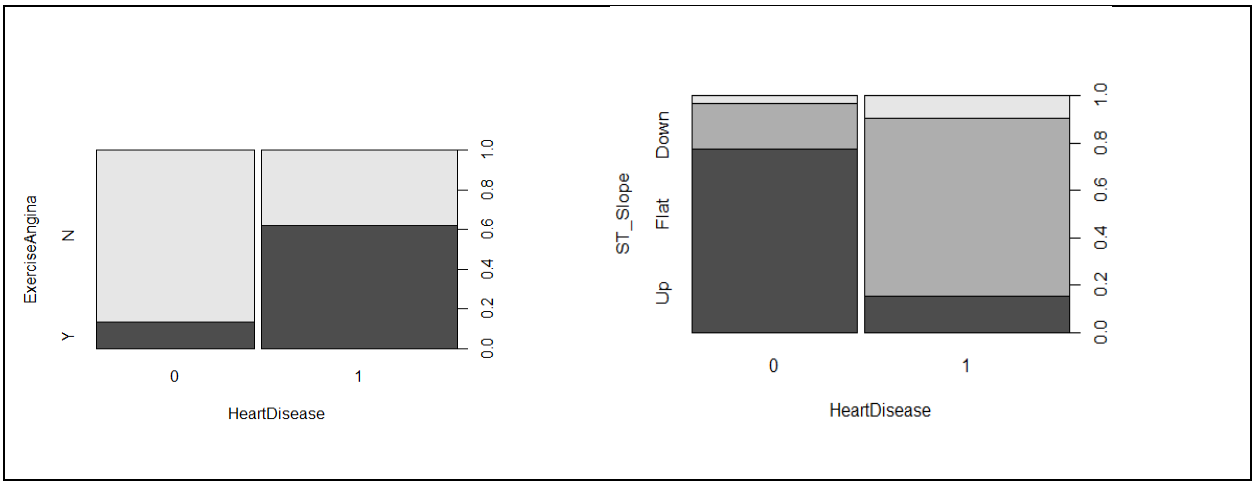


All the histograms are close or similar to a normal distribution except the variable ‘Oldpeak’ which is slightly skewed to the right. In general, the numeric variables are not too skewed hence no transformation is needed.

EDA of Categorical Variables

A chi-square test for association was conducted on each categorical variable against the target variable 'Heart Disease'. From the results, all the p-values were less than 0.05. Thus, all the categorical variables are associated with the target variable. The prevalence of heart disease was also assessed in each category of the categorical variables.





METHODOLOGY

The target variable in this study had a binary response, as such, a classification problem. The goal of this study was to identify the top risk factors of heart disease and also assess model accuracy.

Firstly, a Random Forest classifier was used to analyze the data. Then, the performance of the model was assessed. Thereafter, the variable importance plot was obtained, and the effect of each predictor was examined using partial dependency plots.

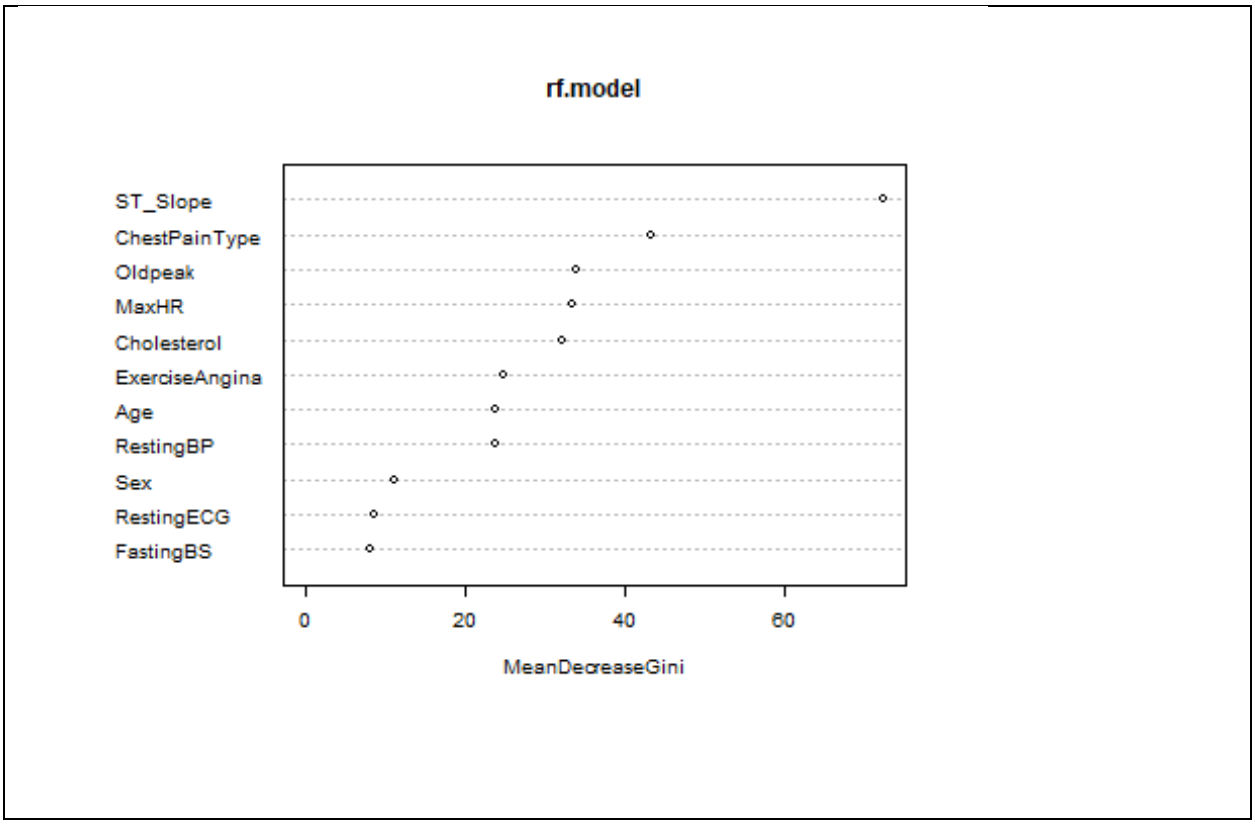
Secondly, a Generalized Additive Model (GAM) was used to analyze the data taking into consideration only the top most variables in order of importance. The performance of the GAM was assessed then a conclusion was drawn.

DATA ANALYSIS

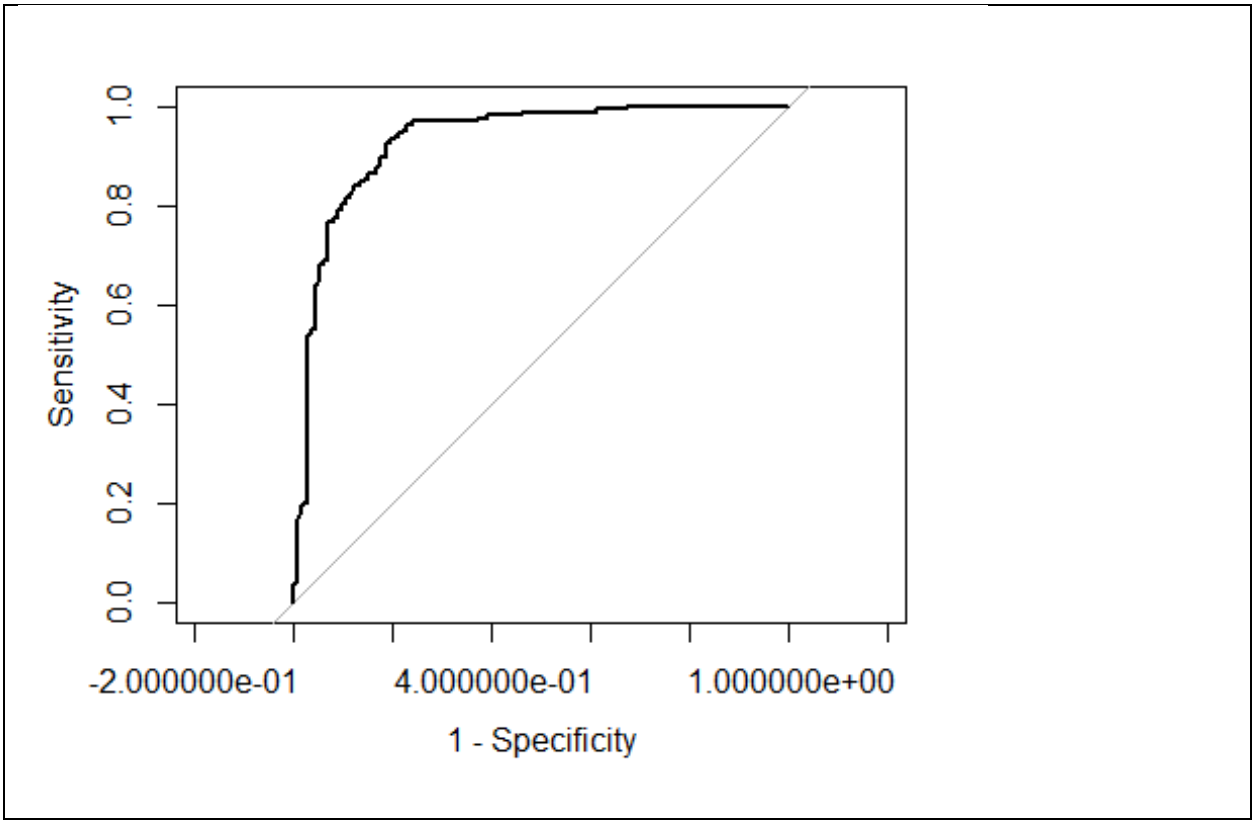
The data was split into training and test sets. The split was done in 70:30 ratios. Hence, 642 observations for training and 276 observations for testing.

Random Forest

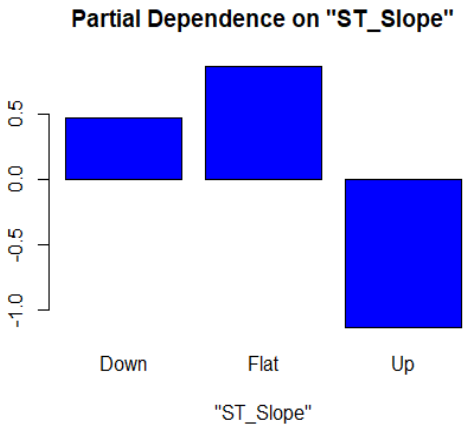
The random forest on the training data produced 11.84% OOB estimate of error rate and a test set error rate of 14.49%. The sensitivity rate on the test set was 0.871 and the specificity rate was 0.832. The variable importance plot was used to identify the top most important variables of which ST_Slope, ChestPainType, and Oldpeak emerged the top three.

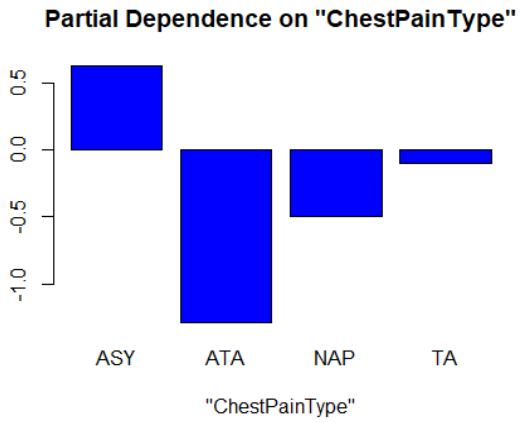


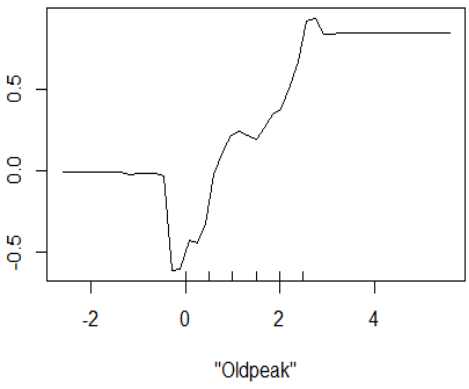
The ROC plot and AUC score were used to assess the performance of the model. An AUC score of 0.93 was obtained.

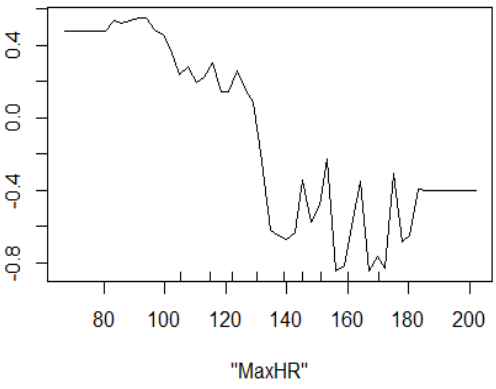


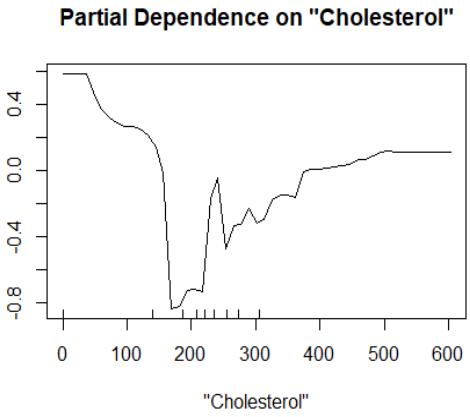
To examine the effects of the predictors on the response, partial dependency plots for each predictor were generated. Now, looking at the plots below, especially the plots of the continuous variables, the effects of these predictors are far from linear.

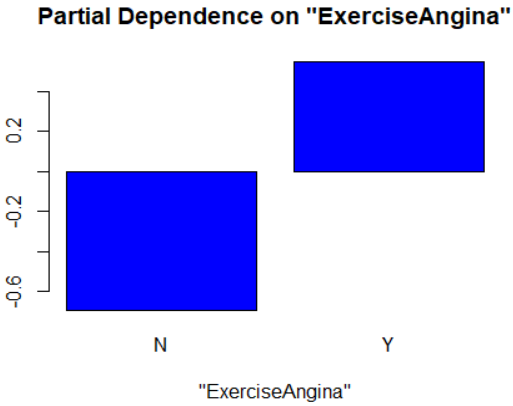

Partial Dependence on "ST_Slope"
0.5
0.0
-0.5
-1.0
Down
Flat
Up
"ST_Slope"

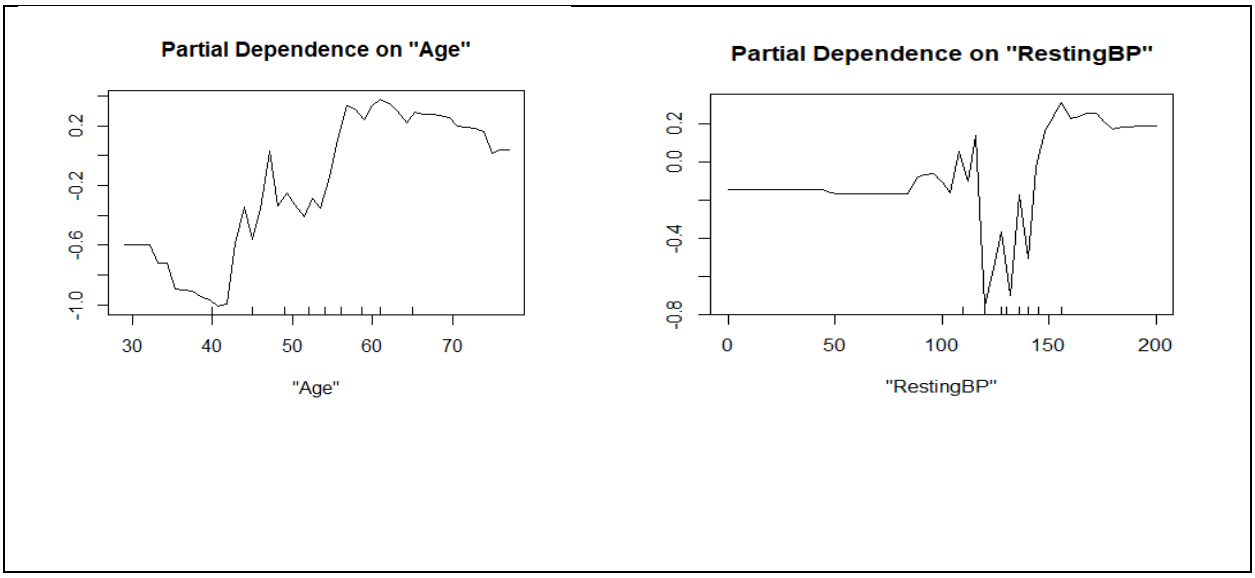

Partial Dependence on "ChestPainType"
0.5
0.0
-0.5
-1.0
ASY
ATA
NAP
TA
"ChestPainType"


Partial Dependence on "Oldpeak"
0.5
0.0
-0.5
-2
0
2
4
"Oldpeak"


Partial Dependence on "MaxHR"
0.4
0.0
-0.4
-0.8
80
100
120
140
160
180
200
"MaxHR"

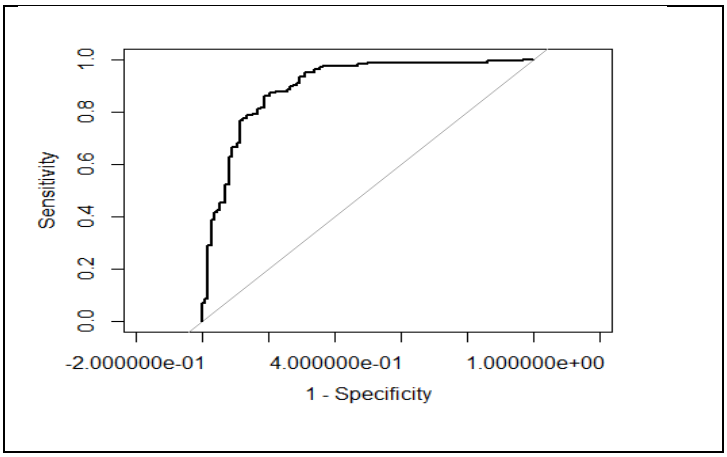

Partial Dependence on "Cholesterol"
0.4
0.0
-0.4
-0.8
0
100
200
300
400
500
600
"Cholesterol"


Partial Dependence on "ExerciseAngina"
0.2
-0.2
-0.6
N
Y
"ExerciseAngina"



Generalized Additive Model (GAM)

Considering the non-linear nature of the predictors, a GAM was used to analyze the top eight important variables and the top five important variables. In both models, smoothing was applied to all the continuous predictors. An ANOVA test was conducted to check the goodness-of-fit of the two GAM models. The second model (model with the top five important variables) appeared to work better and was selected for further analysis. The selected GAM produced a sensitivity rate of 0.844 and a specificity rate of 0.843. The misclassification error rate was 14%. The ROC plot and the AUC score look a little better than for the Random Forest. An AUC score of 0.9338 was obtained.



RESULTS

Below is a summary of all the results from the Random Forest and Generalized Additive Model.

Random Forest using all the predictors in the dataset			
Error rate	Sensitivity	Specificity	AUC score
14.49%	0.871	0.832	0.93
GAM using the 5 top important variables obtained from Random Forest			
Error rate	Sensitivity	Specificity	AUC score
14%	0.844	0.843	0.9338

DISCUSSION

The results from the Random Forest showed that ST_slope is the most important factor in determining whether a patient had heart disease or not. Age and Sex were in the bottom five variables. The AUC score for the Random Forest was 0.93 which is a good AUC score.

The GAM was used for model dimension reduction and interpretability. The GAM also gives a good AUC score of 0.9338 which is a slight improvement from the Random Forest.

CONCLUSION

It was expected to see variables such as Cholesterol, Age and Sex being very important in determining the heart status of a patient. This is because, according to the CDC, high cholesterol and age are very high-risk factors for heart disease. However, the data for this study shows that ST_slope, which is also known as ST-segment or heart rate slope is the most important risk factor. From the partial dependence plot, it can be seen that patients with flat heart rates are highly susceptible to heart disease.

References

WHO. (2020). *Cardiovascular Diseases.* Retrieved from World Health Organization : https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

https://08d104db99.nxcli.net/zkosb/chadox1-ncov-19-update

https://pubmed.ncbi.nlm.nih.gov/31777019