



ST. JOSEPH'S UNIVERSITY
DEPARTMENT OF ADVANCED
COMPUTING

CLASSIFICATION OF HEART DISEASE

PROJECT REPORT

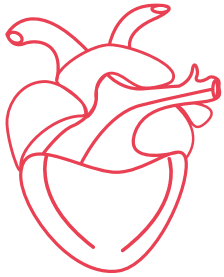
PREPARED BY

AISHWARYA CS	21BDA10
AKHILA TALLAM	21BDA46
HITESH S	21BDA47

CONTENTS

1. Problem Statement	
2. Introduction	
• Heart Disease	
• Dataset	
3. Methodology	
• EDA	
• Correlation Plot	
• Chi-Squared Test	
• Logistic Regression	
• Random Forest Algorithm	
4. Conclusion	

PROBLEM STATEMENT



**Analysis of Heart Disease Classification
Using Logistic Regression And
Random Forest Algorithm.**

INTRODUCTION

● Heart Disease

The Heart is the most important organ of human body. If it does not function properly then it affects other organ of the body. According to a report 7,000,000 die from heart attacks each year. According to WHO report around 17.9 million people die due to CVDS in 2016. 31% of the death of people is due to Heart disease around the globe in every year. The pumping of blood to the human body is the vital function of heart which supply oxygen and nutrients to the human body and also remove other metabolic waste from the body. If there is deficiency of blood in human body then heart doesn't function properly and it stop working which causes the death of human being. Angina occurs when there is temporary loss of blood to the heart causing chest pain.

Cardiovascular disease is of two types.

(1) Heart Attack–It occurs when the heart's blood vessels are suddenly blocked.

(2) Heart Failure–It results from coronary heart disease, hypertension, and cardiomyopathy.

Heart failure is basically when the heart cannot maintain a strong blood flow, resulting in chronic tiredness, resist physical activities, and shortness of breath. Heart failure can be divided into three types.1. right-side heart failure 2. Left-side heart failure 3. congestive heart failure. Right-sided heart failure usually causes left-sided heart failure. In the right-sided heart, failed blood backs up into other tissues such as the liver and the abdomen causing congestion in these areas. As a result of right-sided heart failure, we can have Hepatomegaly and Anciles. In left-sided heart failure, oxygenated blood cannot be pumped out from the heart to the rest of the body. So, blood can backflow. Blood can accumulate in lung veins causing fluid accumulation in the lungs causing shortness of breath and edema.

Living a healthy life style can reduce the effect of heart disease. Drinking plenty of water, eating green vegetables, fat free food, doing exercises, regular check-up of heart, consulting with the doctor if there any family history of heart disease can reduce the effect of heart disease.

• Dataset

The dataset is publically available on the Kaggle website for the analysis.

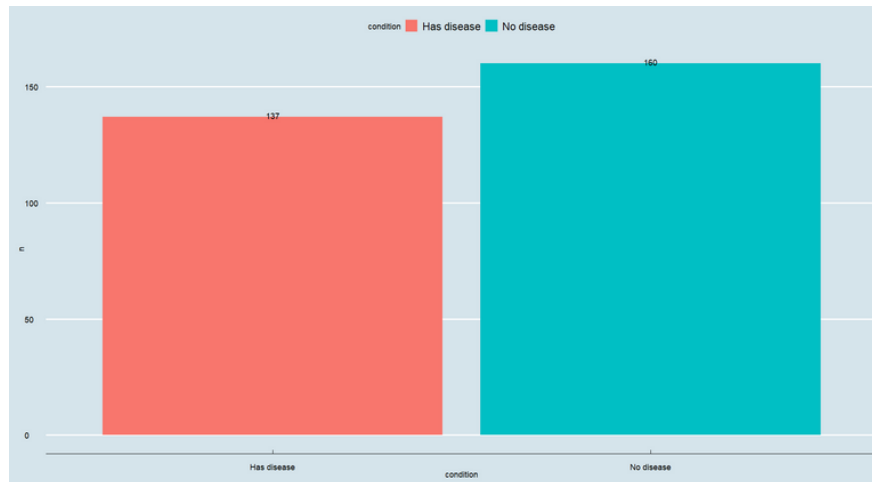
There are 14 attributes in our dataset. They are given below :

1. **age:** age (in years)
2. **sex:** gender (1 = male; 0 = female)
3. **cp :** chest pain type
There are three criteria for classifying different types of angina (chest pain) under three categories (according to this NCBI paper <https://pubmed.ncbi.nlm.nih.gov/20494662/>):
 - Location: Chest pain occurs around the substernal portion of the body
 - Cause: Pain is experienced after induction of emotional/physical stress
 - Relief: The pain goes away after taking nitroglycerine and/or a rest
 - i : typical angina (all criteria present)
 - ii : atypical angina (two of three criteria satisfied)
 - iii : non-anginal pain (less than one criteria satisfied)
 - iv : asymptomatic (none of the criteria are satisfied)
4. **trestbps:** resting blood pressure (in mmHg, upon admission to the hospital)
5. **chol:** serum cholesterol in mg/dL
6. **fbs:** fasting blood sugar > 120 mg/dL (likely to be diabetic) 1 = true; 0 = false
7. **restecg:** resting electrocardiogram results
 - Value 0: normal
 - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
 - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. **thalach:** maximum heart rate achieved
9. **exang:** exercise induced angina (1 = yes; 0 = no)
10. **oldpeak:** ST depression induced by exercise relative to rest (in mm, achieved by subtracting the lowest ST segment points during exercise and rest)
11. **slope:** The slope of the peak exercise ST segment, ST-T abnormalities are considered to be a crucial indicator for identifying presence of ischaemia (according to this research paper on NCBI: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7027664/>)
 - Value 0: upsloping
 - Value 1: flat
 - Value 2: downsloping

12. **ca:** Number of major vessels (0–3) colored by fluoroscopy. Major cardiac vessels are as follows: aorta, superior vena cava, inferior vena cava, pulmonary artery (oxygen-poor blood → lungs), pulmonary veins (oxygen-rich blood → heart), and coronary arteries (supplies blood to heart tissue). A radioactive isotope is introduced to the body followed by x-ray imaging to detect any structural abnormalities present in the heart. The quantity of vessels colored is positively correlated with the presence of heart disease.
13. **thal:** 0 = normal;
1 = fixed defect (heart tissue can't absorb thallium both under stress and in rest);
2 = reversible defect (heart tissue is unable to absorb thallium only under the exercise portion of the test) Thallium testing is a method where the radioactive element thallium (Tl) is introduced to the body through an IV injection, followed by nuclear imaging of the heart with a gamma camera which reveals structural issues and abnormalities of the heart by showing whether the isotope was absorbed by heart tissue under high (exercise) and low (rest) stress conditions.
14. **condition:** 0 = no disease,
1 = disease

METHODOLOGY

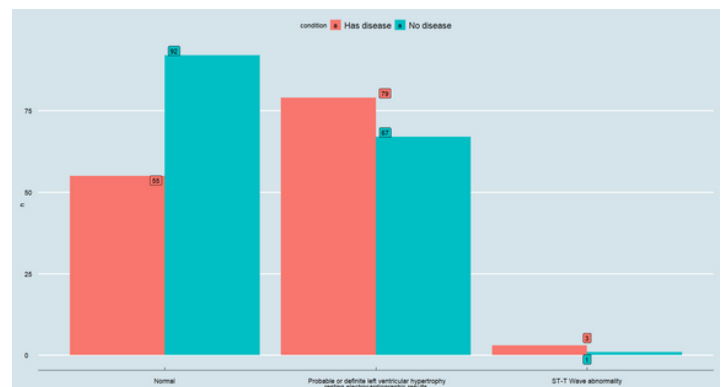
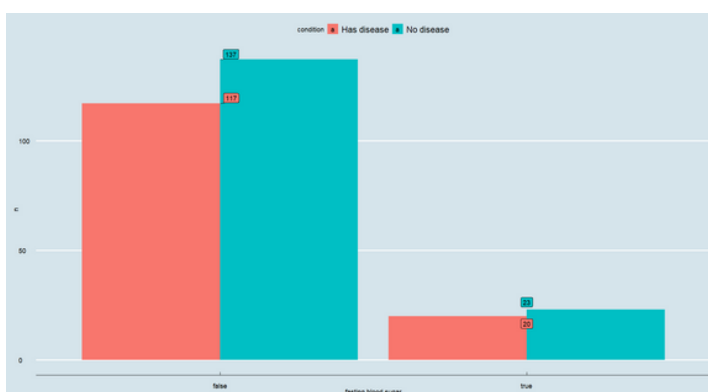
- EDA

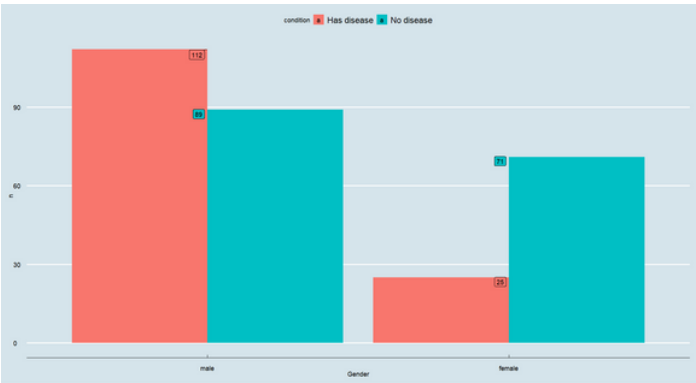
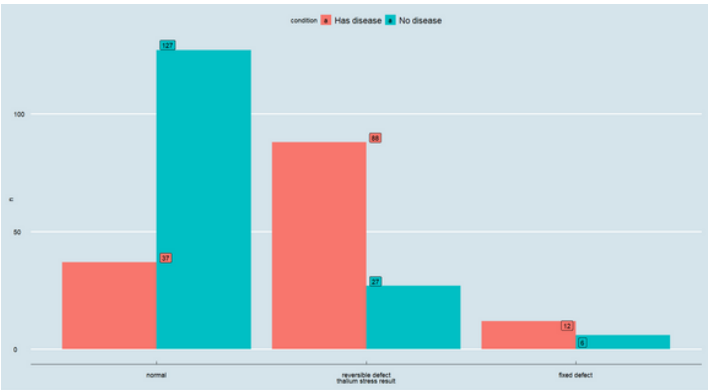
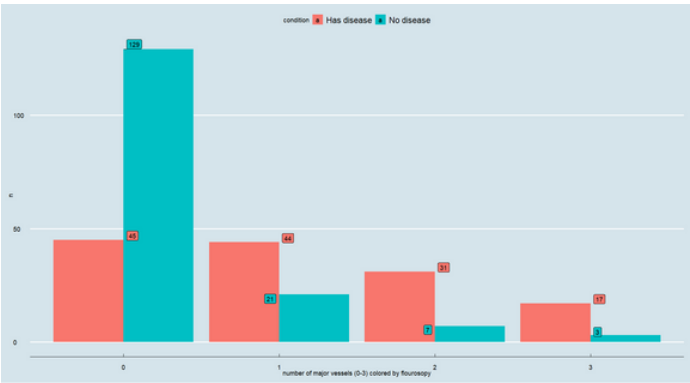
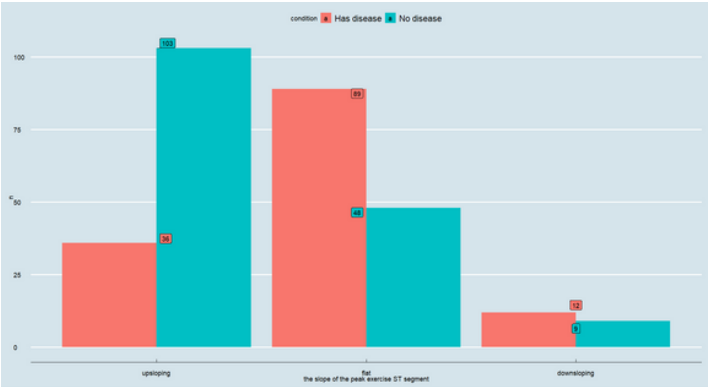
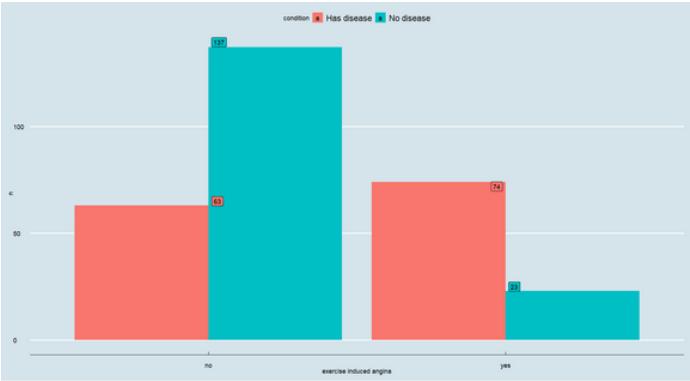


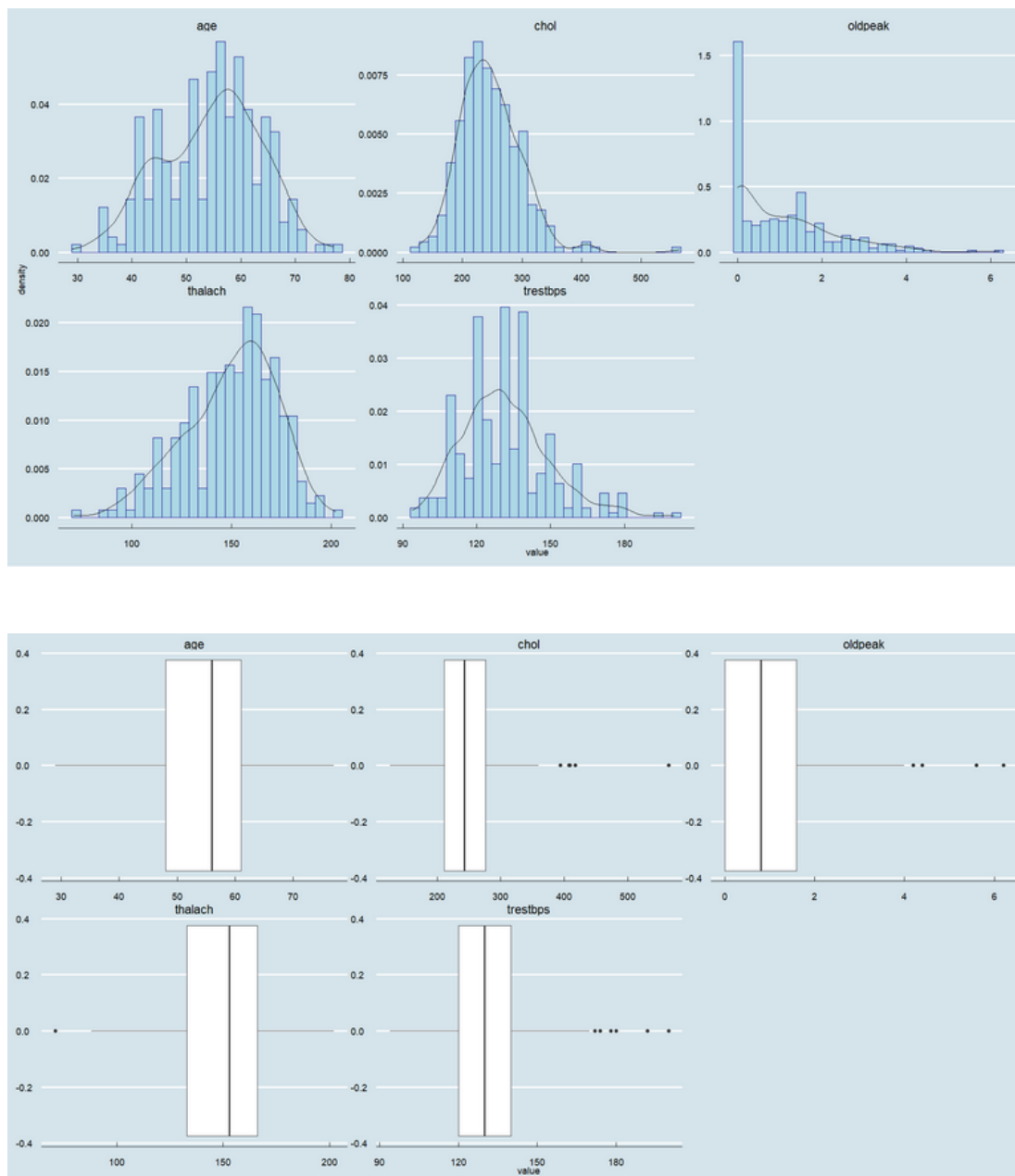
Based on the plot above, it is clear that the dataset is NOT experiencing imbalanced classification which is a supervised learning problem where one class outnumbers other class by a large proportion.

This problem is faced more frequently in binary classification problems than multi-level classification problems.

The term imbalanced refer to the disparity encountered in the dependent (response) variable (in our case the response variable, condition). Thus, an imbalanced classification problem is one in which the response variable has imbalanced proportion of classes. In other words, a data set that exhibits an unequal distribution between its classes is considered to be imbalanced.



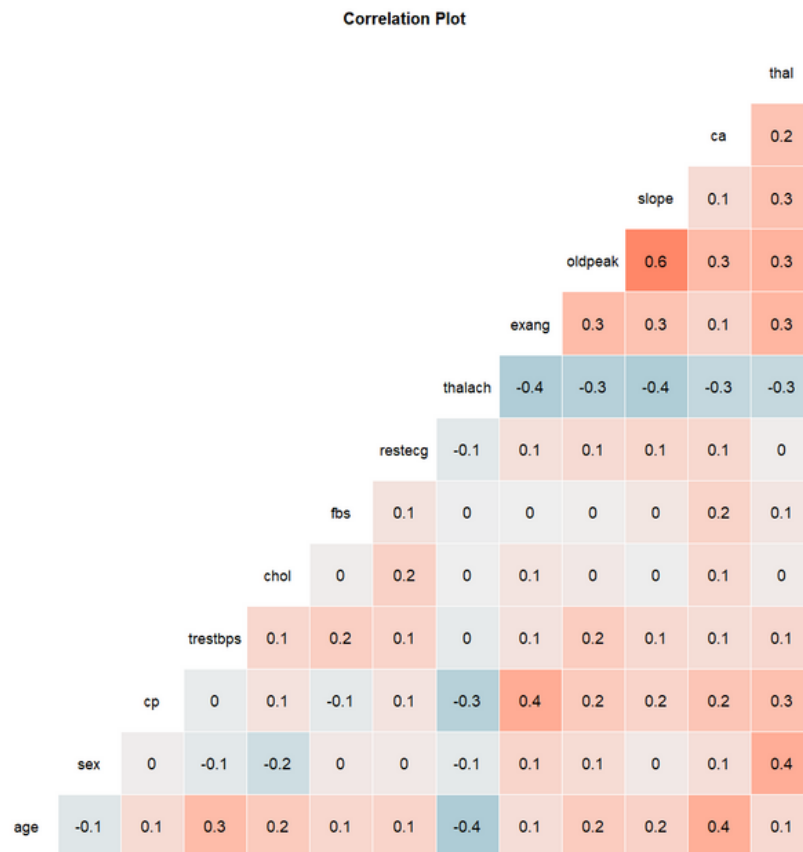




Interpretation of the visualizations above:

- There are fewer people who have heart disease in the dataset.
- From the looks of it, there are fewer people who have fasting blood sugar > 120 mg/dL i.e.(majority is false) than those who have fasting blood sugar > 120 mg/dL (likely to be diabetic).
- People who are showing probable or definite left ventricular hypertrophy by Estes' criteria as their resting electrocardiogram results take the lead when it comes to having heart disease. Followed by the category of normal resting electrocardiogram results. Then those having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) have 3 individuals with the disease. etc

- Correlation Plot



INFERENCE FROM CORRELATION PLOT

From Above, we can see that the pair of oldpeak and slope are the only ones that correlate above 50% i.e 60%. The rest correlate below 50%.

- Chi-Squared Test

Creating a 2-way contingency table(between the predictive variables & the outcome variable, condition)

A hypothesis test called contingency analysis is used to determine whether two categorical variables are independent or not. To put it another way, we're asking, "Can we foretell the value of one variable if we know the value of the other?" If the answer is true, we can conclude that the variables under examination are not independent. If the response is no, the variables under investigation are said to be independent.

The exam is called as 'Contingency Analysis' since it makes use of contingency tables. because the test statistic follows a chi-square distribution, it's also known as the 'Chi-square test of independence.'

The test used to determine if two categorical variables are independent or not. The test's null hypothesis states that the two variables are independent, whereas the alternative hypothesis states that they are not.

We are going to use the built-in `chisq.test()` function to perform Chi-square test of independence.

Firstly, we create a 2-way contingency table.

1.

	Has disease	No disease
female	25	71
male	112	89

```
> chisq.test(table_1)
```

Pearson's Chi-squared test with Yates' continuity co

data: table_1
X-squared = 21.852, df = 1, p-value = 2.946e-06

2.

	Has disease	No disease
asymptomatic	103	39
atypical angina	9	40
non-anginal pain	18	65
typical angina	7	16

```
> chisq.test(table_2)
```

Pearson's Chi-squared test

data: table_2
X-squared = 77.276, df = 3, p-value < 2.2e-16

3.

	Has disease	No disease
false	117	137
true	20	23

```
> chisq.test(table_3)
```

Pearson's Chi-squared test with Yates' continuity correction

data: table_3
X-squared = 1.9997e-31, df = 1, p-value = 1

4.

	Has disease	No disease
Probable or definite left ventricular hypertrophy	79	67
Normal	55	92
ST-T wave abnormality	3	1

```
> chisq.test(table_4)
```

Pearson's Chi-squared test

data: table_4
X-squared = 9.5755, df = 2, p-value = 0.008331

Warning message:
In chisq.test(table_4) : Chi-squared approximation may be incorrect

↓

Pearson's Chi-squared test

data: table_3
X-squared = 0.0029786, df = 1, p-value = 0.9565

5.

	Has disease	No disease
no	63	137
yes	74	23

```
> chisq.test(table_5)
```

Pearson's Chi-squared test with Yates' continuity correction

data: table_5
X-squared = 50.943, df = 1, p-value = 9.511e-13

6.

	Has disease	No disease
downsloping	12	9
flat	89	48
upsloping	36	103

```
> chisq.test(table_6)
```

Pearson's Chi-squared test

data: table_6
X-squared = 43.473, df = 2, p-value = 3.63e-10

7.

	Has disease	No disease
0	45	129
1	44	21
2	31	7
3	17	3

```
> chisq.test(table_7)
```

Pearson's Chi-squared test

data: table_7
X-squared = 72.301, df = 3, p-value = 1.373e-15

~ # that & condition

8.

	Has disease	No disease
fixed defect	12	6
normal	37	127
reversible defect	88	27

```
> chisq.test(table_8)
```

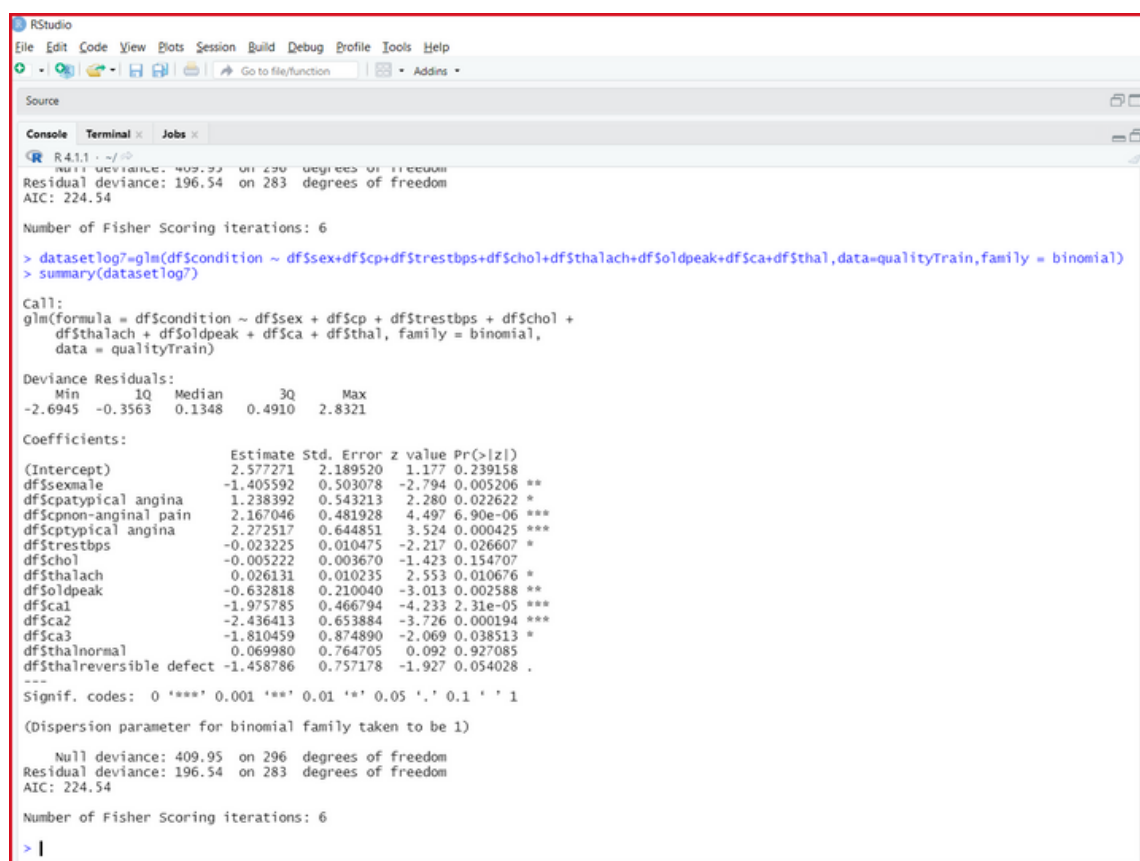
Pearson's Chi-squared test

data: table_8
X-squared = 82.46, df = 2, p-value < 2.2e-16

Interpretations from the above contingency Analysis

1. From the above result, we can see that p-value is less than the significance level (0.05) for all the tables except table 3. Thus, we can reject the null hypothesis and conclude that the two variables (sex & condition (i.e having the disease or not)) are not independent that is there exists an association between the 2 variables.
The same can be said with the variables chest pain, resting electrocardiogram, exercise induced angina, the slope of the peak exercise ST segment, ca and thal all having a p-value of less than 0.05 hence having an association with the outcome variable condition.
2. We can see a warning present itself that is, "Chi-squared approximation may be incorrect", when trying to determine whether the two-variable fasting blood sugar and condition (table_3) are independent or not. `chisq.test` function throws the above warning whenever one of the expected counts is lower than 5.
In this case, we will add the option "correct = FALSE" as the second argument in the `chisq.test()` function to tell R to not do a Yate's correction, which can be overly conservative. Let us see what the output will be when we make this change.
3. In 4th contingency table, we can see that, the previous warning is no longer present and we can see that the p-value is greater than the level of significance 0.05, thus we do not reject the null hypothesis that states that the two variables are independent that is there is no association between fasting blood sugar and the outcome variable condition.

• LOGISTIC REGRESSION MODEL



```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console Terminal Jobs
R 4.1.1 ~ /
Null deviance: 409.95 on 296 degrees of freedom
Residual deviance: 196.54 on 283 degrees of freedom
AIC: 224.54

Number of Fisher Scoring iterations: 6

> datasetlog7=glm(df$condition ~ df$sex+df$cp+df$restbps+df$chol+df$thalach+df$oldpeak+df$ca+df$thal,data=qualityTrain,family = binomial)
> summary(datasetlog7)

Call:
glm(formula = df$condition ~ df$sex + df$cp + df$restbps + df$chol +
    df$thalach + df$oldpeak + df$ca + df$thal, family = binomial,
    data = qualityTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6945 -0.3563  0.1348  0.4910  2.8321

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.577271   2.189520   1.177  0.239158
df$sexmale   -1.405592   0.503078  -2.794  0.005206 **
df$cpatypical angina  1.238392   0.543213   2.280  0.022622 *
df$cpnon-anginal pain  2.167046   0.481928   4.497  6.90e-06 ***
df$cpatypical angina  2.272517   0.644851   3.524  0.000425 ***
df$restbps    -0.023225   0.010475  -2.217  0.026607 *
df$chol       -0.005222   0.003670  -1.423  0.154707
df$thalach    0.026131   0.010235   2.553  0.010676 *
df$oldpeak    -0.632818   0.210040  -3.013  0.002588 **
df$ca1        -1.975785   0.466794  -4.233  2.31e-05 ***
df$ca2        -2.436413   0.653884  -3.726  0.000194 ***
df$ca3        -1.810459   0.874890  -2.069  0.038513 *
df$thalnormal  0.069980   0.764705   0.092  0.927085
df$thalreversible defect -1.458786   0.757178  -1.927  0.054028 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

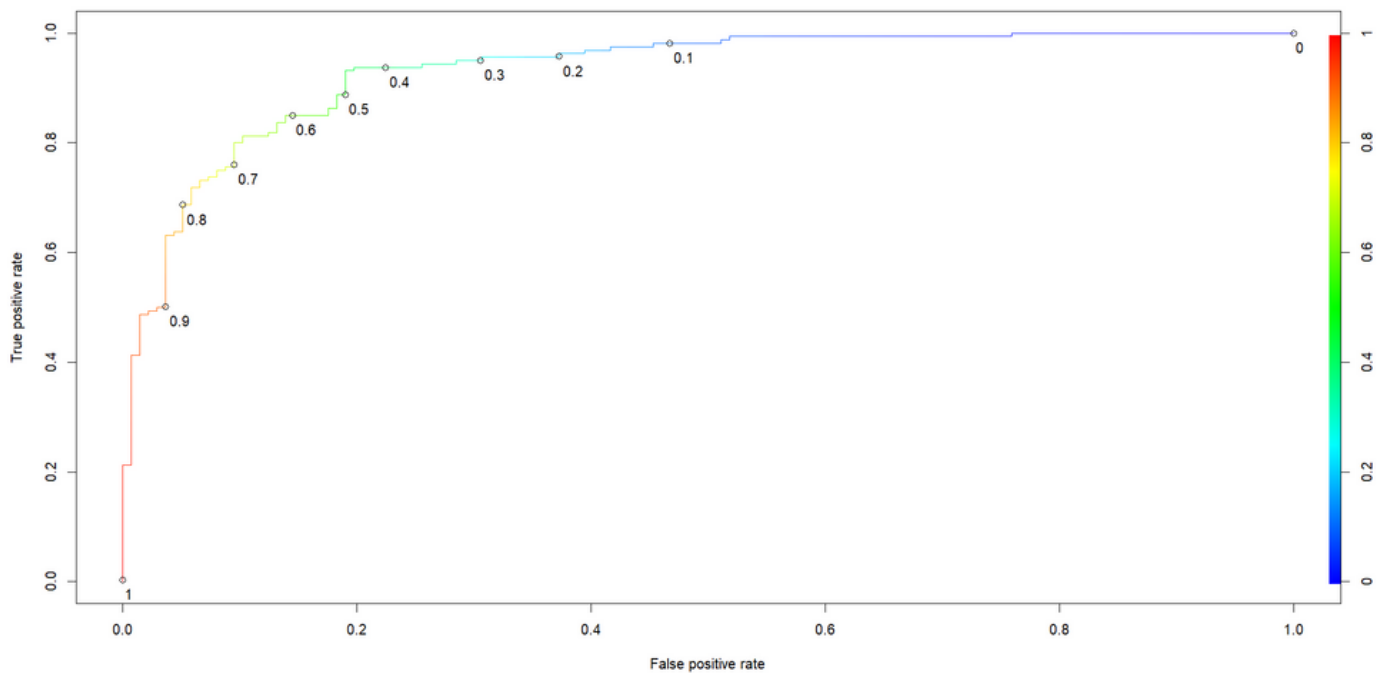
    Null deviance: 409.95 on 296 degrees of freedom
Residual deviance: 196.54 on 283 degrees of freedom
AIC: 224.54

Number of Fisher Scoring iterations: 6

> |
```

Logistic regression model

- we divide our dataset into train and test data 223 observations are used to training the model 74 observations are used to testing the data.
- For building a model we used condition as the dependent variable and other variables are independent.
- To build a good model we removed less significant variable rest ECG, FBS,slope, and EXANG.



ROC CURVE

- From ROCR curve threshold of 0.7 seems to be okay so that true positives are maximized such that maximum number of patients with heart disease are not identified as healthy.

```
> auc = as.numeric(performance(ROCRpred, 'auc')@y.values)
> auc
[1] 0.9314781
>
```

- we can see the value of AUC. Higher the AUC, better the model is at distinguishing between patients with disease and no disease.
- AUC value is 0.92 that means our model is able to distinguish between patients with the disease and no disease with a probability of 0.92. So it is a good value.

```
> table(df$condition, predictTest >= 0.7)

      FALSE  TRUE
Has disease  124   13
No disease   39  121

> #accuracy
> (39+13)/74
[1] 0.7027027
```

Logistic regression model with all the variables and logistic regression model after removing less significant attributes performed best with an accuracy of testing 70%

● RANDOM FOREST ALGORITHM MODEL

```
Console Terminal Jobs x
R 4.1.1 ~ /

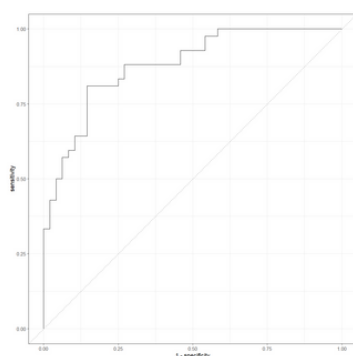
> # Training the final model on the full training data
> final_RF_model <- Final_spec %>% fit(condition ~ ., heart_train)
> final_RF_model
parsnip model object

Ranger result

Call:
ranger::ranger(x = maybe_data_frame(x), y = y, mtry = min_cols(~4,
importance = ~"impurity", num.threads = 1, verbose = FALSE,
x), num.trees = ~364L, min.node.size = min_rows(~5L, x),
seed = sample.int(10^5, 1), probability = TRUE)

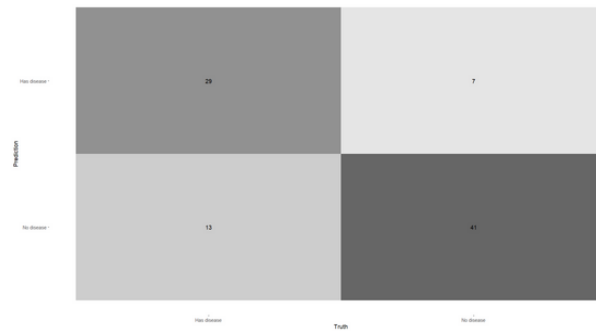
Type:
Number of trees: 364
Sample size: 207
Number of independent variables: 13
Mtry: 4
Target node size: 5
Variable importance mode: impurity
Splitrule: gini
OOB prediction error (Brier s.): 0.1361836
>
```

The present work predicts the suffering rate of a patient from heart disease using random forest algorithm.

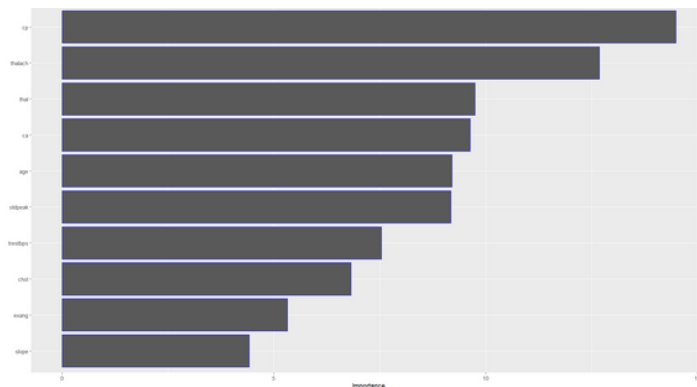


```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>    <chr>         <dbl>
1 roc_auc binary         0.879
```

The ROC curve between true positive rate and false positive rate at different threshold level is plotted. From this we obtain AUC value is 87% that indicates the model is 87% accurately predict whether the patient suffered from heart disease or not



From the confusion matrix ,we get 29 correctly classified negative class and 41 correctly classified positive class.



From this graph we can see the importance of the variable in random forest algorithm model.

CONCLUSION

- Logistic regression model with all the variables and logistic regression model after removing less significant attributes performed best with an accuracy of testing 70%
- From the random forest algorithm we obtained, the sensitivity value is 69%. The specificity value is 85.4% and accuracy value is 87%.
- Comparing the logistic regression model and the random forest algorithm model ,the random forest model as most accuracy model.
- The number of Heart diseases can exceed the control line and reach to the maximum point. Heart disease is complicated and every year lots of people are dying with this disease. By using different systems we can still have some of the major drawbacks. That is these work mainly focus only on the application of classification techniques and algorithms for heart disease prediction, by all these techniques that prepare and build a dataset appropriate for these algorithms. So that we can use logistic regression for predicting if the patient has heart disease or not.