# HEART DISEASE PREDICTION

MMA – 867 Final Project Report

TEAM - ROSEDALE

## 1. Background

According to World Health Organization around 17M[1] people die across the globe due to cardiovascular diseases (CVD's) every year. This figure accounts for 31% of all deaths worldwide annually, making CVD's the number one cause of mortality[2]. The projections made by the American Heart Association show that, by 2035 about 45% of the US adult population will have a form of CVD, costing the economy more than 1 Trillion dollars annually. Fortunately, 80% of all CVD cases are preventable, provided they are detected early and effective preventative measures are taken to minimize the risk[3].

## 2. Objective

Considering the importance of early detection in prevention and treatment of CVD's, it is crucial to devise a high precision diagnostic system that is both efficient and financially less burdensome. The current gold standard for diagnosis of CVD is Coronary Angiography. Though this technique has high accuracy power, it is an invasive procedure with morbidity and mortality rates of 1.5% and 0.15% respectively[4.] As an alternative, a less invasive computer-based CVD risk assessment system can be implemented to mitigate the challenges of Angiography. In an attempt to build this system Machine learning principles especially predictive modelling has the potential to predict the likelihood of a patient to suffer from heart disease .As a first step towards developing this system, this project is aimed towards developing a model that can be implemented for predicting the presence of CVD based on parameters related to a person's physiological condition and medical history.

## 3. Dataset overview

For this project, data has been obtained from Kaggle. The Dataset is based on the information collected by Cleveland clinical foundation for 303 patients. Originally there were 76 attributes in total. In this project, only 14 attributes have been used, referring to the previous clinical research and significant loss of data for many of other variables. These 13 predictor attributes can broadly be classified into three categories. The first set of variables are related to the physiological state of a person. These include Age, Sex, cp (Chest Pain), trestbps(resting blood pressure on admission to hospital), chol(serum cholesterol level), fbs(fasting blood sugar). The second set of variables are based on the tests conducted for diagnosing heart disease. These include restecg(resting electrocardiographic results), ca(number of major vessels coloured by fluoroscopy) and four variables related to the stress test are thalach(maximum heart rate achieved), exang(exercise induced pain), oldpeak(ST depression induced by exercise relative to rest), slope(the slope of the peak exercise ST segment). The third category contains one variable thal, which is related to heredity precondition called Thalassemia. Target is the response variable, which has a binary output with 0 and 1 representing absence and presence of CVD. For detailed description refer to **(*Appendix Table 1*)**

## 4. Preliminary Data Exploration

Introductory graphs are plotted to evaluate the distribution of the response variable for patients and non-patients. This shows approximately 56% are patients and 44% are non-patients, which makes the dataset balanced and minimizes the risk of development of a biased model. Further plots for target distribution with respect to age and gender give more granular insights about the dataset. The youngest patient is in the late twenties and the oldest patient is mid-seventies. The overall distribution of patients and non-patients is proportional for all age brackets. For corresponding plots refer to***Appendix 4.1***

## 4.1 Feature Engineering

All 14 attributes in the dataset are of numeric datatype. For better interpretation, original numeric attributes are kept in their numerical state (Age, trestbps, chol, thalach, oldpeak). The attributes with underlying character base are converted to character datatype (Sex,target). And those attributes which are character datatype with defined ordinality are converted into factors with levels. To visualize the relative influence of all other attributes on the age of patient/non-attribute, a new column is added to the dataset with three age brackets young, middle-aged and old. For detailed technical description refer to R-File submitted with the report.

## 4.2 Principle Component Analysis (PCA)

Graphs for target distribution with respect to gender shows in this sample 75% of females and 35% of males have CVD. Numerous medical studies have shown that CVD manifests in different forms with respect to Gender and Age, resultantly different factors may increase the risk of developing a heart disease[5] disproportionally. Based on this pre-existing medical literature, we evaluated the presence of any sub-groups in the given sample using Principle Component Analysis. Biplot for the first two PC's with respect showed that the presence of CVD in patients of three age brackets may not be affected by all attributes proportionally. Similarly, based on gender, the same attributes do not appear to determine the presence or absence of CVD in males and females. For detailed analysis results refer to ***Appendix 4.2***

## 4.3 Insights from Data Visualization

Metric 1 - Stress test Attributes.

For the given sample higher values for max_heart_rate indicated the presence of heart disease in young and middle-aged adults whereas for older adults both patients and non-patients showed similar test results. Lower values for ST_depression curve implies the presence of disease across all age groups. Higher proportion of young and middle-aged adults with downsloping ST-segments have the disease. For corresponding plots refer to ***Appendix 4.3 a)***

Metric 2 – Cholesterol (and related Attributes)

Cholesterol levels are often linked with the risk of having coronary heart diseases. Plots for cholesterol levels with respect to target and gender generate interesting insights. For middle-aged females, patients have lower cholesterol levels than non- patients., whereas for males both patients and non-patients have similar cholesterol levels. This implies that the cumulative cholesterol levels may not be a reliable attribute to assess the risk for CVD. The diagnostic test, fluoroscopy displays the number of unclogged major bold vessels. As the number of coloured vessels increases, the ratio for non-patients to patients also increases. For corresponding plots refer to ***Appendix 4.3 b)***

Metric 3 - Chest Pain (Angina)

Approximately 80% of females and 75% of males coming to the clinic with symptoms of Atypical and Non-Anginal pain have CVD. All young and middle-aged adults having Atypical Angina have CVD. And 50% of the older adults with symptoms of Atypical and Non-Angina pain have the disease. For corresponding plots refer to ***Appendix 4.3 c)***

Metric 4 – Thalassemia

Targets suffering from irreversible thalassemia defect are almost 100% certain to have cardiovascular disease. For corresponding plots refer to ***Appendix 4.3 d)***

## 5. Model Development

### 5.1 Goal and Approach

Based on the chosen dataset, we have decided to use a binary logistic regression model to predict the heart disease's existence since our response variable is a categorical variable with two levels (1-having disease; 0-not having a disease) and our goal is to make prediction and classify patients into these two categories. Throughout the model development process, we would also like to verify several model assumptions listed below:

1)Observations should be independent from each other. In other words, none of them should come from repeated measurements.

2)There should be little or no multicollinearity issues among the independent variables as this would inflate the variances of our coefficients estimates and result in incorrect statistical inference.

3)Linearity assumption which means independent variables should be linear to our log-odds

And we are going to assess our model performance based on the following metrics:

**1)ROC/AUC:** we will look at ROC of our models and AUC of the ROC plot to see how capable our models are to distinguish between different classes.

**2)Confusion Matrix:** we will calculate accuracy, sensitivity and specificity of all the models based on all the confusion matrices we created and do the comparison.

### 5.2 Data preparation

The first step is to check whether any data is missing in any columns of our dataset and we found our dataset is complete without any missing values.

By performing correlation analysis, we can look at whether there are variables of our dataset that are highly correlated with each other. If the variables are highly correlated (>0.7) with our other variables, then we would probably suffer from multicollinearity. Following the correlation matrix, we can see there is no high correlation between our independent variables. According to the boxplot on all the numerical variables in the dataset, we noticed that there is one observation having a extremely high cholesterol level (over 500) while most of the records are in the range between 200 to 300. In the end, we decided to keep this record in the dataset as this is a realistic number that a person can have, and this is in the scope of our study. We noticed that there are few duplicated records in our dataset which looks like data entry errors. As per the model assumption, all the observations should not come from repeated measurement and observations should be independent to each other, thus we must remove the duplicated records from our sample data. For corresponding plots refer ***Appendix to 5.1 a) b) c) d)***

Since there are several categorical variables in the dataset. To prepare for our model building, we have transformed them into dummy variables so the impact of each level of these variables can be well reflected in the model.

The last step before the model building is to divide our dataset into 'train' and 'test' dataset with a corresponding weight of 80% and 20%. We will train our model on 'train' dataset and assess the model performance on 'test' dataset. Now, we have our train dataset having 242 observations and test dataset having 60 observations.

## 5.3 Model Building

### Model 1 – Baseline Logistic Regression Model

In this first model, we have included all 18 variables in the dataset, we applied the model to our ‘test’ dataset, and we calculated AUC to be 90.04%. While looking at the ROC curve, we decided to choose the decision threshold to be 0.4. The reason to choose this threshold is because this is essentially a medical study so that the risk of providing a false alarm is much smaller than the failure to detect the heart disease from our patients. Concretely, we prefer a relatively higher sensitivity. For corresponding plots refer to ***Appendix 5.2***

According to the confusion matrix, the base model’s accuracy is 84.75%, sensitivity is 90.62% and specificity is 77.78%. The performance is over our expectation as it is not suffering from overfitting issues while having all 18 variables included

| Test | Actual (Y) | Actual (N) |
|---|---|---|
| Predicted (Y) | 29 | 6 |
| Predicted (N) | 3 | 21 |

The next step would be to enhance our current model by reducing the variables numbers in the model based on their t-test’s p-values and determine which variables should be kept in our predictive model.

### Final Model

We used a variable selection method called ‘Stepwise AIC’ to help us dealing with trade-offs between the model’s simplicity and goodness of fit. After Stepwise AIC is used, we cut down our independent variable numbers to 11. Also note that we have decided to manually include ‘age’ in our final model (though it’s p-value is much larger than our significance level 0.1) as we strongly believe ‘age’ has a significant impact on heart disease rate according to numerous medical studies.

Similarly, we chose our decision threshold to be 0.5 because we prefer sensitivity more than specificity. According to the ROC, to further improve the model’s sensitivity, we must shift our threshold to be around 0.1. However, this would largely decrease our specificity. As a result, the final decision was made to keep our threshold as 0.5. According to the confusion matrix, the final model’s accuracy is 86.44%, sensitivity is 93.75%, specificity is 77.78%. and AUC is 90.28%. For corresponding plots refer to ***Appendix 5.2***

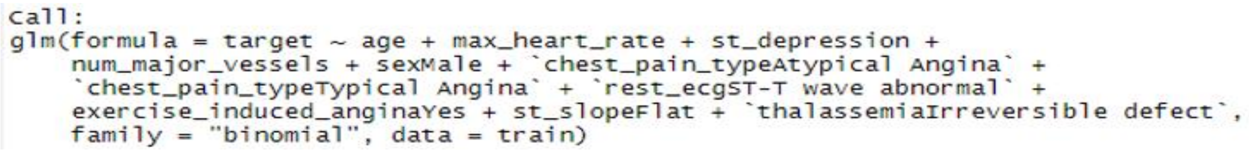
| Train | Actual (Y) | Actual (N) |
|---|---|---|
| Predicted (Y) | 30 | 6 |
| Predicted (N) | 2 | 21 |

## 5.4 Model Validation

After the final model is built, we validate and ensure our model is statistically accurate – it should meet all the model assumptions that we talked about in the previous paragraphs. We perform three tests for testing linearity (graph for numerical variables with respect to log odds), multicollinearity (based on variance inflation factor) and presence of any influential observations (using cook's distance) in the dataset. For corresponding plots and tables refer to ***Appendix 5.3***

## 5.5 Model Interpretation

Below are all 11 variables in the final model.



```
Call:
glm(formula = target ~ age + max_heart_rate + st_depression +
    num_major_vessels + sexMale + `chest_pain_typeAtypical Angina` +
    `chest_pain_typeTypical Angina` + `rest_ecgST-T wave abnormal` +
    exercise_induced_anginaYes + st_slopeFlat + `thalassemiaIrreversible defect`,
    family = "binomial", data = train)
```

Coefficient Interpretation

Every one-year age increase will raise the heart disease log odds by 2%. One additional max heart rate (#/min) will also increase the heart disease log odds by 2%. The person with ST depression will decrease heart disease log odds by 48%. One additional major vessel the person has will decrease the heart disease log odds by 58%. If other attributes are the same, Male are 76% less likely to have heart disease comparing to Female. For corresponding details refer to Table 2 ***Appendix 5.4***

## 6 Applications

Healthcare systems across the world generate a huge amount of data. Consequently, the potential for machine learning in healthcare is immense and it can generate insights to allow for the adoption of preventive measures against diseases, improve the discovery of new treatments and make the existing ones more effective. Some of the most important applications of the model are discussed below:

| Healthcare | Medical Research | Public Health Administration | Insurance companies |
|---|---|---|---|
| 1)For Early diagnosis of CVD<br>2)For integrating with health tracking systems. | 1)To evaluate the effectiveness of new medicines and drugs in clinical trials<br>2)For selecting observed groups for medical trials | 1)For Prediction of CVD in a population.<br>2)To select target groups for preventative steps | 1)To predict policy holder's risk of having CVD.<br>2)To forecast number of claims resulting from CVD. |

## 6.1 Future Improvements and Applications

As a future extension of this model, we will be looking at collecting more data from varied geographical areas and including more attributes to make the current model more inclusive and versatile. We will also be experimenting with advanced algorithms like Decision Trees. As an augmentation, a heart health tracker app can be designed by integrating this model which will allow users to track their heart health and measure the risk of getting heart diseases. The app can also suggest preventive measures, healthy food and lifestyle changes based on the results of the model. For corresponding empirical design refer to ***Appendix 6.1 a) b)***

## APPENDIX

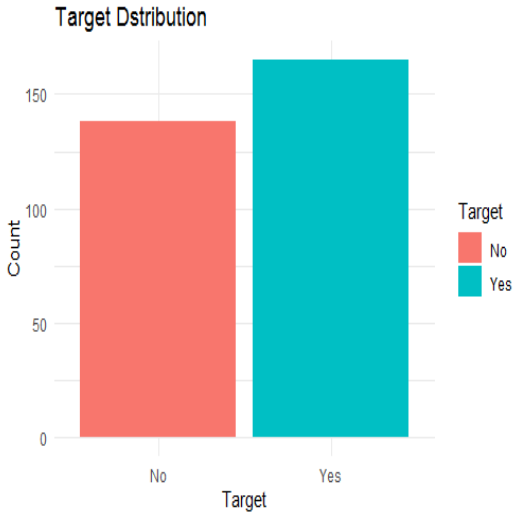
### 3. Dataset Overview

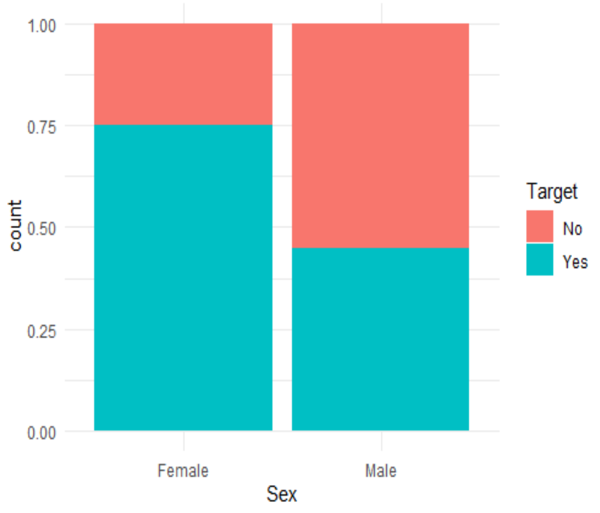
**Table - 1**

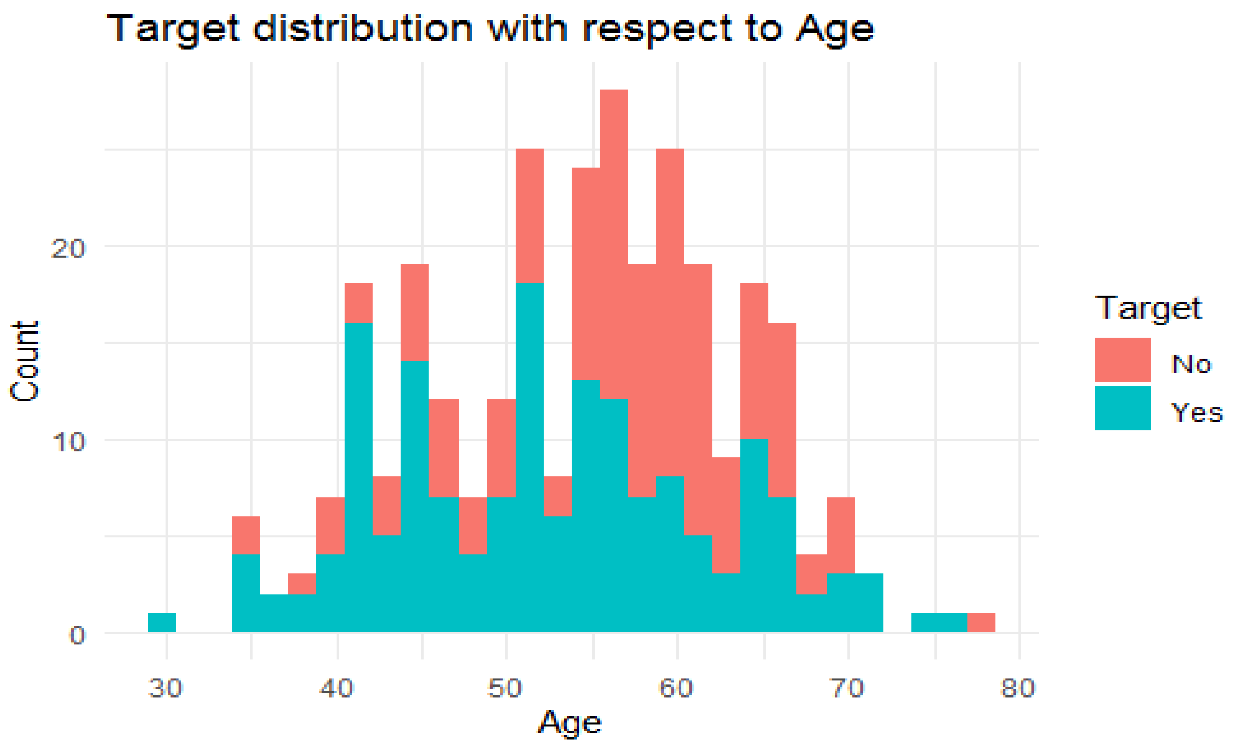
| S.no | Original Attribute Name | Changed Column Name | Original Datatype | Transformed Datatype |
|---|---|---|---|---|
| 1. | age | age | Numeric | Numeric |
| 2. | sex | sex | Numeric<br>Values (1,0) | Character<br>1-Male<br>0-Female |
| 3. | cp | chest_pain_type | Numeric<br>Values(0,1,2,3) | Character<br>0-Typical Angina<br>1-Atypical Angina<br>2-Non- Anginal Pain<br>3- Asymptotic |
| 4. | trestbps | resting_blood_ptessure | Numeric | Numeric |
| 5. | chol | cholesterol | Numeric | Numeric |
| 6. | fbs | fasting_blodd_sugar | Numeric<br>Values(0,1) | Character<br>1- >120mg/dl<br>0-<120mg/dl |
| 7. | restecg | rest_ecg | Numeric<br>Values(0,1,2) | Character<br>0-Normal<br>1-ST-T wave abnormal<br>2-probable oe definite LVH |
| 8. | thalach | max_heart_rate | Numeric | Numeric |
| 9. | exang | exercise_induced_angina | Numeric<br>Values(0,1) | Character<br>0-No<br>1-Yes |
| 10.. | oldpeak | st_depression | Numeric | Numeric |
| 11. | slope | st_slope | Numeric<br>Values(0,1,2) | Character<br>0-Upsloping<br>1-Falt<br>2-Downsloping |
| 12. | ca | um_major_vessels | Numeric | Numeric |
| 13. | thal | thalassemia | Numeric<br>Values(0,1) | Character<br>0-No<br>1-Yes |
| 14. | target | target | Numeric<br>Values(0,1) | Character<br>0-No, 1-Yes |

## 4. Preliminary Data Exploration

### 4.1 Target distribution with respect to Sex and Age

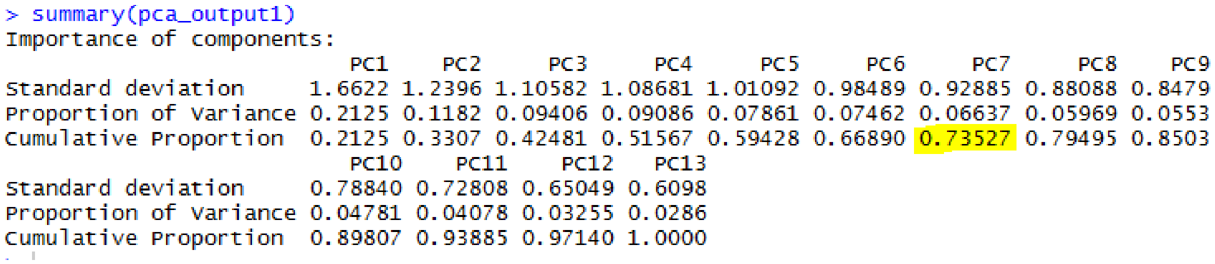






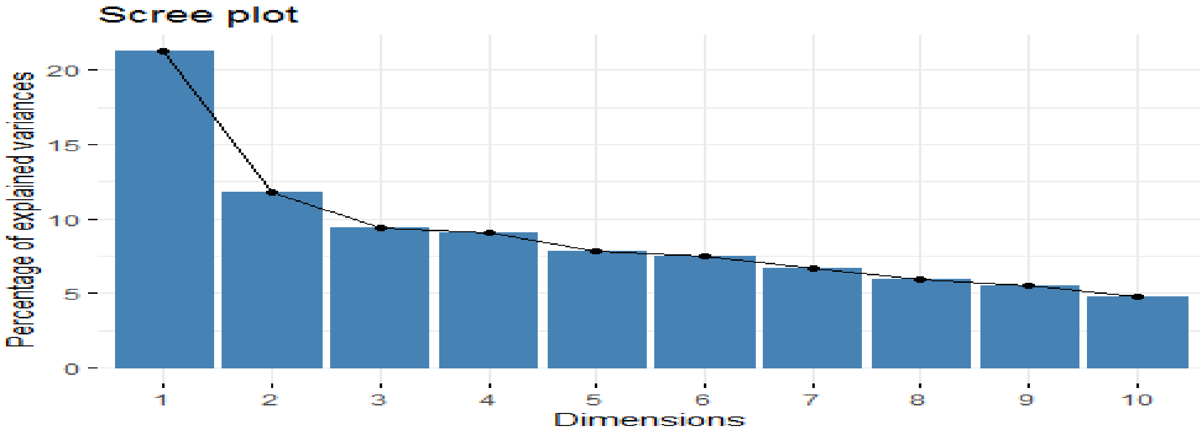
## 4.2 PCA Results for dataset

First 7 PC’s contain approximately 74% for variance/information

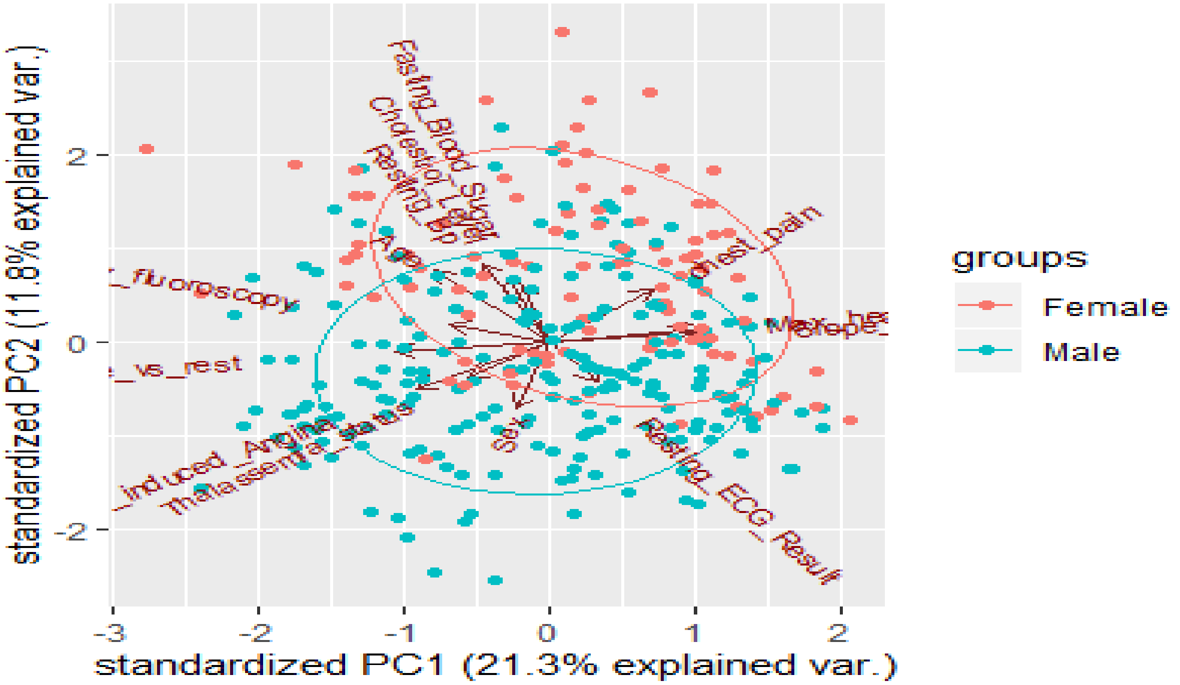


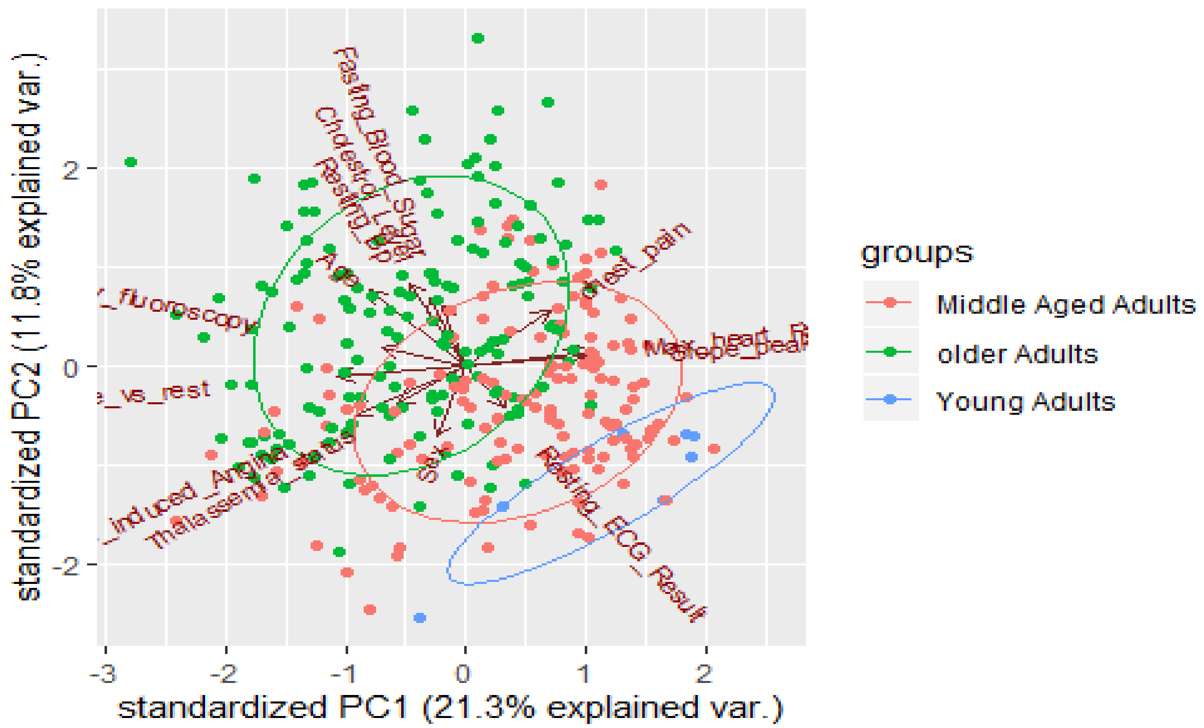
```
> summary(pca_output1)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8    PC9
Standard deviation     1.6622 1.2396 1.10582 1.08681 1.01092 0.98489 0.92885 0.88088 0.8479
Proportion of Variance 0.2125 0.1182 0.09406 0.09086 0.07861 0.07462 0.06637 0.05969 0.0553
Cumulative Proportion  0.2125 0.3307 0.42481 0.51567 0.59428 0.66890 0.73527 0.79495 0.8503
                          PC10    PC11    PC12   PC13
Standard deviation     0.78840 0.72808 0.65049 0.6098
Proportion of Variance 0.04781 0.04078 0.03255 0.0286
Cumulative Proportion  0.89807 0.93885 0.97140 1.0000
```

Scree-plot for first 10 PC’s

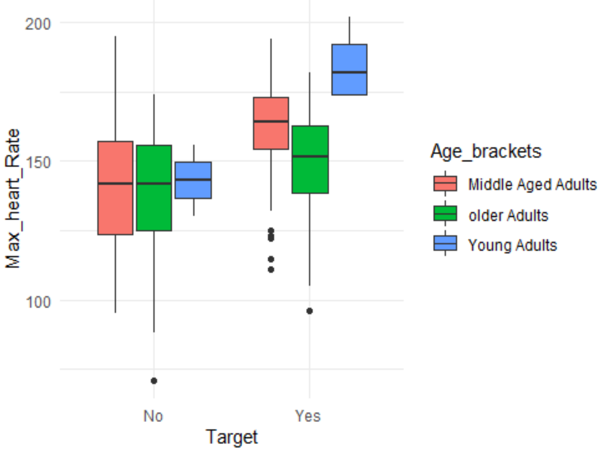


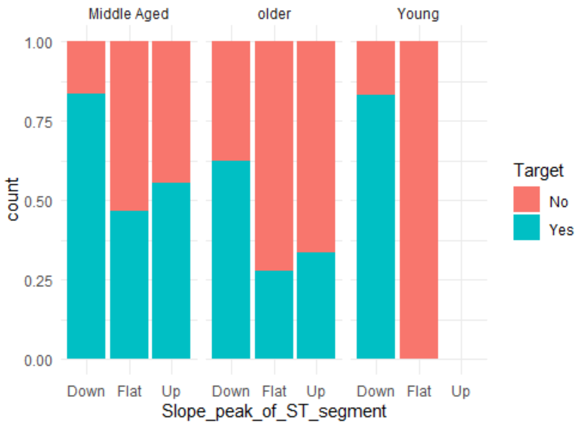
Bi-plots for PC1 and PC2 for Sex and Age-Brackets showing groups in the sample

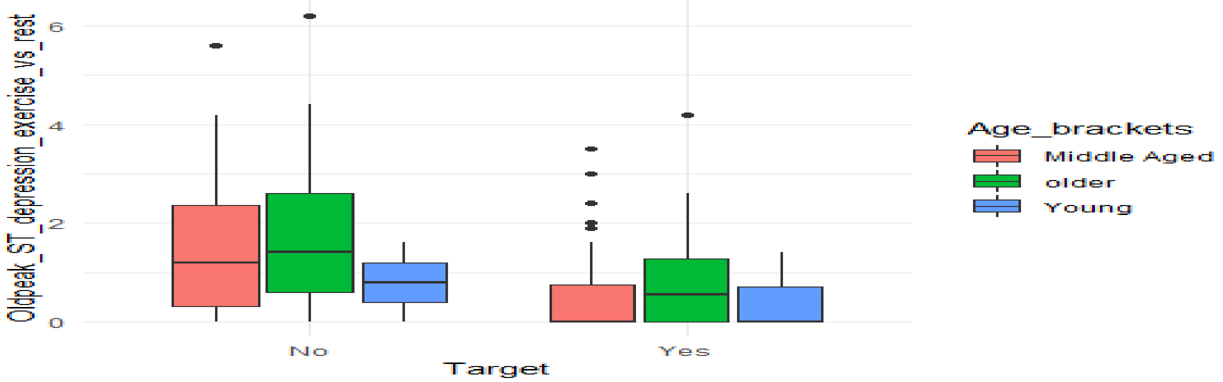




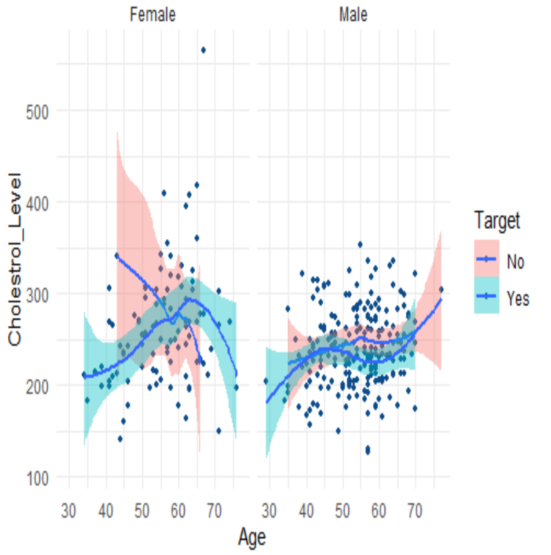
## 4.3 a) Plots for Metric 1 – Stress Test Attributes

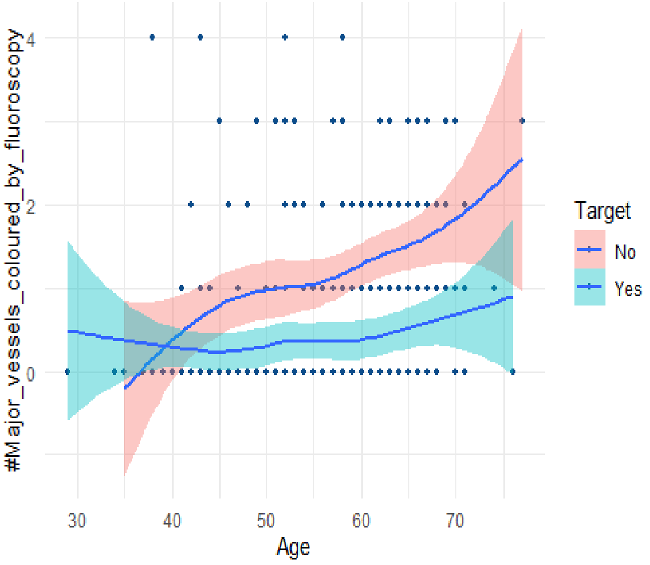




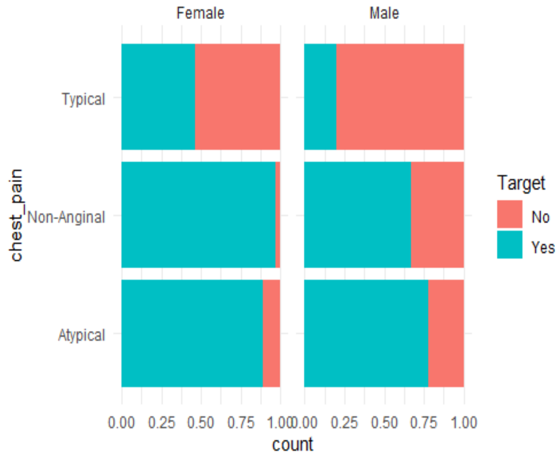


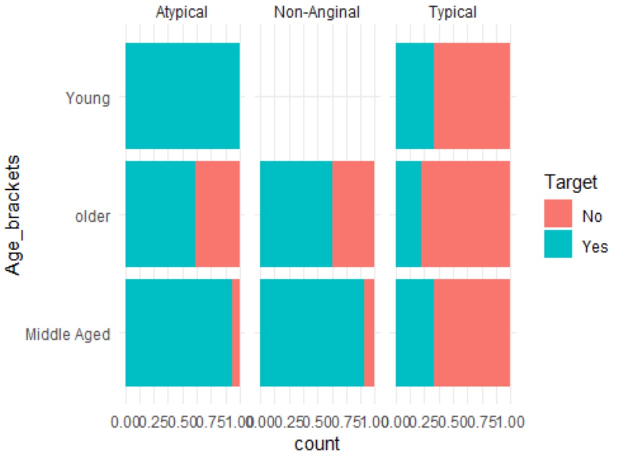
## b) Plots for Metric 2 – Cholesterol (and related Attributes)



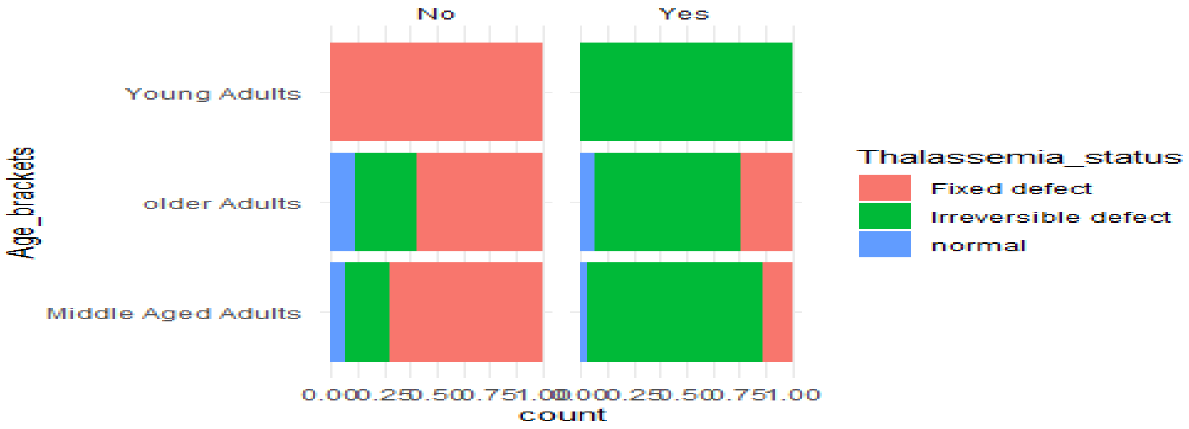


## c) Plots for Metric 3 – Chest Pain (Angina)

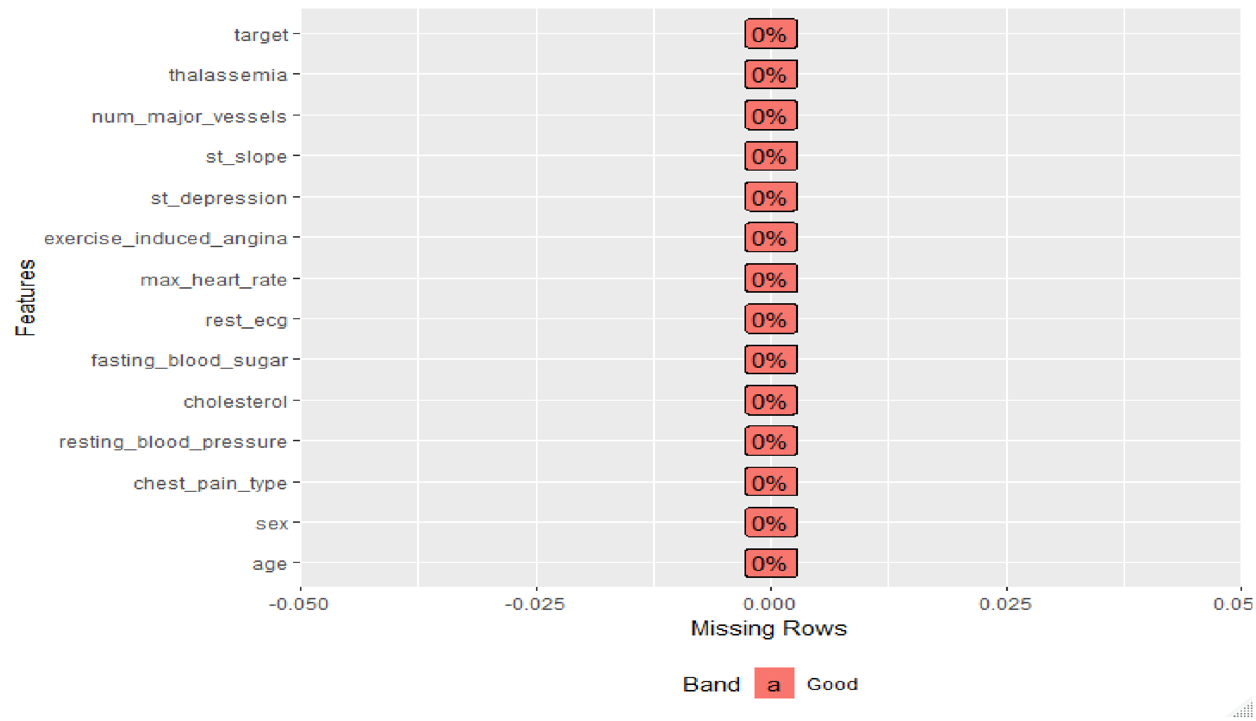




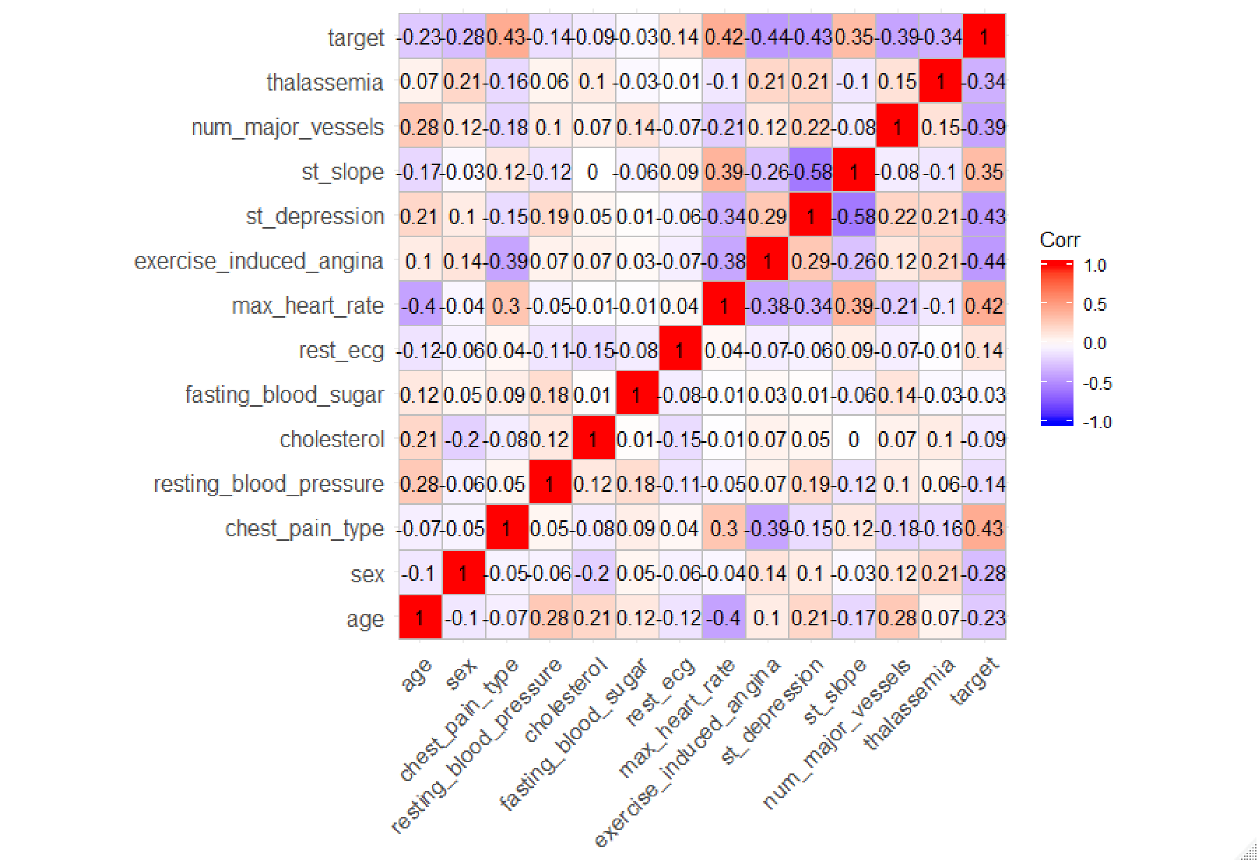
## d) Plot for Metric 4 – Thalassemia



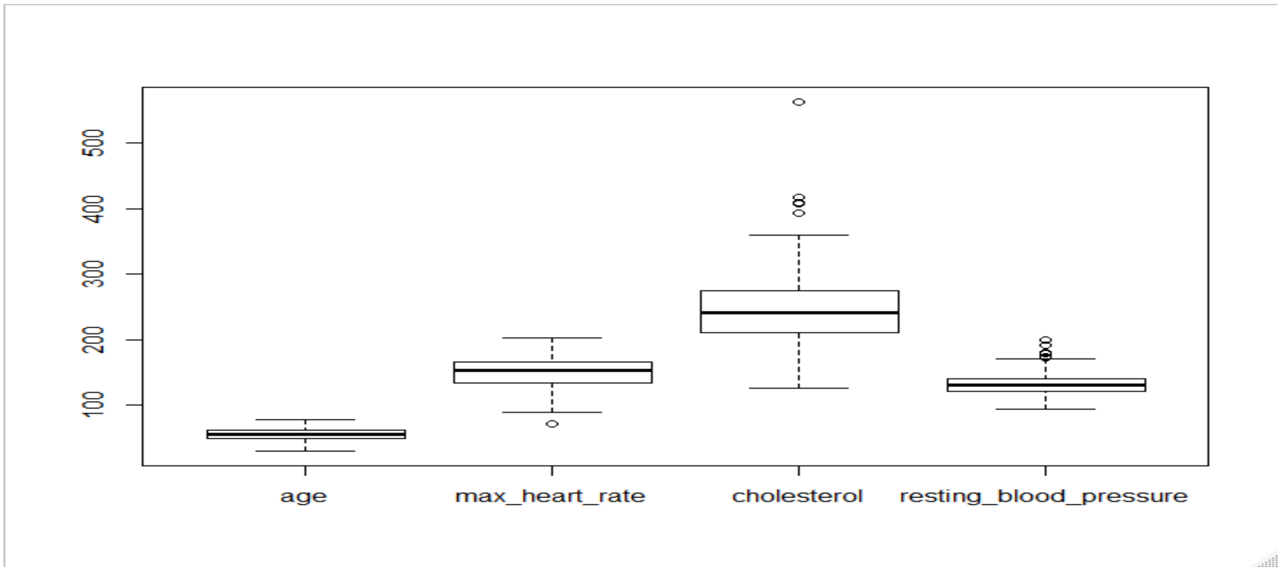
## 5.1 a) Plot showing Missing Values



## b) Correlation Matrix



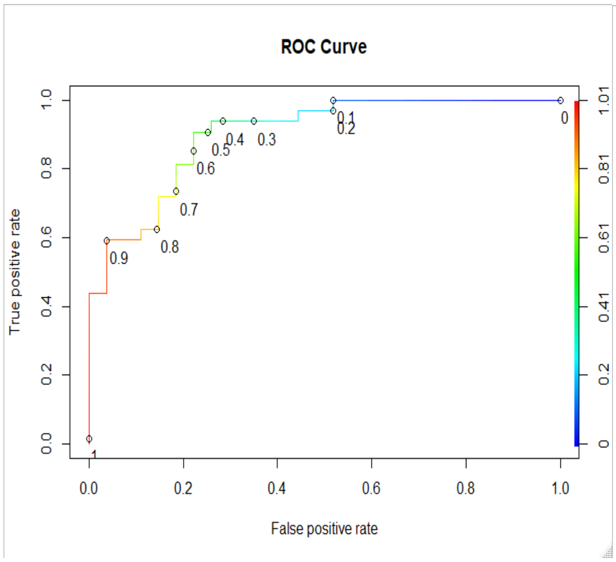
### c) Outlier Analysis



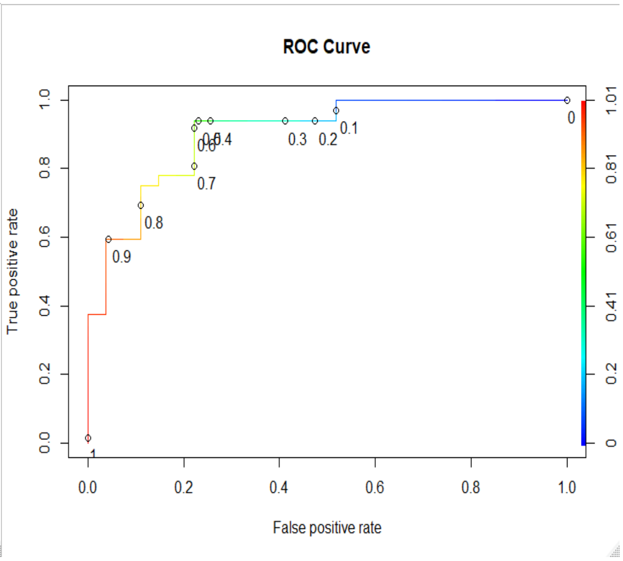
### d) Duplicated Data

| obs | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 164 | 38 | 1 | 2 | 138 | 175 | 0 | 1 | 173 | 0 | 0 | 2 | 4 | 2 | 1 |
| 165 | 38 | 1 | 2 | 138 | 175 | 0 | 1 | 173 | 0 | 0 | 2 | 4 | 2 | 1 |

### 5.2 ROC curves



Base Model

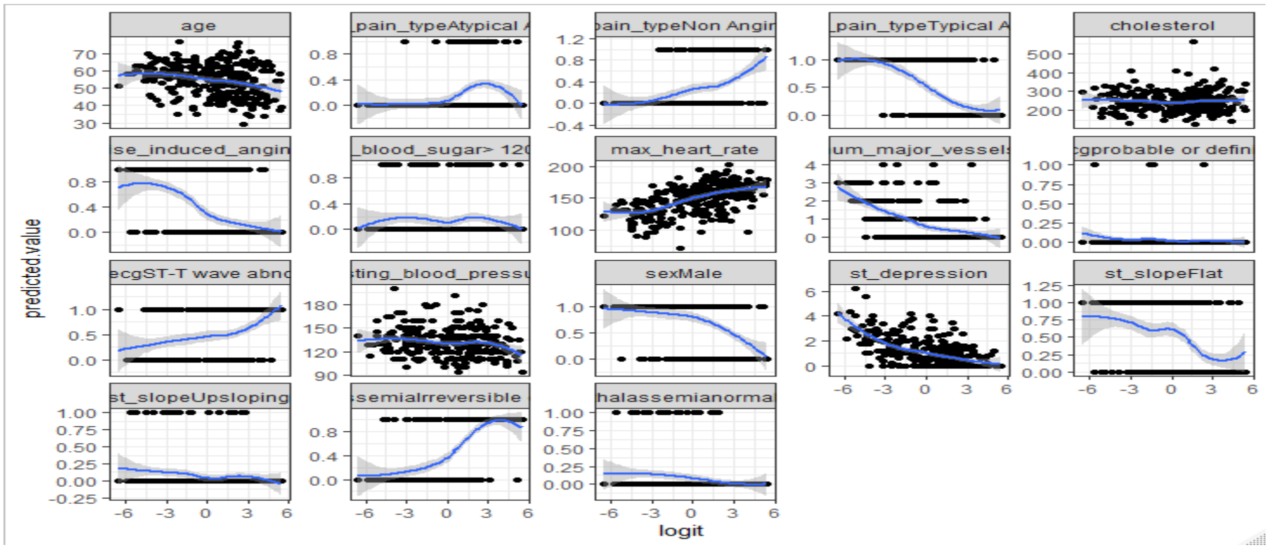


Final Model

### 5.3 Model Validation – Post hoc Tests

#### Linearity Test

We analysis the relationship between each independent variable and our log odds to see if linearity assumption is satisfied. The graph, for all the numerical variables are linear to our log odds.



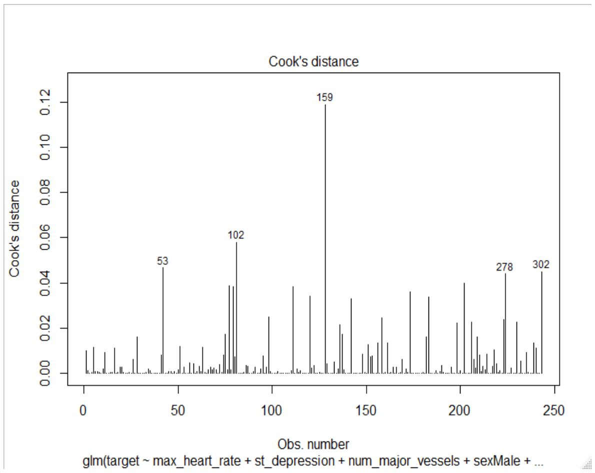
#### Multicollinearity test

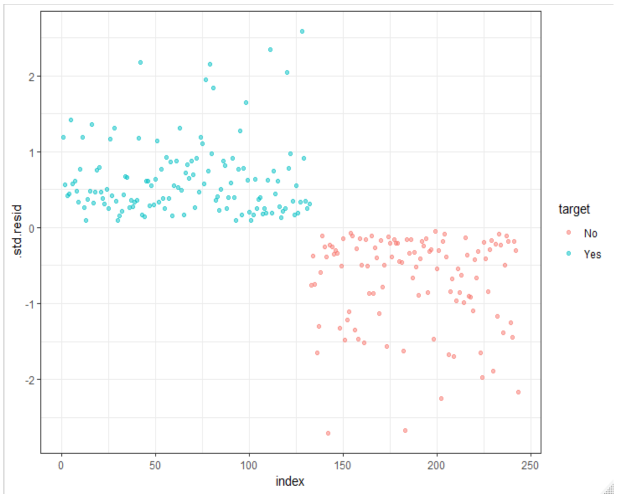
We calculated the Variance Inflation Factor (VIF) for each variable in our final model. According to the general rule, if VIF is bigger than 5, then we should be worried about multicollinearity issues in the model. Based on the table, we have no variables having its VIF bigger than 2. Therefore, we have no evidence to show that our model suffers from multicollinearity issue.

| Variables | VIF |
|---|---|
| age | 1.27 |
| max_heart_rate | 1.24 |
| st_depression | 1.23 |
| num_major_vessels | 1.08 |
| sexMale | 1.34 |
| chest_pain_type:Atypical Angina | 1.44 |
| chest_pain_type:Typical Angina | 1.44 |
| rest_ecg: ST-T wave abnormal | 1.06 |
| exercise_induced_agina:Yes | 1.11 |
| st_slope:Flat | 1.36 |
| thalassemia: Irreversible defect | 1.22 |

### Influential Observations

We also investigated to see if there are influential observations in our dataset that might affect model's accuracy since they could be potential multivariate outliers. By checking cook's distance plot and studentized residual plot, we didn't notice highly influential observations in our model as the highest studentized residual is less than 3.

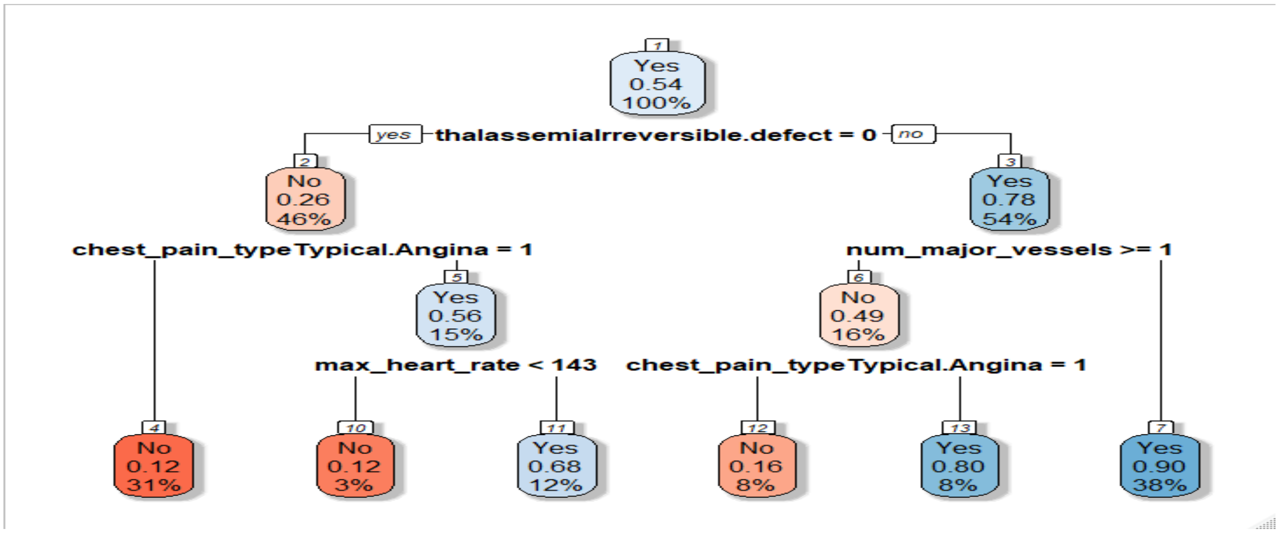




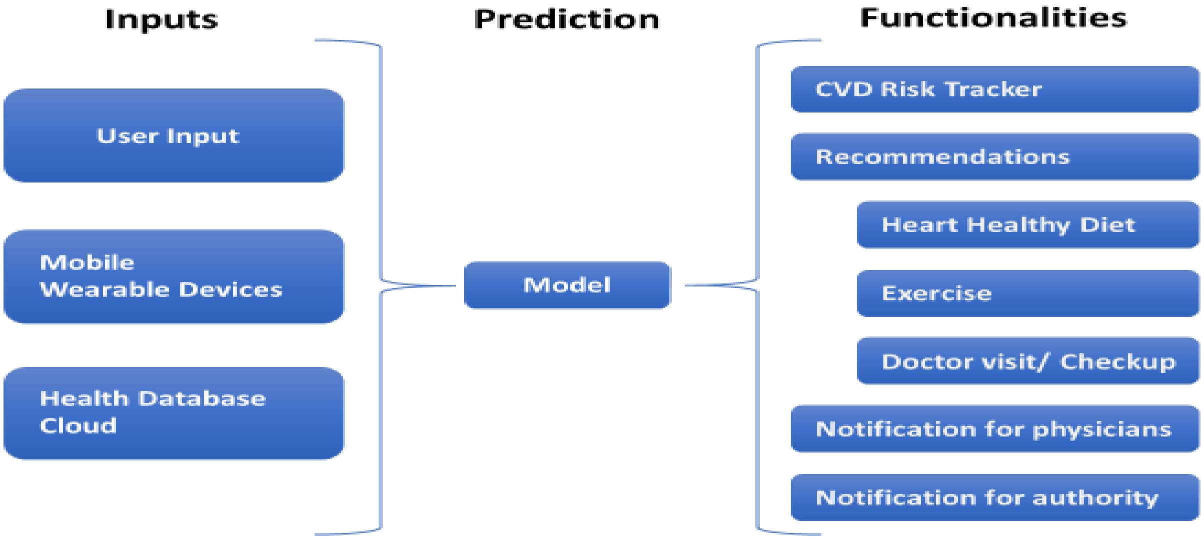
### 5.4 Table – 2 (Below are all 11 variables in our final model along with their coefficients.)

| Variables | Exp(Coefficient) |
|---|---|
| intercept | 0.44 |
| age | 1.02 |
| max_heart_rate | 1.02 |
| st_depression | 0.52 |
| num_major_vessels | 0.42 |
| sexMale | 0.24 |
| chest_pain_type:Atypical Angina | 0.37 |
| chest_pain_type:Typical Angina | 0.18 |
| rest_ecg: ST-T wave abnormal | 2.07 |
| exercise_induced_agina:Yes | 0.48 |
| st_slope:Flat | 0.50 |
| thalassemia: Irreversible defect | 3.40 |

### 6.1 a) Possible Application of Decision Trees



### b) Empirical Design for healthcare App



## REFERENCES

1.“Cardiovascular Diseases.” World Health Organization, World Health Organization, https://www.who.int/health-topics/cardiovascular-diseases/.

2.“CDC Prevention Programs.” Www.heart.org, https://www.heart.org/en/get-involved/advocate/federal-priorities/cdc-prevention-programs

3.“CDC Prevention Programs.” Www.heart.org, https://www.heart.org/en/get-involved/advocate/federal-priorities/cdc-prevention-programs

4.The Lesson from the Complications of Coronary Arteriography. https://journal.chestnet.org/article/S0012-3692(16)52943-2/fulltext

5. Möller-Leimkühler, Anne Maria. “Gender differences in cardiovascular disease and comorbid depression.” *Dialogues in clinical neuroscience* vol. 9,1 (2007): 71-83.

6. Harvard Health Publishing. “Gender Matters: Heart Disease Risk in Women.” *Harvard Health*, https://www.health.harvard.edu/heart-health/gender-matters-heart-disease-risk-in-women.