University of Nottingham
UK | CHINA | MALAYSIA

Cardiovascular diseases Analysis and Prediction using Statistical Modelling and Machine Learning

by Kritin Pandita, MSc

Submitted to The University of Nottingham
in September 2022
in partial fulfilment of the conditions for the award of the degree of
Master of Science in Data Science

I declare that this dissertation is all my own work, except as indicated in the text

Table of Contents

Abstract

Machine Learning has continued to evolve since it commenced seven decades ago and subsequently has found itself an important application in the medical field. Cardio-vascular diseases are claimed to be a major global health problem in modern medicine that is leading to almost 17.9 million deaths a year as of 2019 which constitutes 32 percent of total deaths[4]. Prediction of cardio-vascular diseases at a preliminary stage to prevent further complications and possibly, death, hence becomes a necessity. Due to the convoluted nature of the factors and attributes involved in this disease, there exist a lot of possibilities of human fallacies and misconceptions. So, there is a dire need for technological advancement that can help medical researchers in the early detection of this lethal disease and hence can recommend certain lifestyle changes to the patients to stop the inevitable development of these cardiovascular diseases. Recent healthcare research and studies have made a virtuous use of various cutting-edge techniques and advanced scientific methods to accurately diagnose and predict the disease based on the data received from the person. One such technology is Machine Learning which helps us in making smarter and more accurate predictions on the data. The major contribution of this research project is to find the machine learning classification approach that can give the best classification accuracy when it comes to the prediction of this disease. This research work is conducted on different machine learning classification algorithms such as Logistic Regression, Naive Bayes Classifier, SVM, and Random Forest Algorithms. The dataset that we have based this research project on has been taken from Kaggle. The results and analysis of each stage of data modelling and classification are present in this study.

Acknowledgements

First off, I want to express my gratitude to The University of Nottingham for giving me access to all the materials I needed for my dissertation. Next, I want to thank Prof. M Hubbard, my dissertation supervisor, for helping me during the entire process. He is incredibly encouraging, and provided me with a lot of insightful suggestions that helped me improve the contents of my work. Then I would also like to thank all the teachers and academic staff at The University of Nottingham who also assisted me throughout my journey here in Nottingham. I hope that this adventure will allow me to learn a lot. Finally a huge thank you to my parents, who placed their faith in me and helped me through difficult periods on this path.

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Research Background

The human heart is a hollow muscular organ that uses rhythmic contraction and dilation to move blood through the circulatory system and is located in the chest, slightly to the left of the centre [1] . The heart beats around 100,000 times a day, pumping almost 8 pints of blood throughout the body 24/7 [2]. A heart is made up of the epicardial, the myocardial, and the endocardial layers which are surrounded by the pericardial thin outer layer. The heart is also made up of 4 chambers: 2 atria and 2 ventricles. The right atrium receives oxygen-poor blood from the body and pumps it to the right ventricle which pumps the oxygen-poor blood to the lungs. The lungs, in turn, pump the oxygen-rich blood to the left atrium which subsequently pumps it towards the left ventricle[2][3]. The left ventricle pumps the oxygen-rich blood to the rest of the body[2][3]. Heart diseases are also called cardiovascular diseases(CVDs)[4]. Cardiovascular diseases are the leading cause of death globally except in Africa[5]. According to the World Health Organisation(WHO), together CVDs have resulted in 17.9 million deaths(32.1% globally) in 2015, a substantial increase from 12.3 million(25.8%globally) in 1990[4].

Cardiovascular diseases develop when the heart and blood vessels malfunction[5].

They are commonly associated with the accumulation of fatty deposits inside the arteries (atherosclerosis) and an increased risk of blood clots[6]. It is also linked to artery damage in organs like the brain, heart, kidneys, and eyes. The term "cardiovascular disease" refers to a group of disorders such as coronary heart disease, cerebrovascular disease, rheumatic heart disease, and so on[5]. The underlying cause of this disease is associated with plaques of atheroma that form in the walls of blood vessels[6]. Many risk factors accelerate the development of atheroma, which can be classified as modifiable or non-modifiable[7]. Non-modifiable risk factors include age, gender, a family history of CVDs, and ethnic background, whereas modifiable risk factors include smoking, a low HDL (high-density lipoprotein)

cholesterol level, a high non-HDL cholesterol level, a lack of physical activity, an unhealthy diet, obesity, and poorly controlled diabetes.[7] The exponential rise in CVDs may be attributed to a lack of early detection of these diseases.[7] This global issue can be mitigated if these issues are identified at an early stage, which can be accomplished with the help of technological advancements such as Machine Learning.

1.2 Aims and Objectives

This research project aims to estimate the risk of cardio-vascular diseases based on their corresponding medical information using statistical modelling and machine learning algorithms. Due to the complexity and variability of the factors involved in this disease, the margin of error is minute and hence various modern professionals are relying on data classification and mining techniques to make decisions that are as accurate and precise as possible. Although prediction modelling is not an alternative to clinical treatments, it can serve as a first-hand tool to be aware of any type of disease and be prepared for it. The correct prediction of heart attacks using analysis and classification techniques can hence prevent life threats. The main objective of this research project would be to classify whether the patient is prone to CVDs in the future based on various medical attributes that are compiled in a dataset available to us using various classification algorithms. This early classification can be used by various medical researchers to comment on certain lifestyle changes that the patient needs to inculcate to stop the progression of this disease.

1.3 Research Objectives and Scope of the Research

This research attempts to answer the following questions centred around the dataset in particular :

1. Investigating the dataset that we have extracted from Kaggle and using appropriate graphical techniques to check how all the 13 data attributes such as age,
sex, Cholesterol, etc. are distributed with respect to each other and the target variable ,

analysing the data accordingly and forming an unbiased inference from the same using Exploratory Data Analysis.

2. Performing correlation analysis on the data using various techniques such as Principal Component Analysis, Correlation plots, and Boruta Techniques to comment about the strength of how strongly various attributes are related to each other and the outcome variable to comment on the importance of variables in a hierarchical manner.

3. Performing machine learning classification algorithms on the dataset and comparing metrics such as AUC values and classification accuracies of each algorithm to determine which algorithm is the best for heart attack classification in particular so that it can be fine-tuned even further to ameliorate the accuracy of the classifier.

1.4 Structure of the Dissertation

The remaining part of the dissertation is expected to contain four more chapters.

Chapter 2 of the thesis discusses the related work of the researchers that have worked on this topic in the past and a critical review of the literature.

Chapter 3 discusses the research methodology ,the takeaways of the research work, extensive data exploration, data pre-processing steps and their respective implementations required for the study.

Chapter 4 provides a detailed discussion on the results and analysis and discussion about the results and analysis of the chosen machine learning models.

Chapter 5 includes the conclusion, limitations and the future scope of the project

Chapter 2 : Literature Review

2.1 Introduction

It is necessary to start with the background study before we dive into our work to gain an understanding of the existing research and debates as well as discuss the gaps and limitations that come with our field of research. This chapter will hence, extensively discuss the Machine learning classification algorithms like Logistic Regression, Naïve Bayes Classification, Support Vector Machine, and Random Forest Classification Algorithms. Using more than one algorithm allows us to compare them and comment on classification accuracy to determine which algorithm is best suited for this dataset. Subsequently, related work from the past will also be reviewed to get a better understanding of what was done, how it was done, the methodologies used for the processed results, and the accuracies of the system.

2.2 Machine Learning

Machine Learning(ML) is a type of Artificial Intelligence(AI) that enables system to learn and improve from experience without being explicitly programmed. It focuses on developing computer programs that can access data and use it to learn for themselves.[8]

A formal definition of ML was proposed by Tom Mitchell[9] using a well-posed learning problem , stating that a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E[10].

To associate this definition with our research, we want to develop a heart attack prediction system and analogously the task T of our system is to estimate the risk of cardiovascular disease by predicting the occurrence of heart diseases based on the medical history of the patients. The performance measure P is the classification accuracy of our classification model and the experience E, in this case, is the already processed clinical trial datasets. The system will get more experienced as we increase the clinical datasets and hence increase the precision and accuracy of the model.

As with any method, there are different types of machine learning algorithms with their own advantages and disadvantages. To understand the advantages and disadvantages, it is important to have a discussion about the data that will be fed to the machine learning

algorithm on the basis of which we will infer the type that is suited to our usage.

Machine learning has three types of learning models :

1)Supervised Machine Learning[11].

2)Unsupervised Machine Learning[12].

3)Reinforcement Machine Learning[13].

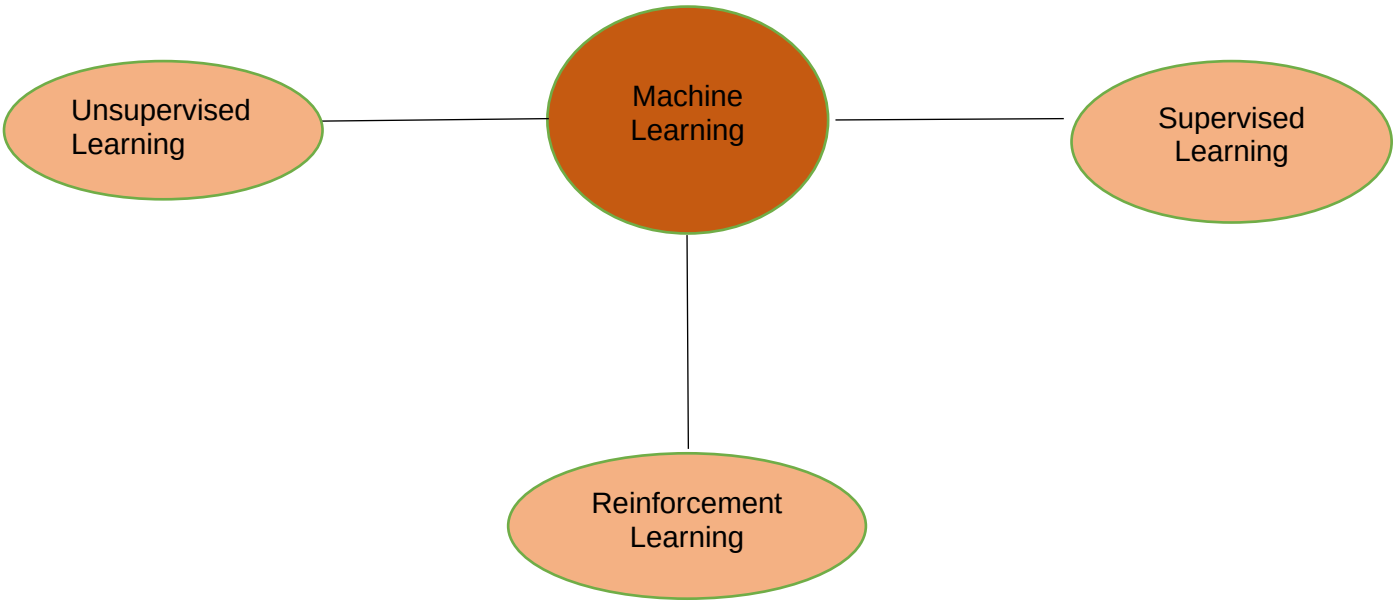


Fig2.1-Types of Machine Learning

2.2.1 Supervised Machine Learning

Supervised Learning, also known as Supervised Machine Learning, is a subcategory of Machine Learning and Artificial Intelligence that is defined by the use of labelled datasets to train algorithms that are used to accurately classify data and predict outcomes based on

the classification of this data[11]. Supervised learning uses a training dataset that is specifically fed to the classification model to discover and learn patterns that simultaneously train the classifier. It then uses the testing dataset that is also derived from the same data sample for prediction and forecasting purposes. There are generally two supervised learning algorithms[14]:

- Regression: Regression is used to understand the relationship between dependent and independent variables and hence, is generally used to make projections and forecasts. Linear Regression, Polynomial Regression, and Logistic Regression are all types of Regressions that are used. An example of Regression could be predicting the sales price of a house based on several parameters[15], [16].
- Classification: Classification uses algorithms that will specifically assign a data point to a specific category based on the fact that it tries to recognize data points and tries to derive certain specific conclusions on how those entities should be grouped or labelled. Common classification algorithms are linear classifiers, Support Vector Machines, K- nearest neighbour, etc. An example of classification could be the prediction of heart diseases based on the medical history of patients[17].

2.2.2 Unsupervised Machine Learning

Unsupervised Machine Learning uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms are designed to discover hidden patterns and similarities between data points without the need for human intervention. Its ability to discover similarities and differences in information makes it the ideal solution for anomaly detection, visualization, and exploratory data analysis. The goal is to find hidden structures and patterns behind the data and bring similar data points together as a cluster[12].

2.2.3 Reinforcement Machine Learning

Reinforcement Machine Learning is the training of machine learning models to make a sequence of decisions. The agent learns to achieve a goal in an uncertain, potentially complex environment. The computer employs a trial and error to come up with a solution to the problem. To get the machine to do what the programmer wants , the artificial intelligence gets either rewards or penalties for the actions it performs. The objective of the machine is to perform in an optimal way so as to minimise the punishments and maximize the rewards. Although the designer sets the rules of the game, he gives the model no hints or suggestions on how to solve the game. By leveraging the power of search and many trials, reinforcement learning is currently the most effective way to hint the machine’s creativity. Figure 2.2 shows

the basic framework behind reinforcement machine learning[13]

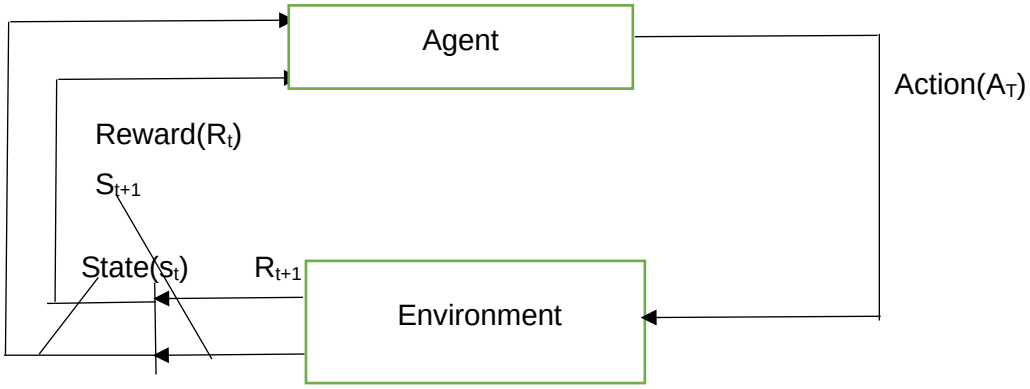


Figure2.2-Reinforcement Learning

2.3 Logistic Regression

Logistic Regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a dataset. A logistic regression model predicts the values of a dependent data variable by analysing the relationship between one or more independent variables[18]. The variables considered here are in binary format, as 0 or 1 or it can also be a Boolean value, as True or False. If the output variables or labels are three or more than three, then the logistic regression is considered to be multinominal logistic regression.[19], [20]

The main advantages of choosing Logistic Regression as a classification algorithm is that it is easy to implement, interpret and easy to train. It can interpret model coefficients as indicators of feature importance. Hence even though it is widely used it comes with its own set of limitations and disadvantages such as it assumes linearity between dependent and independent variables and hence it can't deal with non-linear problems properly. Moreover, there are more compact and powerful algorithms such as Neural Networks that can easily outperform this algorithm.[20]

The basic principal function behind logistic regression is the logit function which can be defined as the natural logarithms of ratios of the probabilities of success. Thus given a categorical variable Y and n number of predictors the logit function can be mathematically

defined as [15]

$$logit(Y) = log[(p)/(1-p)] = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

where β_0 is the intercept and β_i are the regression coefficients for each predictor X_i where i can be 1,2….n

Data must often be divided into many pieces when using supervised machine learning techniques like Logistic Regression in order to train, validate, and ultimately test classifiers. This separation of test data poses a challenging trade-off between having more statistical power to estimate generalisation performance vs better parameter selection and model fitting when there is a lack of data.[11]

Ambrish G et al(2022)[21] used Logistic Regression technique for the prediction of cardiovascular disease. The logistical regression classifier is tested using the UCI repository dataset with five different ratios associated with the testing and training dataset. It was observed that maintaining a 90:10 split ratio for training and testing resulted in the highest classification accuracy of 87.10%. The minimum classification accuracy was observed when the splitting ratio was observed to be 50:50 and was 81.58%[21].This research specifically deals with the logistic regression to predict the binary output(0= less chance of heart attack 1= more chance of heart attack) based on the features that are present as the medical history of the patient . We will apply logistic regression on these features to predict the binary output.

2.4 Random Forest Classification

Random Forest Classification Algorithm[22] is a machine learning algorithm that uses the combined results of various decision tree to form a single result. It is fairly easy to use and has a flexible nature which has resulted in its usage for regression and classification purposes. It makes use of ensemble learning[23] which is technique that involves an amalgamation of a lot of classification techniques to provide solutions to complex problems.

Random Forest Algorithms require the initialization of three main parameters which are to be set before it is used for training a dataset. The parameters in discussion are the size of the node, the number of trees and the number of features that are to be dealt with. A random forest is preferred to decision tree algorithms since it overcomes the limitations that are faced in the latter algorithm by increasing accuracy and precision and overcoming overfitting of datasets[22].

In the case of classification because of Random Forest , the training data is fed to train various decision trees. The dataset consists of observations and features that are selected

randomly during splitting of the trees. Every decision tree consists of leaf , decision and root nodes. The leaf node is the final output that is produced by that decision tree and that takes place for every single decision tree present in the forest. In this case, the output chosen by the majority of the decision trees is chosen as the final output of that particular random forest[22].

The main algorithm behind this classifier like the name implies is made up of a large set of discrete decision trees that work together as an ensemble[23]. Each individual and uncorrelated tree in the random forest produces a class prediction, and the class with the most votes becomes the final prediction of our model. The algorithm works on the principle that a significant number of fairly statistically independent models (trees) acting as a working group will outclass any of the constituent models individually[22].

Using random combinations of the hyperparameters[24], the optimal solution for the created model is found using the random search approach[25]. The objective function receives random inputs that are generated and evaluated by the random search algorithm. It works well because it makes no assumptions on the objective function's structure. This can be useful for issues when there is a lot of domain knowledge that could bias the optimization method or impact it, leading to the discovery of counterintuitive solutions. Ashir Javeed et al. (2019)[26] designed an Intelligent Learning System which was based on Random Search Algorithm and an optimised version of Random Forest Model for improved heart disease detection in which they conducted two sets of experiments. In the first set only random forest algorithm is used and in the second set the proposed optimised version of the random forest that involves Random Search Algorithm is used. The latter is a conventionally better performer since it produces a higher classification accuracy than the former by a significant factor of 3.3 percent[26].

Figure2.3-Random Forest Classification[17]

2.5 Support Vector Machine Classification(SVM)

A support vector machine is a machine learning algorithm that is extensively used for classification, regression and outlier detection. The aim of this classification technique is to create a boundary line(which is known as a hyper-plane), that can divide a n-dimensional space into number of classes so that we can easily put the new data point in the correct dimensional space in the future[27].

SVM decides the location of the hyper-plane by taking the most extreme-points into consideration. These data-points or vectors that are the closest to the hyper plane and hence affect the position of the boundary line are called support vectors since their function is to support the hyper-plane. The dimensions of the hyper plane are dependent on the attributes that are taken into consideration. If there are more than two attributes present, then instead of a linear plane, a 2D plane will be considered[27]. Since our research is based on a binary classification problem(0= less chance of heart attack 1= more chance of heart attack) , we will be primarily focusing on linear SVM.

The procedure of the SVM algorithm can be divided into certain steps for further understanding:

- The classes are forecasted by the SVM algorithm. The first of the classes is labelled as 1, while the other is labelled as -1[27].
- The problem is converted into an optimisation problem in which a function called as hinge loss function is computed which is used to find the maximum margin between the support vectors. The loss is defined with the following formula[28]

$$l(y)=max(0,1-t\ .\ y)$$

  where t is the outcome (1 or -1) and y is the classifier output
- The loss function is also referred to as the cost function which is zero when there are no terms that are incorrectly predicted. But in the case that there are some terms that

are incorrectly predicted we have to add a penalty regularization factor which is mathematically shown as follows

$$\min_w (0.5\sum_{j=1}^{n} w_j^2)+\sum_{k=1}^{m} max(0,1-t_k . yk)$$

where w is the weight that are optimised with the help of advanced calculus operations[28].

SVM can be explained more clearly using an example. Suppose we have an animal that actually looks like a tiger but also has some features of a lion, so if we want to accurately identify whether the animal is a tiger or a lion, we can do that by creating a support vector machine. The SVM model will first be trained using a training dataset that will consist of a lot of images of lions and tigers which would train it to correctly understand the different features of lions and tigers. After the training period , we will use the algorithm on our testing data which in this case is the strange animal. The SVM will create a decision boundary between these two data(tigers and lions) and also will compute where the support vectors, which will be decided on the basis of the extreme cases. On the basis of the support vectors it will correctly classify it to be a tiger.

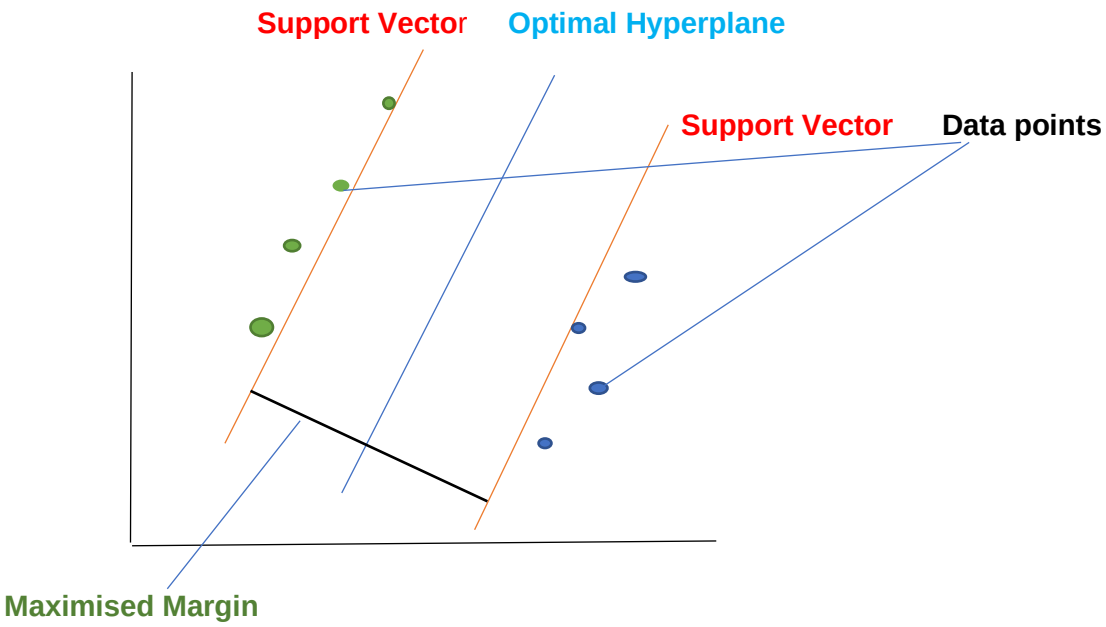


Figure2.4-Support Vector Machine Outline

In a research published in 2016[29], S.Radhimeenakshi employed SVM to analyse cardiac attacks. She employed the Cleveland database and the stat-log database and used the Support Vector Machine, Artificial Neural Networks, and Decision Tree Algorithm to compare the outcomes of the two databases. Using the Support Vector Machine technique, she was 84% accurate. She used a linear kernel classifier that can be mathematically represented in the form of F(x)=WI+bias where W is the weight and I is the input vector. The hyperplane was fixed at F(x)=0 which meant that points could be classified depending on whether the value of F(x)>0 or F(x) <0[30] .

We divided the data points in our study into two groups, diseased and normal patients based on feature variables such as cholesterol, age, sex and other medical information that is given to us , the data points are divided into two groups. Similar to the study described above, we employed a support vector machine in this research to linearly classify the data, divide the datapoints into two groups depending on the characteristics by defining a hyperplane f(x)=0, and then classify the data into those groups.

2.6 Naive Bayes Classification

Naïve Bayes Classification is a fairly straightforward and the most commonly used classification algorithm from the Bayesian network[11]. There is a very strong assumption associated with the Naïve Bayes Classification which is that all the random variables are independent to each other[31]. To understand this classification technique, it is important to understand what The Bayes Theorem is initially[32]. The Bayes theorem provides you a method to calculate conditional probability and even though it is generally used as a tool to calculate probability in statistics it can also play a major part in Machine Learning for predictive modelling using Bayes Classifier and Naïve Bayes. This theorem gives a way to calculate posterior probability P(a/b) for the probability class P(a) given that certain other probabilities are given to us. Hence, mathematically[32],

$$P(A/B) = (P(A)\ P(B/A)) / P(B)$$

here,
P(A/B) is the probability that A will happen given that B has already taken place
P(A) is the probability of the event A
P(B/A) is the probability that event B will happen given that A has already taken place
P(B) is the probability of the event B

Naïve Bayes Classifier uses Bayes Theorem to predict membership probabilities for each probability class such as the probability that a given record belongs to a particular class. The class that has the highest probability is considered to be the outcome class. As we established earlier that Naïve Bayes always assumes that all features are unrelated and independent to each other so the presence or absence of a feature has negligible influence towards how the Naïve Bayes classifier will work when it comes to finding the outcome class[31].

Particle Swarm Optimisation[33] is a feature extraction technique that is used to remove the redundant features to fine tune the accuracy of the classifier. Uma. N Dulhare(2018)[34] designed a prediction system using Naïve Bayes and Particle Swarm Optimisation(PSO). In this experiment the Naïve Bayes Classifier's accuracy is improved using PSO as a feature subset selection pre-processing algorithm. In PSO optimisation process the result is obtained when the number of iterations reaches in a state of convergence which in turn gives a classification accuracy of 87.91 %. The predictive model with a standalone Naïve Bayes with no PSO optimisation gave an accuracy of 79.12 % which is 8.79 % less than the existing system[34].

2.7 Conclusion of Literature Review

The literature study contributes to a thorough understanding of the statistical modelling and machine learning field. Many individuals and researchers have employed various strategies and machine learning techniques. Despite the fact that other methods and strategies have been suggested, this research will concentrate on a thorough analysis of the Machine Learning methods included in the literature review. The literature review increased our knowledge of the many algorithms that were covered there.

Chapter 3 : Methodology

3.1 Introduction

This chapter discusses the research techniques and methods that are used in this research project extensively. The dataset on which the methodologies are performed is in the form of a tabular data that has been extracted from Kaggle. The methodology starts with Feature extraction, followed by feature reduction which makes the data feasible for further data modelling and analysis. The methodology of the data begins with data exploration, data pre-processing[35] , data normalisation[36], feature selection and extraction[37], training and prediction of the results.

3.2 Network Design of The Methodology

This section of this chapter illustrates the network flow of the methodology. It describes the actual flow of the entire network design. The flowchart shown below explains the sequence involved in the network design. In principle, dataset is collected and downloaded from the source website. After preliminary data exploration, it is then pre-processed[35] to solve for the problem of missing values and dealing with redundant features by feature extraction techniques[37]. The data is hence feasible for classification which leads to the next stage that involves building, training and testing of the datasets.

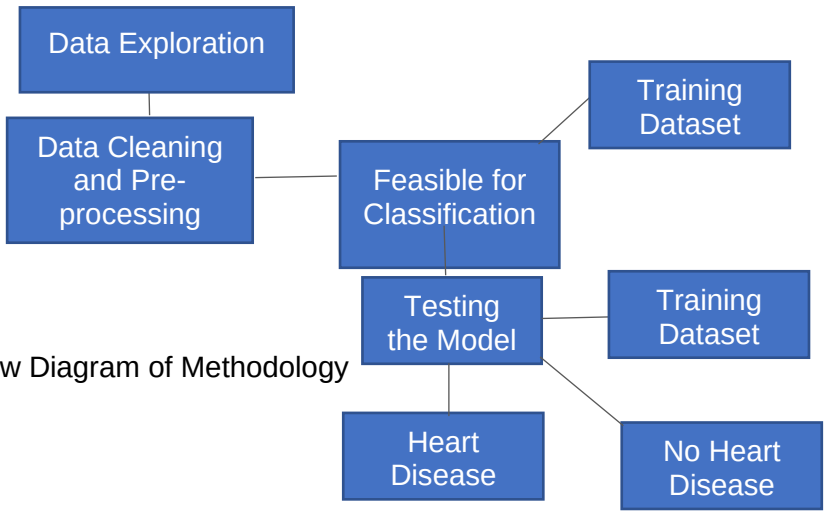


Figure3.1 Network Flow Diagram of Methodology

3.3 Data Exploration and Description

The dataset[38] used for this research project has been extracted from Kaggle , a subsidiary of Google LLC, is an online community that allows users to find and publish datasets, explore and build models in a web-based data science environment. The csv file[31] found on Kaggle supposedly originated from The University of California, Irvine, Machine Learning Repository[39]. It consists of 303 samples with 14 attributes ; 13 numeric input attributes named age, sex, chest pain type, cholesterol, fasting blood sugar, resting ecg, maximum heart rate, exercise induced angina, old peak, slope, number of vessels coloured, thal and one outcome variable named target variable (which has the values ranging from 0-1). The table 3.1 showcases information about the values and their respective ranges. The sample dataset along with the columns is shown in figure 3.1

Attribute	Brief Description	Domain of the Values
Age	The age of the patient in years. Since, generally heart attacks increase with increasing age, the age range has been kept from 35 onwards.	29-77
Sex	The gender of the patient, The nominal values of 0 and 1 are assigned respectively to facilitate evaluation.	0-1 0-Female 1-Male
CP	The most important risk factor when it comes to cardiovascular diseases is chest pains or chest angina. Chest Angina acts as an initial diagnosis for the presence of a particular problem.	Chest Pain Type: Value 1-Typical Angina Value 2-Atypical Angina Value 3-Non-anginal Pain Value 4- Asymptotic Pain
TrtBps	The Blood Pressure contributes to the systematic functioning of the heart. If you have heart failure, there is also a good chance that you have high blood pressure which is also called hypertension	94-200 mm Hg
Chol	Cholesterol is a fat-like substance called a lipid that's found naturally in the blood. Lipids are vital for the functioning of the body.	126-564 mg dL^{-1}
Fbs	Blood Glucose is a disease that occurs when your blood sugar is too high. For this dataset, blood sugar is distinguished by whether the patient's blood sugar is higher than 120 mg DL^{-1}	Sugar >120 mg dL^{-1}: Value 0: False Value 1: True

RestECG	Resting ECG results. The Electrocardiogram results are accepted as the current standard for heart evaluation. The ECG or EKG tests are often used to assess the heart rate and rhythm and detect diseases like heart attacks, abnormal heart rhythms etc.	Resting Electrocardiographic results Value 0: Normal Value 1: Having ST-T Wave abnormality Value 2: LV Hypertrophy
Thalach	Maximum Heart Rate Achieved. The dataset showcases that the maximum heart rate achieved depends on the age of the patient. So generally speaking the youngest person in the dataset is expected to have the maximum heart rate	71-202
Exang	Exercise induced Angina is the attribute that tells us about whether there has been pain caused due to exercise or not.	0-No,1-Yes
Oldpeak	The resting stress test segment depression is the marker for cardiac events	0-6.2
Slope	Slope of peak exercise ST segment can help us in judging whether the patient has heart disease. The depression of the ST segment could help in determining a heart disease	Value 1-Upsloping Value 2-Flat Value 3-Downsloping
Caa	Number of vessels coloured by Fluoroscopy	0-4
Thall	Thallium Heart Rate Label	Value 1- Normal Value 2-Fixed Defect Value 3- Reversible Defect
Output	Chances of the heart disease	Value 0- Less chance of a heart attack Value 1- More chance of a heart attack

Table3.1-Data Dictionary

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
6	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
7	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
8	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1

Figure3.2-Dataset before Pre-Processing

3.4 Exploratory Data Analysis

Exploratory Data Analysis(EDA) refers to the preliminary investigation techniques that are taken place on the datasets to gain insights about any hidden patterns or anomalies using graphical representations and interactive visualisations. EDA allows us to get a clear idea about the dataset. In the case of this research project it is important to establish preliminary research questions that are centred around the dataset and use exploratory data analysis as a tool to answer them clearly[40].
To get started we would like to know the distribution of the data for the comparison of the presence and the absence of the disease

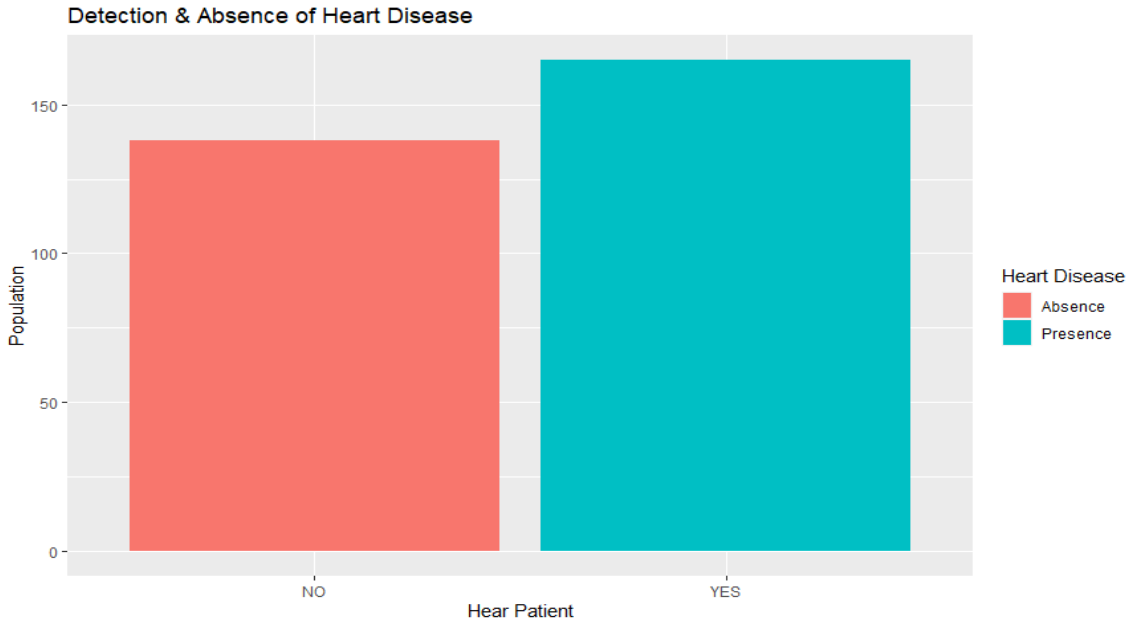


Figure3.3-Distribution of Detection and Absence of Heart Diseases

As we can see it would be safe to infer from the data that there are more people that are having a heart disease than the ones that don't have any heart problems. After quantifying the data observed in the bar graph, we observe that 54.45% of the population are diagnosed with heart diseases, whereas 45.54% of the population are not having any types of heart diseases. Hence making the development of an accurate predictor even more essential.

3.4.1 Cumulative Analysis for Age

Now to look at this at an even granular level we fill try to find which age group is prone to the most heart diseases.

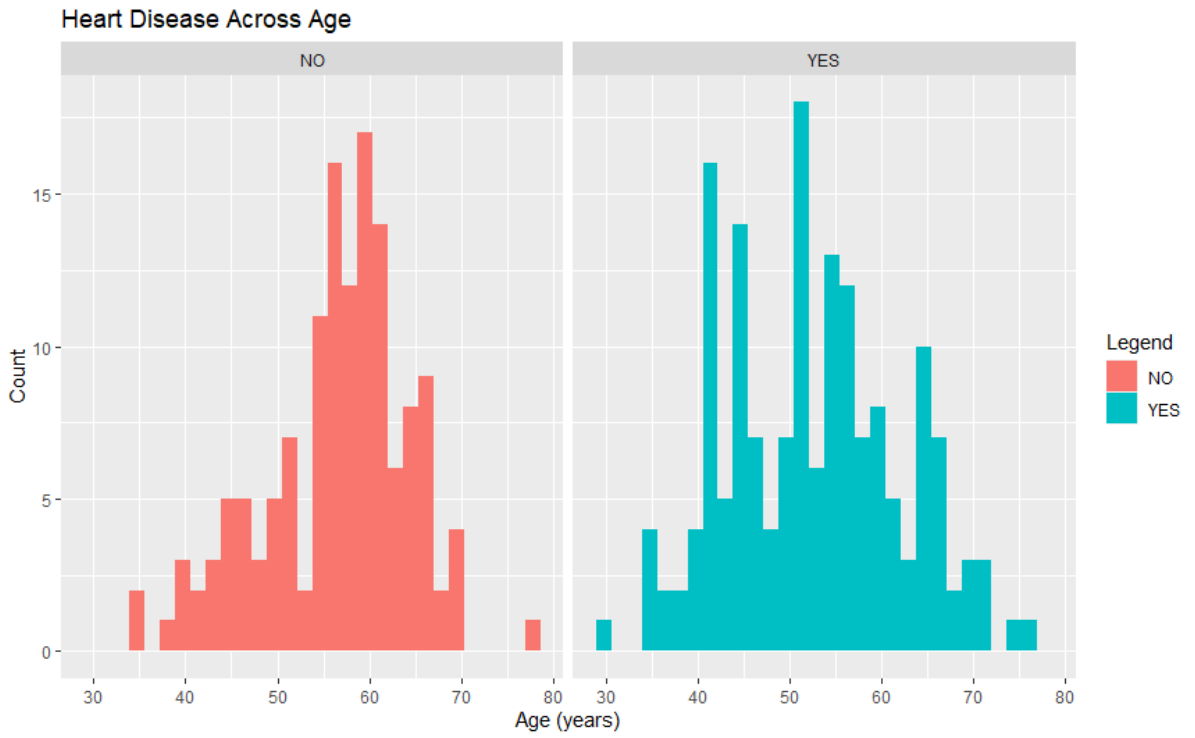


Figure3.4-Age Frequencies Distribution

Age is one of the more important risk factors in developing cardiovascular diseases[41]. As we can see from the frequency-age distribution as the age increases so does the frequencies of heart diseases. This could be supported by the fact that as we grow older our heart can no longer beat fast enough due to the build up of fatty deposits in the arteries over the years. This build up often leads to pain which is called angina pain and hence also results in other medical conditions such as stiffening of the arteries which leads to increasing blood pressure and hypertension.

3.4.2 Cumulative Analysis for Chest Pains and Resting Blood Pressure

Suppose we want to look at how gender is distributed with respect to the resting blood pressure along with the respective chest pains.
From the figure 3.5 , we are eligible to make the following conclusions about the dataset :

- We can see that for asymptotic chest pains, females are having higher blood pressures than 180 which is far above the normal range[42]. Males also have higher blood pressures but the maximum is less than 180.

- In the case of Atypical Angina, both males and females generally are having fairly normal readings when it comes to the blood pressure. There are a few outliers in the case of some males that have fairly high blood pressures and one case whose BP has reached catastrophic levels
- For Non-Anginal Pain, the average resting blood pressure for both males and females is observed to be less than 150 irrespective of a few outliers present in the male category where we notice a case where the BP is almost 90 and two cases where the blood pressure is very high[42].

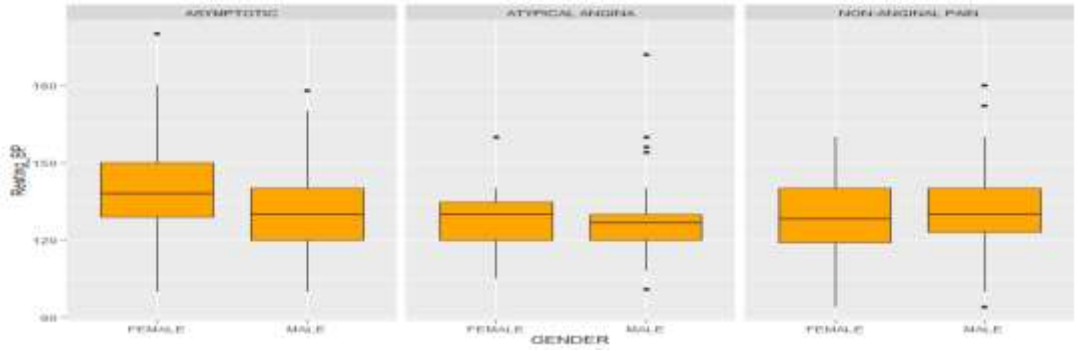


Figure3.5 Gender Blood Pressure Chest Pain Box Plot Distribution

The figure 3.6 shows a bar chart that tells us about the distribution of the types of chest pains that are seen in patients. Surprisingly, asymptomatic chest pain has the highest prevalence of heart disease, while typical angina pain has the lowest. There are more people without heart disease who have atypical or typical angina chest pain than people with heart disease.

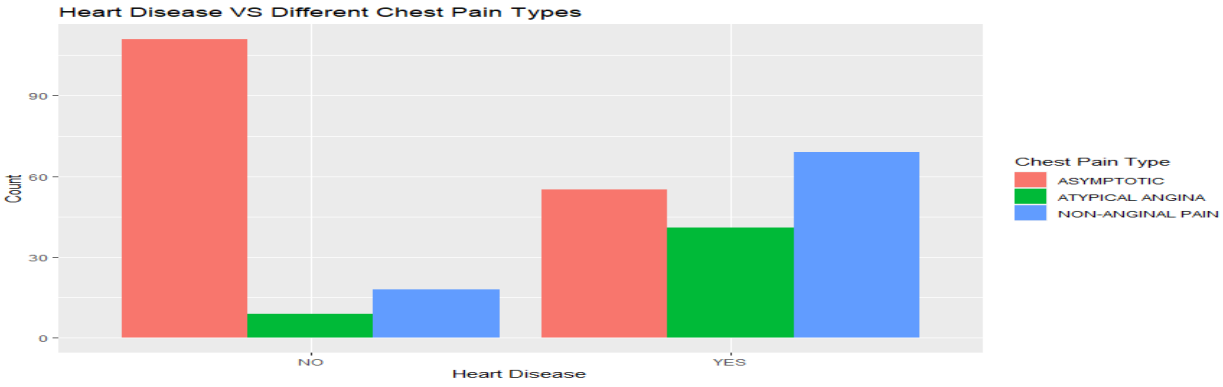


Figure3.6 Types of Chest Pains

3.4.3 Cumulative Analysis for Cholesterol

Cholesterol is a wax-like substance that is found in your bloodstream. Although your body requires cholesterol to build healthy cells, high cholesterol levels can increase your likelihood of developing disease[43]. High cholesterol deposits can cause fatty deposits in your arteries and blood vessels. According to WebMD[44] :

- The ideal serum cholesterol level is less than 200mg/dL.
- The Borderline High of serum cholesterol level is: 200-239 mg/dL
- A high serum cholesterol level is defined as more than 240 mg/dL. (being at this stage may double the risk of heart disease)

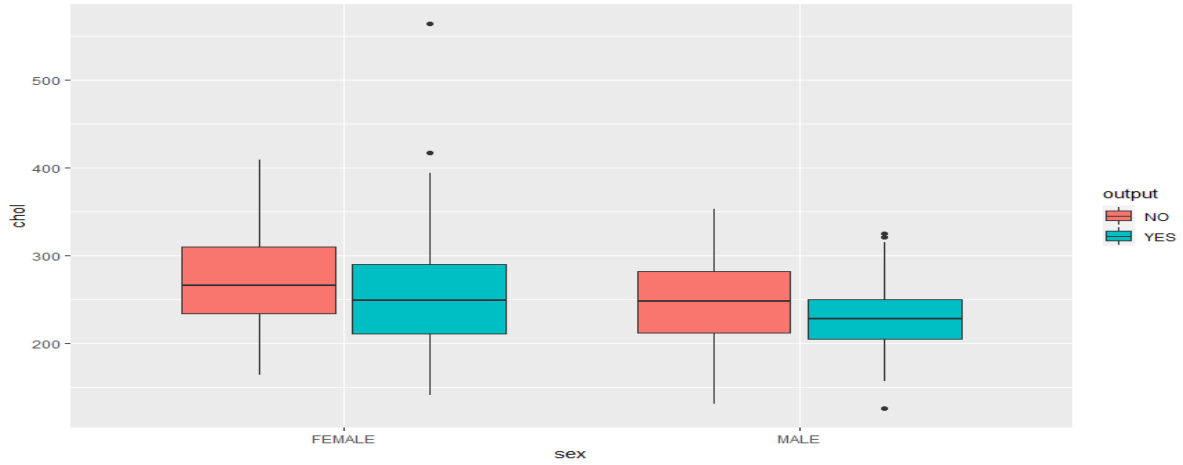


Figure 3.7 Cholesterol Sex Target Variable Boxplot

- From the figure 3.7,we can see that those who are not diagnosed with any cardiovascular diseases have average cholesterol levels that are ranging near the 250 mark but there are no outliers in this case so there are no cases which have shooting cholesterol levels.
- In the case of the presence of cardiovascular disease , we can infer that even though the average male and female do have cholesterol levels that are less than 250, there are a few cases in both genders that show that shooting cholesterol levels could also be a risk factor when it comes to these diseases.

As shown in figure 3.8, the average cholesterol levels of those with and without the heart disease fall under the “high serum cholesterol level” category[44]. We can hence do a welch hypothesis test[45] to check whether the two averages are different by checking the statistical significance using a t-test.

output	average_chol	min	max
NO	251.09	131	409
YES	242.23	126	564

Figure 3.8-Serum Cholesterol Statistics

Our null hypothesis states that the average serum cholesterol levels of a patient with heart disease and a patient without heart disease are the same. As we can see in figure 3.9, the p-value of 0.136 is larger than the threshold significance level of 0.005 which would make impossible to reject the null hypothesis.

```
        Welch Two Sample t-test

data:  heartattack2$chol by heartattack2$output
t = 1.4948, df = 298.03, p-value = 0.136
alternative hypothesis: true difference in means between group NO and group YES is not equal to 0
95 percent confidence interval:
 -2.803241 20.516548
sample estimates:
 mean in group NO mean in group YES
         251.0870          242.2303
```

Figure 3.9-Hypothesis Testing

3.4.4 Cumulative Analysis for Sex

As it can be seen in the figure 3.10 , the incidence of having heart diseases is higher in men than woman but at the same time there are also more men that do not have heart diseases.

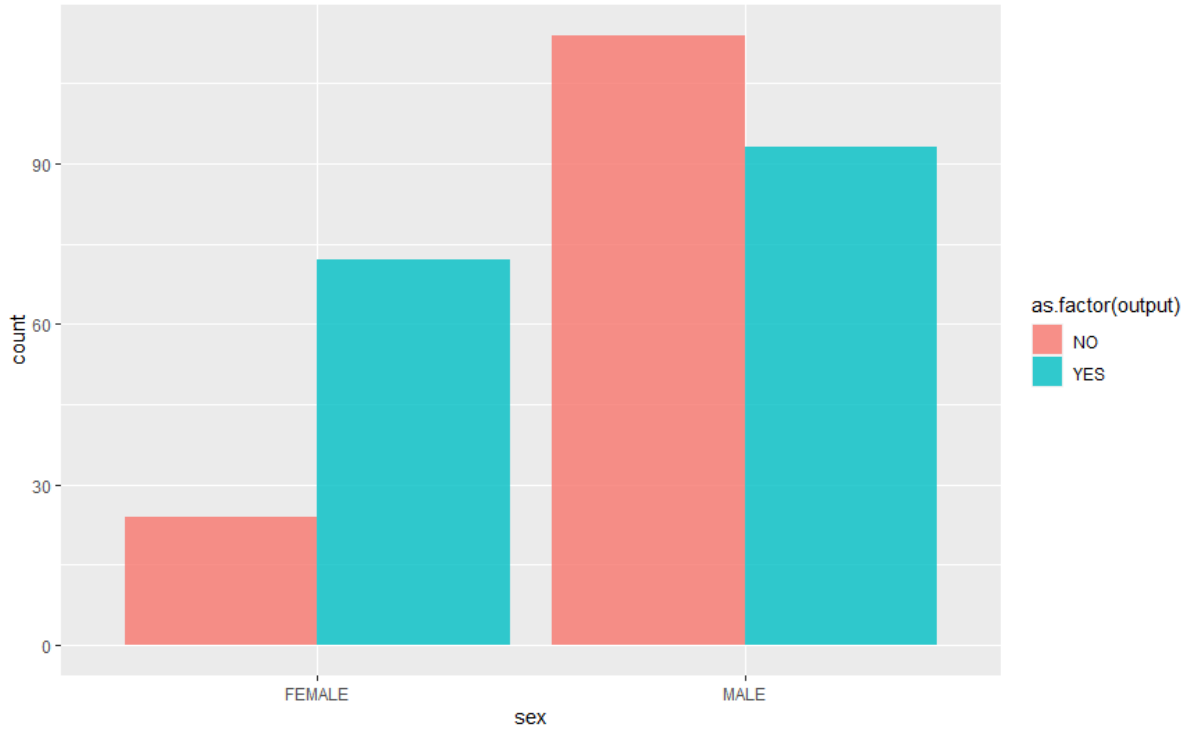


Figure3.10-Sex- Target Variable Boxplot

3.4.5 Cumulative Analysis of Resting Blood Sugar

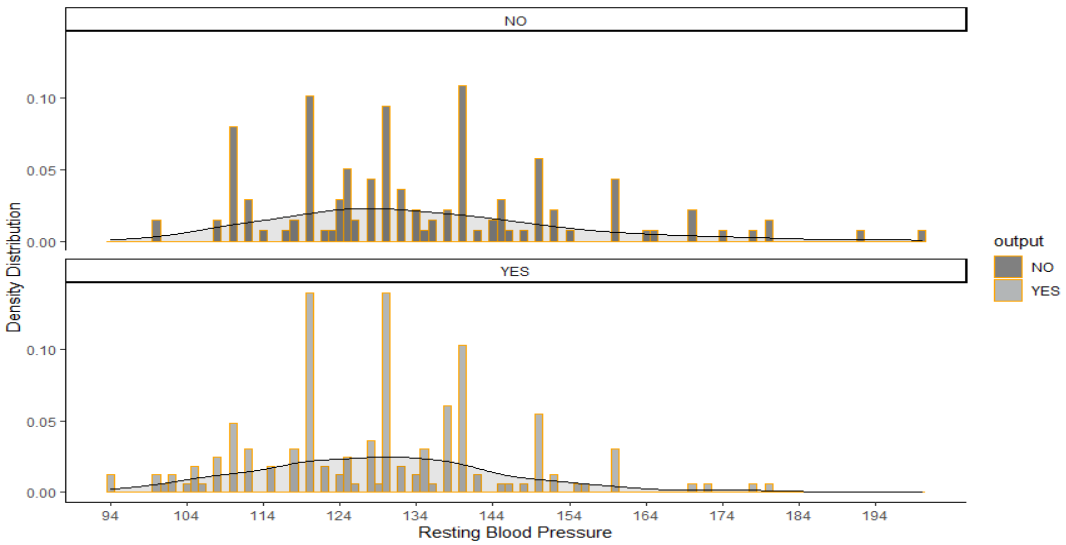


Figure3.11-Resting Blood Pressure- Target Variable Density Distribution

While it is well documented that high blood pressure can harm the arteries that supply the heart over time[46], the distribution of the two subgroups here seems to be extremely similar. It does not reveal any pattern suggesting that the disease-ridden group has elevated blood pressure aside from the emergence of a few outliers.

3.4.6 Cumulative Analysis of Fasting Blood Sugar

According to figure 3.12 the bar-plot shows us that fasting levels of blood glucose do not appear to be associated with heart disease, as there would seem to be a similar number of individuals with both high and low fasting blood sugar levels who have heart disease.

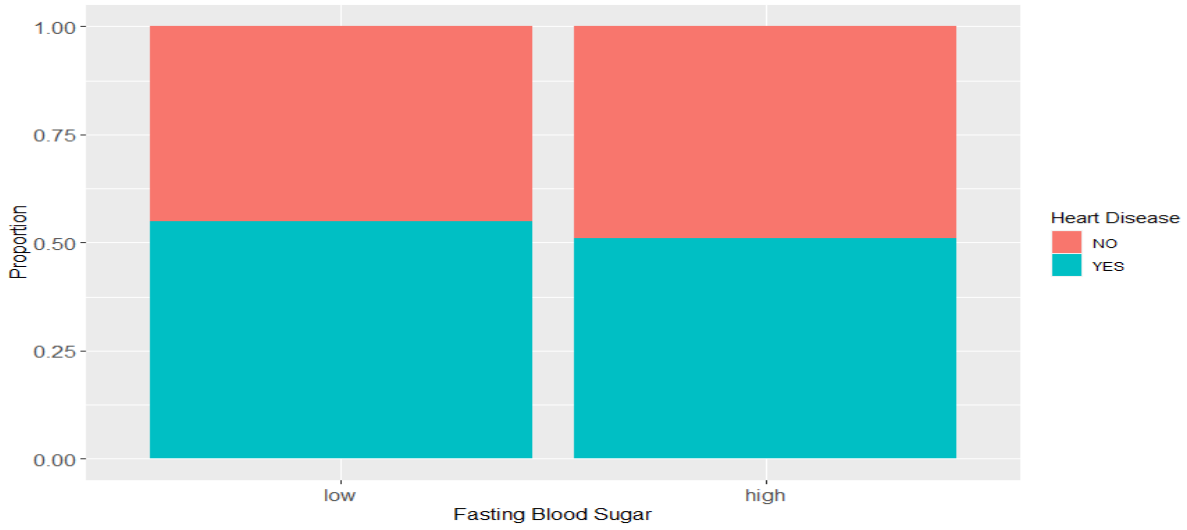


Figure3.12-Fasting Blood Sugar Histogram

3.4.7 Cumulative Analysis of Resting Electrocardiogram Feature

A electrocardiogram test is a simple test that can be used to check the physical and electrical conditioning of the human heart. Typically, each heartbeat is triggered by an electrical signal that starts at the top of your heart and travels to the bottom and when a heart is showing signs of disease, it affects its electrical activity[47].

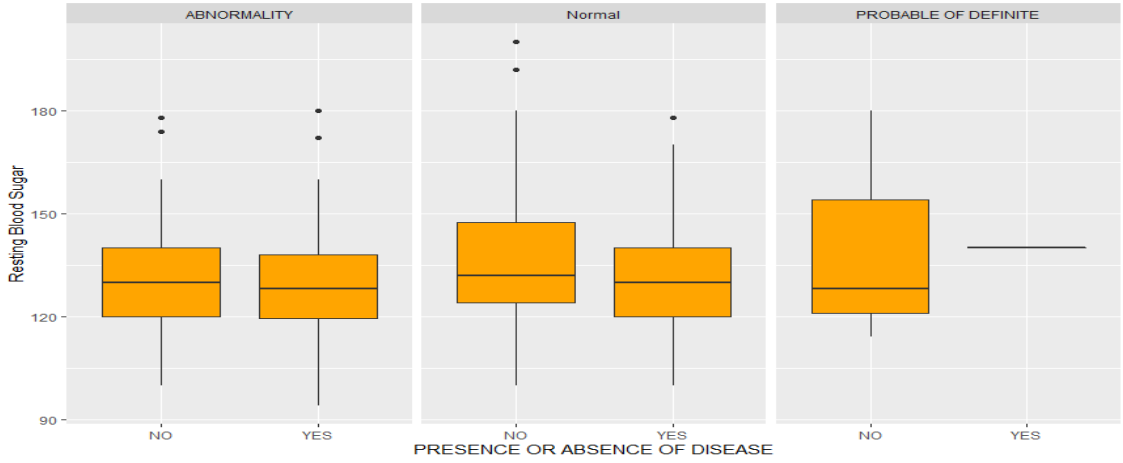


Figure3.13-Resting ECG-Resting BS-Outcome boxplot

The following conclusions can be made from the visualisations presented in Figure 3.13 :

1. From the figure we can observe that majority of the patients that have abnormal ECG and normal ECG show no signs of high resting blood sugar irrespective of the fact that there are a few outliers in both categories.

2. The figure also shows that in the case of LV Hypertrophy(probable of definite where value of ECG = 2), there is only one case where the patient has the disease and is showing that particular type of ECG. The resting blood sugars in these cases are also not particularly shooting up.

3.4.8 Cumulative Analysis of Maximum Heart Rate Feature

Even though according to doctors, elevated heart rate is shown to not be a risk factor when it comes to heart diseases, studies have shown that an increase in heart rate by 10 beats per minute was associated with an increased risk of cardiac death by at least 20%, and this increase shown is similar to the one observed with an increase in heart rate by 10 mm Hg[48], [49]. When it comes to this dataset, we can show the relationship between the two attributes “thalach” and “output” with the help of a boxplot since one of the variables involved is categorical and one is continuous. The figure that shows the visualisation is shown below along with our unbiased takeaways regarding the same.

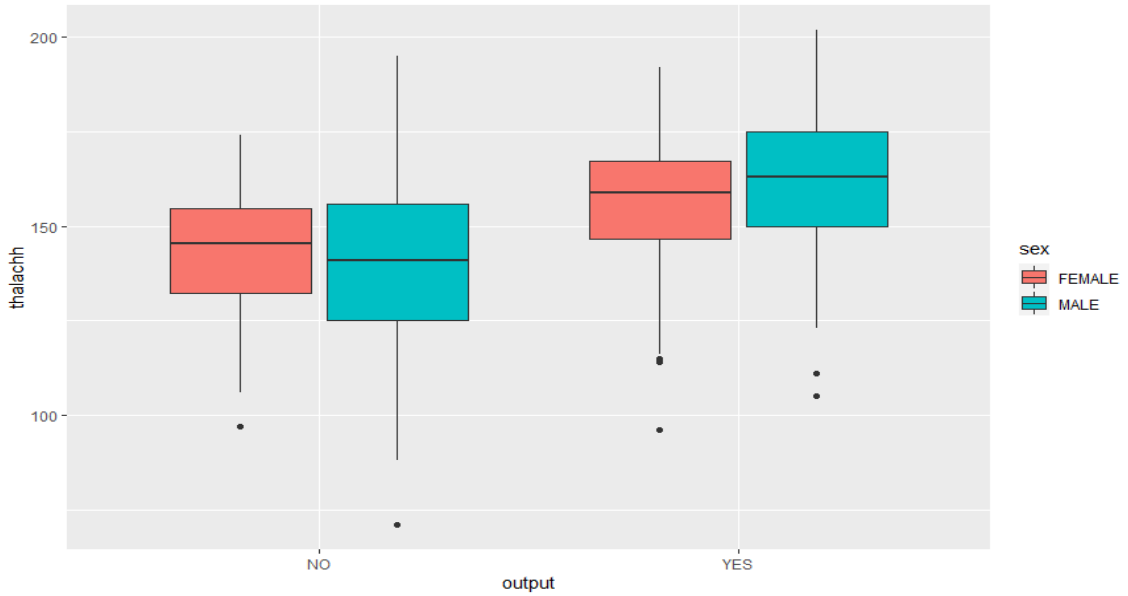


Figure3.14-Maximum Heart Rate – Output Distribution

From the figure 3.14 visualisation, we can infer the following points:

1. We can see that the maximum heart rate achieved in the case of female and male patients that are not diagnosed with heart diseases are lower than 150 which is fairly normal even though there are a few outliers present in both categories where the heart rate achieved is less than 100 which would probably mean that they are in good shape.

2. In the case of females and males suffering from cardiovascular diseases the maximum heart rate achieved is closer to 165 which is fairly higher than the other case. Similarly there are a few outliers present here as well which show fairly lower maximum heart rates.

There are multiple age-predicted maximal heart rate (APMHR) equations that can be used for researching the average maximum heart rate. HRmax = 220 - Age is the most commonly used APMHR equation . Hence, we keep this equation in mind as principle for the analysis of our data. As we can see from figure 3.15, we see a pattern emerging that shows that as we get older ,the maximum heart rate starts to decrease. Although the stated standard deviation for this equation is 10–12 bpm and its prediction accuracy is constrained, "it is nonetheless utilised in clinical settings and published in resources by well-established organisations in the area." Therefore, when analysing our data, we will use this equation as a guide.[50]

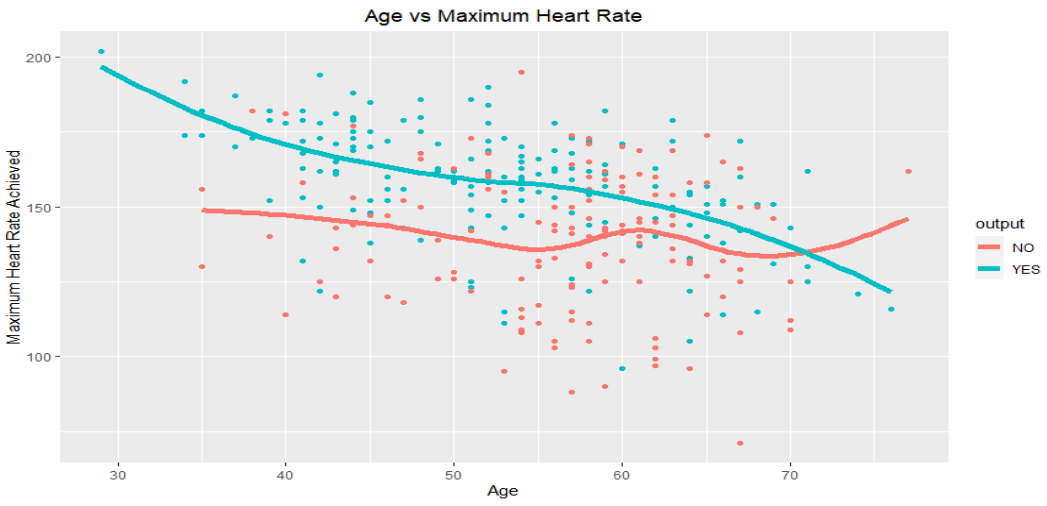


Figure3.15-Age vs Maximum Heart Rate

3.4.9 Cumulative Analysis of Calcified Major Vessels

Vascular calcifications are mineral deposits on the walls of your arteries and veins. The calcification of arteries makes it difficult for the arteries to expand and contract which puts the patient in a vulnerable situation in which they are prone to cardiovascular issues. Since, the contraction and expansion of the vessels is difficult, less blood reaches to the heart muscles. Calcium deposits in your arteries have little to do with your nutritional plan or any other external supplements that you are ingesting. They generally happen because the blood cells in your arteries are not working properly and can hence be a leading cause for heart diseases or simply ageing. The figure 3.16 shows the stacked bar-graph that tells you the relationship between calcified vessels and the output variable. According to the dataset majority of the patients that have heart diseases have zero calcified vessels.[51]

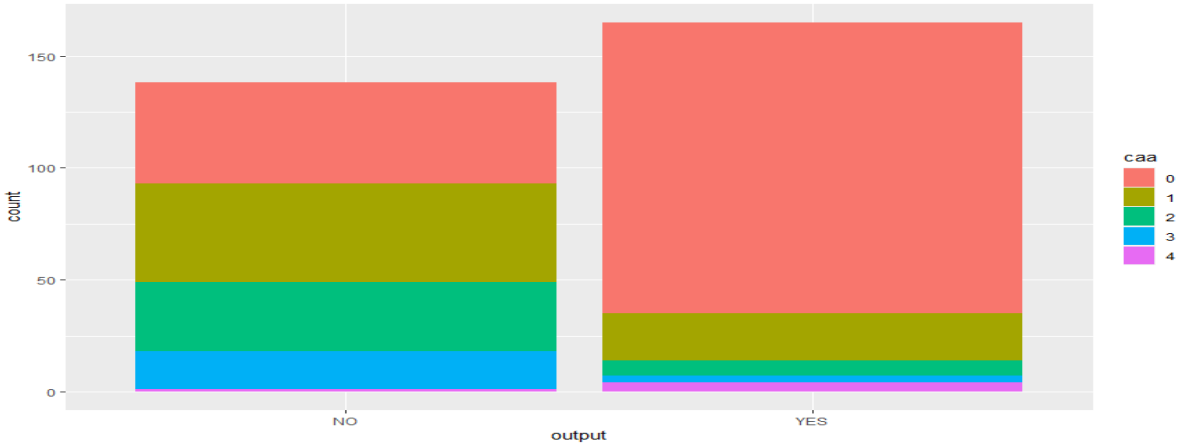


Figure3.16-Calcified Arteries vs Output Variable

3.4.10 Cumulative Data Analysis of Blood Thalassemia

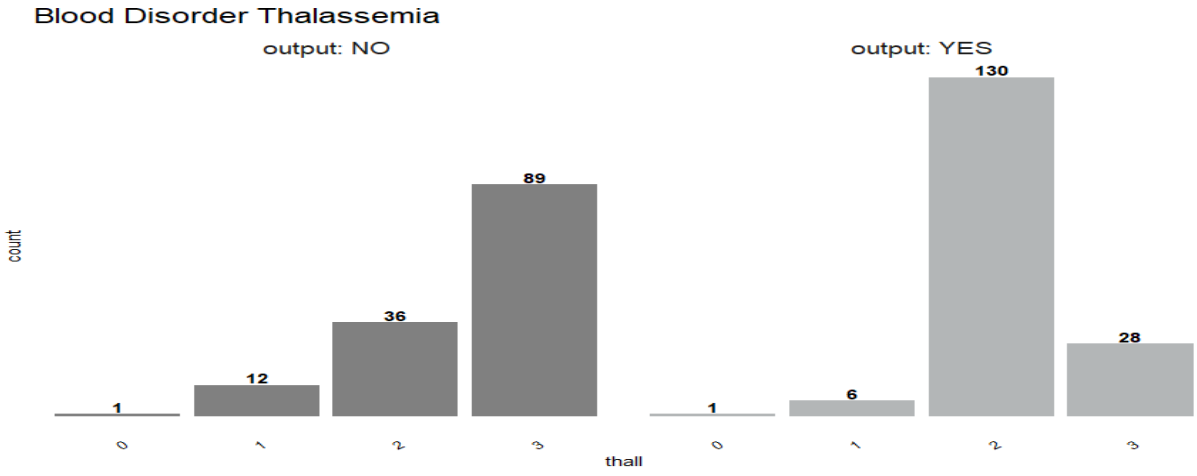


Figure3.17-Blood Thalassemia Type Vs Target Variable Bar Plot

Figure 3.17 shows that majority of the patients that suffer from heart diseases also suffer from a type of Blood disorder Thalassemia which would mean that it is fairly highly indicative of a potential cardiovascular disease. It can also be noted that most of the patients that suffer from cardiovascular diseases have a type-2 fixed defect blood thalassemia disorder.

3.4.11 Important Observations of the Cumulative Data Analysis

This sub section will help in summarising the most significant insights that we have extracted from the data using exploratory data analysis that will help us in classification eventually :

- Despite it being touted as a risk factor when it comes to cardiovascular diseases by various publications[43] , serum cholesterol surprisingly seems to not be a good indicator within this dataset for prediction and classification purposes because of its insignificant p-value.
- Age does seem to have a relationship with heart disease. The presence of disease is skewed to the left, whereas in the case of its absence we can infer that it is much more normally distributed.
- There also appears to be a link between chest pain types and heart attacks. It appears that there are more people that don't have the disease suffering from asymptomatic chest pains than the people that have the disease. Chest pains are one of the most common symptoms when it comes to heart-attacks in general[52] so this result is suspect and warrants further investigation. Seeing as heart disease is a silent disease, it is critical to conduct regular screenings.
- It is easier for younger patients to reach higher maximum heart rates when compared to older patients for whom it is much more difficult to achieve the same ratings. This would mean that younger patients are much more vulnerable and hence, have a higher probability when it comes to developing a heart disease in the future.
- Fasting Blood Sugar seems to be fairly insignificant as it is observed to have

negative correlation with the output variable as seen in the exploratory data analysis.

- Females are bound to have a lower risk of contracting a cardiovascular disease.
- Thalassemia Blood Disorder also looks to be a fair indicator for the presence and absence of the heart diseases.

3.5 Data Pre-Processing

Data pre-processing[35] is also called as data cleaning. It is one of the most important steps in acquiring the most out of dataset. This is a method that eliminates data inconsistencies such as missing values, out of range values, unformatted data, and noise. The process is usually time-consuming because it involves a lot of trial and error. Our pre-processing includes missing value handling[53], correlation analysis[54], data standardization and normalisation[36] and techniques for feature extraction and selection for feature reduction[37](figure 3.18).

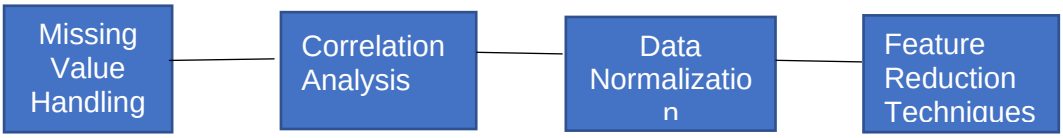


Figure3.18-Data Pre-Processing Network Flow

3.5.1 Missing value Handling

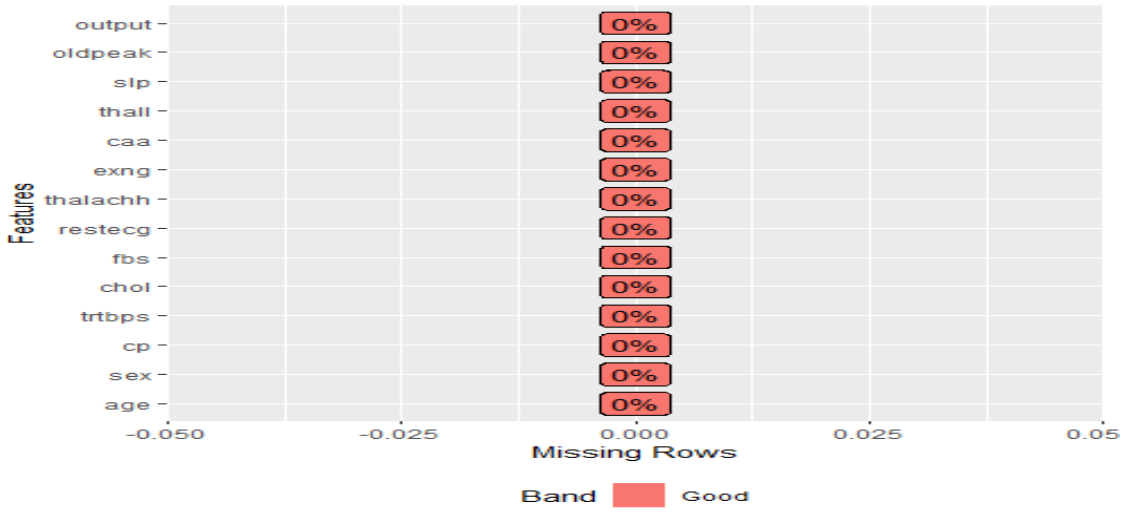


Figure3.19-Missing Data Percentage Plot

In real-world data, a specific element may be missing for a variety of reasons, including corrupt data, failure to load the information, or incomplete extraction. Handling missing values is one of the most difficult challenges that analysts face, because making the right decision on how to handle it results in robust data models[53]. The Kaggle dataset that we are using for our research has been thoroughly checked for missing and inconsistent values and we have seen zero null values. The only change we have done to the dataset is internal data conversion that will be helpful in data visualisation. We have changed the boolean inputs of the categorical variables such as sex,fbs,exng,cp,restecg to their analogous information that is given in the domain section of table 3.1. The tabular data after pre-processing changes can be seen in the figure 3.20

Original Dataset Source	**Data Points**	**Data Attributes (Columns)**	**Null Sets**	**Total Features**
Kaggle	303 observations	14	Zero	14

Table3.20-Missing Value Table

	output	sex	fbs	exng	cp	restecg	slp	caa	thall	age	trtbps	chol	thalachh	oldpeak
1	YES	MALE	>120	NO	ASYMPTOTIC	Normal	0	0	1	63	145	233	150	2.3
2	YES	MALE	<=120	NO	NON-ANGINAL PAIN	ABNORMALITY	0	0	2	37	130	250	187	3.5
3	YES	FEMALE	<=120	NO	ATYPICAL ANGINA	Normal	2	0	2	41	130	204	172	1.4
4	YES	MALE	<=120	NO	ATYPICAL ANGINA	ABNORMALITY	2	0	2	56	120	236	178	0.8
5	YES	FEMALE	<=120	YES	ASYMPTOTIC	ABNORMALITY	2	0	2	57	120	354	163	0.6
6	YES	MALE	<=120	NO	ASYMPTOTIC	ABNORMALITY	1	0	1	57	140	192	148	0.4
7	YES	FEMALE	<=120	NO	ATYPICAL ANGINA	Normal	1	0	2	56	140	294	153	1.3
8	YES	MALE	<=120	NO	ATYPICAL ANGINA	ABNORMALITY	2	0	3	44	120	263	173	0.0
9	YES	MALE	>120	NO	NON-ANGINAL PAIN	ABNORMALITY	2	0	3	52	172	199	162	0.5

Figure3.20-Tabular Dataset after Pre-Processing

3.5.2 Correlation Analysis

The quantitative approach of correlation analysis assesses the extent of a link between the two numerically measured continuous variables (e.g. height and weight). When a

researcher wishes to determine whether there may be connections between variables, this particular form of analysis is helpful. Correlation analysis evaluates the correlation coefficient, which tells you how much one variable changes when the other does.

The correlation coefficient is a measure that determines how tightly two variables' movements have been associated. The Pearson product-moment correlation, which creates the most common correlation coefficient, is used to measure the correlation between two variables. This correlation coefficient, however, may not always be an appropriate option of dependence in a non-linear relationship. The range of the correlation coefficient lies between 1 and -1. A correlation of 1.0 represents a perfect positive correlation , -1.0 represents a perfect negative correlation and a correlation of 0 means that the two attributes have no correlation between them[54].

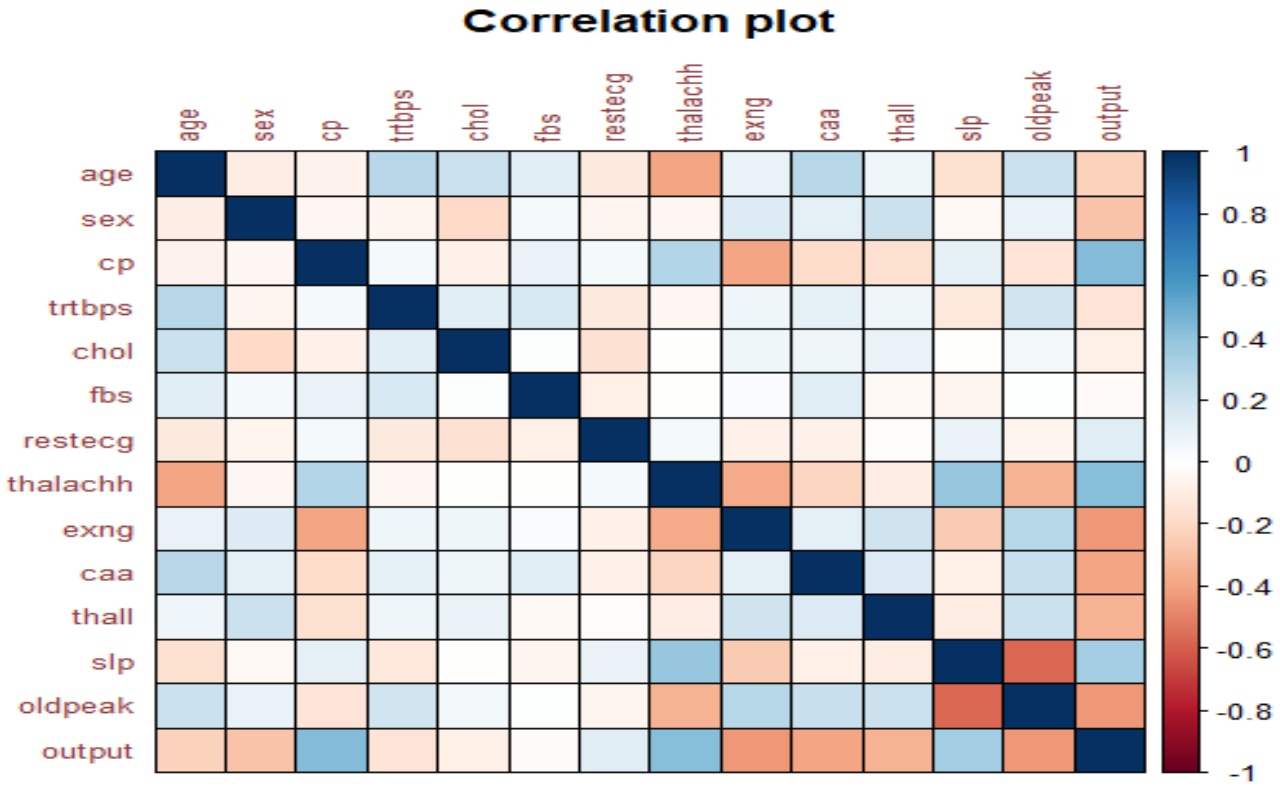


Figure3.21 Correlation Plot of the dataset

From the figure 3.21, we can make the following conclusions about the dataset:

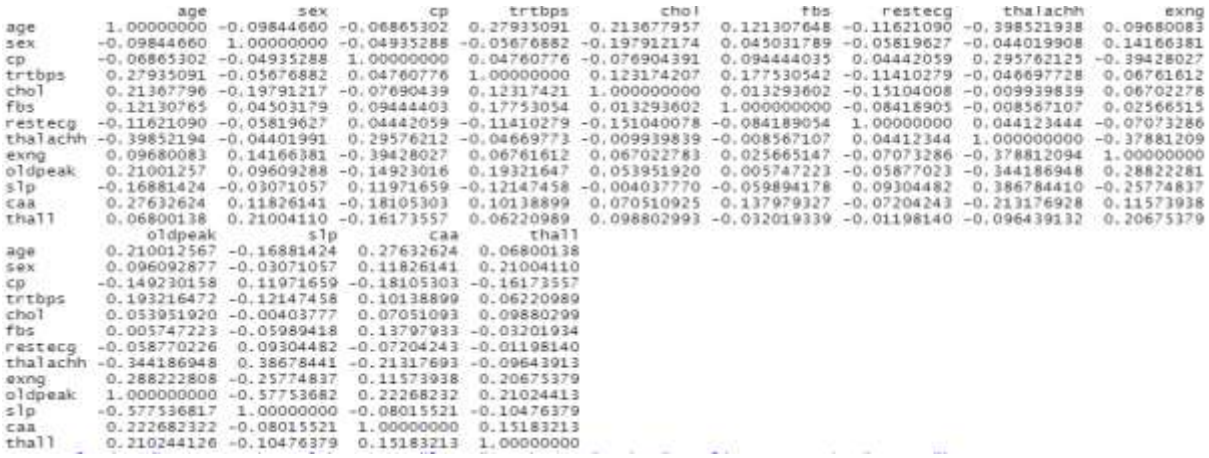
- Age is positively correlated to cholesterol, old peak , blood pressure , caa(major vessels), exercise angina, blood glucose and negatively correlated to maximum heart rate, slope, rest-ecg, chest pain. According to this correlation plot , the chances of getting a heart disease is also negatively correlated to age. Women are

also more likely to have a heart problem than men(based on ratio).

- Chest pain is the most positively correlated attribute to the output variable which would mean it has the highest chance of being the risk factor for the occurrence of the disease. It is also very positively correlated to maximum heart rate which would mean that maximum heart rate could easily lead to chest pains.
- Cholesterol has a positive correlation with age and no correlation with slope.
- Old-peak (ST depression induced by exercise relative to rest) is highly negatively correlated to the chances of having a heart attack , slope and maximum heart rate.
- People with the presence of exercise induced Angina are less likely to have a heart attack. It is also positively correlated to thall and old-peak.
- Maximum heart rate has a high positive correlation with slope and output variable which would mean the higher the heart rate the more chances of getting a heart attack which would mean younger people are more prone to heart attacks.
- People with lower number of blood vessels have more chance of having an Heart attack. Blood vessels are also positively correlated to age which would mean that older people have more blood vessels.
- Slope has a positive correlation with heart attack and maximum heart rate.

3.5.3 Data Scaling and Normalisation

Scaling is a data normalisation technique used to compare data attributes that are not measured in the same way. The dataset is scaled using mean and standard deviation and is very frequently applied to vectors and columns in a particular data frame. Scaling in particular is very important for feature reduction techniques like Principal Component Analysis (PCA)[55] because of the way the principal components are computed. In this research in particular, we have 14 attributes that are a combination of numeric and categorical variables with different ranges. The scaled values of the dataset can be seen in figure 3.22



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng
age	1.00000000	-0.09844660	-0.06865302	0.27935091	0.213677957	0.121307648	-0.11621090	-0.398521938	0.09680083
sex	-0.09844660	1.00000000	-0.04935288	-0.05676882	-0.197912174	0.045031789	-0.05819627	-0.044019908	0.14166381
cp	-0.06865302	-0.04935288	1.00000000	0.04760776	-0.076904391	0.094444035	0.04442059	0.295762125	-0.39428027
trtbps	0.27935091	-0.05676882	0.04760776	1.00000000	0.123174207	0.177530542	-0.11410279	-0.046697728	0.06761612
chol	0.21367796	-0.19791217	-0.07690439	0.12317421	1.000000000	0.013293602	-0.15104008	-0.009939839	0.06702278
fbs	0.12130765	0.04503179	0.09444403	0.17753054	0.013293602	1.000000000	-0.08418905	-0.008567107	0.02566515
restecg	-0.11621090	-0.05819627	0.04442059	-0.11410279	-0.151040078	-0.084189054	1.00000000	0.044123444	-0.07073286
thalachh	-0.39852194	-0.04401991	0.29576212	-0.04669773	-0.009939839	-0.008567107	0.04412344	1.000000000	-0.37881209
exng	0.09680083	0.14166381	-0.39428027	0.06761612	0.067022783	0.025665147	-0.07073286	-0.378812094	1.00000000
oldpeak	0.21001257	0.09609288	-0.14923016	0.19321647	0.053951920	0.005747223	-0.05877023	-0.344186948	0.28822281
slp	-0.16881424	-0.03071057	0.11971659	-0.12147458	-0.004037770	-0.059894178	0.09304482	0.386784410	-0.25774837
caa	0.27632624	0.11826141	-0.18105303	0.10138899	0.070510925	0.137979327	-0.07204243	-0.213176928	0.11573938
thall	0.06800138	0.21004110	-0.16173557	0.06220989	0.098802993	-0.032019339	-0.01198140	-0.096439132	0.20675379

	oldpeak	slp	caa	thall
age	0.210012567	-0.16881424	0.27632624	0.06800138
sex	0.096092877	-0.03071057	0.11826141	0.21004110
cp	-0.149230158	0.11971659	-0.18105303	-0.16173557
trtbps	0.193216472	-0.12147458	0.10138899	0.06220989
chol	0.053951920	-0.00403777	0.07051093	0.09880299
fbs	0.005747223	-0.05989418	0.13797933	-0.03201934
restecg	-0.058770226	0.09304482	-0.07204243	-0.01198140
thalachh	-0.344186948	0.38678441	-0.21317693	-0.09643913
exng	0.288222808	-0.25774837	0.11573938	0.20675379
oldpeak	1.000000000	-0.57753682	0.22268232	0.21024413
slp	-0.577536817	1.00000000	-0.08015521	-0.10476379
caa	0.222682322	-0.08015521	1.00000000	0.15183213
thall	0.210244126	-0.10476379	0.15183213	1.00000000

Figure3.22-Scaling of the data frame

3.5.4 Feature Reduction Techniques

Feature or dimensionality reduction techniques[37] is the subsequent transformation of a dataset from a high dimensional space to a lower dimensional space in such a way that the lower dimensional space retains the meaningful properties of the original data. Working in high-dimensional spaces can be inconvenient for a variety of reasons; raw data is recurrently sparse as a result of the curse of dimensionality, and analysing the data is typically computationally intractable (hard to control or deal with). Dimensionality reduction is common in fields such as signal processing, speech recognition, neuro-informatics, and computational biology that deal with large numbers of observations and/or variables. Feature reduction approaches can be subsequently divided into feature reduction and feature extraction techniques that help us in facilitating other analysis. In this research project, we have made an extensive use of Principal Component Analysis(PCA)[55], Boruta Feature Selection Algorithms for feature selection and extraction.

3.5.4.1 Principal Component Analysis(PCA)

The principal component analysis is a simple method for reducing the dimensions of a given dataset without having to lose most of the information stored within. Principal component analysis, the main linear technique for dimensionality reduction, maps the data linearly in a

lower dimensional space in such a way that the variance is maximised. The covariance and eigen vectors are computed for the calculation of the principal components. The largest eigen values that also are known as the principal components hence can be used to construct the majority of the data variance in a lower dimensional space.

After calculation of the eigen values, it is important to estimate the right amount of principal components that should be chosen. Even though there is no unique way to comment about the estimation of the principal components, the widely used method is elbow method in which we examine the scree plot(figure 3.23). We look for the principal component after which the variance subsequently starts to drop off.[55]

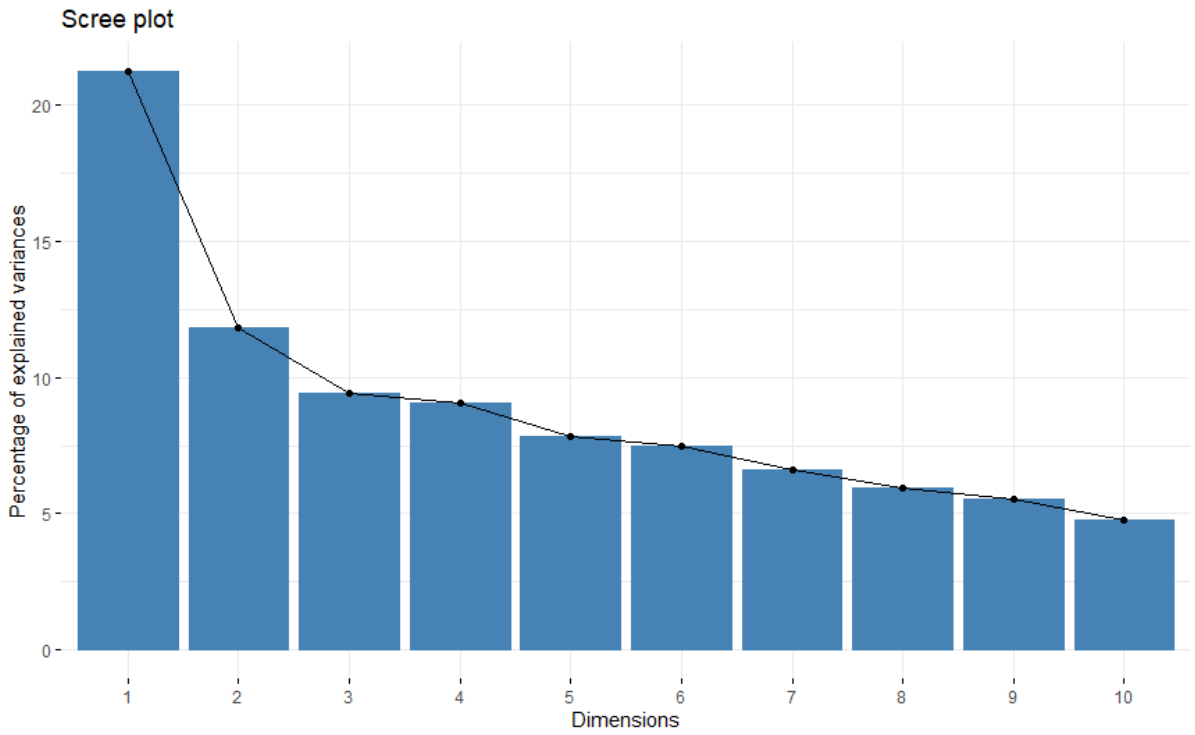


Figure3.23 Scree Plot of Principal Components

In this case we can see that 8 principal components explain 80% of the variation of the data even though there are a lot of eigen vectors that have less than 10 percent of cumulative variance. The figures 3.24 and 3.25 show the description and contribution of the components involved.

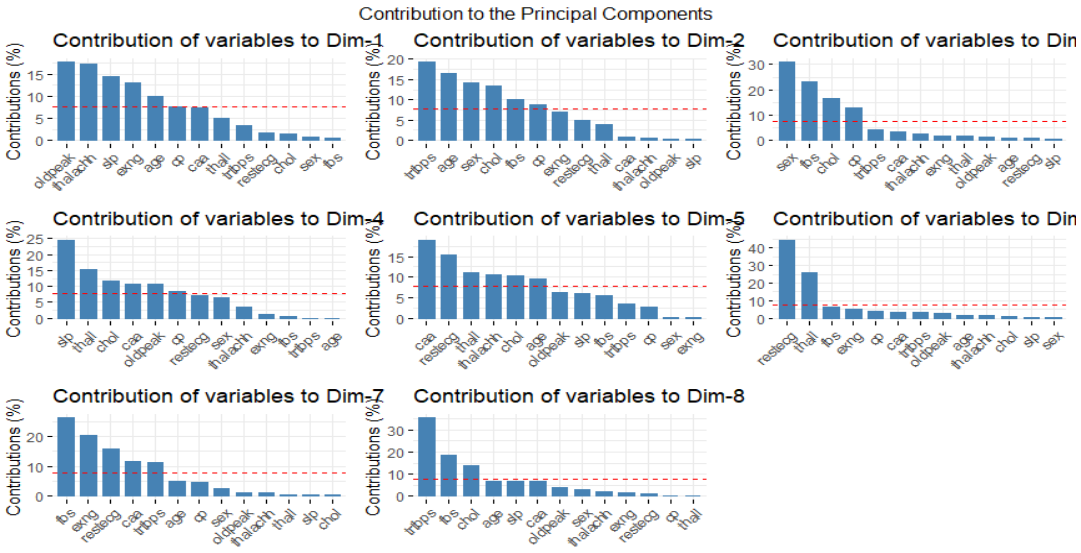


Figure3.24-Contribution of the variables to the 8 chosen principal components

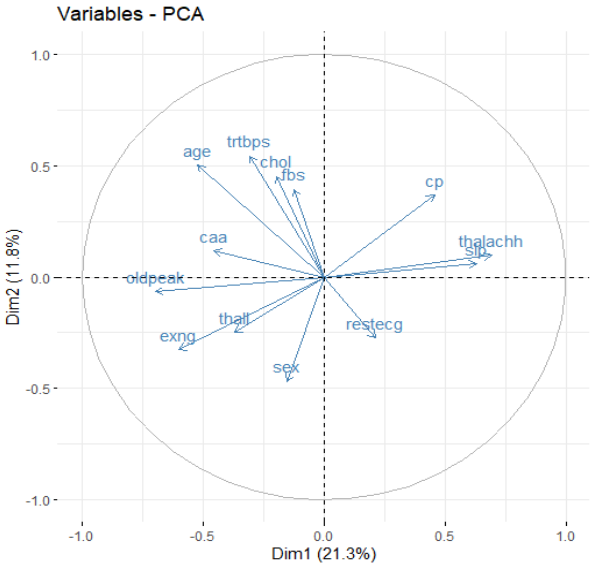


Figure3.25-Description of the generated Components (PC1 vs PC2)

3.5.4.2 Boruta Feature Selection Technique

Boruta Feature Selection Technique is a wrapper algorithm closely associated with the Random Forest Classifier. Its optimal usage can be seen in a case when we have a Classification problem and the dataset that we are working upon has a lot of attributes that make the model complicated and make it difficult for the model accuracy to reach its maximum potential. Boruta employs an all-relevant feature selection algorithm, which

captures all features that are relevant to the outcome variable in some circumstances. In contrast, most feature extraction algorithms use a minimal optimal method in which they rely on a small subset of features to produce the lowest error on a chosen classifier[56].

The figure 3.26 shows the Boruta Plot that tell us about the status of the attributes after the algorithm is fully iterated. The blue boxplots correspond to the minimum, maximum and average Z scores of the shadow attributes. The green, yellow and red boxplots signify the acceptance, holding and rejection of the attributes on the basis of the Z scores. As we can see, we have one attribute(trtbps) that has been put in the tentative category. A decision can be taken for that particular attribute by comparing its median Z score with that of the best shadow attribute's median Z score[56]. The figure 3.27 shows us that the feature that was put in the tentative category before has been fully accepted. So our final analysis leaves us with 3 attributes that are rejected which are cholesterol, resting electrocardiogram and Fasting Blood Sugar.

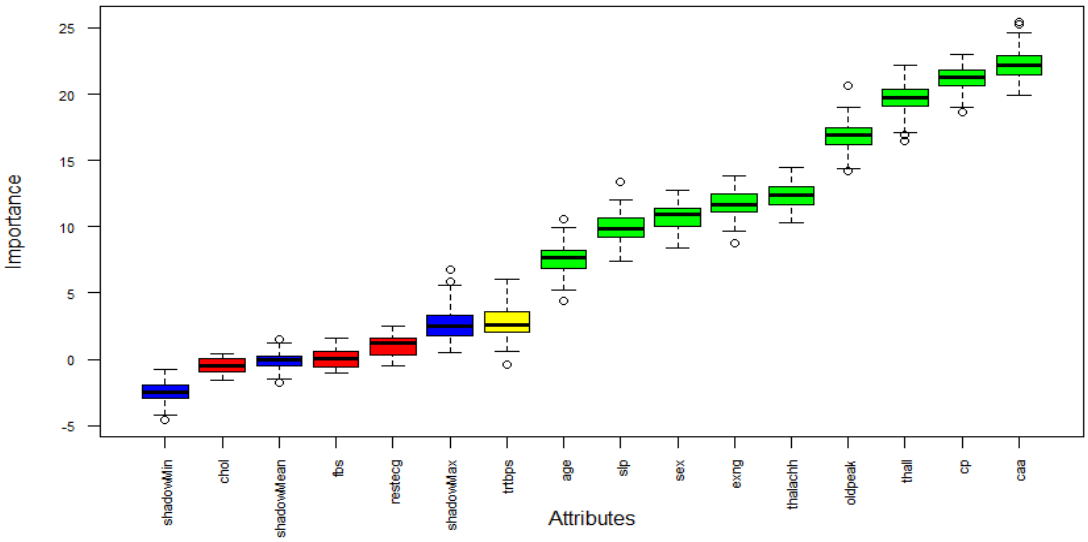


Figure3.26-Boruta Plot Shows TrtBps Attribute in the Tentative Category Before Tentative Rough Fixing Process

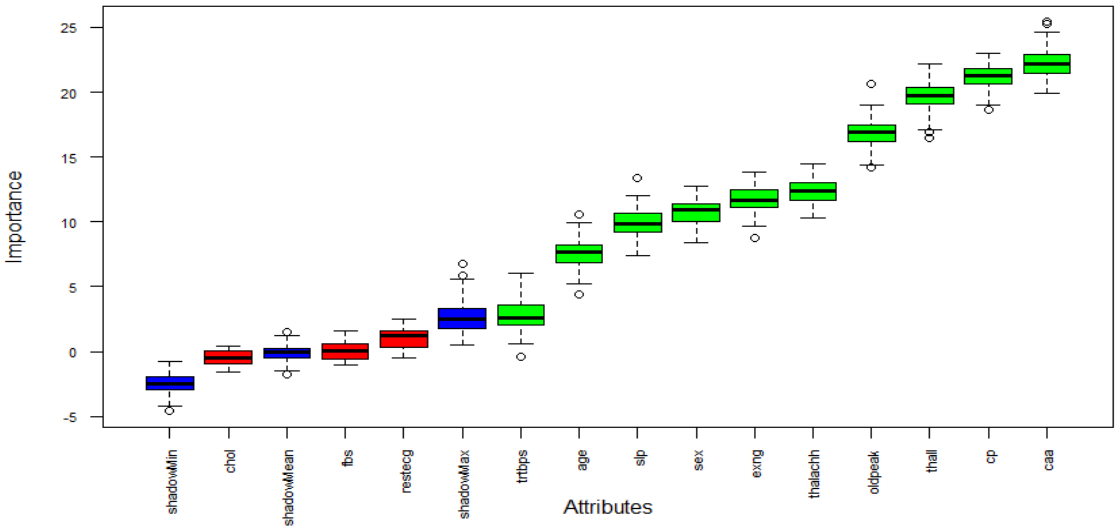


Figure3.27-Boruta Plot After Tentative Rough Fixing

3.6 Summary of the Feature Importance Techniques

- After conducting Principal Component Analysis, we found that 8 principal components of the data explain about 80% of the variation despite the fact that many eigen vectors have less than 10% cumulative variance.
- The Boruta Feature Selection Technique was used as a feature reduction algorithm that helps us in building predictive models that are free from correlated variables and noise in general. It was seen that cholesterol, resting electrocardiogram and Fasting Blood Sugar were the three attributes that were rejected. We also have evidence to support that Cholesterol surprisingly had no correlation with the likelihood of having heart-disease when we conducted a hypothesis test to check whether the two averages are different by checking the statistical significance using a t-test (figure 3.9).
- Hence the variables that are not significant to the dataset are resting Electro-cardiogram, fasting blood-sugar, cholesterol.

Chapter 4 : Results, Analysis and Discussion

4.1 Introduction

This chapter showcases and analyses the results obtained from our chosen statistical Modelling and well-known machine learning techniques i.e. Logistic Regression, Naïve Bayes Classification, Random Forest Classification and Support Vector Machine Classification Algorithms. The aim of this chapter is to capture the underlying relationships between our outcome variable – the dependent variable that tells us about the status of the patient and the independent data features such as Cholesterol, Age, Sex, etc…. and thereby build the best classification algorithms that fully encapsulate the information given to us and give accurate predictions about the health status of the patients. We will begin this section by briefly discussing about the intrinsic performance metrics that will be used to determine the optimal classifier.

4.2 Performance Metrics in Machine Learning

A variety of metrics can be used to evaluate the performance of Machine Learning prediction, classification, and regression algorithms. The metrics that will be used to comment on the performance of the classifier should be chosen with care because the performance of the algorithm is entirely dependent on the metric, and it is equally possible that a particular metric is not suitable for measuring the performance of the classifier. Some of the different metrics that are used for performance evaluation are mentioned below :

- **Confusion Matrix** :

  When it comes to statistical classification in the field of machine learning, we frequently rely on a confusion matrix (or error matrix), that comments about the performance of a supervised machine learning algorithm. The confusion matrix(table 3.3) is made up of two rows and two columns, the first row contains the number of true positives(TP), that indicates the number of positive samples classified accurately and true negatives(TN),which indicates the number of negative samples classified accurately while the false positives(FP), comment about the actual

negative classes that are classed as positive and false negatives(FN), talk about the samples that are actually positive but falsely classed as negative are contained in the second row. The name derives from the fact that it is simple to determine whether the system is confusing two classes. It hence becomes important to construct this matrix as it is directly responsible in the computation of the metrics that are responsible for performance evaluation of the classifier.[57]

	ACTUAL VALUES: POSITIVE	ACTUAL VALUES: NEGATIVE
PREDICTED VALUES: POSITIVE	True Positives(TP)	False Positives(FP)
PREDICTED VALUES: NEGATIVE	False Negatives(FN)	True Negatives(TN)

Table 4.1-Confusion Matrix

- **Classification Accuracy** :

  It is the performance metric that is used to determine how effectively can a classifier measure what it is supposed to measure. Hence, Classification accuracy is a metric that summarises a classification model's performance by dividing the number of correct predictions by the total number of predictions. It is simple to calculate and understand, making it the most commonly used metric for evaluating classifier models. Mathematically, the accuracy of the classifier is given by [57]:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- **Recall or Sensitivity :** The recall metric is defined as the percentage of correctly classified diseased patients (TP) divided by the total number of patients with the disease. It is hence, also called as the True Positive Rate(TPR).The term sensitivity is derived from statistics as a measure of the performance of a binary classification, whereas recall is more closely related to the Information Engineering domain. Mathematically, the sensitivity of the classifier is given by[57] :

$$\frac{TP}{TP + FN}$$

- **Specificity :** The specificity metric of the classifier measures a model's ability to predict true negatives in each available category. It is hence, also called as the True Negative Rate(TNR). For example, the proportion of patients with no heart disease who are correctly identified as having 'No Heart Disease'. Mathematically, the specificity of the classifier is given by[57] :

$$\frac{TN}{FP + TN}$$

- **Precision :** Precision is the indicator of a machine learning model that talks about the quality of the positive prediction made by the classifier. Precision is calculated by dividing the number of true positives by the total number of positive predictions, in which case the mathematical representation is as follows[57] :

$$\frac{TP}{TP + FP}$$

- **F-Score :** The harmonic mean of precision and recall is used to calculate the F1 score. It is used to rate performance statistically. In other words, an F1-score (from 0 to 1, with 0 being the lowest and 1 being the highest) is a mean of a person's performance based on two factors: precision and recall. The formula for the calculation of the f-score[57] :

$$\frac{2*(Precision * Recall)}{Precision + Recall} = \frac{TP}{TP + 0.5\,(FP+FN)}$$

- **AUC-ROC Curves :** The area under the receiver operating characteristic curve, or AUC-ROC curve, is a graphical plot that compares sensitivity and FPR and represents a binary classification algorithm's diagnostic ability. Classifiers that produce curves closer to the top-left corner perform better. To begin graphically plotting the AUC curve, it is necessary to first create a baseline. Any random classifier is expected to produce points plotted diagonally to the sensitivity and specificity curves. A significantly better classifier will deviate from this baseline diagonal plot, making it more accurate. We will use the AUC (Area Under Curve) of the ROC curve to evaluate the performance of our classifier; this is a measure of discrimination or diagnostics. As a result, the higher the AUC, the better the model distinguishes between, say, heart-attack vs non-heart attack patients following an accident. As a result, an excellent classifier has AUC that is closer to 1, whereas a poor classifier has AUC that is closer to 0[58][59].

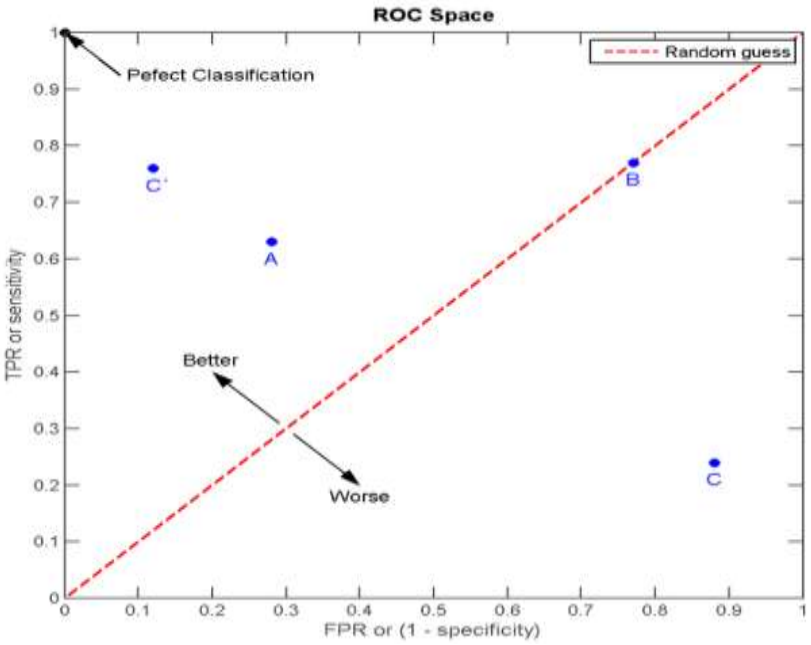


Figure 4.1 ROC Curve[59]

4.3 Analysis and Results – Logistic Regression

This section of the chapter deals with choosing the appropriate model for logistic regression classification followed by the partitioning of the data into training and testing datasets which will be used as an input to perform the classification. The classification will be further metricized using performance statistics to judge its performance.

4.3.1 Choosing the appropriate model for Logistic Regression

After establishing the various performance metrics that will be used to comment about how the algorithms will perform and ensuring that the dataset is properly encoded and structured we will be running a regression analysis on our original model that has 13 features against the output variable to comment about the predictors which are statistically significant with a p value of less than 0.05. The regression output is shown in figure 4.2 and the variables that are significant are highlighted using various significance codes. As we can see predictor variables such as age, cholesterol, fasting blood sugar and resting electrocardiogram have fairly high p-values which would not make them statistically significant. This claim has also been previously been backed by our feature selection technique(Boruta Technique-Figure 3.25 and 3.26) that we used in the pre-processing section of our research. Sex, Chest Pain and Calcified Vessels are the three predictors that have very low p-values and hence have maximum significance. The Akaike Information Criterion (AIC)[60] is 239.44 which even

though is fairly low can be reduced further if we remove the variables that are not statistically significant. The lower the AIC, the better is the model.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.450472   2.571479   1.342 0.179653
age         -0.004908   0.023175  -0.212 0.832266
sex         -1.758181   0.468774  -3.751 0.000176 ***
cp           0.859851   0.185397   4.638 3.52e-06 ***
trtbps      -0.019477   0.010339  -1.884 0.059582 .
chol        -0.004630   0.003782  -1.224 0.220873
fbs          0.034888   0.529465   0.066 0.947464
restecg      0.466282   0.348269   1.339 0.180618
thalachh     0.023211   0.010460   2.219 0.026485 *
exng        -0.979981   0.409784  -2.391 0.016782 *
caa         -0.773349   0.190885  -4.051 5.09e-05 ***
thall       -0.900432   0.290098  -3.104 0.001910 **
slp          0.579288   0.349807   1.656 0.097717 .
oldpeak     -0.540274   0.213849  -2.526 0.011523 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 211.44  on 289  degrees of freedom
AIC: 239.44
```

Figure 4.2 Regression Analysis of Full Dataset

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.423391   1.916185   1.265 0.205980
sex         -1.588807   0.433237  -3.667 0.000245 ***
cp           0.854141   0.180253   4.739 2.15e-06 ***
caa         -0.755279   0.183549  -4.115 3.87e-05 ***
thall       -0.916021   0.279482  -3.278 0.001047 **
exng        -0.947169   0.400644  -2.364 0.018073 *
slp          0.604485   0.341428   1.770 0.076651 .
trtbps      -0.021043   0.009876  -2.131 0.033110 *
oldpeak     -0.531327   0.207396  -2.562 0.010410 *
thalachh     0.022843   0.009320   2.451 0.014251 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 215.77  on 293  degrees of freedom
AIC: 235.77

Number of Fisher Scoring iterations: 6
```

Figure 4.3 Regression Analysis of Processed Dataset

The figure 4.3 illustrates the regression analysis that is done after we have removed all the predictor variables that were not statistically significant. We can see all the variables have a representation in the significance code row even though there are a few predictors (sex ,chest pain ,caa) that are more significantly important than others (slope). The AIC of the model has also reduced to 235.77 which is an optimal figure when you consider the number of features involved and the size of the dataset. Further we decided to partition the dataset

into training and testing dataset to test the classification of the machine . We decided to use 75 % of the dataset to train and the 25 % of the dataset to test to get the confusion matrix along with the performance metrics as an output in figure 4.4

4.3.2 Model Classification Testing

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 24  3
         1 10 39

               Accuracy : 0.8289
                 95% CI : (0.7253, 0.9057)
    No Information Rate : 0.5526
    P-Value [Acc > NIR] : 3.447e-07

                  Kappa : 0.6471

 Mcnemar's Test P-Value : 0.09609

            Sensitivity : 0.7059
            Specificity : 0.9286
         Pos Pred Value : 0.8889
         Neg Pred Value : 0.7959
             Prevalence : 0.4474
         Detection Rate : 0.3158
   Detection Prevalence : 0.3553
      Balanced Accuracy : 0.8172

       'Positive' Class : 0
```

Figure 4.4-Confusion Matrix of Dataset

The order of the confusion matrix has been reversed because it takes the “positive class” to be zero. For simplicity, we can summarize this output in a tabular form as we have done in table 4.2

Total Columns of Data = 76	Actual: No Heart Disease	Actual: Heart Disease
Predicted: No Heart Disease	TN=39	FN=10
Predicted: Heart Disease	FP=3	TP=24

Table 4.2 Tabular Confusion Matrix – Logistic Regression

From the table 4.2, we can also comment about the performance metrics :

- **Accuracy** : This is the performance metric that is used to determine how effectively can a binary classifier measure what it is supposed to measure. In this case accuracy is calculated by (**T P + T N) / Total** = (**24 + 39) / 76 = 0.8289** i.e. the classifier is accurate 82.89% of the time.

- **Precision** : Precision is calculated by dividing the number of true positives by the total number of positive predictions. In this case precision can be calculated by **T P / (T P + F P) = 24 / (24 + 3) = 0.888** i.e. the classifier is precise 88.8 percent of the time.

- **Recall** : The recall metric is defined as the percentage of correctly classified diseased patients (TP) divided by the total number of patients with the disease. It is calculated by **T P /(T P + F N) = 24 / 34 = 0.705** i.e. the recall of the classifier is 70.5 %.

- **Specificity** : The specificity metric of the classifier measures a model's ability to predict true negatives(TN) in each available category. It is calculated by **T N /(T N + F P) = 39 / 42 = 0.9286** i.e. the specificity of the classifier is 92.86 %

- **F-score** : The F score is calculated by (**T P) / (T P + 0.5 (F P+F N)) = 24/24 +0.5(13)=0.78**

- **AUC Score** : After plotting the ROC Curve for the testing dataset and computing the Area Under Curve Score for this classifier to comment about its diagnostic abilities, we found that the AUC Score of this classifier is **0.89**. Higher the AUC of the curve, the farther it is from the diagnostic line that divides the ROC Plane into two halves. According to a fairly rough classifying system an AUC Score between 0.8-0.9 is considered to be very good. The figure 4.5 shows the ROC curve having an AUC score of 0.89.

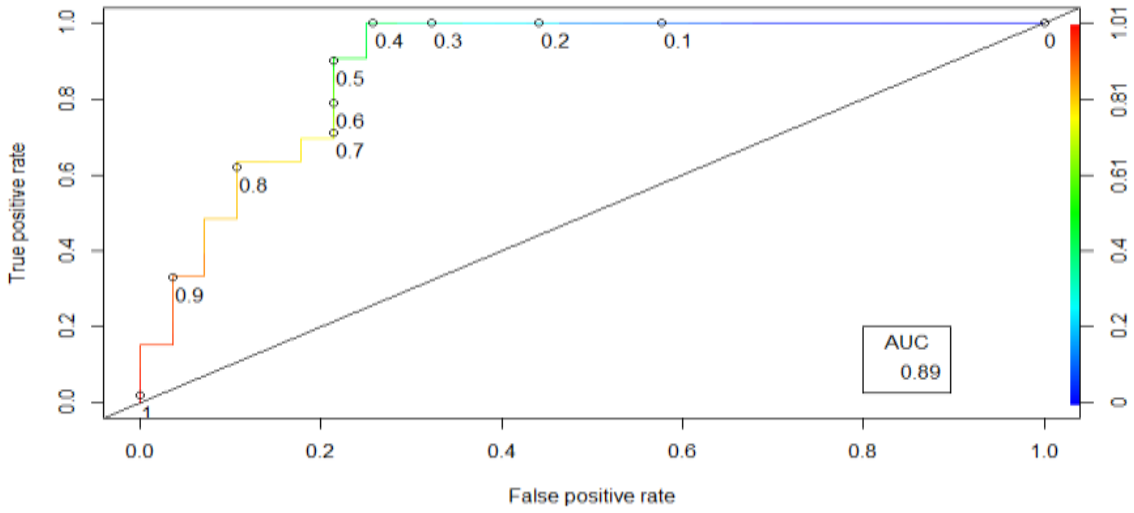


Figure 4.5-AUC Curve.

4.3.3 Tabular Summary

The table 4.3 summarises the performance metrics that we have been computed for this classification technique

Name of the classifier	Accuracy	Sensitivity	Specificity	F-Score	Precision	Area Under Curve
Logistic Regression	82.89%	70.5%	92.86%	0.78	88.8%	0.89

Table 4.3-Model Testing Summary for The Logistic Regression Classifier

4.4 Analysis and Results – Random Forest Classification

This section of the chapter deals with choosing the appropriate model for Random Forest Classification. We will be using Random Forest's feature importance algorithm followed by the partitioning of the data into training and testing datasets which will be used as an input to perform the classification. The classification will be summarised using performance statistics to judge its performance.

4.4.1 Feature Importance Plot

This is an underlying random forest result that shows how important each variable is in the data classification process. The Mean Decrease Accuracy plot shows how much correctness is lost when each variable is removed from the model. The greater the accuracy loss, the more important the variable is for successful classification. The variables are listed in order of decreasing importance. The Gini index's average decreasing trend is an indicator of how each variable relates to the uniformity of the random forest's nodes and leaves. The greater the mean decrease accuracy or Gini score, the greater the importance of the variable in the model[61].

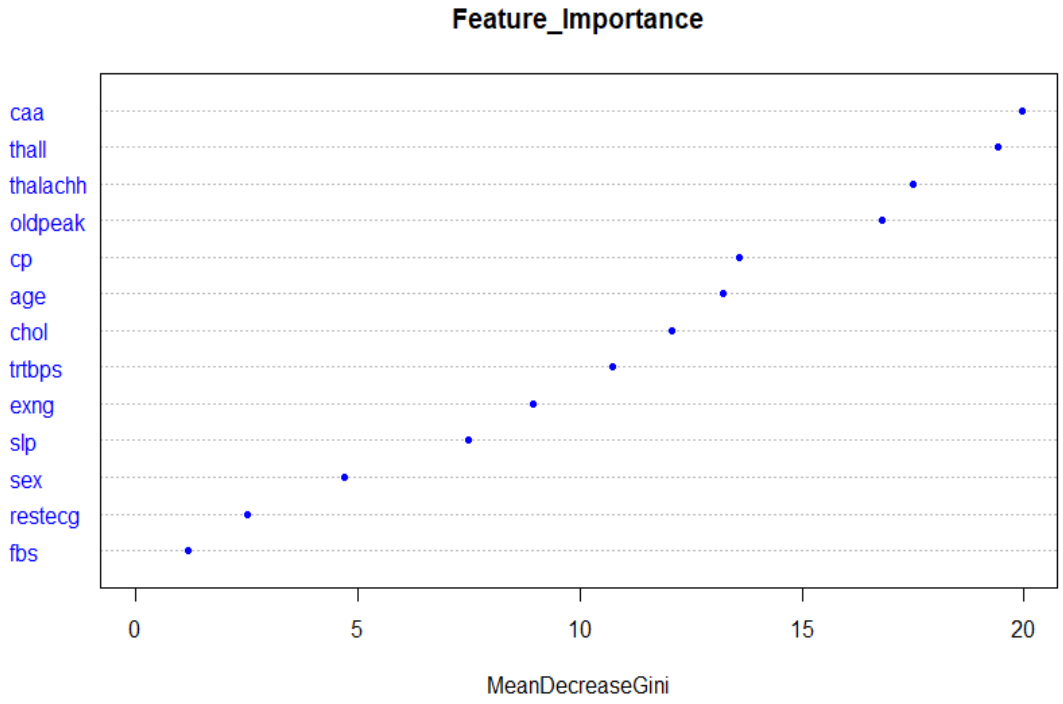


Figure4.6-Feature Importance Plot – Random Forest

The feature importance plot in figure 4.6 shows that the attributes rest-ecg, sex and fbs all have very low Gini Indexes which means that if they are removed from the model the correctness of the model does not deviate by a lot[61], indicating that they are the least significant variables in our dataset. They may be considered redundant, and we may remove them from the dataset prior to the classification stage.

4.4.2 Model Classification Testing

```
p2    0  1
  0 31  9
  1  4 32
> conf.mat<-table(p2,test$output)
> accuracy.val <- round(sum(diag(conf.mat))/sum(conf.mat),2)
> accuracy.val
[1] 0.83
>
> #calculating Precision
> precision.val <- conf.mat[4] / sum(conf.mat[4], conf.mat[2])
> precision.val
[1] 0.8888889
>
> #Calculating Sensitivity
> sensitivity.val <- conf.mat[4] / sum(conf.mat[4], conf.mat[3])
> sensitivity.val
[1] 0.7804878
>
> #calculating specificity
> specificity.val <- conf.mat[1] / sum(conf.mat[1], conf.mat[2])
> specificity.val
[1] 0.8857143
```

Figure4.7-Confusion Matrix

To generate the confusion matrix, we decided to use 75% of the dataset for training and 25% for testing. For simplicity, we can summarise the confusion matrix in a tabular format

Total Columns of Data = 76	Actual: No Heart Disease	Actual: Heart Disease
Predicted: No Heart Disease	TN=31	FN=9
Predicted: Heart Disease	FP=4	TP=32

Table 4.4 Tabular Confusion Matrix – Random Forest

From the table 4.2 we can comment about the performance metrics as follows :

- Accuracy : In this case the accuracy is calculated by **T P + T N / Total** = **32 + 31 / 76 = 0.83** i.e. the classifier is accurate 83% of the time.

- Precision : In this case precision can be calculated by **T P / T P + F P = 32 / 32 +4 = 0.89** i.e. the classifier is precise 89 percent of the time.

- Recall : The recall metric is calculated by **T P / T P + F N = 32 / 41 = 0.78** i.e. the sensitivity of the classifier is 78 %.

- Specificity : The specificity metric of the classifier is calculated by **T N / T N + F P = 31 / 35 = 0.885** i.e. the specificity of the classifier is 88.5 %

- F-score : The F score is calculated by **T P / T P + 0.5 (F P+F N) = 32/32 +0.5(13)=0.83**

- AUC Score : We discovered that the AUC Score of this classifier is 0.83 after plotting the ROC Curve for the testing dataset and computing the Area Under Curve Score for this classifier to comment on its diagnostic abilities. The greater the AUC of the curve, the greater the distance from the diagnostic line that divides the ROC Plane into two halves. An AUC Score of 0.8-0.9 is considered very good by a fairly crude classification system. Figure 4.7 depicts the ROC curve with its AUC .

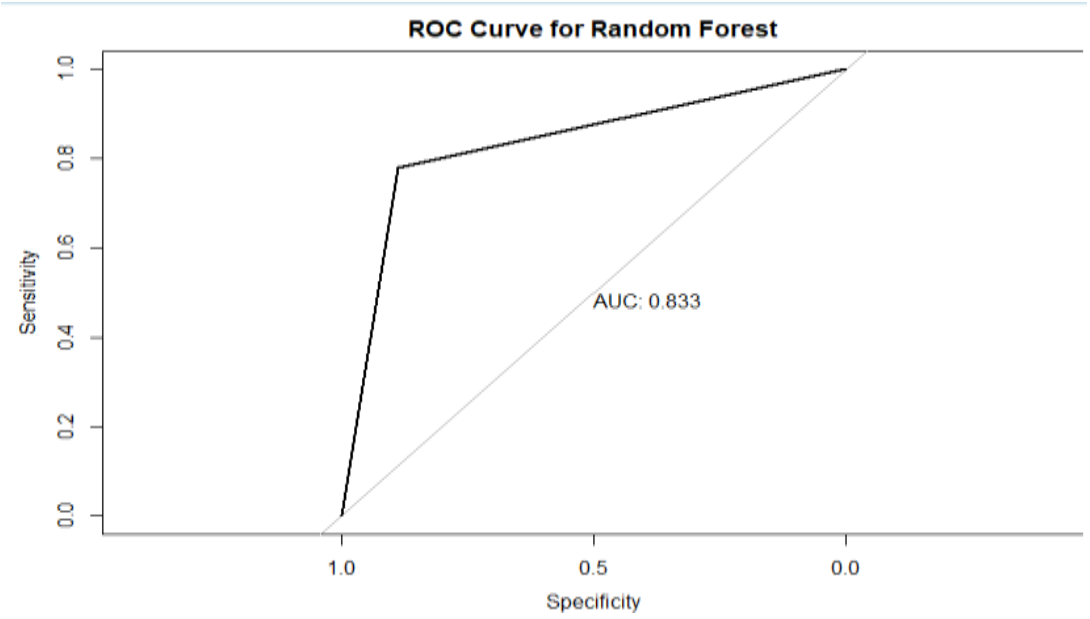


Figure 4.8 AUC-ROC Curve

4.4.3 Tabular Summary

Table 4.5 summarises the performance metrics computed for this classification technique.

Name of the classifier	Accuracy	Sensitivity	Specificity	F-Score	Precision	Area Under Curve
Random Forest	83.00%	78%	88.5%	0.83	89.0%	0.833

Table 4.5-Model Testing Summary for The Random Forest Classifier

4.5 Analysis and Results – Support Vector Machine

This section of the chapter deals with choosing the appropriate model for Support Vector Machine Classification. We will be fitting the SVM model to the data using a 75/25 training split and using optimised fine-tuning parameters such as cost and gamma parameters and performing 10 fold cross-validation on the original heart disease data. The classification will be summarised using performance statistics to judge its performance and will be compared with other techniques at the end of the chapter.

4.5.1 Fitting and Fine Tuning The SVM Model

In the specific case of heart attack analysis we are dealing with data that is supposed to be linearly separable i.e. the patients are either diagnosed with the disease or they are not. Hence, the kernel that we will be using for this model will be “linear” in which case the data-points will be categorised into two halves by a hyperplane.

```
Call:
svm(formula = output ~ ., data = train_heart, kernel = "linear",
    cost = 10, scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  10

Number of Support Vectors:  89

 ( 41 48 )


Number of Classes:  2

Levels:
 0 1
```

Figure 4.9-Model Fitting-SVM

The summary of the fitted model in the figure shows us that we have 89 support vectors – 41 in the first class and 48 in the second class. We will now be fine tuning the model using the tune() which is used to obtain the optimal value of hyperparameters such as cost, gamma and coef0 built in the e1071 library that contains functions for performing Support Vector Machines in R and choosing the model with the best set of cost parameters.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
  cost
   0.1

- best performance: 0.1713439

- Detailed performance results:
   cost      error dispersion
1 1e-03 0.4308300 0.13648928
2 1e-02 0.1800395 0.08113445
3 1e-01 0.1713439 0.08134598
4 1e+00 0.1887352 0.07720309
5 5e+00 0.1754941 0.07676084
6 1e+01 0.1754941 0.07676084
7 1e+02 0.1754941 0.07676084
```

Figure 4.10-Cost Tuning - SVM

How far an SVM should be permitted to "bend" with the data is determined by the cost parameter i.e. changing the cost parameter will change the orientation of the SVM model according to our preferences. We strive for a smooth decision surface for a low cost and more accurate classification of more points for a higher cost. After fine tuning the model we find that the best cost parameter is 0.1 since it has the smallest error rate attached to it.

4.5.2 Model Classification Testing

This section will deal with the computation of the confusion matrix that has been formulated on the testing dataset. After formation of the confusion matrix we have calculated various performance metrics to determine the classifier's performance. To generate the confusion matrix, we decided to use 75% of the dataset for training and 25% for testing. For simplicity, we can summarise the confusion matrix in a tabular format that we have mentioned in Table 4.6

```
confusionmatrix  0  1
                0 22  2
                1 12 40
> #Calculating Accuracy
> accuracy.val <- round(sum(diag(conf.mat))/sum(conf.mat),2)
> accuracy.val
[1] 0.82
>
> #Calculating Precision
> precision.val <- conf.mat[4] / sum(conf.mat[4], conf.mat[2])
> precision.val
[1] 0.7692308
>
> #Calculating Sensitivity
> sensitivity.val <- conf.mat[4] / sum(conf.mat[4], conf.mat[3])
> sensitivity.val
[1] 0.952381
>
> #Calculating Specificity
> specificity.val <- conf.mat[1] / sum(conf.mat[1], conf.mat[2])
> specificity.val
[1] 0.6470588
> #Calculating F-Score
> fscore <- (2 * (sensitivity.val * precision.val))/(sensitivity.val + precision.val)
> fscore
[1] 0.8510638
```

Figure 4.11-Confusion Matrix - SVM

Total Columns of Data = 76	Actual: No Heart Disease	Actual: Heart Disease
Predicted: No Heart Disease	TN=22	FN=2
Predicted: Heart Disease	FP=12	TP=40

Table 4.6-Tabular Confusion Matrix- Support Machine Vector

From the table 4.6 we can comment about the performance metrics as follows :

- **Accuracy** : Accuracy is calculated by (**TP + TN) / Total** = (**40+22) / 76 = 0.815** i.e. the classifier is accurate 81.5% of the time.
- **Precision** : Precision can be calculated by **TP /(TP + FP) = 40 / (40+12) = 0.769** i.e. the classifier is precise 76.9% of the time.
- **Recall/Sensitivity** : The sensitivity is calculated by **TP/ (TP+FN) = 40 / 40+2 = 0.952** i.e. the sensitivity of the classifier is 95.2 %.

- **Specificity** : Specificity of the classifier is calculated by **TN /(TN+FP) = 22 / (22+12) =** i.e. the specificity of the classifier is 64.7 %
- **F-score** : The F score is calculated by **T P /(T P+0.5(F P+F N)) = 40/(40 +0.5(14))=0.851**

- **AUC Score** : In order to assess this classifier's diagnostic capabilities, we plotted t the ROC Curve for the testing dataset and computed the Area Under Curve Score, which we found to be 0.81. The distance from the diagnostic line that splits the ROC Plane in half increases with the AUC of the curve. A rather rudimentary categorization algorithm rates an AUC Score of 0.8-0.9 as excellent. The ROC curve is shown in Figure 4.7 along with its AUC

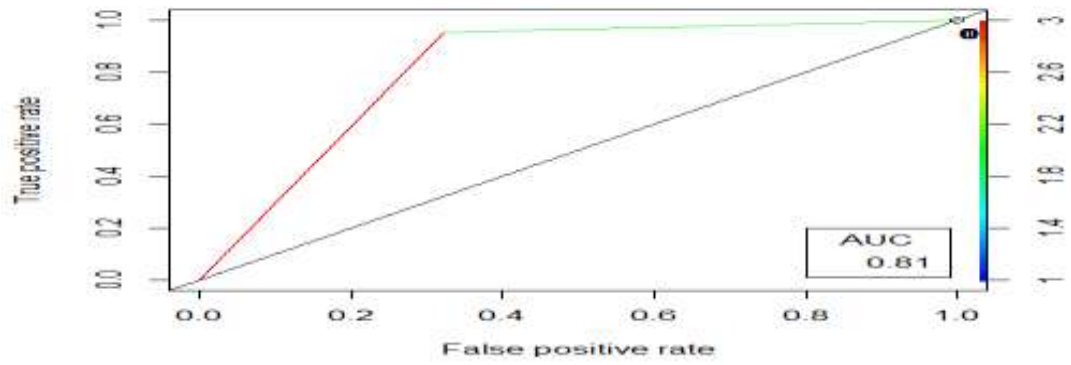


Figure 4.12 AUC Curve – SVM

4.5.3 Tabular Summary

Table 4.7 summarises the performance metrics computed for this classification technique.

Name of the classifier	Accuracy	Sensitivity	Specificity	F-Score	Precision	Area Under Curve
Linear Support Vector Machine	81.50%	95.2%	64.7%	0.851	76.9%	0.81

Table 4.7 Model Testing Summary for Support Vector Machine

4.6 Analysis and Results – Naïve Bayes Classifier

This section of the chapter deals with choosing the appropriate model for Naïve Bayes Classification. We will be fitting the Naïve Bayes model to the data using a 75/25 training split and summarising the classification using performance statistics to judge its performance and will be compared with other techniques at the end of the chapter.

4.6.1 Fitting the Naïve Bayes Model

This section talks about how we have chosen the model that has been used for the model fitting and prediction purposes. The Naïve Bayes function in R uses the Bayes theorem to calculate conditional posterior probabilities of a categorical variable that is dependent on various independent predictor variables.[31]

```
================================== Naive Bayes ==================================

- Call: naive_bayes.formula(formula = output ~ ., data = train_heart,      usekernel = T)
- Laplace: 0
- Classes: 2
- Samples: 227
- Features: 13
- Conditional distributions:
    - KDE: 13
- Prior probabilities:
    - 0: 0.4537
    - 1: 0.5463

---------------------------------------------------------------------------------
```

Figure 4.13-Model Fitting – Naïve Bayes

The use-Kernel feature is attributed the value of true which means that Kernel Density Estimation is applied to each numeric predictor. Kernel density estimation (KDE), a non-parametric approach to estimate the probability density function of a random variable based on kernels as weights, is the application of kernel smoothing for probability density estimation. Kernel based Densities perform better when the numerical variables are not normally distributed[62]. The model takes the training dataset as an input and outputs two prior probability classes. We see that about 45.3% of the datapoints belong to the category where the output variable is “0” and 54.6% of the datapoints belong to the category where the output variable is “1”. We will create predictions on testing and training data using the predict function and store the predictions of the testing dataset in a confusion matrix.

4.6.2 Model Classification Testing

This section will deal with the computation of the confusion matrix that has been formulated

on the testing dataset. After formation of the confusion matrix we have calculated various performance metrics to determine the classifier's performance.

```
p2    0  1
  0 29 10
  1  6 31
> #Calculating Accuracy
> accuracy.val <- round(sum(diag(conf.mat))/sum(conf.mat),2)
> accuracy.val
[1] 0.79
>
> #Calculating Precision
> precision.val <- conf.mat[4] / sum(conf.mat[4], conf.mat[2])
> precision.val
[1] 0.8378378
>
> #Calculating Sensitivity
> sensitivity.val <- conf.mat[4] / sum(conf.mat[4], conf.mat[3])
> sensitivity.val
[1] 0.7560976
>
> #Calculating Specificity
> specificity.val <- conf.mat[1] / sum(conf.mat[1], conf.mat[2])
> specificity.val
[1] 0.8285714
>
> fscore <- (2 * (sensitivity.val * precision.val))/(sensitivity.val + precision.val)
> fscore
[1] 0.7948718
```

Figure 4.14 Confusion Matrix – Naïve Bayes

To make it more readable we will summarise the confusion matrix in a tabular format as follows :

Total Columns of Data = 76	**Actual: No Heart Disease**	**Actual: Heart Disease**
Predicted: No Heart Disease	**TN=29**	**FN=10**
Predicted: Heart Disease	**FP=6**	**TP=31**

Table 4.8 Tabular Confusion Matrix- Naïve Bayes Classification

A manual calculation of the performance metrics can be also commented on as follows:

- **Accuracy** : Accuracy is calculated by (**TP + TN) / Total** = (**31+29) / 76 = 0.789** i.e. the classifier is accurate 78.9% of the time.

- **Precision** : Precision can be calculated by **TP /(TP + FP) = 31 / (31+6) = 0.837** i.e. the classifier is precise 83.7% of the time.

- **Recall/Sensitivity** : The sensitivity is calculated by **TP/ (TP+FN) = 31 / (31+10) = 0.756** i.e. the sensitivity of the classifier is 75.6 %.

- **Specificity** : Specificity of the classifier is calculated by **TN /(TN+FP) = 29 / (29+6) =0.828** i.e. the specificity of the classifier is 82.8 %

- **F-score** : The F score is calculated by **TP/(TP+0.5(F P+F N))= 31/(31 +0.5(16))=0.794**

- **AUC Score** : In order to assess this classifier's diagnostic capabilities, we plotted the ROC Curve for the testing dataset and computed the Area Under Curve Score, which we found to be 0.866.

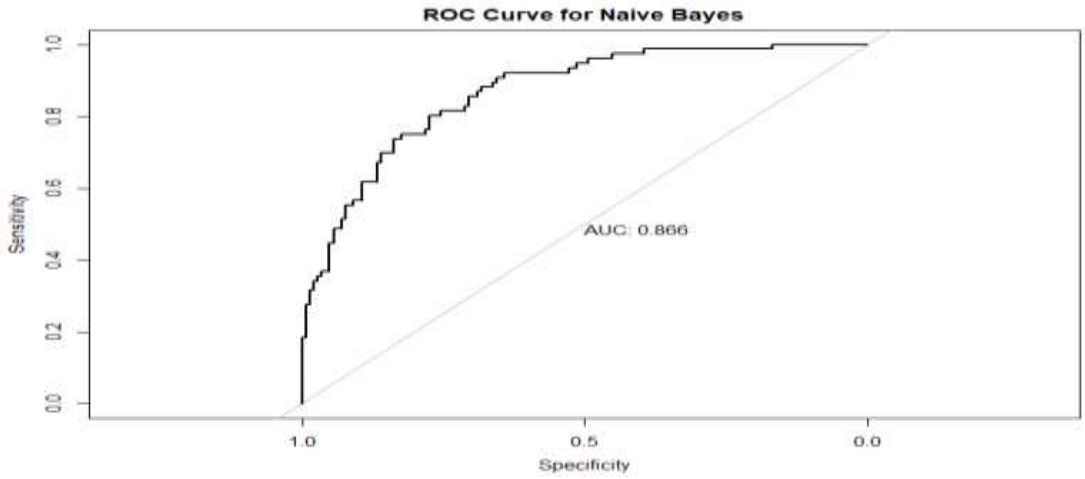


Figure 4.15-AUC-ROC Curve Naïve Bayes

4.6.3 Tabular Summary

Table 4.9 summarises the performance metrics computed for this classification technique.

Name of the classifier	Accuracy	Sensitivity	Specificity	F-Score	Precision	Area Under Curve
Naïve Bayes	78.90%	75.6%	82.8%	0.794	83.7%	0.866

Table 4.9-Model Testing Summary for Naïve Bayes Classifier

4.7 Comparison between different Classification Approaches

Table 4.10 summarises the performance metrics computed for every single classification technique that we have used to perform prediction and classification.

Name of the Classifier	Classification Accuracy	Sensitivity	Specificity	F-Score	Precision	Area Under Curve
Logistic Regression	82.89%	70.5%	**92.86%**	0.78	88.8%	**0.89**
Random Forest	**83.00%**	78%	88.5%	0.83	**89.0%**	0.833
Linear Support Vector Machine	81.50%	**95.2%**	64.7%	**0.851**	76.9%	0.81
Naïve Bayes Classifier	78.90%	75.6%	82.8%	0.794	83.7%	0.866

Table 4.10-Performance metrics comparison between different classification approaches

4.7.1 Discussion about the comparison between metrics

In this section, we will be commenting about the performance of every single classification approach with respect to all the performance metrics in detail and also will talk about the various reasons behind that performance of that particular metric.

4.7.1.1 Comparing classification accuracies of different classifiers

As we can see from the figure 4.16, the bar-plot clearly shows the relationships between the accuracy percentages and the classifiers. From this plot we can infer that even though there is negligible difference of almost 0.11% between Logistic Regression(82.89) and Random Forest(83.00) as shown in Table 4.10, both of these classifiers are comfortably the best machine learning algorithms when it comes to the accurate classification of the patients into one of the two categories. The reasons for this can be credited to the extensive model fitting algorithm that commented about the significance of various attributes and the feature importance plots that allowed us to reduce the amount of features significantly.[37], [61] Naïve Bayes (78.9) performed comparatively worse in this case. This would make sense as

this algorithm assumes independence between predictors which is generally not applicable in real-life and hence limits its functionality to some extent.[32]

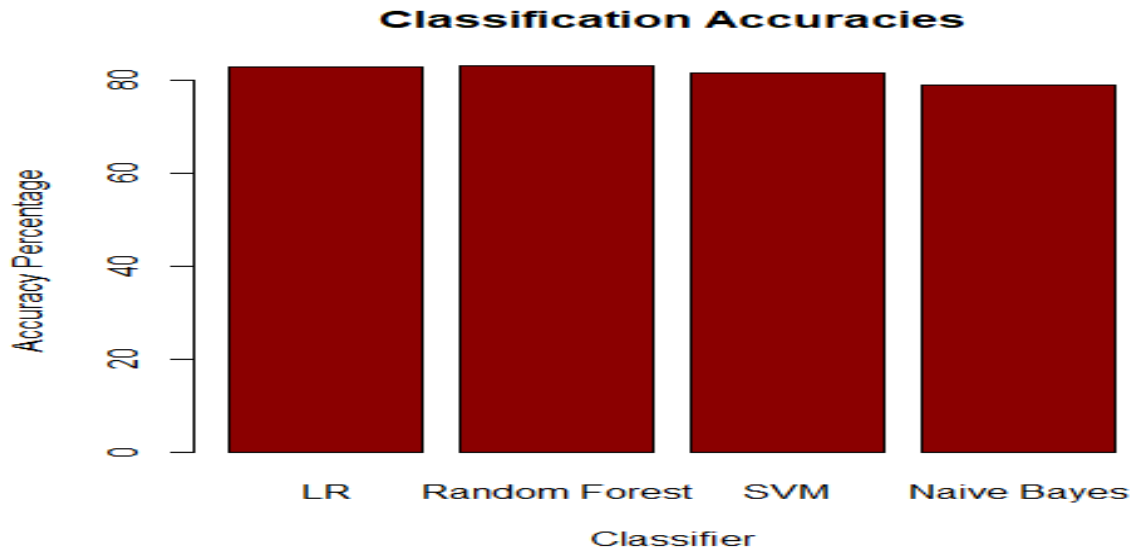


Figure4.16-Accuracy Comparison

4.7.1.2 Comparing sensitivity/recall of different classifiers

The figure 4.17 shows the relationship between the classifiers and the magnitudes of their sensitivities. From this plot, we can conclude that SVM(95.2) is the most sensitive classifier followed by Random Forest(78) and Naïve Bayes(75.6). We see that Logistic Regression (70.5) is the least sensitive classifier in our experiment. A high sensitivity level for Support Vector Machine means that it is the best classification algorithm used in our experiment when it comes to correctly identifying the patients with the diseases. So in the case of SVM, it has a 95.2% chance of identifying patients that have the disease but at the same time will miss 4.8% of the patients who are suffering from the disease. High sensitivity tests typically have low specificity[63]. In other words, they have a high risk of false positives in addition to being effective at detecting illness cases. If the test's objective is to detect every case of a certain ailment, its sensitivity must be high in order to minimise the likelihood of false negative results. In other words, it should be quite likely that the test will identify someone as having the illness. When the consequences of not treating the ailment are severe and/or the medication is highly effective and has few adverse effects, this is extremely significant. In the case of our research, it would be safe to infer that sensitivity is more important than specificity because a more sensitive test means a negative result is less likely to indicate the presence of a disease, therefore increasing its negative predictive value.

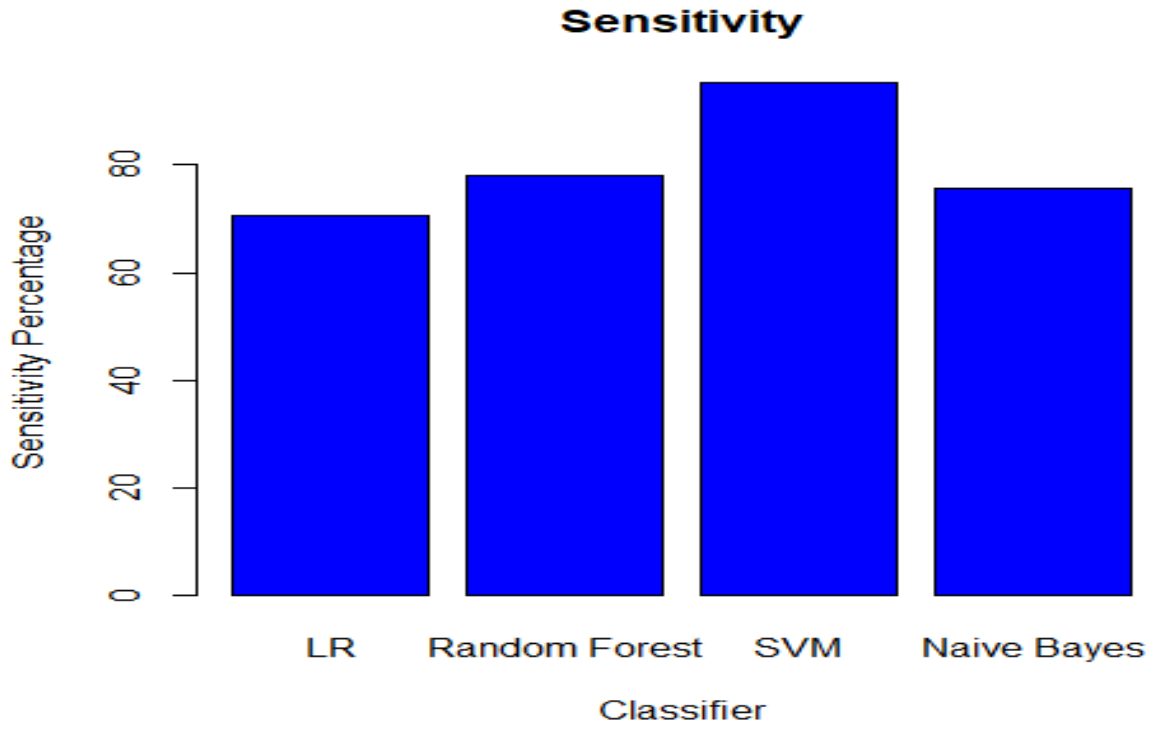


Figure 4.17-Comparing Sensitivity

4.7.1.3 Comparing specificity of different classifiers

The bar-plot in figure 4.18 shows us the relationship between the classifier and the magnitudes of the specificity percentages. We can clearly see that Logistic Regression (92.86) is the most specific classifier followed by the Random Forest(88.5) and Naïve Bayes Classifiers(82.8). We hence see that the classification algorithm that had the most sensitivity, Support Vector Machines(64.7) is the least specific. So in the case of the Logistic Regression classifier , it will successfully identify 92.86 % of the total patients that do not have the disease but at the same time will not be able to identify about 7.14 % of the patients who have the disease. A high specificity is necessary if the test's objective is to properly identify those who do not have the disease while minimising the number of false positives. That is, it should be quite likely that the test will reject those who do not have the illness. The low specificity yet the high accuracy of the SVM can be attributed to the optimal parametric tuning that helped us decide the best values for the cost and gamma parameters. Generally optimisation of the parameters values will result in better results.

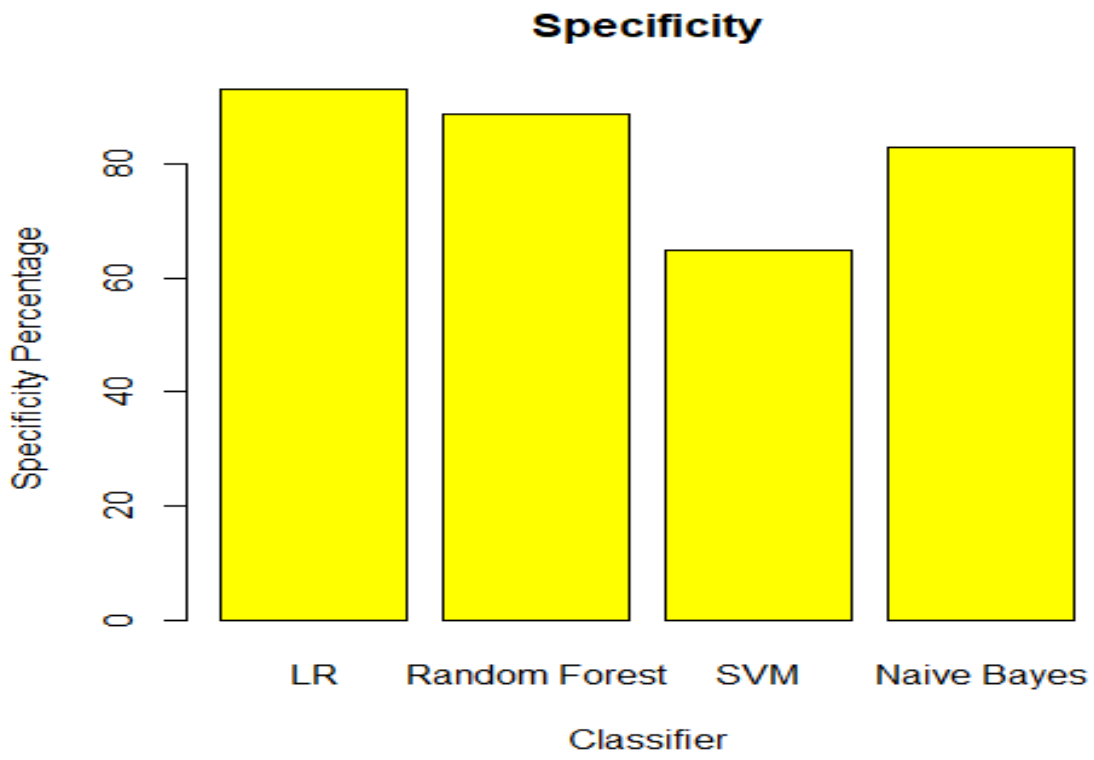


Figure 4.18-Comparing Specificity

4.7.1.4 Comparing Precision of different classifiers

From the bar-plot shown in the figure 4.19 we can infer that Random Forest(89) just edges Logistic Regression(88.8) as the most precise classifier. Support Vector Machine(76.9) is the least precise algorithm out of all four algorithms which would suggest that it has the highest amount of false positive instances. However, if a false positive is inexpensive, this might not be so terrible. Even though precision and accuracy follow similar trends when it comes to our testing classifiers since both Logistic Regression and Random Forest have high accuracy and high precision, accuracy and precision are independent of each other. High precision indicates that, given same conditions, the results of repeated measurements of a known value will be very consistent. The best quality scientific observations are both accurate and precise which are in-fact observed in Logistic Regression and Random Forest.

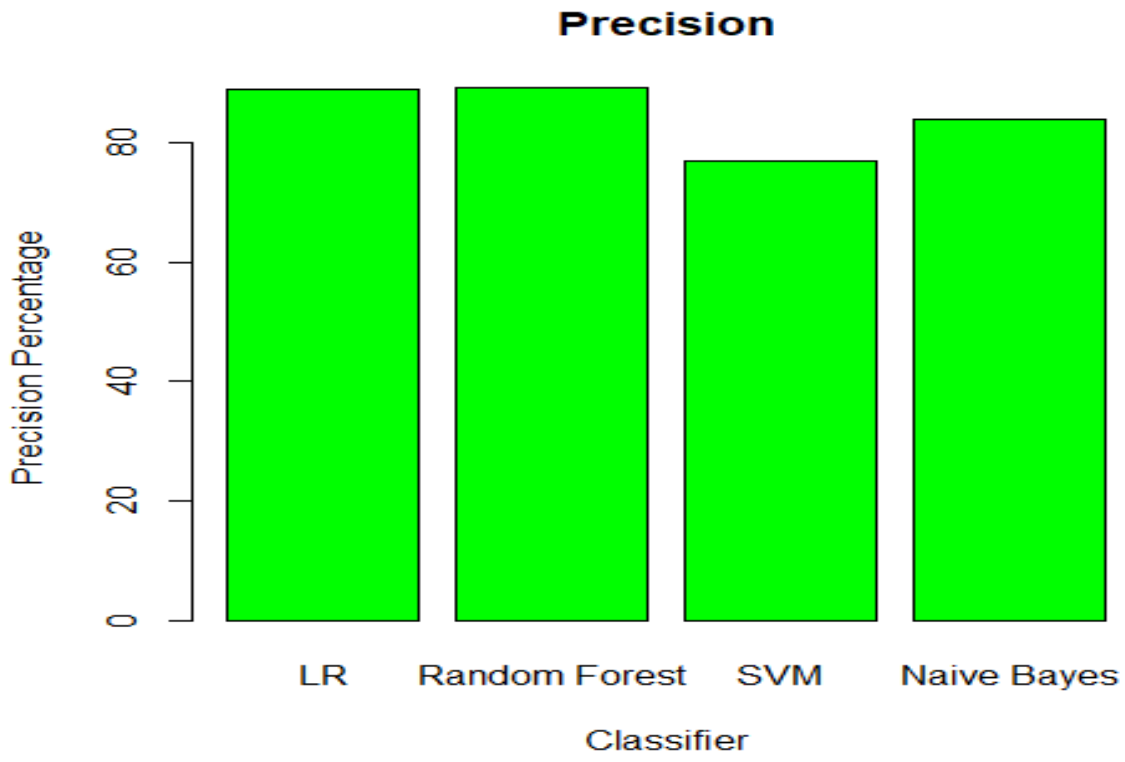


Figure 4.19-Comparing Precision

4.7.1.5 Comparing F-scores of different classifiers

We can conclude from the bar-plot shown in the figure 4.20 that the F-score of the Support Vector Machine(0.85) is the highest followed by the F-scores of Random Forest(0.83) ,Naïve Bayes(0.794) and Logistic Regression(0.78) Classifiers. Since the F-score is considered to be the harmonic mean of the precision and recall metrics of the classifiers, the idea behind the F-measure is that both measurements are equally important and that a good F-measure can only be obtained by combining excellent precision and good recall. Hence we can say that SVM does the best job out of all the classifiers when it comes to capturing the properties of recall and precision in a balanced ratio. In the case when we have to lend more importance to a particular metric involved in the f-score we can define another classification metric knows as the “F-beta score” which is the weighted harmonic mean of precision and recall[57]. The beta parameter talks about the weight assigned to sensitivity in the combined score. If the value of beta < 1 the priority is given to the precision otherwise if the value of beta > 1, the priority is given to the recall metric.

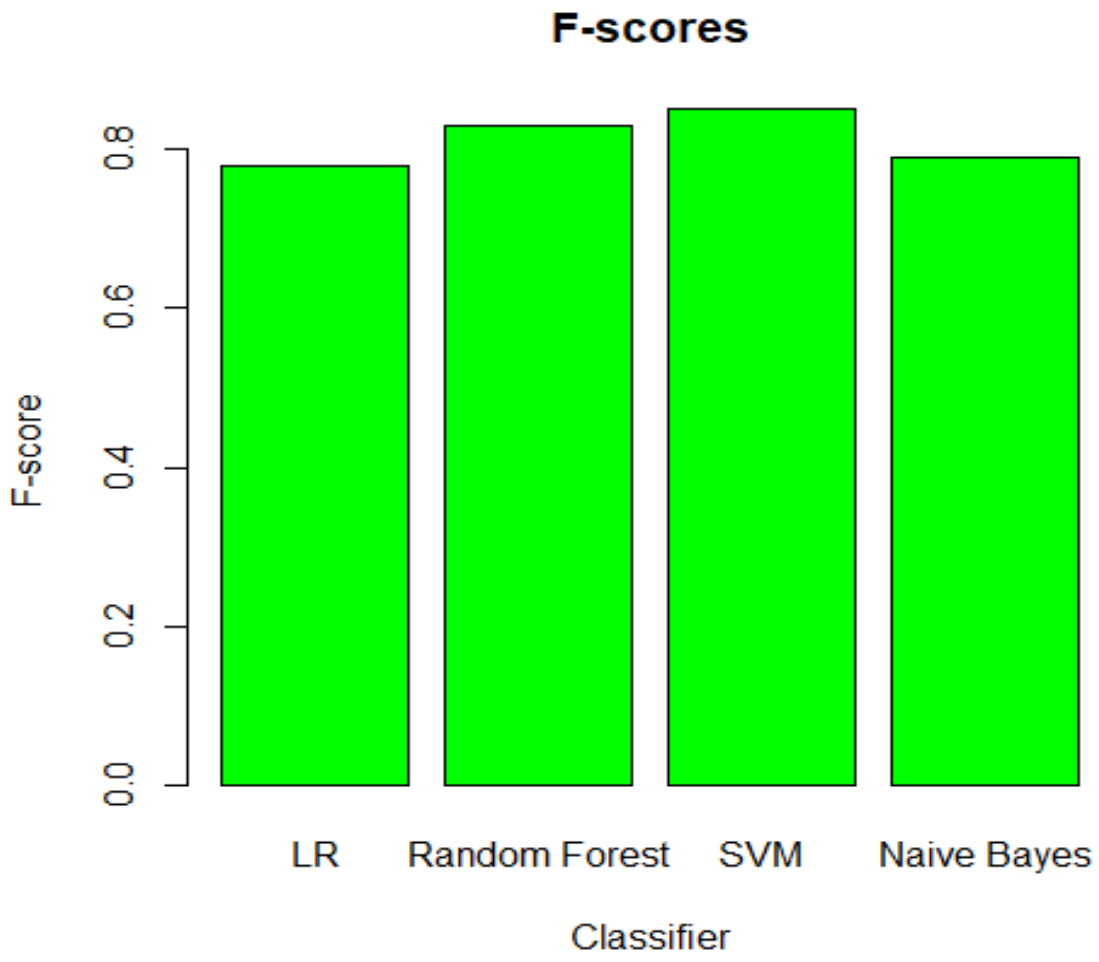


Figure 4.20-Comparing F-Scores

4.7.1.6 Comparing Area-Under-Curve Values of different classifiers

This bar-plot shown in the figure 4.21 compares different AUC Scores of the classifiers and we can conclude that Logistic Regression(0.89) is the classifier that has the largest AUC value followed by Naïve Bayes(0.866) , Random Forest(0.833) and SVM(0.81). The higher the AUC score of the classifier the better separability characteristics it possesses and the easier it is for the classifier to differentiate between the two classes. So we can conclude by saying that since the area under the ROC curve of the logistic regression classifier is the highest, it makes it the best classifier that we have used when it comes to distinguishing between the patients that have the diseases and the patients who don't. Having said that it is important to establish the fact that the AUC and the accuracy of the classifier are two independent metrics and both of them measure entirely different things[58]. Accuracy is used to find the fractions between correctly assigned positive and negative classes while

AUC is used to determine the optimal threshold to find a balance between sensitivity and FPR.

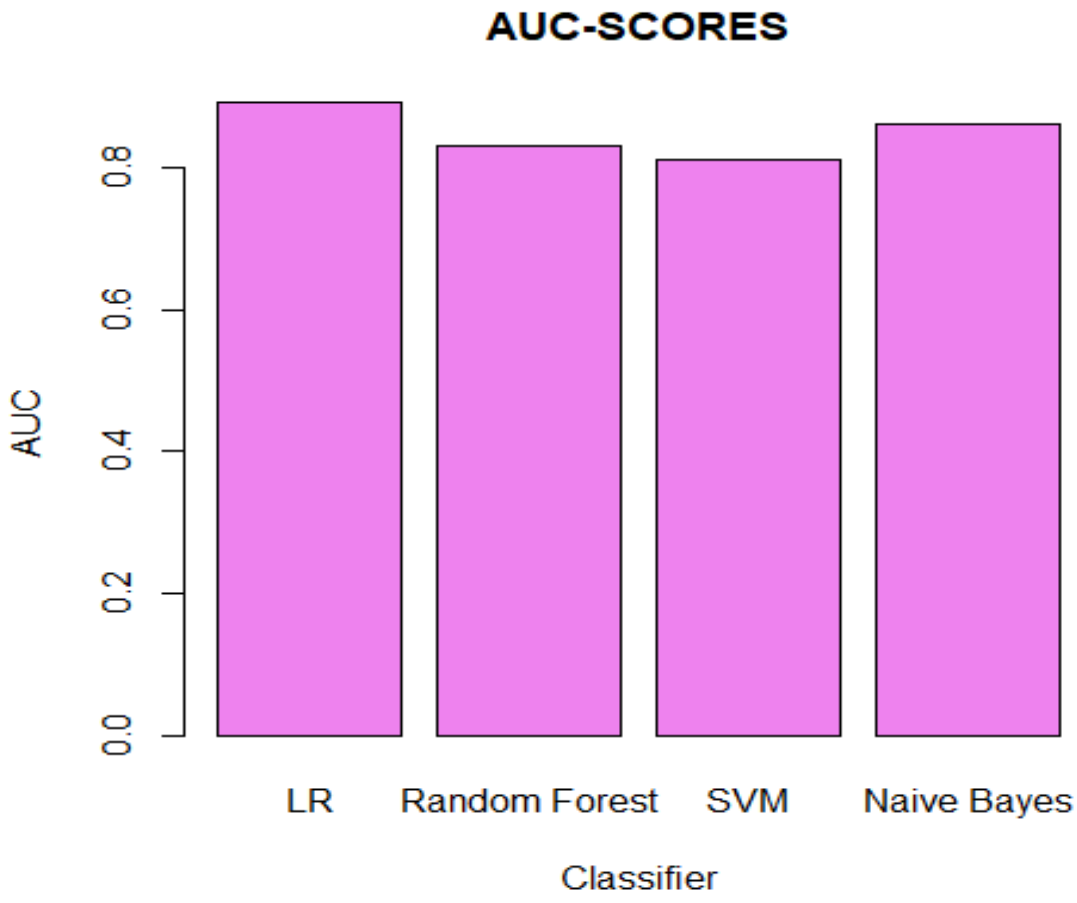


Figure 4.21-Comparing AUC Scores

Chapter 5 : Conclusion, Limitations and Future Scope

5.1 Conclusion

The main objective of this research work is to examine a dataset related to heart diseases and comment about the best techniques and approaches that can predict cardio-vascular diseases with the best accuracy. After we conducted thorough exploratory data analysis, we found a few deficiencies in the dataset that we solved by proposing various pre-processing and feature reduction techniques that helped in the proliferation of the accuracies of the classifiers. Age, cholesterol, and fasting blood sugar were found to be insignificant variables and so they were not included in the final model. Further we analysed the dataset and determined that what can be described as the best classifier has a lot to do with the purpose of our classification. We can say that because different classifiers were proficient at different performance metrics. Random Forest was the classifier that could be described as the most accurate and precise whereas Logistic Regression had the best AUC score which would make it the algorithm with the best diagnostic ability when it comes to setting an optimal threshold between TPR and FPR. It was also the most “specific” classifier when it came to properly identify those who do not have the disease. Similarly Support Vector Machines was the most “sensitive” classifier when it came to correctly identifying the patients with the diseases. So for instance if we would want a classifier that would be good at identifying the maximum of true positives i.e. the patients that tested positive and actually have the disease then we can use SVM for the classification as it has the highest sensitivity.

5.2 Limitations of the Research

The main goal of this research was to estimate the risk of cardiovascular diseases based on the corresponding medical information of the patients and accurately predict the person's health condition by training various machine learning models with training data to predict the test data and provide accurate results. The main challenge that arose during the research was achieving the highest level of accuracy, and the major constraint that stood in our way is the size of the dataset. The dataset is small in size, and adding more useful information to the dataset such as the smoking habits, obesity and family history of cardio-vascular diseases could not only help in predicting the results on a larger scale but could also change our current results by a substantial margin and increasing sample size would also require a lot of time and money.

5.3 Future Scope of The Research

Since this research work is primarily focused on the classification of patients into two categories depending on their medical history, by increasing the sample size, the research project may be improved to increase accuracy of the classification models. It has been clearly proven that the better the learning model is trained, the more accurate the results will be. Future work on this project will also include creating a web application based on this framework. Many users will have convenient access to a web application that can assist the patient in predicting their condition if they are worried about changes in their health problems.

References

[1]“Lumen Learning-The Circulatory System”, Retrieved from **https://courses.lumenlearning.com/wm-biology2/chapter/circulatory-system/**

[2]“British Heart Foundation” Retrieved from ***https://www.bhf.org.uk/informationsupport/how-a-healthy-heart-works#:~:text=Each%20day%2C%20your%20heart%20beats,organs%20and%20muscles%20work%20properly.***

[3]“U.S. Department of Health & Human Services-Congenital Heart Defects” Retrieved from ***https://www.cdc.gov/ncbddd/heartdefects/howtheheartworks.html***

[4]“World Health Organisation-CVDs” Retrieved From ***https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)***

[5]***“***Cardiovascular Diseases- Wikipedia” Retrieved from **https://en.wikipedia.org/wiki/Cardiovascular_disease#:~:text=Cardiovascular%20diseases%20are%20the%20leading,million%20(25.8%25)%20in%201990.**

[6]”Atherosclerosis-NHS” Retrieved from ***https://www.nhs.uk/conditions/atherosclerosis/***

[7] “What are The Risk Factors -National Institute For Health and Care Excellence (2020) Retrieved From ***https://cks.nice.org.uk/topics/cvd-risk-assessment-management/background-information/risk-factors-for-cvd/***

[8] Y. Baştanlar and M. Özuysal, ‘Introduction to Machine Learning’, in *miRNomics: MicroRNA Biology and Computational Analysis*, vol. 1107, M. Yousef and J. Allmer, Eds. Totowa, NJ: Humana Press, 2014, pp. 105–128. doi: 10.1007/978-1-62703-748-8_7.

[9] T. Mitchell, *Machine Learning*. : McGraw-Hill Science/Engineering/Math, 1997.

[10] T. Mitchell, *The Discipline of Machine Learning*. 2006.

[11] T. Jiang, J. L. Gradus, and A. J. Rosellini, ‘Supervised Machine Learning: A Brief Primer’, *Behavior Therapy*, vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/j.beth.2020.05.002.

[12] R. Gentleman and V. J. Carey, ‘Unsupervised Machine Learning’, in *Bioconductor Case Studies*, New York, NY: Springer New York, 2008, pp. 137–157. doi: 10.1007/978-0-387-77240-0_10.

[13] L. P. Kaelbling, M. L. Littman, and A. W. Moore, ‘Reinforcement Learning: A Survey’, *jair*, vol. 4, pp. 237–285, May 1996, doi: 10.1613/jair.301.

[14] G. Mrukwa, ‘Supervised and Unsupervised Machine Learning - Types of ML’, Oct. 08, 2018. https://www.netguru.com/blog/supervised-machine-learning

[15] S. Rong and Z. Bao-wen, ‘The research of regression model in machine learning field’, *MATEC Web Conf.*, vol. 176, p. 01033, 2018, doi: 10.1051/matecconf/201817601033.

[16] S. Ray, ‘7 Regression Techniques you should know!’, Aug. 14, 2015. https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

[17] I. G. Maglogiannis, Ed., *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*. Amsterdam ; Washington, DC: IOS Press, 2007.

[18] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, ‘Prediction of Cardiovascular Disease Using Machine Learning Algorithms’, in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, Mar. 2018, pp. 1–7. doi: 10.1109/ICCTCT.2018.8550857.

[19] D. G. Kleinbaum, M. Klein, and E. R. Pryor, *Logistic regression: a self-learning text*, 3rd ed. New York: Springer, 2010.

[20] A. THANDA, ‘What is Logistic Regression? A Beginner’s Guide’, May 24, 2022. https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

[21] A. G, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, ‘Logistic regression technique for prediction of cardiovascular disease’, *Global Transitions Proceedings*, vol. 3, no. 1, pp. 127–130, Jun. 2022, doi: 10.1016/j.gltp.2022.04.008.

[22] Y. Liu, Y. Wang, and J. Zhang, ‘New Machine Learning Algorithm: Random Forest’, in *Information Computing and Applications*, vol. 7473, B. Liu, M. Ma, and J. Chang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 246–252. doi: 10.1007/978-3-642-34062-8_32.

[23] O. Sagi and L. Rokach, ‘Ensemble learning: A survey’, *WIREs Data Mining Knowl Discov*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/widm.1249.

[24] A. Chauhan, ‘Random Forest Classifier and its Hyperparameters’, Feb. 23, 2021. https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

[25] P. Liashchynskyi and P. Liashchynskyi, ‘Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS’. arXiv, Dec. 12, 2019. Accessed: Sep. 14, 2022. [Online]. Available: http://arxiv.org/abs/1912.06059

[26] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, ‘An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection’, *IEEE Access*, vol. 7, pp. 180235–180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

[27] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines: and other kernel-based learning methods*. Cambridge ; New York: Cambridge University Press, 2000.

[28] S. Luo, 'Loss Function(Part III): Support Vector Machine', Oct. 15, 2018. https://towardsdatascience.com/optimization-loss-function-under-the-hood-part-iii-5dff33fa015d

[29] M. N. Hoda, *INDIACom-2016 proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development: (16th-18th March, 2016)*. Piscataway, NJ: IEEE, 2016.

[30] S. Ghosh, A. Dasgupta, and A. Swetapadma, 'A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification', in *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, Tamilnadu, India, Feb. 2019, pp. 24–28. doi: 10.1109/ISS1.2019.8908018.

[31] F.-J. Yang, 'An Implementation of Naive Bayes Classifier', in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, Dec. 2018, pp. 301–306. doi: 10.1109/CSCI46756.2018.00065.

[32] E. Eells, 'Review: Bayes's Theorem', *Mind*, vol. 113, no. 451, pp. 591–596, Jul. 2004, doi: 10.1093/mind/113.451.591.

[33] R. Poli, J. Kennedy, and T. Blackwell, 'Particle swarm optimization: An overview', *Swarm Intell*, vol. 1, no. 1, pp. 33–57, Oct. 2007, doi: 10.1007/s11721-007-0002-0.

[34] U. N. Dulhare, 'Prediction system for heart disease using Naive Bayes and particle swarm optimization', *biomedicalresearch*, vol. 29, no. 12, 2018, doi: 10.4066/biomedicalresearch.29-18-620.

[35] S. García, F. Herrera, and J. Luengo, *Data Preprocessing in Data Mining*, 1st ed. 2015. Cham: Springer International Publishing : Imprint: Springer, 2015. doi: 10.1007/978-3-319-10247-4.

[36] D. Singh and B. Singh, 'Investigating the impact of data normalization on classification performance', *Applied Soft Computing*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[37] A. Jovic, K. Brkic, and N. Bogunovic, 'A review of feature selection methods with applications', in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2015, pp. 1200–1205. doi: 10.1109/MIPRO.2015.7160458.

[38] " Heart Attack| Analysis and Prediction Dataset" Retrieved from ***https://www.kaggle.com/code/chaitanya99/heart-attack-analysis-prediction/data***

[39] "UCI Repository Dataset" Retrieved from **https://archive.ics.uci.edu/ml/datasets/heart+disease**

[40] S. Morgenthaler, 'Exploratory data analysis', *WIREs Comp Stat*, vol. 1, no. 1, pp. 33–44, Jul. 2009, doi: 10.1002/wics.2.

[41] H.-J. Priebe, ‘The aged cardiovascular risk patient’, *British Journal of Anaesthesia*, vol. 85, no. 5, pp. 763–778, Nov. 2000, doi: 10.1093/bja/85.5.763.

[42] S. T. Hardy *et al.*, ‘Maintaining Normal Blood Pressure Across the Life Course: The JHS’, *Hypertension*, vol. 77, no. 5, pp. 1490–1499, May 2021, doi: 10.1161/HYPERTENSIONAHA.120.16278.

[43] G. N. Levine, J. F. Keaney, and J. A. Vita, ‘Cholesterol Reduction in Cardiovascular Disease — Clinical Benefits and Possible Mechanisms’, *N Engl J Med*, vol. 332, no. 8, pp. 512–521, Feb. 1995, doi: 10.1056/NEJM199502233320807.

[44] K. Miller, ‘Understanding Your Cholesterol Report’, Aug. 20, 2022. https://www.webmd.com/cholesterol-management/understanding-your-cholesterol-report#:~:text=Optimal%3A%20Less%20than%20100%20mg,%3A%20160%2D189%20mg%2FdL

[45] R. M. West, ‘Best practice in statistics: Use the Welch *t* -test when testing the difference between two groups’, *Ann Clin Biochem*, vol. 58, no. 4, pp. 267–269, Jul. 2021, doi: 10.1177/0004563221992088.

[46] Mayo Clinic Staff, ‘High blood pressure dangers: Hypertension’s effects on your body’. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868#:~:text=High%20blood%20pressure%20can%20cause%20many%20heart%20problems%2C%20including%3A,arrhythmias)%20or%20a%20heart%20attack.

[47] M. S. A. Jensen, ‘Electrocardiogram interpretation in general practice’, *Family Practice*, vol. 22, no. 1, pp. 109–113, Nov. 2004, doi: 10.1093/fampra/cmh601.

[48] C. Perret-Guillaume, L. Joly, and A. Benetos, ‘Heart Rate as a Risk Factor for Cardiovascular Disease’, *Progress in Cardiovascular Diseases*, vol. 52, no. 1, pp. 6–10, Jul. 2009, doi: 10.1016/j.pcad.2009.05.003.

[49] P. Palatini and S. Julius, ‘Elevated Heart Rate: A Major Risk Factor for Cardiovascular Disease’, *Clinical and Experimental Hypertension*, vol. 26, no. 7–8, pp. 637–644, Jan. 2004, doi: 10.1081/CEH-200031959.

[50] Shookster D, Lindsey B, Cortes N, Martin JR. Accuracy of Commonly Used Age-Predicted Maximal Heart Rate Equations. Int J Exerc Sci. 2020 Sep 1;13(7):1242-1250. PMID: 33042384; PMCID: PMC7523886.

[51] JacquelineC. M. Witteman, FransJ. Kok, JanL. C. M. Van Saase, and HansA. Valkenburg, ‘AORTIC CALCIFICATION AS A PREDICTOR OF CARDIOVASCULAR MORTALITY’, *The Lancet*, vol. 328, no. 8516, pp. 1120–1122, Nov. 1986, doi: 10.1016/S0140-6736(86)90530-1.

[52] J. P. Potokar and D. J. Nutt, ‘Chest pain: panic attack or heart attack?’, *Int J Clin Pract*, vol. 54, no. 2, pp. 110–114, Mar. 2000.

[53] I. Ezzine and L. Benhlima, ‘A Study of Handling Missing Data Methods for Big Data’, in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, Marrakech, Oct. 2018, pp. 498–501. doi: 10.1109/CIST.2018.8596389.

[54] R. Taylor, ‘Interpretation of the Correlation Coefficient: A Basic Review’, *Journal of Diagnostic Medical Sonography*, vol. 6, no. 1, pp. 35–39, Jan. 1990, doi: 10.1177/875647939000600106.

[55] H. Abdi and L. J. Williams, ‘Principal component analysis: Principal component analysis’, *WIREs Comp Stat*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[56] M. B. Kursa and W. R. Rudnicki, ‘Feature Selection with the **Boruta** Package’, *J. Stat. Soft.*, vol. 36, no. 11, 2010, doi: 10.18637/jss.v036.i11.

[57] S. Narkhede, ‘Understanding Confusion Matrix’, May 09, 2018. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[58] J. Muschelli, ‘ROC and AUC with a Binary Predictor: a Potentially Misleading Metric’, *J Classif*, vol. 37, no. 3, pp. 696–708, Oct. 2020, doi: 10.1007/s00357-019-09345-1.

[59] Wikipedia, the free encyclopedia, ‘Receiver operating characteristic’. https://www.wikiwand.com/en/Receiver_operating_characteristic

[60] J. E. Cavanaugh and A. A. Neath, ‘The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements’, *WIREs Comp Stat*, vol. 11, no. 3, May 2019, doi: 10.1002/wics.1460.

[61] J. Rogers and S. Gunn, ‘Identifying Feature Relevance Using a Random Forest’, in *Subspace, Latent Structure and Feature Selection*, vol. 3940, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 173–184. doi: 10.1007/11752790_12.

[62] M. Majka, ‘Package “naivebayes”’, Mar. 08, 2020. https://cran.microsoft.com/web/packages/naivebayes/naivebayes.pdf

[63] T.-W. Loong, ‘Understanding sensitivity and specificity with the right side of the brain’, *BMJ*, vol. 327, no. 7417, pp. 716–719, Sep. 2003, doi: 10.1136/bmj.327.7417.716.