HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

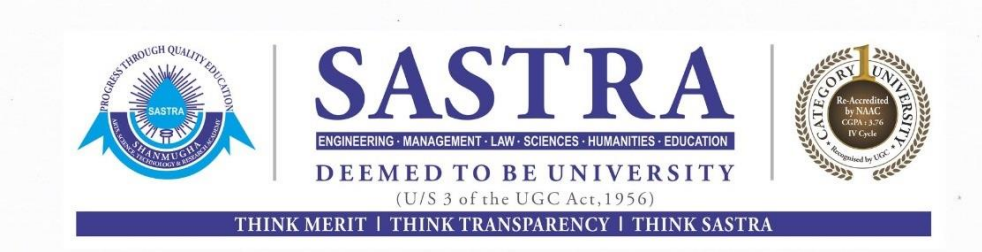
Thesis submitted to the SASTRA Deemed to be University
in partial fulfillment of the requirements
for the award of the degree of

M.Sc. Mathematics

Submitted by

M. KOWSALYA
(Reg. No.: 223045018)

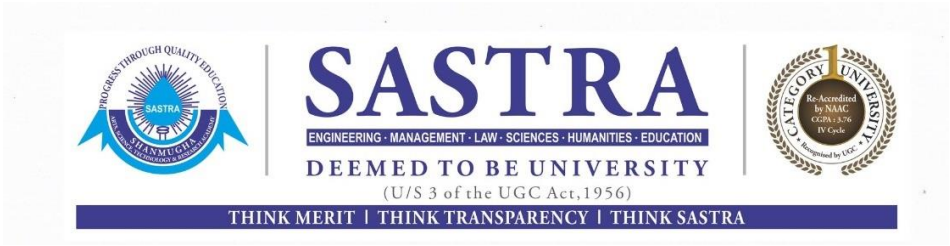
MAY 2023



DEPARTMENT OF MATHEMATICS

SRINIVASA RAMANUJAN CENTRE

KUMBAKONAM, TAMIL NADU, INDIA – 612 001



DEPARTMENT OF MATHEMATICS

SRINIVASA RAMANUJAN CENTRE

KUMBAKONAM – 612 001

BONAFIDE CERTIFICATE

This is to certify that the thesis titled "**Heart Disease Prediction Using Machine Learning Techniques**" submitted in partial fulfillment of the requirements for the award of the degree of M.Sc. Mathematics to the SASTRA Deemed to be University, is a bona-fide record of the work done by **Ms. M. KOWSALYA** (Reg. No. 223045018) during the final semester of the academic year 2022-2023, in the **Department of Mathematics, Srinivasa Ramanujan Centre**, under my supervision. This thesis has not formed the basis for the award of any degree, diploma, associate-ship, fellowship or other similar title to any candidate of any University.

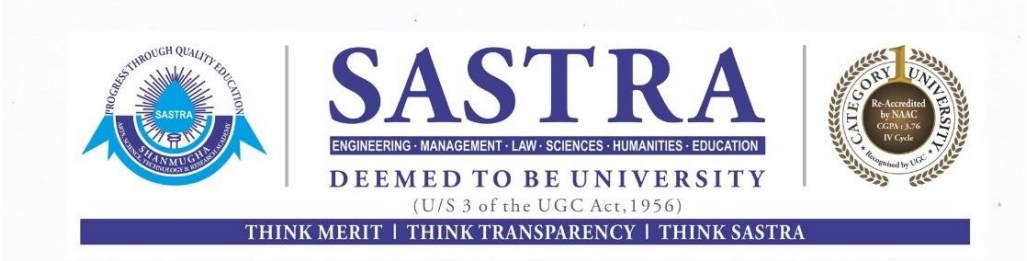
Signature of Project Supervisor :

Name with Affiliation : Dr. K. KANNAN
Senior Assistant Professor
Department of Mathematics

Date :

Project *Vivavoc*e held on 30/05/2023

Examiner 1 **Examiner 2**



DEPARTMENT OF MATHEMATICS

SRINIVASA RAMANUJAN CENTRE

KUMBAKONAM – 612 001

DECLARATION

I declare that the thesis titled "**Heart Disease Prediction Using Machine Learning Techniques**" submitted by me is an original work done by me under the guidance of **Dr. K. Kannan, Senior Assistant Professor, Department of Mathematics, Srinivasa Ramanujan Centre, SASTRA Deemed to be University** during the final semester of the academic year 2022-2023, in the **Department of Mathematics**. The work is original and wherever Ihave used materials from other sources, I have given due credit and cited them in the text of the thesis. This thesis has not formed the basis for the award of any degree, diploma, associate-ship, fellowship or other similar title to any candidate of any University.

Signature of the candidate :

Name of the candidate : M. KOWSALYA

Date :

Acknowledgements

I would like to express my sincere thanks to our honourable Vice-Chancellor **Dr. S. Vaidhyasubramaniam,** for providing me the opportunity to undergo this programme at Srinivasa Ramanujan Centre, Kumbakonam.

I would like to express my thanks to **Dr. R. Chandramouli,** Registrar, for this earnest support.

I would like to express my gratitude to **Dr. S. Swaminathan,** Dean, Planning and Development, for his support and constant encouragement.

I would like to extend my thanks to **Dr. V. Ramaswamy,** Dean, Srinivasa Ramanujan Centre, Kumbakonam for providing the necessary facilities.

I wish to express my heartfelt thanks to **Dr. D. Narasimhan,** Head, Department of Mathematics, Srinivasa Ramanujan Centre, Kumbakonam for his valuable comments and ideas to enhance the quality of the project.

I would like to express my sincere thanks to my guide **Dr. K. Kannan,** Department of Mathematics for his valuable guidance, motivation and continuous support for the completion of this project.

I would like to express my thanks to my Project Coordinators **Dr. S. Venkatesh** and **Dr. K. Parameswari,** Department of Mathematics for their continuous guidance and support for this project.

My thanks are due to all the staff members of the Department of Mathematics for their encouragement.

(**M. KOWSALYA)**

Abstract

Chapter 1 is concerned with a brief introduction to heart disease and some basic definitions of machine learning techniques and dataset information are mentioned.

Chapter 2 gives an explanation of methods in machine learning algorithms like KNN, SVM, random forest, LDA and naive bayes as well as methodology is explained.

Chapter 3 deals with a case study of a heart disease prediction on heart dataset, in which basic descriptive statistics are calculated and using some machine learning algorithms, heart disease is predicted.

Chapter 4 consists of a case study of heart disease dataset using a hybrid approach. We have done parameter reduction based on the soft set and results are compared. The best model is identified.

Chapter 5 Finally, the conclusion of the present project is given.

Table of Contents

Chapter 1

INTRODUCTION

Heart disease is the most deadly disease in the world and one of them is a heart attack. It has become the most common disease among any age group in the past few decades. It's also known as an acute myocardial infarction (AMI) and is one of the most dangerous conditions in the cardiovascular disease category. When blood flow to the heart is impeded or blocked, a heart attack happens. Blood clotting is usually caused by the buildup of cholesterol and other substances in the arteries. Plaque is fat and contains cholesterol. Many heart diseases are on the rise, which can lead to lifestyle changes, work stress and a poor diet.

According to estimates, 17.9 million individuals died from cardiovascular causes in 2019, which accounted for 32% of all fatalities worldwide. Many factors, including heredity, age, gender, tobacco use, inactivity, autoimmune disorders, high blood pressure, alcohol consumption, stress level, cholesterol and existing heart problems, Physiological signs may affect the prevalence of cardiac illnesses around the world[3]. Researchers are working to create an efficient method for the timely detection of cardiac diseases because current methods for diagnosing heart illness are ineffective at early detection for a number of reasons, including accuracy and computing time. Predicting cardiovascular disease is challenging, especially in developing nations. where there is a lack of trained medical workers,

diagnostic tools and additional resources needed to identify and treat people with heart conditions [1].

The body experiences a variety of symptoms, including weariness and shortness of breath, if the heart's pumping motion is ineffective and the heart muscle is unable to supply the blood and oxygen the body needs. As your heart is not functioning properly, this condition is known as heart failure. While treatments, including surgery, typically cannot cure heart failure, they can lead to longer lifespans and fewer negative consequences for individuals. Heart bypass surgery is also known as coronary artery bypass grafting (CABG). Surgery is done to treat clogged or constricted arteries supplying your heart muscle. This major heart procedure is frequently done to treat patients with angina and coronary artery disease.[20]

A language and environment for statistical computing and graphic design is called R and it is effective in data analysis. The R language has gained popularity because it makes it simple for researchers to include many machine learning approaches into a single programme. Also, it gives researchers an easy way to exchange codes[18]. R packages are collections of assembled code, test data, and R functions. In the R environment, packages are kept in a directory called library. There are numerous bundles of packages available. For example, over 10,000 packages for the R programming language are available in the CRAN repository and the number of packages is constantly increasing.

The field of machine learning is extremely broad and diverse, with its applications and breadth expanding day by day. Artificial intelligence's machine learning field employs algorithms to harvest data and then forecast future trends. In the current data science era, machine learning algorithms are continually being

employed across many fields to obtain insightful understanding and use the data to make judgements. In this period of time, data is necessary for industries and organizations. Nowadays, data stored in hospitals is in electronic format and contains information like medical history, mortality rate and symptoms. To forecast and assess the reliability of the provided data and machine learning uses a range of classifiers, including supervised, unsupervised and ensemble learning. Machine learning has the potential to increase accuracy by taking advantage of complicated connections between risk factors[8].

Machine learning (ML) is a sort of artificial intelligence (AI) that enables AI programmes to forecast outcomes more correctly without being given specific guidance to do so. Machine learning techniques employ historical data as input to predict new output values. Due to the fact that machine learning is typically a better predictor than traditional approaches, machine learning prediction is chosen because it employs algorithms. Machine learning is able to spot patterns and connections that people cannot.

The supervised learning uses labelled datasets to train algorithms to correctly categorise data or predict outcomes and until it is well fitted, the model modifies its weights. This occurs as part of the review process to prevent the model from becoming overfit.

A heart disease diagnosis can be made using machine learning technologies before a person sustains a serious injury. The science and technology field of machine learning is expanding and it may be used to diagnose cardiac problems[4]. The most popular strategy for predicting heart attacks is the naive bayes method. Data is gathered from a variety of sources, after being properly categorised and then evaluated to gather the required data. We can conclude that machine

learning is particularly well suited to heart disease prediction.

1.1 Literature Review

At the Johns Hopkins Hospital, William Osler (1849–1919), chief physician and professor of clinical medicine, conducted considerable research on angina and was among the first to propose that it was more of a syndrome than a particular disease. According to the Association for Cardiovascular Angiography and Treatments, in the 1960s and 1970s therapies for heart disease included bypass surgery and percutaneous balloon angioplasty[19].

R was created in 1993 by Ross Ihaka and Robert Gentleman and includes linear regression, machine learning techniques, statistical inference, time series and other features.

Frank Rosenblatt was a great psychologist known for his work on machine learning. He invented the perceptron, a machine learning method, in 1957. The Perceptron algorithm was one of the first to use artificial neural networks, which are now widely used in machine learning.

Zoccali [17] in 2006, patients with end-stage kidney disease (ESKD) have a very significant mortality risk and also cardiovascular disease and both of these outcomes are significantly related to renal function and in populations with heart problems.

Many studies and beneficial investigations have been conducted in relation to the medical profession approach to predicting diseases using machine learning models. A technique that emphasises the value of feature selection in cardiac disease prediction was put forward by M.A. Jabbar et al.[2] in 2013, For feature selection, they used a genetic algorithm and for KNN later. They got good

results. For the purpose of predicting cardiac disease, several researchers have also used deep learning algorithms.

Sonam Nikhar et al.[15], in 2016 this investigation aimed to provide a in-depth analysis of the Naive Bayes and decision tree classifiers used in our evaluation, especially in the diagnosis of cardiac problems. The results of certain studies indicate that decision trees outperform bayes classification systems when future data mining techniques are applied to a comparable dataset.

Using a public dataset of 573 records, Karthiga et al.[14], successfully completed a study to forecast the development of coronary artery disease in 2017. The authors processed the dataset, using the decision tree and naive bayes classifier by the MATLAB data analysis tool and then they generated accuracy results to evaluate the performance of the models. According to the results reported, decision trees have higher accuracy than naive bayes.

In the same year, Sophie H Bots et al.[16], While fatalities from heart disease and stroke declined dramatically for both genders between 1980 and 2010. Some data suggested that men experienced more age-specific reductions than women. Males continue to die from CHD and stroke more frequently than women in their old age.

Dejun Zhang et al. (2018) used principal component analysis for the adaboost algorithm to predict clinical outcomes in breast cancer. In a study to predict heart disease, Tarawneh et al.[10] in 2019, used hybrid data mining classifier techniques. Datasets were retrieved from the UCI repository, which comprises 303 entries and 76 variables. For 14 attributes, model training and testing were done. The classification algorithms that assessed the preciseness of the prediction of heart disease include KNN, SVM, RF and NB. The SVM and NB anticipated

cardiac disease more accurately 89.2% and produced better outcomes.

In the year 2019, Mamatha Alex et al.[11], this study makes use of artificial neural networks, random forests, KNN and support vector machine techniques. In contrast to the previously mentioned classification methods for data mining, the artificial neural network predicts the highest accuracy for diagnosing heart disease.

Edwar Macias et al.[7] in 2020, researched a machine learning approach for enhancing fatality forecast in end-stage renal disease and while all the factors contributed to the highest performance, those discovered by random forest had more anticipating power than those selected with knowledge from experts. There are some machine learning methods to treat chronic kidney disease (CKD).

In the same year, Davide Chicco et al.[5], discovered that machine learning models and analysis can be used in the field of heart failure and it happens when the heart becomes less able to effectively pump blood all over the body. This is the ultimate fate for many people. Heart failure affects about 2% of people in wealthy countries and jumps to 6%-10% among those aged 65 and above.

In the year 2021, Harshit Jindal et al.[8], used logistic regression and KNN to create a prediction system. The suggested model has demonstrated increased accuracy. Mohammed Khalid Hossen [12], finds that the accuracy rate of the logistic regression model will be 95.7% in 2022, indicating that machine learning algorithms will be considered a predefined method to seek cardiac illnesses in the near future.

1.2 Genesis of the Project

Cardiovascular diseases (CVD) affect numerous number of people and have even killed people worldwide. The presence of cardiovascular disease can be detected using machine learning by accounting for parameters such as chest pain, cholesterol levels, age and other parameters. When forecasting heart attacks, machine learning (ML) can be very helpful because it can take into account a variety of risk factors, such as high blood pressure, high cholesterol, an abnormal pulse rate, diabetes, etc. One type of machine learning is supervised learning, which makes it easy to find heart disease. Now, the project goal is to determine a person's risk of heart disease using machine learning methods in R software. As a result, it will be easier to provide patients with proper care while minimising harmful effects. Several people were affected by heart disease and the mortality rate increased because of that, we have chosen to study heart disease prediction and its root causes.

1.3 Data Collection

A organised dataset of individuals was chosen, accounting for their history of cardiovascular disease in addition to other medical conditions. Heart disease includes the various illnesses that harm the heart. According to the World Health Organisation (WHO), cardiovascular diseases are the main reason why people in their middle years pass away. We use a data source that contains the medical histories of 289 different patients from various age groups. In the study, we collected the data set from [13]. This data comprises 13 medical attributes and the last column represents output, which are age, sex, chest pain (cp), blood pressure

(trtbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic (restecg), heart rate (thalachh), angina (exng), thaillum stress test (thall), old peak, slope, number of major vessels and output. The last column target value denotes the patient's absence or presence of disease, expressed by a binary of 0 or 1 respectively. The dataset contains numeric values in csv file format.

age	sex	cp	trtbps	chol	fbs	restecg	thallach	exng	oldpeak	slp	caa	thall	output
35	1	2	130	250	0	1	187	0	3.5	0	0	2	1
43	1	0	120	177	0	0	120	1	2.5	1	0	3	0
60	1	0	145	282	0	0	142	1	2.8	1	2	3	0
76	0	2	140	197	0	2	116	0	1.1	1	0	2	1
37	0	2	120	215	0	1	170	0	0	2	0	2	1
40	1	0	110	167	0	0	114	1	2	1	0	3	0
55	1	1	130	262	0	1	155	0	0	2	0	2	1
44	1	0	112	290	0	0	153	0	0	2	1	2	0

Table 1.1: Sample Data

Chapter 2

METHODS AND METHODOLOGY

The approaches utilized in machine learning research are described in this section, along with the techniques that were employed. A discussion of the ML approach and its associated difficulties is followed by a description of a few methods.

In descriptive statistics, basic mean, median, skewness, kurtosis and basic data visualization are calculated. The accuracy measures of various classification techniques, including LDA, CART, KNN, SVM, Naive Bayes classifier and Random Forest are compared.

First, we choose a dataset and divide it into training 80% and testing 20% datasets. Next, we have to remove every null value from the data set. The dataset was then visualised using simple descriptive statistics. Then, we used many algorithms in R software, including LDA, CART, KNN, SVM, Naive Bayes and random forest. Then, we assessed the performance of each technique and based on the following performance, we evaluated the accuracy and based on that outcome is displayed.

2.1 Support Vector Machine (SVM)

One of the most popular supervised learning algorithms is the Support Vectotr Machine, which may be used to solve regression as well as classification problems. The SVM algorithm's objective is to identify the best decision line that can divide the space in n dimensions into subspaces and immediately classify the input points. The aim is to obtain the highest marginal hyperplane. This is mostly used for handwriting detection[9].

2.2 Classification and Regression Tree (CART)

A variant of the decision tree method is called CART (Classification and Regression Tree). In order to achieve this, CART algorithm looks for the subnodes best homogeneity and the root node is used as the training set. The CART algorithm, which is used in Machine learning, demonstrates how the target variable's values can be anticipated based on other variables.

2.3 K-Nearest Neighbour (KNN)

The K-nearest Neighbours algorithm employs proximity to generate classifications or forecasts about how a certain data point will be classified. It is a supervised learning technique. It is mostly used for non-parametric classifier and it can be called as lazy algorithm because we have trained it many times. While it can be utilised to address both classification and regression issues. In terms of detection and recognition, it is very expensive and KNN works best when there are a few attributes[9].

2.4 Naive Baiyes (NB)

The Naive Bayes classifier, a supervised machine learning approach and is used for tasks involving classification, like text classification. It is the simplest and most efficient of the classification techniques and it derives from the Bayes theorem. The Spam filtration and article classification mainly use naive bayes. An extremely fast and scalable model can be constructed and scored using the naive bayes method.

2.5 Linear Discriminant Analysis (LDA)

The term "linear discriminant analysis" refers to a technique for minimising the number of dimensions. Ronald A. Fisher developed the original method in 1936 and it functions as a pre-processing stage in applications for machine learning and pattern categorization. It resolves two problems with classification and is frequently used for reduction. As a dimensionality reduction approach, LDA can be used to decrease the number of features in a dataset.

2.6 Random Forest (RF)

Applications of random forests include classification and regression. It uses the greater part of the decision tree for classification and it uses the averages of every decision tree for regression. This model builds decision trees based on data samples, then gets the forecast from each one until choosing the best choice that offers the maximum accuracy compared to other decision tree techniques. The accuracy is maintained even in the absence of missing data.

2.7 Decision Tree

The approach of supervised learning known as a decision tree is frequently used to solve categorization issues. It is a classifier that uses a tree-like structure, with internal nodes representing the features of a dataset, branches representing the decision-making process and every node in the leaf representing the classification outcome. Root nodes are used to produce decisions and have many branches, whereas leaf nodes are the outcomes of decisions and it doesn't have additional branches.

Chapter 3

RESULTS AND DISCUSSION

3.1 Descriptive statistics

Descriptive statistics is the study of numerical and graphical methods to characterise and present your data. In this, we have described the minimum value for all attributes and the maximum value for all attributes. The first quartile is 1/4 th value of each attribute, third quartile is 3/4th of the value of each attribute and the median is the middle value in attributes (ex: age starts from min 29 to max 77 and its mid value is 54).

The mean is the average of attributes (ex: the mean of age is 54) and the range was calculated by maximum minus minimum (ex: for age, the max value is 77 and the min value is 29, range = 77-29 = 48, the range of age is 48). Further, standard deviation is high means risk is high , if it is low means risk is low .

Skewness is a measure of the symmetry or asymmetry of the distribution of data. In normal distribution, kurtosis determines whether the data is heavy- or light-tailed and it can be positive skewed (curve pushed towards the right side) or negative skewed (curve pushed towards the left side).

Responent No	age	sex	cp	trtbps	chol	fbs	restecg
Minimun	29	0	0	94	126	0	0
First Quartile	47	0	0	120	212	0	0
Median	54	1	1	130	243	0	1
Mean	54	0.6782	1.021	131.4	248	0.1453	0.5156
Third Quartile	60	1	2	140	276	0	1
Maximum	77	1	3	200	564	1	2
Range	48	1	3	106	438	1	2
Standard Deviation	9.13	0.47	1.03	17.52	51.60	0.35	0.51
Skewness	-0.14	-0.76	0.40	0.73	1.17	2.00	0.09
kurtosis	-0.56	-1.43	-1.23	0.97	4.55	2.02	-1.63

Responent no	thalachh	exng	oldpeak	slp	caa	thall
Minimun	71	0	0	0	0	0
First Quartile	136	0	0	1	0	2
Median	154	0	0.600	1	0	2
Mean	150.2	0.3183	1.008	1.419	0.7128	2.315
Third Quartile	168	1	1	2	1	3
Maximum	202	1	6.200	2	4	3
Range	131.0	1.0	6.2	2.0	4.0	3.0
Standard Deviation	22.90	0.47	1.13	0.61	1.02	0.60
Skewness	-0.56	0.78	1.30	-0.54	1.35	-0.44
kurtosis	-0.10	-1.40	1.79	-0.63	0.93	0.42

Table 3.1: Descriptive Statistics of Overall Data

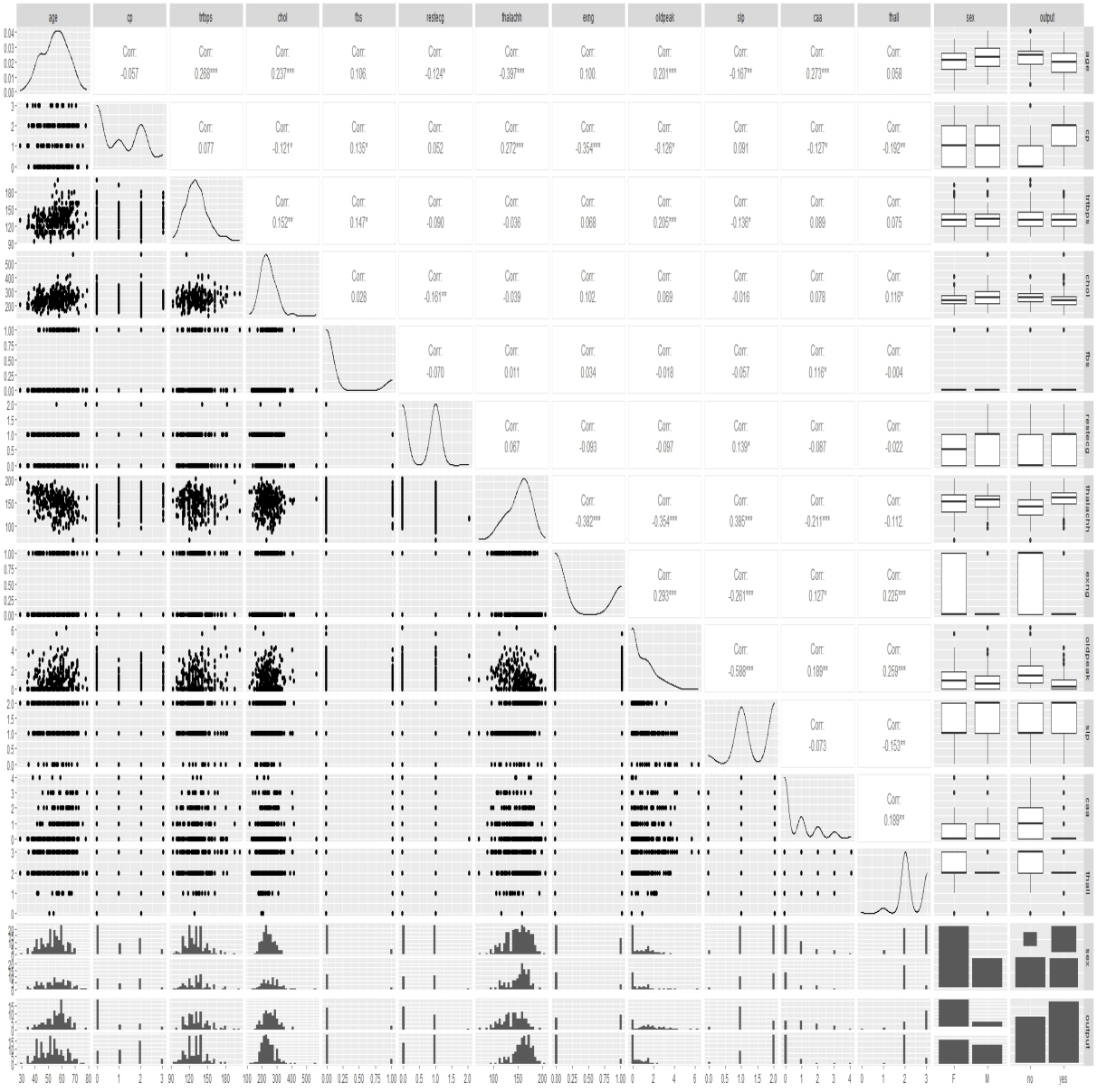


Figure 3.1: Descriptive statistics

In order to provide us with some of the evidence that was in the data, we first generated the descriptive statistics (Table 3.1).

In the above (Figure 3.1) scattered plot is displayed on the left side of the

plot and correlation values are displayed on the right side. Females have higher heart disease rates than males and dotted lines represent scattered plots.

Out of 289 patients, 165 are at risk of heart disease, where 93 are female (32%) and 72 are male (24.91%).

Further, age increases then risk of heart disease increases and cholesterol is positively correlated at 0.237, so this attribute is important for risk factors. The trtbps is blood pressure, is positively correlated with 0.268. The thalachh is heart rate is positively correlated with 0.272. The old peak is positively correlated with 0.201. The chest pain, thallium stress test and slope are negatively correlated.

3.2 Support Vector Machine (SVM)

The non-linear support vector machine is used for supervised learning. The kappa value is 0.5408 and the sensitivity for svm is 0.7200 and its specificity is 0.8182, respectively. The accuracy of the test data is 0.7759 with a confidence interval of (0.6859, 0.9013). The outcome is displayed in (Fig.3) and a confusion matrix is given below.

	Reference	
prediction	No	Yes
No	18	6
Yes	7	27

3.3 Classification and Regression Trees (CART)

The error at the root node for the training data is 124/289 = 0.42907 and misclassification rate of cart is 1-0.9394 = 0.0606. The accuracy for cart is 0.9138

and its kappa value is 0.8234. The positive predicted value is 0.8800 and negative predicted value is 0.9394. The outcome is displayed in (Figure 3.2). This section includes the confusion matrix.

	Reference	
prediction	No	Yes
No	22	3
Yes	2	31

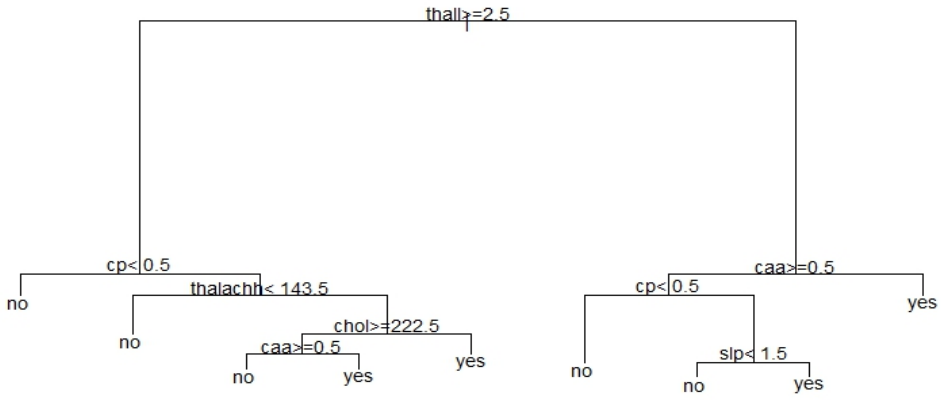


Figure 3.2: CART

3.4 K-Nearest Neighbours (KNN)

The ROC, sensitivity and specificity for different values of k in the KNN model for the training data set are obtained and shown in (Table 3.2). The square root of the training data value k = 11 is the final value with an accuracy of 0.7931 and

95% confidence interval of (0.6665, 0.8883) and its balanced accuracy is 0.7794. The outcome was displayed in (Fig-1) and this section includes the confusion matrix.

	Reference	
prediction	No	Yes
No	17	4
Yes	8	29

k	ROC	Sensitivity	Specificity
5	0.8811427	0.7477778	0.9069597
7	0.8859290	0.7377778	0.9199634
9	0.8936121	0.7381481	0.9102564
11	0.9010969	0.7218519	0.9228938
13	0.9012098	0.7007407	0.9373626
15	0.9051272	0.6981481	0.9521978
17	0.9078449	0.6811111	0.9523810
19	0.9076496	0.6866667	0.9547619
23	0.9149746	0.6903704	0.9549451

Table 3.2: ROC, sensitivity and specificity for training data

3.5 Naive Bayes Classifier (NB)

The naive bayes method is a supervised learning. The testing accuracy is 0.7759, the sensitivity is 0.7600 and specificity is 0.7879. The balanced accuracy for naive bayes is 0.7739.

In the naive bayes algorithm attributes thallach, caa, oldpeak, thall and cp are all important attributes and restecg, chol, trtbps and fbs are the least important attributes. This section includes the confusion matrix and importance graph (Figure 3.3).

	Reference	
prediction	No	Yes
No	19	7
Yes	6	26

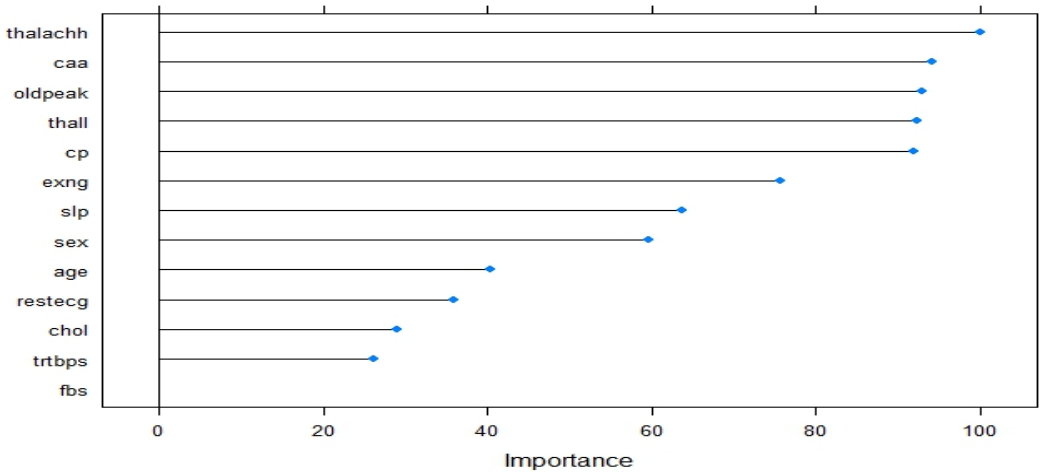


Figure 3.3: Naive Bayes

3.6 Linear Discriminant Analysis (LDA)

A classification method is linear discriminant analysis. The test accuracy is 0.7586 with a kappa value of 0.5031 and its confidence interval is (0.6283, 0.8613) and its balanced accuracy is 0.7491. The output is displayed in (Fig-4 and 3.4).

This section includes the confusion matrix.

	Reference	
prediction	No	Yes
No	17	6
Yes	8	27

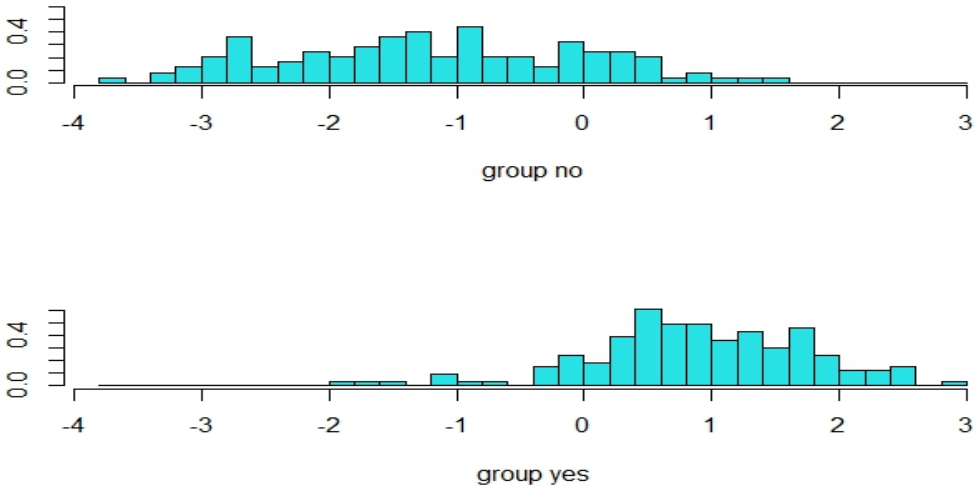


Figure 3.4: LDA Plot

3.7 Random Forest (RF)

Out-of-Bag (OOB-misclassification rate) is 17.3% and it accurately predicts 95 people without heart disease with class error 0.2338710 and 144 patients with heart disease correctly with class error 0.1272727. The accuracy for random forest is 0.8339. The sensitivity is 0.833 and specificty is 0.8333. Its ROC curve is displayed in (Fig-2). This section includes the confusion matrix.

prediction	Reference	
	No	Yes
No	95	29
Yes	19	146

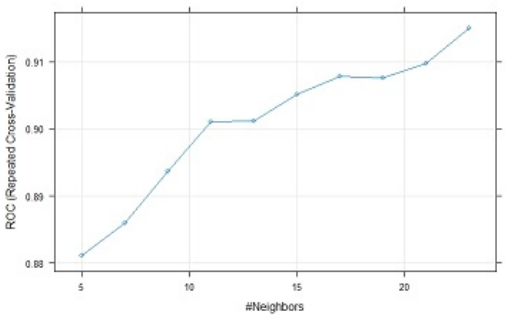


Fig 1: KNN

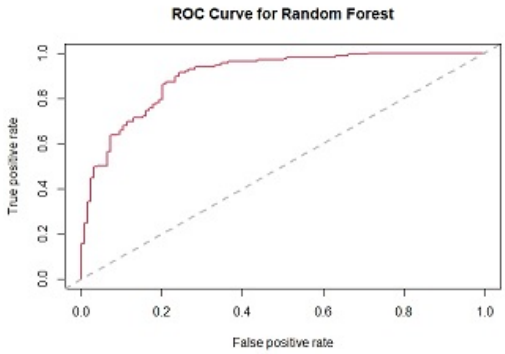


Fig 2: RF

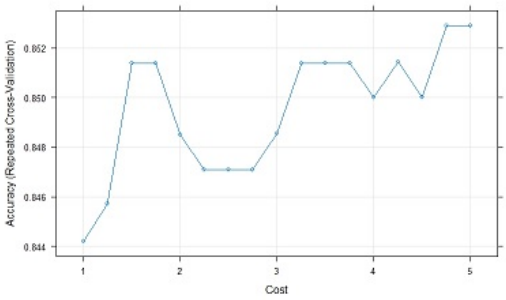


Fig 3: SVM

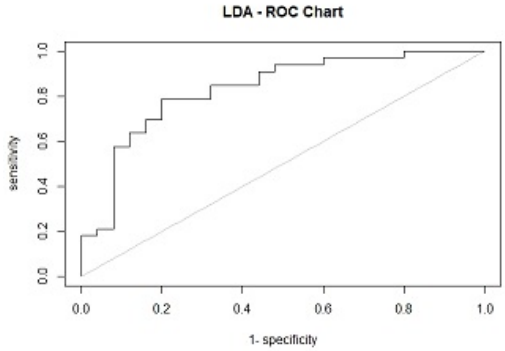


Fig 4: LDA

3.8 Decision Tree

The Method of supervised learning Problems involving regression and classification can be solved with decision trees. It has two nodes: a branch node

and a leaf node. Additionally, it maintains accuracy even when a high proportion of the data is missing and is very effective at estimating missing data.

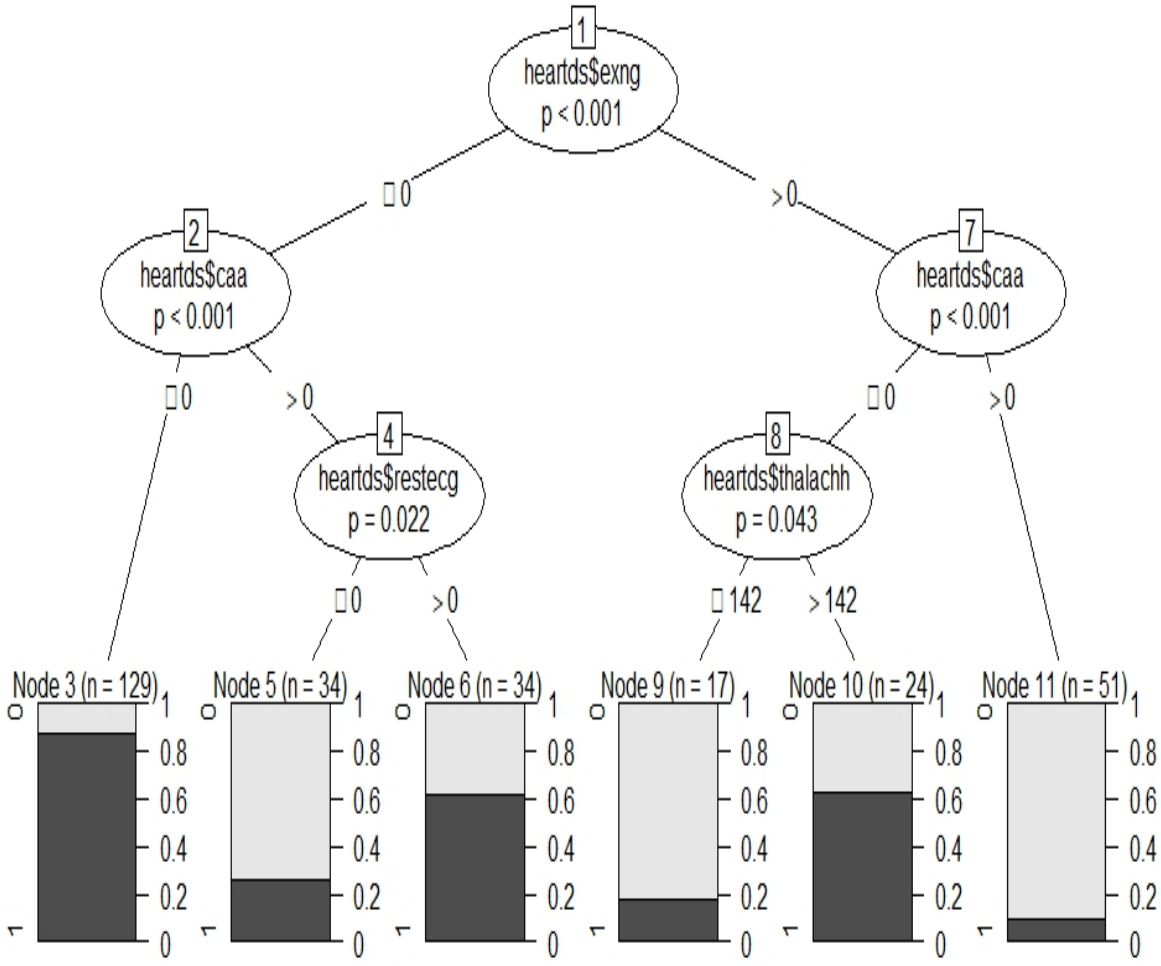


Figure 3.5: Decision tree

3.9 Effect Plots

This model effect map displays the anticipated probabilities for output for each gender, with males having the highest expected probabilities of risk with 0.8451681 and females having the lowest expected probability of risk with 0.4274244. The likelihood of risk for cp ranges from 0.366454 to 0.8893285 and varies from 0 to 3. The trtbps impact plot shows that risk is negatively correlated with trtbps and the range of the likelihood of risk from 0.7510306 to 0.2610423 is from 94 to 200 and The effect plot, which shows that risk is inversely related to cholesterol and that varies from 100 to 600, shows that the likelihood of risk for cholesterol is 0.8017338 to 0.1044568.

The probability of risk for fbs ranges from 0.5902977 to 0.5611900, where risk is constant and the likelihood of risk for restecg ranges from 0.5011814 to 0.7918080, where risk is directly correlated to restecg and ranges from 0 to 2. The range of the risk probability for thalachh is 0.2097212 to 0.8021313 and the risk is directly correlated to thalachh. The impact plot for exng from 0.6677334 to 0.4008912 varies from 0 to 1 and exng is inversely correlated with risk.

The likelihood of risk ranges from 0.73164318 to 0.05225514 and varies from 0 to 6, according to the impact plot for oldpeak and risk is inversely correlated to oldpeak and the probability of risk for slp ranges from 0.4629457 to 0.6344278 varies from 0 to 2 and hazard is directly correlated to slp. The expected value of risk, which ranges from 0.7254944 to 0.0737926 goes from 0 to 4, according to the impact plot for caa and risk is inversely related to caa and risk is inversely related to thall and has an expected value that ranges from 0 to 3 for values between 0.9070780 and 0.4443529.

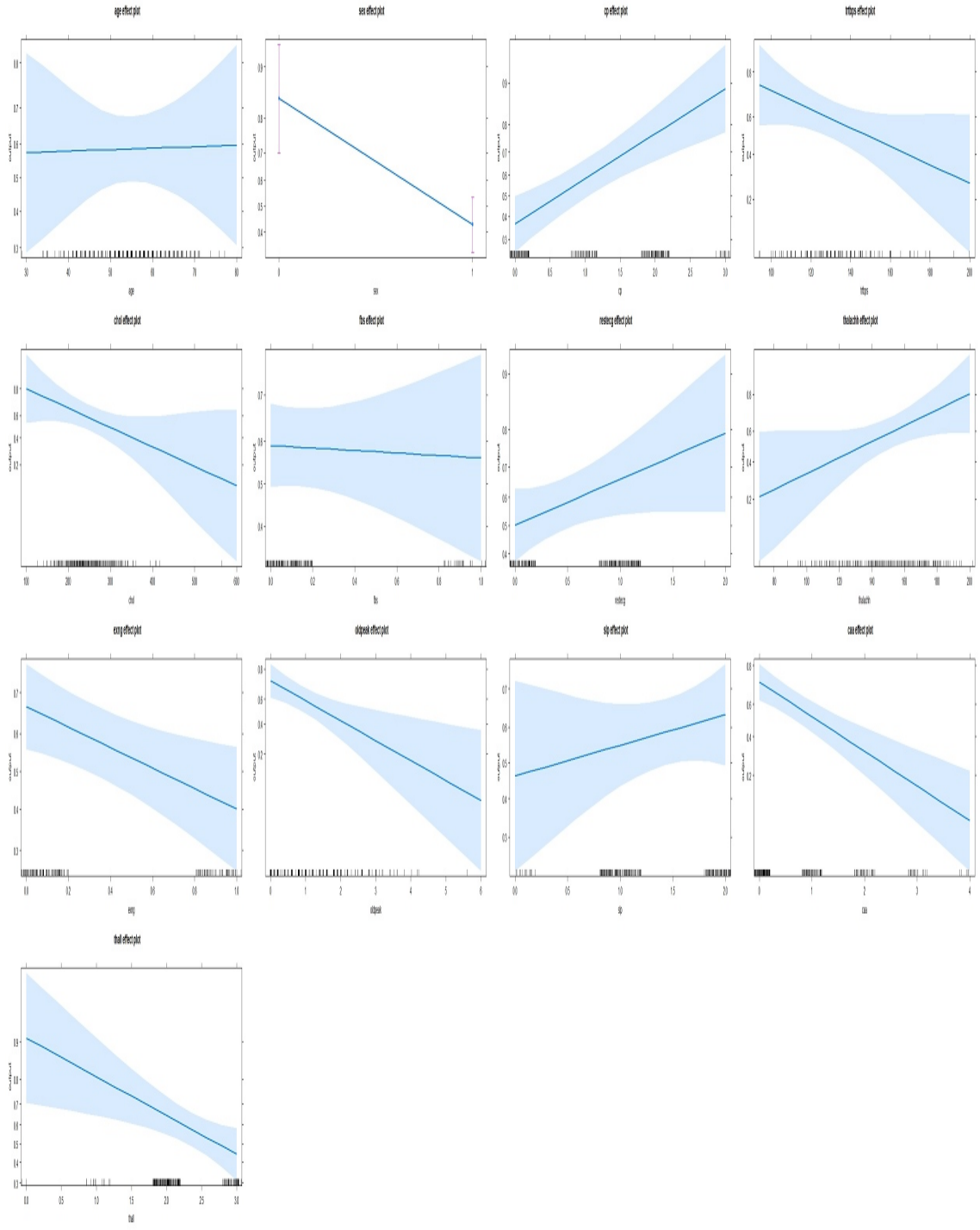


Figure 3.6: Effect plots

Chapter 4

RESULTS ON PARAMETER REDUCTION

Dimensionality reduction is another name for parameter reduction. Data visualisation typically makes use of dimensionality reduction techniques, which aim to reduce the number of characteristics in a dataset. It is mostly used because large number of dataset will perform poorly in machine learning. Using soft sets, the number of parameters is reduced and a core parameter is found for simple processing. By performing feature selection to eliminate "irrelevant" features that are not important to the classification problem and by removing the least significant variables from the model, dimensionality reduction reduces the number of variables.

We have taken values a,b,c,d,e,f,g,h,i,j,k and l as age, sex, chest pain (cp), blood pressure (trtbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic (restecg), heart rate (thalach), angina (exng), thaillum stress test (thall), old peak, slope and number of major vessels (caa).

In this, we have done a parameter reduction using soft sets to analyse all different parameters by hybrid approch and they are displayed as possible parameters related to heart disease and it will be compared with all accuracy to get the best parameter.

S.No	No.of parameter	parameter	Accuracy					
			NB	CART	KNN	SVM	LDA	RF
1	12	abcdefghijkl	0.7759	0.91.38	0.7931	0.7759	0.7586	0.8339
2	8	adefijl	0.8276	0.8103	0.569	0.7241	0.7241	0.7759
3	6	adefil	0.8103	0.8276	0.5517	0.8103	0.7759	0.8448
4	6	abcejl	0.7586	0.7586	0.6552	0.8276	0.7931	0.8448
5	6	abdehl	0.7069	0.8103	0.6379	0.7069	0.7586	0.7241
6	6	abdejl	0.6552	0.7414	0.569	0.7069	0.7414	0.7759
7	6	abdekl	0.7586	0.7931	0.569	0.7759	0.7931	0.7759
8	6	abdhil	0.7931	0.7759	0.6379	0.7586	0.7414	0.7931
9	6	abdijl	0.7586	0.7414	0.5	0.7414	0.7586	0.7759

S.No	No.of parameter	parameter	Accuracy					
			NB	CART	KNN	SVM	LDA	RF
10	6	abdikl	0.7241	0.7759	0.5172	0.8448	0.7586	0.8448
11	6	acdehl	0.7241	0.7414	0.6379	0.7759	0.7414	0.7759
12	6	acdeil	0.7759	0.7241	0.569	0.7414	0.7586	0.7069
13	6	acdekl	0.7414	0.7759	0.569	0.7414	0.7931	0.8103
14	6	acdhil	0.7414	0.7414	0.6207	0.7931	0.7586	0.7931
15	6	acdijl	0.7586	0.7586	0.5172	0.7241	0.7241	0.7931
16	6	acdikl	0.7759	0.7931	0.5862	0.8103	0.7759	0.8276
17	6	acefjl	0.7241	0.7586	0.6379	0.7586	0.7069	0.8448
18	6	acegil	0.7759	0.7586	0.6552	0.8103	0.7414	0.8621

S.No	No.of parameter	parameter	Accuracy					
			NB	CART	KNN	SVM	LDA	RF
19	6	acehil	0.7759	0.7414	0.6552	0.8103	0.7586	0.8276
20	6	acehjl	0.7414	0.7586	0.6552	0.7241	0.6897	0.7759
21	6	aceijl	0.7241	0.7586	0.6552	0.8103	0.7414	0.8621
22	6	aceikl	0.8103	0.7931	0.6207	0.7759	0.7759	0.8793
23	6	acejkl	0.7586	0.7586	0.6379	0.7931	0.7241	0.8448
24	6	aeghil	0.8103	0.7759	0.6379	0.7931	0.7414	0.7931
25	6	aegijl	0.7586	0.7414	0.6379	0.7931	0.7759	0.8448
26	6	aegikl	0.7586	0.8276	0.6207	0.8103	0.7759	0.8448
27	5	bcejl	0.7759	0.7586	0.6379	0.8276	0.7931	0.7931
28	5	acejl	0.7586	0.7586	0.6552	0.8103	0.7241	0.8448

Table 4.1: Overall Accuracy for All Parameters

From the above (Table 4.1) is the values of accuracy. We have done a parameter reduction using soft sets of that have got 27 parameters as output that are adefijl, adefil, acdehl, aeghil, acejkl, aceikl, aceijl, acehjl, acehil, acegjl, abdikl, abdijl, acefjl, abdhil, acdikl, acdijl, acdhil, abcejl, acdekl, acdejl, aegikl, aegijl, bdekl, abdejl, abdehl, bcejl and acejl. Using that, we have calculated accuracy for all 27 parameters. On the basis of the highest accuracy of all 27 parameters, we choose adefil as the best parameter.

The adefil represents age(a), trtbps(d), chol(e), fbs(f), exng(i) and caa(l) and these attributes give the best accuracy to predict heart disease. So, it is considered the best parameter.

prediction	Reference											
	NB		CART		KNN		SVM		LDA		RF	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No	15	1	23	8	11	12	17	3	16	4	20	4
Yes	10	32	2	25	14	21	8	30	9	29	5	29

Table 4.2: Confusion matrix for best parameter

In the above confusion matrix (Table 4.2) Naive Bayes has predicted a low false positive rate, likewise KNN has predicted a high false positive rate. The true positive rate for CART is high. likewise, for KNN it has a low true positive rate. The CART has a low number of false negative rate . The KNN has in high number of predicting false negative rate . For, predicting the true negative rate, naive bayes predicted as higher and CART has predicted low in number.

Method	Accuracy	95%C.I	No Information rate	p-value	Kappa
NB	0.8103	(0.6859, 0.9013)	0.569	$9.434e^{-}0.5$	0.5957
CART	0.8276	(0.7057, 0.9141)	0.569	$2.735e^{-}05$	0.6584
KNN	0.5517	(0.4154, 0.6826)	0.569	0.6563	0.0771
SVM	0.8103	(0.6859, 0.9013)	0.569	$9.434e^{-}05$	0.6037
LDA	0.7759	(0.6473,0.8749)	0.569	0.000826	0.5317
RF	0.8448	(0.7258, 0.9265)	0.569	$7.081e^{-}06$	0.6821

Method	Mcnermar's Test	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
NB	0.01586	0.6000	0.9697	0.9375	0.7619
CART	0.1138	0.9200	0.7576	0.7419	0.9259
KNN	0.8445	0.4400	0.6364	0.4783	0.6000
SVM	0.2278	0.6800	0.9091	0.8500	0.7895
LDA	0.267257	0.6400	0.8788	0.8000	0.7632
RF	1	0.8000	0.8788	0.4783	0.6000

Method	Prevalence	Detection Rate	Detection Prevalence	Balanced Accu	Positive Class
NB	0.4310	0.2586	0.2759	0.2586	NO
CART	0.4310	0.3966	0.5345	0.8388	NO
KNN	0.4310	0.1897	0.3966	0.5382	NO
SVM	0.4310	0.2931	0.3448	0.7945	NO
LDA	0.4310	0.2759	0.3448	0.7594	NO
RF	0.4310	0.3448	0.4138	0.8394	NO

Table 4.3: Total value for best parameter

In the above (Table 4.3) we used the following methods Naive Bayes, classification and regression tree (CART), K-nearest neighbour (KNN), support vector machine (SVM), linear discriminant analysis (LDA) and random Forest (RF). In this we calculated an accuracy for which is RF gives the best accuracy of 0.8448 and KNN gives the worst accuracy of 0.5517. The confidence interval is calculated to check whether the accuracy calculated was correct or not and the no information rate is for the hypothesis test to calculate the overall accuracy if it should be greater than the no information rate. The probablity value and

kappa value are calculated and represent the strength of agreement.

For the classification problem, we used Mcnermar's test and Cart gives the best sentivity, the true positive rate is 0.9200 and the best specificity is NB. The positively predicted value, which is actual positive, has the highest value of 0.9375 as NB and the lowest value of RF as 0.4783, the negatively predicted value is actual negative, has the highest value. The KNN and RF has lowest value.

The probability of having a disease is prevalance and it can be calculated by actual positive divided by all values in the confusion matrix. Then, the detection rate is a sample of proportions detected correctly. The term balanced accuracy means balanced data in the dataset and RF has the highest balanced accuracy of 0.8394 and we have taken the positive class as No for all methods.

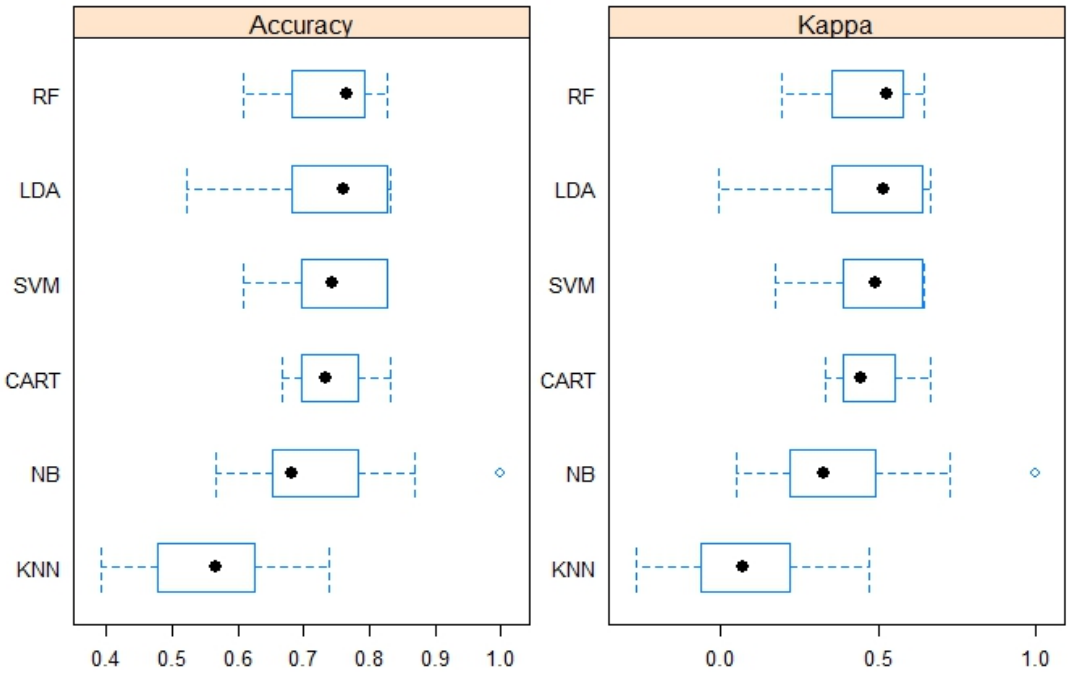


Figure 4.1: Comparison of accuracy and kappa value

In the overall comparison, the accuracy for support vector machine (svm) with 14 parameters is 0.7759 and the accuracy for 6 parameters (i.e., adefil) is 0.8103. The accuracy for adefil is increasing faster than the other 14 parameters. Likewise, the linear discriminant analysis (LDA) algorithm for 14 parameters is 0.7586 and the reduced parameter (i.e., adefil) is 0.7759, which is slightly higher than the original parameter. The accuracy for random forest (RF) in 14 parameters is 0.8339, the accuracy for reduced parameters is 0.84448 and the accuracy for random forest in reduced parameters is increasing. Likewise, the accuracy for k-nearest neighbour for 14 parameters is 0.7931 and for reduced parameters, it is 0.5517 and its accuracy is decreased compared to 14 parameters. Then the accuracy for classification and regression tree (CART) for 14 parameters is 0.9138 and for reduced parameters (i.e., adefil). It is 0.8276, which is lower than 14 parameters. Likewise, accuracy for the naive bayes (NB) algorithm for 14 parameters is 0.7759 and accuracy for reduced parameters is 0.8103, which is higher than 14 parameters.

Chapter 5

CONCLUSION

In this project, machine learning techniques were used to calculate accuracy. The output prediction accuracy using machine learning algorithms provides us with knowledge about the optimal method to use and estimating future output predictions based on historical data will always be beneficial. As a result of the comparison, classification and regression tree provides the best accuracy of 0.9193, while the k-nearest neighbour method provides the worst accuracy of 0.5517 and best parameter is adefil. Therefore, the analysis described above will be useful for future forecasting approaches on the basis of which decisions may be made about how to help people stay out of a risk zone.

References

[1] Abdul Saboor et al, A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms, Hindawi, Article ID 1410169, 2022.

[2] M.Akhil jabbar, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, International Conference on Computational Intelligence: Modeling Techniques and Applications, 2013.

[3] Amanda H. Gonsalves et al, Prediction of Coronary Heart Disease, 2019.

[4] Anupama Yadav, Levish Gediya, Adnanuddin Kazi, Heart Disease Prediction Using Machine Lerning, 2021.

[5] Davide Chicco et al, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, BMC Medical Informatics and Decision Making, 2020

[6] Dejun Zhang, Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer, IEE Access, 2018.

[7] Edwar Macias et al, Mortality prediction enhancement in end-stage renal disease: A machine learning approach, Informatics in Medicine Unlocked, 2020.

[8] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath, Heart disease prediction using machine learning algorithms, 2021.

[9] K. Kannan, A. Menaga, Risk Factor Prediction by Naive Bayes Classifier, Logistic Regression Models, Various Classification and Regression Machine Learning Techniques, 2021

[10] Monther Tarawneh, Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques, ACTA SCIENTIFIC NUTRITIONAL HEALTH, 2019.

[11] Mamatha Alex P and Shaicy P Shaji, "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique", International Conference on Communication and Signal Processing, 2019.

[12] Mohammed Khalid Hossen, Heart Disease Prediction Using Machine Learning Techniques, American Journal of Computer Science and Technology, 2022.

[13] Neha Nandal et al, Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis, 2022.

[14] A.Sankari Karthiga et al, Early Prediction of Heart Disease Using Decision Tree Algorithm, International Journal of Advanced Research in Basic Engineering Sciences and Technology, 2017.

[15] Sonam Nikhar et al, Prediction of Heart Disease Using Machine Learning Algorithms, International Journal of Advanced Engineering, Management and Science (IJAEMS), ISSN : 2454-1311, 2016.

[16] Sophie H Bots et al, Sex differences in coronary heart disease and stroke mortality, 2017.

[17] C.Zoccali, Traditional and emerging cardiovascular and renal risk factors: An epidemiologic perspective, International Society of Nephrology, 2006.

[18] https://www.r-project.org/about.html

[19] https://www.healthline.com/health/heart-disease/history

[20] https://www.nhsinform.scot/illnesses-and-conditions/heart-and-blood-vessels/conditions/common-heart-conditions.

MSC PROJECT 22-23 KOWSALYA M

by Kannan K

Submission date: 29-May-2023 06:24AM (UTC+0530)
Submission ID: 1165596830
File name: 223045018_kowsalya1m.pdf (2.26M)
Word count: 6247
Character count: 30368

HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Thesis submitted to the SASTRA Deemed to be University
in partial fulfillment of the requirements
for the award of the degree of

M.Sc. Mathematics

Submitted by

M. KOWSALYA
(Reg. No.: 223045018)

MAY 2023

DEPARTMENT OF MATHEMATICS

SRINIVASA RAMANUJAN CENTRE

KUMBAKONAM, TAMIL NADU, INDIA – 612 001

Chapter 1

INTRODUCTION

Heart disease is the most deadly disease in the world and one of them is a heart attack. It has become the most common disease among any age group in the past few decades. It's also known as an acute myocardial infarction (AMI) and is one of the most dangerous conditions in the cardiovascular disease category. When blood flow to heart is impeded or blocked, a heart attack happens. The blood clotting is usually caused by the buildup of cholesterol and other substances in the arteries. Plaque is fat and contains cholesterol. Many heart diseases are on the rise, which can lead to lifestyle changes, work stress and poor diet.

According to estimation, 17.9 million individuals died from cardiovascular causes in 2019, which accounted for 32% of all fatalities worldwide. Many factors, including heredity, age, gender, tobacco use, inactivity, autoimmune disorders, high blood pressure, alcohol consumption, stress level, cholesterol and existing heart problems, Physiological signs may affect the prevalence of cardiac illnesses around the world[3].Researchers are working to create an efficient method for the timely detection of cardiac diseases because current methods for diagnosing heart illness are ineffective at early detection for a number of reasons, including accuracy and computing time. Predicting cardiovascular Disease is challenging, especially in developing nations. where there is a lack of trained medical workers,

diagnostic tools and additional resources needed to identify and treat persons with heart conditions [1].

The body experiences a variety of symptoms, including weariness and shortness of breath, if the heart's pumping motion is ineffective and the heart muscle is unable to supply the blood and oxygen the body needs. As your heart is not functioning properly, this condition is known as heart failure. While treatments, including surgery, typically cannot cure heart failure, they can lead to longer lifespans and fewer negative consequences for individuals. Heart bypass surgery is also known as coronary artery bypass grafting (CABG). Surgery is done to treat clogged or constricted arteries supplying your heart muscle. This major heart procedure is frequently done to treat patients with angina and coronary artery disease.[20]

A language and environment for statistical computing and graphic design is called R. and it is effective in data analysis. The R language has gained popularity because it makes it simple for researchers to include many machine learning approaches into a single programme. Also, it gives researchers an easy way to exchange codes.[18].R packages are collections of assembled code, test data, and R functions. In the R environment, packages are kept in a directory called library. There are numerous bundles of packages available. For example, Over 10,000 packages for the R programming language are available in the CRAN repository, as well as the amount of packages is constantly increasing.

The field of machine learning is extremely broad and diverse, with its applications and breadth expanding day by day. Artificial intelligence's machine learning field employs algorithms to harvest data and then forecast future trends. In the current data science era, machine learning algorithms are continually being

employed across many fields to obtain insightful understanding and use the data to make judgments. In this period of time, data is necessary for industries and organizations. Nowadays, data stored in hospitals is in electronic format and contains information like medical history, mortality rate and symptoms. To forecast and assess the reliability of the provided data, Machine learning uses a range of classifiers, including supervised, unsupervised and ensemble learning. Machine learning has the potential to increase accuracy by taking advantage of complicated connections between risk factors [8].

Machine learning (ML) is a sort of artificial intelligence (AI) which enables AI programmes to forecast outcomes more correctly without being given specific guidance to do so. Machine learning techniques employ historical data as input to predict new output values. Due to the fact that machine learning is typically a better predictor than traditional approaches, machine learning prediction is chosen because it employs algorithms. Machine learning is able to spot patterns and connections that people cannot.

The supervised learning uses labelled datasets to train algorithms to correctly categorise data or predict outcomes and Until it is well fitted, the model modifies its weights. This occurs as part of the review process to prevent the model from becoming overfit.

A heart disease diagnosis can be made using machine learning technologies before a person sustains serious injury. The science and technology field of machine learning is expanding and it may be used to diagnose cardiac problems. [4]. The most popular strategy for predicting heart attacks is the naive bayes method. Data is gathered from a variety of sources, After being properly categorised and then evaluated to gather the required data. We can conclude that machine

learning is particularly well suited to heart disease prediction.

1.1 Literature Review

At the Johns Hopkins Hospital, William Osler (1849–1919), chief physician and professor of clinical medicine, conducted considerable research on angina and was among the first to propose that it was more of a syndrome than a particular disease. According to the Association for Cardiovascular Angiography and Treatments, in the 1960s and 1970s, therapies for heart disease included bypass surgery and percutaneous balloon angioplasy[19].

R was created in 1993 by Ross Ihaka and Robert Gentleman and includes linear regression, machine learning techniques, statistical inference, time series, and other features.

Frank Rosenblatt was a great psychologist known for his work on machine learning. He invented the perceptron, a machine learning method, in 1957. The Perceptron algorithm was one of the first to use artificial neural networks, which are now widely used in machine learning.

Zoccali [17] in 2006, patients with end-stage kidney disease (ESKD) have a very significant mortality risk and also cardiovascular disease and both of these outcomes are significantly related to renal function and in populations with heart problems.

Many studies and beneficial investigations have been conducted in relation to the medical profession approach for predicting diseases using machine learning models. A technique that emphasizes the value of feature selection in cardiac disease prediction was put forward by M.A. Jabbar et al.[2] in 2013, For feature selection, they used a genetic algorithm and for KNN later. They got good

results. For the purpose of predicting cardiac disease, several researchers have also used deep learning algorithms.

Sonam Nikhar et al.[15], in 2016 this investigation aims to provide a in-depth analysis of the Naive Bayes and decision tree classifiers used in our evaluation, especially in the diagnosis of cardiac problems. The results of certain studies that Decision trees outperform Bayes classification systems when future data mining techniques are applied to a comparable dataset.

Using a public dataset of 573 records, Karthiga et al. [14], successfully completed study to forecast the development of coronary artery disease in 2017. The authors processed the dataset. using the decision tree and Naive Bayes classifier by MATLAB data analysis tool, and then they generated accuracy results to evaluate the performance of the models. According to the results reported, decision trees have higher accuracy than naive bayes.

In the same year, Sophie H Bots et al.[16], While fatalities from heart disease and stroke declined dramatically for both genders between 1980 and 2010, Some data suggested that men experienced more age-specific reductions than women. Males continue to die from CHD and stroke more frequently than women their old age.

Dejun Zhang et al. (2018) used principal component analysis for the ADABOOST algorithm to predict clinical outcomes in breast cancer. In a study to predict heart disease, Tarawneh et al.[10] in 2019, used hybrid data mining classifier techniques. Datasets were retrieved from the UCI repository, which comprises 303 entries and 76 variables. For 14 attributes, model training and testing were done. The classification algorithms assessed the preciseness of the prediction of

heart disease include KNN, SVM, RF and NB. SVM and NB anticipated cardiac disease more accurately 89.2% and produced better outcomes.

In the year 2019, Mamatha Alex et al.[11], this study makes use of artificial neural networks, random forests, KNN and support vector machine techniques. In contrast to the previously mentioned classification methods for data mining, the artificial neural network predicts the highest accuracy for diagnosing heart disease.

Edwar Macias et al.[7] in 2020, researched a machine learning approach for enhancing fatality forecast in end-stage renal disease and while all the factors contributed to the highest performance, those discovered by random forest had more anticipating power than selected with knowledge from experts. The are some machine learning methods to treat chronic kidney disease (CKD).

In the same year, Davide Chicco et al.[5], Machine learning models and analysis can be discovered in the field of heart failure and it happens when the heart becomes less able to effectively pump blood all over the body. This is the ultimate fate for many people. Heart failure affects about 2% of persons in wealthy countries, and jumps to 6%-10% among those aged 65 and above.

In the year 2021, Harshit Jindal et al.[8], used logistic regression and KNN to create a prediction system. The suggested model has demonstrated increased accuracy. Mohammed Khalid Hossen [12], finds that the accuracy rate of the logistic regression model to be 95.7% in 2022, indicating that machine learning algorithms will be considered as a predefined method to seek cardiac illnesses in the near future.

1.2 Genesis of the Project

Cardiovascular diseases (CVD) affects numerous number of people and even killed people worldwide. The presence of cardiovascular disease can be detected using machine learning by accounting for parameters such as chest pain, cholesterol levels, age and other parameters. When forecasting heart attacks, machine learning (ML) can be very helpful because they can take into account a variety of risk factors, such as high blood pressure, high cholesterol, an abnormal pulse rate, diabetes, etc. One type of machine learning is supervised learning, which makes it easy to find heart disease. Now, the project goal is to determine a person's risk of heart disease using machine learning methods in R software. As a result, it will be easier to provide patients with proper care while minimising harmful effects. Several people were affected by heart disease and the mortality rate increased. Because of that, we have chosen to study heart disease prediction and its root causes.

1.3 Data Collection

A organised dataset of individuals was chosen, accounting for their history of cardiovascular disease in addition to other medical conditions. Heart disease includes the various illnesses that harm the heart. According to the World Health Organisation (WHO) cardiovascular diseases are the main reason why people in their middle years pass away. We use a data source that contains the medical histories of 289 different patients from various age groups. In the study, we collected the data set from [13]. This data comprises 13 medical attributes and last column represents output, which are age, sex, chest pain (cp),

blood pressure (trtbps), cholesterol (Chol), fasting blood sugar (Fbs), resting electrocardiographic (restecg), heart rate (Thalach), angina (Exng), thaillum stress test (Thall), old peak, slope, number of major vessels and output. The last column target value denotes the patient's absence or presence of disease, expressed by a binary of 0 or 1 respectively. The dataset contains numeric values in CSV file format.

age	sex	cp	trtbps	chol	fbs	restecg	thallach	exng	oldpeak	slp	caa	thall	output
35	1	2	130	250	0	1	187	0	3.5	0	0	2	1
43	1	0	120	177	0	0	120	1	2.5	1	0	3	0
60	1	0	145	282	0	0	142	1	2.8	1	2	3	0
76	0	2	140	197	0	2	116	0	1.1	1	0	2	1
37	0	2	120	215	0	1	170	0	0	2	0	2	1
40	1	0	110	167	0	0	114	1	2	1	0	3	0

Table 1.1: sample Data

Chapter 2

METHODS AND METHODOLOGY

The approaches utilized in machine learning research (MLR) are described in this section, along with the techniques that were employed. A discussion of the ML approach and its associated difficulties is followed by a description of a few methods.

In descriptive statistics, basic mean, median, skewness, kurtosis and basic data visualization are calculated. The accuracy measures of various classification techniques, including LDA, CART, KNN, SVM, Naive Bayes classifier and Random Forest are compared.

First choosing a dataset, we have dividing it into training80% and testing20% datasets. Next, we have to remove every value that is blank values from the data set. The dataset was then visualised using simple descriptive statistics. Then, we used many algorithms in R software, including LDA, CART, KNN, SVM, naïve bayes and random forest. Then, we assessed the performance of each technique and based on the Following performance, we evaluate the accuracy and based on that outcome is displayed.

2.1 Support Vector Machine (SVM)

One of the most popular supervised learning algorithms is the SVM and it may be used to solve regression as well as classification problems. The SVM algorithm's objective is to identify the best decision line that can divide the space in n dimensions into subspaces and immediately classify the input points. The aim is to obtain the highest marginal hyperplane. This is mostly used for handwriting dectection. [9].

2.2 Classification and Regression Tree (CART)

A variant of the decision tree method is called CART (Classification and Regression Tree). In order to achieve this, the CART algorithm looks for the subnodes best homogeneity and root node is used as the training set. The CART algorithm, which is used in ML, demonstrates how the target variable's values can be anticipated based on other variables.

2.3 K-Nearest Neighbour (KNN)

The K-nearest Neighbours algorithm employs proximity to generate classifications or forecasts about how a certain data point will be categorised. It is a supervised learning technique. It is mostly used for non-parametric classifier and it can be called as lazy algorithm because we have train it many times. While it can be utilised to address both classification and regression issues. In terms of detection and recognition, it is very expensive, and KNN works best when there are a few attributes [9].

2.4 Naive Baiyes (NB)

The Naive Bayes classifier, a supervised machine learning approach, is used for tasks involving classification like text classification. It is the simplest and most efficient for classification techniques and it derives from the Bayes theorem. Spam filtration and article classification mainly use Naive Bayes. An extremely fast and scalable model can be constructed and scored using the naive bayes method.

2.5 Linear Discriminant Analysis (LDA)

The term "linear discriminant analysis" refers to a technique for minimising the number of dimensions. Ronald A. Fisher developed the original method in 1936, and it functions as a pre-processing stage in applications for machine learning and pattern categorization. It resolves two problems with classification and is frequently used for reduction. As a dimensionality reduction approach, LDA can be used to decrease the amount of features in a dataset.

2.6 Random Forest (RF)

Applications of random forests include classification and regression. It uses the greater part of the decision tree for classification, and it uses averages of every decision tree for regression. This model builds decision trees based on data samples, then gets the forecast from each one until choosing the best choice that offers the maximum accuracy compared to other decision tree techniques. The accuracy is maintained even in lack of missing data.

2.7 Decision Tree

The approach of supervised learning known as a decision tree is frequently used to solve categorization issues. It is a classifier that uses a tree-like structure, with internal nodes representing the features of a dataset, branches representing the decision-making process, and every node in the leaf representing the classification outcome. Root nodes are used to produce decisions and have many branches, whereas leaf nodes are the outcomes of decisions and it doesn't have additional branches.

Chapter 3

RESULT AND DISCUSSION

3.1 Descriptive statistics

Descriptive statistics is the study of numerical and graphical methods to characterise and present your data. In this, we have described the minimum value for all attributes and the maximum value for all attributes. The first quartile is 1/4 th value of each attribute, third quartile is 3/4th of the value of each attribute and the median is the middle value in attributes (ex: age starts from min 29 to max 77 and its mid value is 54).

The mean is the average of attributes (ex: the mean of age is 54) and the range was calculated by maximum minus minimum (ex: for age, the max value is 77 and the min value is 29, range = 77-29 = 48, the range of age is 48). Further, to calculate the standard deviation is $\sigma = \sqrt{\frac{\sum x_i + \mu}{N}}$ then standard deviation is high means risk is high , if it is low means risk is low .

Skewness is a measure of the symmetry or asymmetry of the distribution of data. In normal distribution, kurtosis determines whether the data is heavy- or light-tailed and it can be positive skewed (curve pushed towards the right side) or negative skewed (curve pushed towards the left side).

Responent No	age	sex	cp	trtbps	chol	fbs	restecg
Minimun	29	0	0	94	126	0	0
First Quartile	47	0	0	120	212	0	0
Median	54	1	1	130	243	0	1
Mean	54	0.6782	1.021	131.4	248	0.1453	0.5156
Third Quartile	60	1	2	140	276	0	1
Maximum	77	1	3	200	564	1	2
Range	48	1	3	106	438	1	2
Standard Deviation	9.13	0.47	1.03	17.52	51.60	0.35	0.51
Skewness	-0.14	-0.76	0.40	0.73	1.17	2.00	0.09
kurtosis	-0.56	-1.43	-1.23	0.97	4.55	2.02	-1.63

Responent no	thalachh	exng	oldpeak	slp	caa	thall
Minimun	71	0	0	0	0	0
First Quartile	136	0	0	1	0	2
Median	154	0	0.600	1	0	2
Mean	150.2	0.3183	1.008	1.419	0.7128	2.315
Third Quartile	168	1	1	2	1	3
Maximum	202	1	6.200	2	4	3
Range	131.0	1.0	6.2	2.0	4.0	3.0
Standard Deviation	22.90	0.47	1.13	0.61	1.02	0.60
Skewness	-0.56	0.78	1.30	-0.54	1.35	-0.44
kurtosis	-0.10	-1.40	1.79	-0.63	0.93	0.42

Table 3.1: Descriptive Statistics of Overall Data

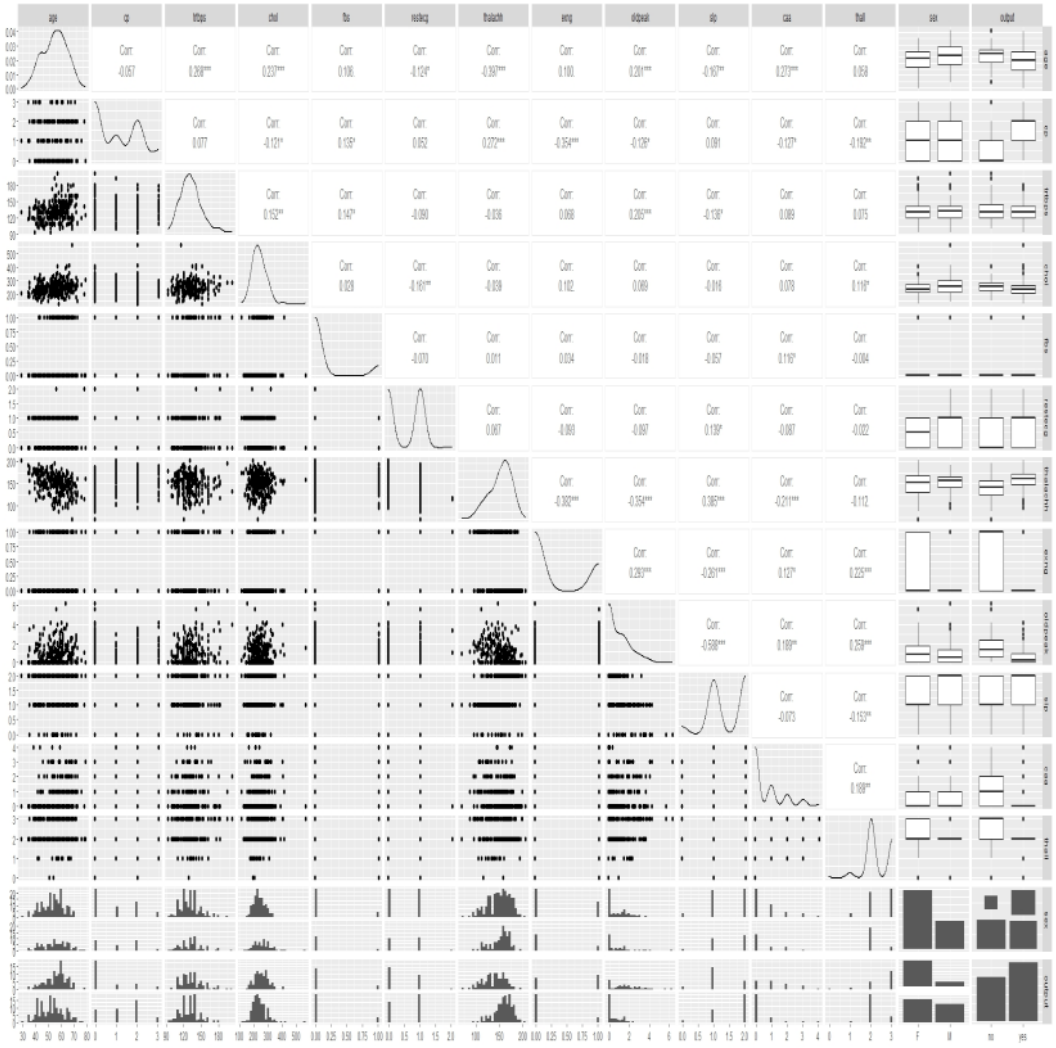


Figure 3.1: Descriptive statistics

In order to provide us with some of the evidence that was in the data, we first generated the descriptive statistics (Table 3.1).

In the above (Figure 3.1) scattered plot is displayed on the left side of the

plot and correlation values are displayed on the right side. Females have higher heart disease rates than males, and dotted lines represent scattered plots.

Out of 289% patients, 165 are at risk of heart disease, where 93 are female (32%) and 72 are male (24.91%).

Further, as age increases, the risk of heart disease increases and cholesterol is positively correlated at 0.237, so this attribute is important for risk factors. The trtbps is blood pressure, is positively correlated with 0.268. The thalachh is heart rate is positively correlated with 0.272. The old peak is positively correlated with 0.201. The chest pain, thallium stress test, and slope are negatively correlated.

3.2 Support Vector Machine (SVM)

The non-linear support vector machine is used for supervised learning. The kappa value is 0.5408 and the sensitivity for svm is 0.7200 and its specificity is 0.8182, respectively. The accuracy for the test data is 0.7759 with a confidence interval of (0.6859, 0.9013).The outcome is displayed in (Fig.3) and confusion matrix is given below.

	Reference	
prediction	No	Yes
No	18	6
Yes	7	27

3.3 Classification and Regression Trees (CART)

The error at the root node for the training data is 124/289 = 0.42907 and misclassification rate of the cart is 1-0.9394 = 0.0606. The accuracy for cart

is 0.9138 and its kappa value for cart is 0.8234. The positive predicted value is 0.8800 and negative predicted value is 0.9394. The outcome is displayed in (Figure 2.2). This section includes the confusion matrix

	Reference	
prediction	No	Yes
No	22	3
Yes	2	31

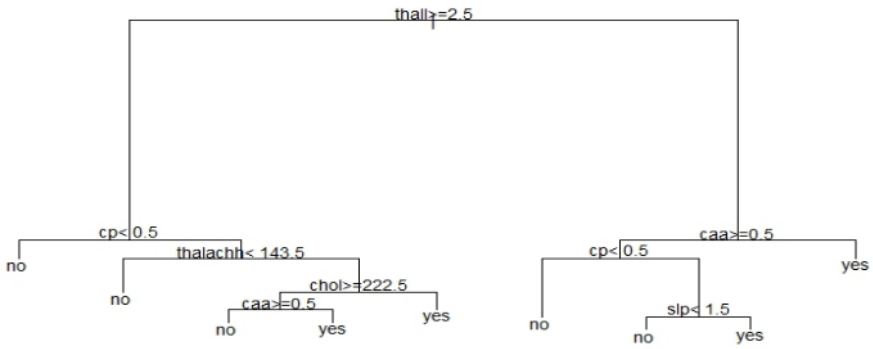


Figure 3.2: CART

3.4 K-Nearest Neighbours (KNN)

The ROC, sensitivity and specificity for different values of k in the KNN model for the training data set are obtained and shown in (Table 2.2). The square root of the training data value k = 11 is the final value with an accuracy of 0.7931 and

95% confidence interval of (0.6665, 0.8883) and its balanced accuracy is 0.7794. The outcome was displayed in (Fig-1) and This section includes the confusion matrix.

	Reference	
prediction	No	Yes
No	17	4
Yes	8	29

k	ROC	Sensitivity	Specificity
5	0.8811427	0.7477778	0.9069597
7	0.8859290	0.7377778	0.9199634
9	0.8936121	0.7381481	0.9102564
11	0.9010969	0.7218519	0.9228938
13	0.9012098	0.7007407	0.9373626
15	0.9051272	0.6981481	0.9521978
17	0.9078449	0.6811111	0.9523810
19	0.9076496	0.6866667	0.9547619
23	0.9149746	0.6903704	0.9549451

Table 3.2: ROC, sensitivity and specificity for training data

3.5 Naive Bayes Classifier (NB)

Using naive bayes method the testing accuracy is 0.7759, the sensitivity is 0.7600 and specificity is 0.7879. The balanced accuracy for naive bayes is 0.7739.

In naive bayes algorithm attributes thallach, caa, oldpeak, thall and cp are

all important attributes and restecg, chol, trtbps and fbs are the least important attributes. This section includes the confusion matrix and importance graph (Figure 3.3).

	Reference	
prediction	No	Yes
No	19	7
Yes	6	26

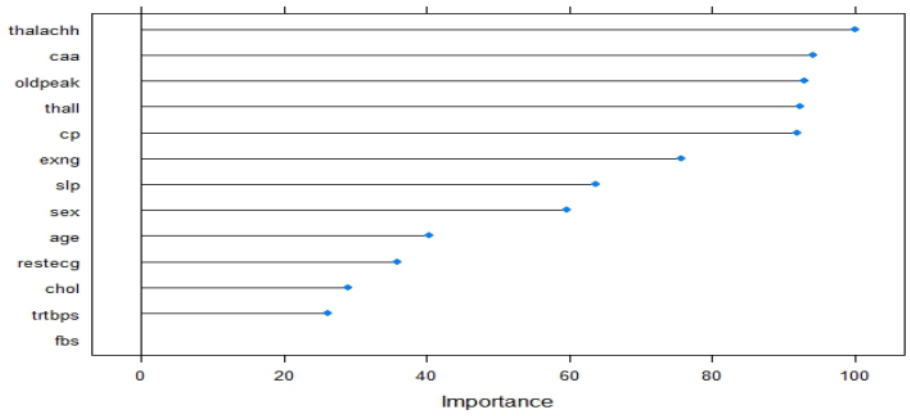


Figure 3.3: Naive Bayes

3.6 Linear Discriminant Analysis (LDA)

A classification method is linear discriminant analysis and its test accuracy is 0.7586 with a kappa value of 0.5031 and its confidence interval is (0.6283, 0.8613) and its balanced accuracy is 0.7491. The output is displayed in (Fig-4). This section includes the confusion matrix.

prediction	Reference	
	No	Yes
No	17	6
Yes	8	27

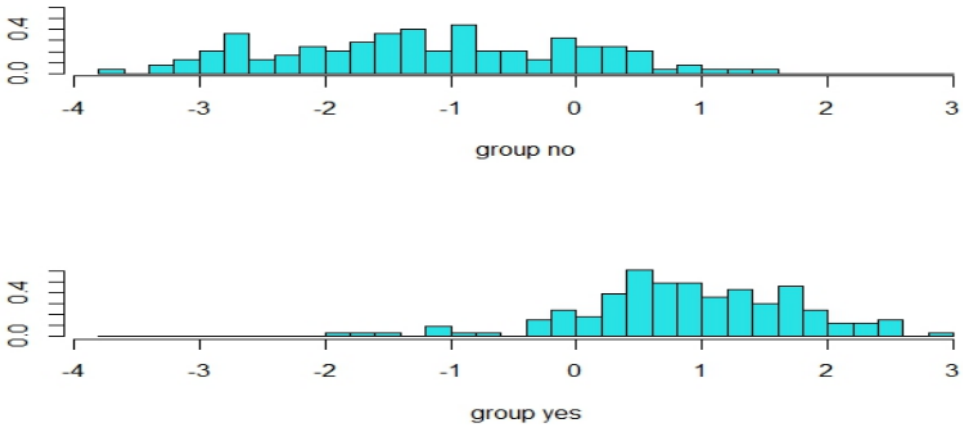


Figure 3.4: LDA Plot

3.7 Random Forest (RF)

Out-Of-Bag (OOB-misclassification rate) is 17.3% and it accurately predicts 95 people without heart disease with class error 0.2338710 and 144 patients with heart disease correctly with the class error 0.1272727. The accuracy for random forest is 0.8339 and its ROC curve is displayed in (Fig-2). This section includes the confusion matrix.

prediction	Reference	
	No	Yes
No	95	29
Yes	19	146

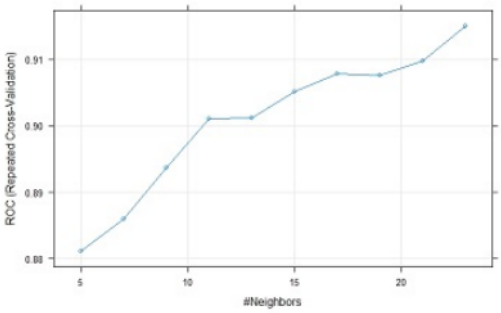


Fig 1: KNN

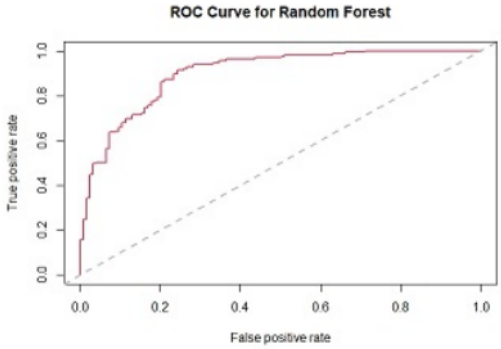


Fig 2: RF

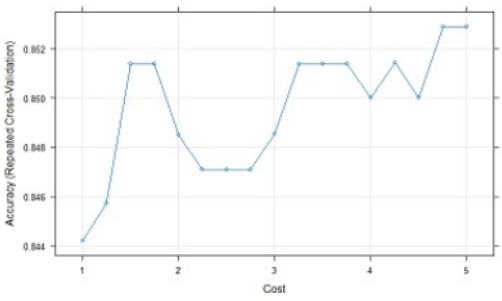


Fig 3: SVM

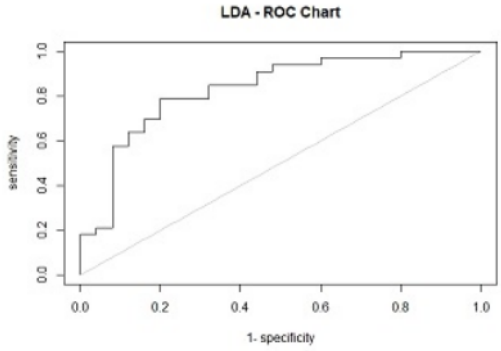


Fig 4: LDA

3.8 Decision Tree

The Method of supervised learning Problems involving regression and classification can be solved with decision trees. It has two nodes: a branch node and a leaf node. Additionally, it maintains accuracy even when a high proportion of the data is missing and is very effective at estimating missing data.

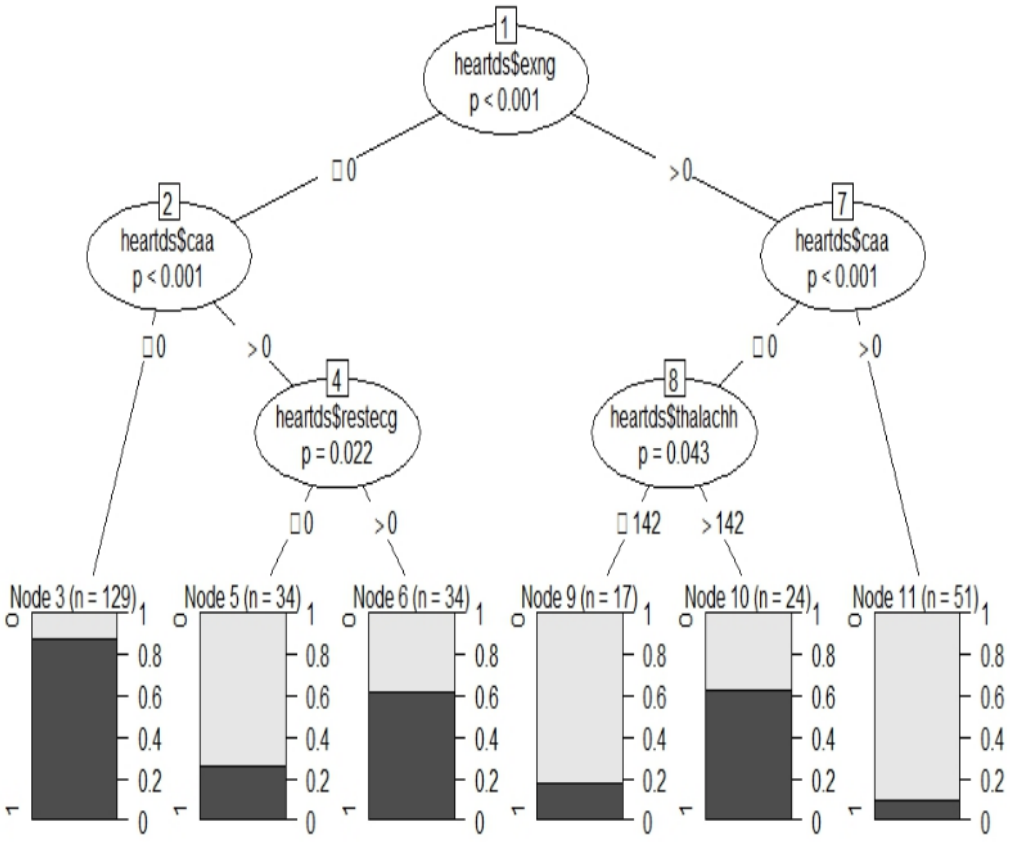


Figure 3.5: Decision tree

3.9 Effect Plots

This model effect map displays the anticipated probabilities for output for each gender, with males having the highest expected probabilities of risk with 0.8451681 and females having the lowest expected probability of risk with 0.4274244.

The effect plot, which shows that risk is inversely related to cholesterol and that it varies from 100 to 600, shows that the likelihood of risk for cholesterol is 0.8017338 to 0.1044568. The likelihood of risk for cp ranges from 0.366454 to 0.8893285 and varies from 0 to 3, according to the effect plot, while the chance of risk for thalachh ranges from 0.2097212 to 0.8021313 and goes from 71 to 200, also according to the effect plot.

The trtbps impact plot shows that risk is negatively correlated with trtbps and that the range of the likelihood of risk from 0.7510306 to 0.2610423 is from 94 to 200 and The probability of risk for fbs ranges from 0.5902977 to 0.5611900, where risk is constant, and the likelihood of risk for restecg ranges from 0.5011814 to 0.7918080, where risk is directly correlated to restecg and ranges from 0 to 2.The range of the risk probability for thalachh is 0.2097212 to 0.8021313 and the risk is directly correlated to thalachh. The impact plot for exng from 0.6677334 to 0.4008912 varies from 0 to 1 and exng is inversely correlated with risk.

The likelihood of risk ranges from 0.73164318 to 0.05225514 and varies from 0 to 6, according to the impact plot for oldpeak, and hazard is inversely correlated to oldpeak and the probability of risk for slp ranges from 0.4629457 to 0.6344278 varies from 0 to 2 and hazard is directly correlated to slp. The expected value of risk, which ranges from 0.7254944 to 0.0737926, goes from 0 to 4, according to

the impact plot for caa, and risk is inversely related to caa and Risk is inversely related to thall and has an expected value that ranges from 0 to 3 for values between 0.9070780 and 0.4443529.

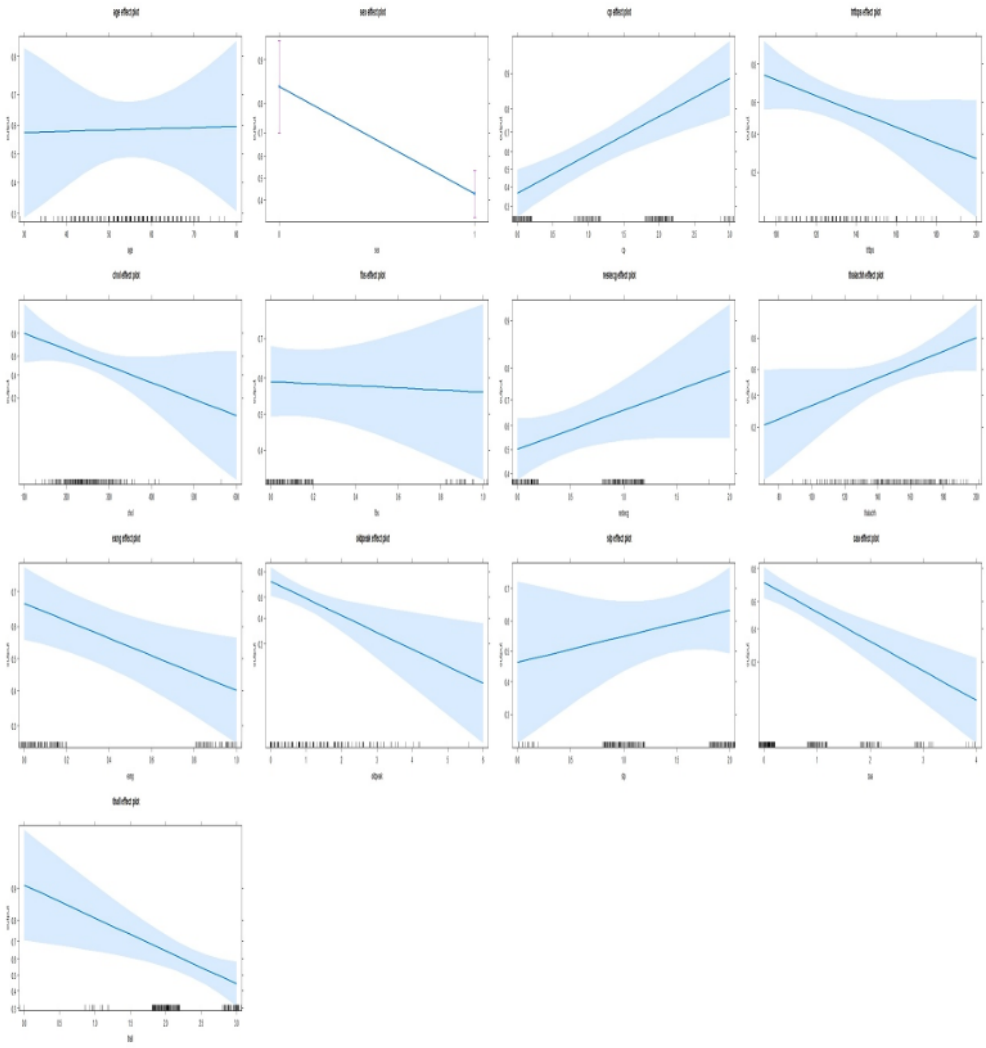


Figure 3.6: Effect plots

Chapter 4

PARAMETER ANALYSIS

Dimensionality reduction is another name for parameter reduction. Data visualisation typically makes use of dimensionality reduction techniques, which aim to reduce the amount of characteristics in a dataset. It is mostly used because large number of dataset will perform poorly in machine learning. Using soft sets, the number of parameters is reduced and a core parameter is found for simple processing. By performing feature selection to eliminate "irrelevant" features that are not important to the classification problem and by removing the least significant variables from the model, dimensionality reduction reduces the number of variables.

We have taken values a,b,c,d,e,f,g,h,i,j,k and l as age, sex, chest pain (cp), blood pressure (trtbps), cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic (restecg), heart rate (thalach), angina (exng), thaillum stress test (thall), old peak, slope and number of major vessels(caa).

In this, we have done a parameter reduction using soft sets to analyse all different parameters, and they are displayed as possible parameters related to heart disease. To increase accuracy, it will be compared with all parameters to get the best accuracy.

S.No	No.of parameter	parameter	Accuracy					
			NB	CART	KNN	SVM	LDA	RF
1	12	abcdefghijkl	0.7759	0.91.38	0.7931	0.7759	0.7586	0.8339
2	8	adefijl	0.8276	0.8103	0.569	0.7241	0.7241	0.7759
3	6	adefil	0.8103	0.8276	0.5517	0.8103	0.7759	0.8448
4	6	abcejl	0.7586	0.7586	0.6552	0.8276	0.7931	0.8448
5	6	abdehl	0.7069	0.8103	0.6379	0.7069	0.7586	0.7241
6	6	abdejl	0.6552	0.7414	0.569	0.7069	0.7414	0.7759
7	6	abdekl	0.7586	0.7931	0.569	0.7759	0.7931	0.7759
8	6	abdhil	0.7931	0.7759	0.6379	0.7586	0.7414	0.7931
9	6	abdijl	0.7586	0.7414	0.5	0.7414	0.7586	0.7759

S.No	No.of parameter	parameter	Accuracy					
			NB	CART	KNN	SVM	LDA	RF
10	6	abdikl	0.7241	0.7759	0.5172	0.8448	0.7586	0.8448
11	6	acdehl	0.7241	0.7414	0.6379	0.7759	0.7414	0.7759
12	6	acdeil	0.7759	0.7241	0.569	0.7414	0.7586	0.7069
13	6	acdekl	0.7414	0.7759	0.569	0.7414	0.7931	0.8103
14	6	acdhil	0.7414	0.7414	0.6207	0.7931	0.7586	0.7931
15	6	acdijl	0.7586	0.7586	0.5172	0.7241	0.7241	0.7931
16	6	acdikl	0.7759	0.7931	0.5862	0.8103	0.7759	0.8276
17	6	acefjl	0.7241	0.7586	0.6379	0.7586	0.7069	0.8448
18	6	acegil	0.7759	0.7586	0.6552	0.8103	0.7414	0.8621

S.No	No.of parameter	parameter	Accuracy					
			NB	CART	KNN	SVM	LDA	RF
19	6	acehil	0.7759	0.7414	0.6552	0.8103	0.7586	0.8276
20	6	acehjl	0.7414	0.7586	0.6552	0.7241	0.6897	0.7759
21	6	aceijl	0.7241	0.7586	0.6552	0.8103	0.7414	0.8621
22	6	aceikl	0.8103	0.7931	0.6207	0.7759	0.7759	0.8793
23	6	acejkl	0.7586	0.7586	0.6379	0.7931	0.7241	0.8448
24	6	aeghil	0.8103	0.7759	0.6379	0.7931	0.7414	0.7931
25	6	aegijl	0.7586	0.7414	0.6379	0.7931	0.7759	0.8448
26	6	aegikl	0.7586	0.8276	0.6207	0.8103	0.7759	0.8448
27	5	bcejl	0.7759	0.7586	0.6379	0.8276	0.7931	0.7931
28	5	acejl	0.7586	0.7586	0.6552	0.8103	0.7241	0.8448

Table 4.1: Overall Accuracy for All Parameters

From the Table 2.3 age, trtbps, chol, fbs, exng and caa (ie.adefil) and these attributes give the best accuracy to predict heart disease. so, it is considered as best parameter.

prediction	Reference											
	NB		CART		KNN		SVM		LDA		RF	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No	15	1	23	8	11	12	17	3	16	4	20	4
Yes	10	32	2	25	14	21	8	30	9	29	5	29

Table 4.2: Confusion matrix for best parameter

In this above confusion matrix (Table 4.2) Naive bayes has predicted low false positive rate likewise KNN has predicted high false positive rate. The true positive rate for CART is high likewise for KNN it has low true positive rate. The CART has low number of false negative rate likwise KNN has in high number. For, predicting the false negative rate naive bayes has higher and CART has predicted low in number.

Method	Accuracy	95%C.I	No Information rate	p-value	Kappa
NB	0.8103	(0.6859, 0.9013)	0.569	9.434e^{-}0.5	0.5957
CART	0.8276	(0.7057, 0.9141)	0.569	2.735e^{-}05	0.6584
KNN	0.5517	(0.4154, 0.6826)	0.569	0.6563	0.0771
SVM	0.8103	(0.6859, 0.9013)	0.569	9.434e^{-}05	0.6037
LDA	0.7759	(0.6473,0.8749)	0.569	0.000826	0.5317
RF	0.8448	(0.7258, 0.9265)	0.569	7.081e^{-}06	0.6821

Method	Mcnermar's Test	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value
NB	0.01586	0.6000	0.9697	0.9375	0.7619
CART	0.1138	0.9200	0.7576	0.7419	0.9259
KNN	0.8445	0.4400	0.6364	0.4783	0.6000
SVM	0.2278	0.6800	0.9091	0.8500	0.7895
LDA	0.267257	0.6400	0.8788	0.8000	0.7632
RF	1	0.8000	0.8788	0.4783	0.6000

Method	Prevalence	Detection Rate	Detection Prevalence	Balanced Accu	Positive Class
NB	0.4310	0.2586	0.2759	0.2586	NO
CART	0.4310	0.3966	0.5345	0.8388	NO
KNN	0.4310	0.1897	0.3966	0.5382	NO
SVM	0.4310	0.2931	0.3448	0.7945	NO
LDA	0.4310	0.2759	0.3448	0.7594	NO
RF	0.4310	0.3448	0.4138	0.8394	NO

Table 4.3: Total value for best parameter

Method	CART	LDA	SVM	KNN	RF	NB
CART		0.009437	-0.005287	0.178541	-0.003969	0.016930
LDA	1.00000		-0.014723	0.169104	-0.013406	0.007493
SVM	1.00000	1.00000		0.183827	0.001318	0.022217
KNN	0.01314	0.04687	0.06096		-0.182510	-0.161611
RF	1.00000	1.00000	1.00000	0.04251		0.020899
NB	1.00000	1.00000	1.00000	0.27430	1.00000	

Table 4.4: Comparison of Kappa

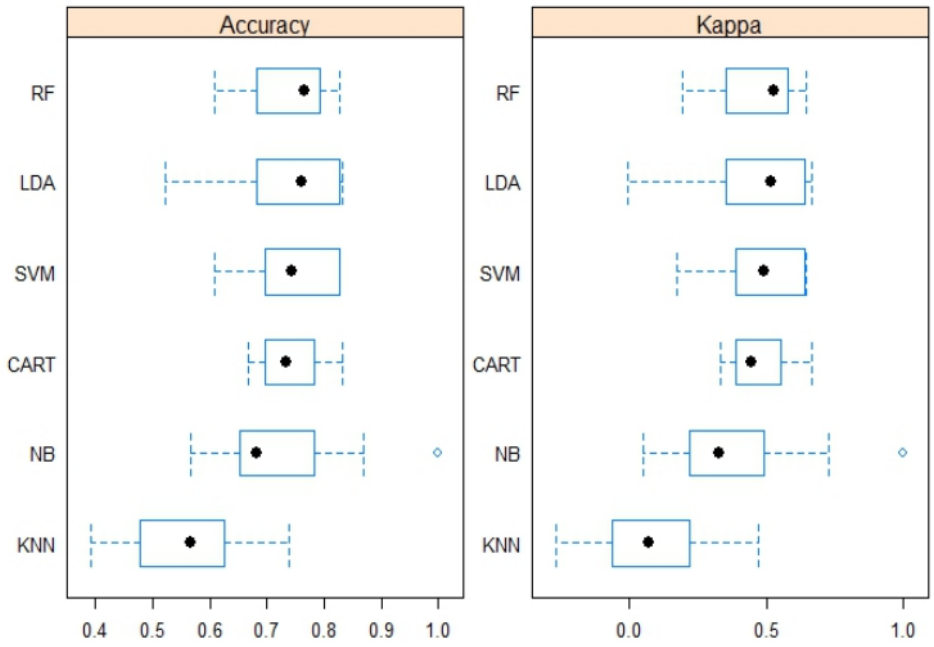


Figure 4.1: Comparison of accuracy and Kappa value

In the overall comparison, the accuracy for support vector machine (svm) with 14 parameters is 0.7759 and the accuracy for 6 parameters (i.e., adefil) is 0.8103. The accuracy for adefil is increasing faster than the other 14 parameters. Likewise, the linear discriminant analysis (LDA) algorithm for 14 parameters is 0.7586 and the reduced parameter (i.e., adefil) is 0.7759, which is slightly higher than the original parameter. The accuracy for random forest (RF) in 14 parameters is 0.8339, the accuracy for reduced parameters is 0.84448 and the accuracy for random forest in reduced parameters is increasing. Likewise, the accuracy for k-nearest neighbour for 14 parameters is 0.7931 and for reduced

parameters, it is 0.5517 and its accuracy is decreased compared to 14 parameters. Then the accuracy for classification and regression tree (CART) for 14 parameters is 0.9138 and for reduced parameters (i.e., adefil). It is 0.8276, which is lower than 14 parameters. Likewise, accuracy for the Naive Bayes (NB) algorithm for 14 parameters is 0.7759 and accuracy for reduced parameters is 0.8103, which is higher than 14 parameters.

Chapter 5

CONCLUSION

In this project, machine learning techniques were used to calculate accuracy. The output prediction accuracy using machine learning algorithms provides us with knowledge about the optimal method to use and estimating future output predictions based on historical data will always be beneficial. As a result of the comparison, the random forest method provides the best accuracy of 0.8448, while the k-nearest neighbour method provides the worst accuracy of 0.5517. Therefore, the analysis described above will be useful for future forecasting approaches on the basis of which decisions may be made about how to help people in staying out of a risk zone.

References

[1] Abdul Saboor et al, A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms, Hindawi, Article ID 1410169, 2022.

[2] M.Akhil jabbar, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, International Conference on Computational Intelligence: Modeling Techniques and Applications, 2013.

[3] Amanda H. Gonsalves et al, Prediction of Coronary Heart Disease, 2019.

[4] Anupama Yadav, Levish Gediya, Adnanuddin Kazi,Heart Disease Prediction Using Machine Lerning, 2021.

[5] Davide Chicco et al, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, BMC Medical Informatics and Decision Making, 2020

[6] Dejun Zhang, Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer, IEE Access, 2018.

[7] Edwar Macias et al, Mortality prediction enhancement in end-stage renal disease: A machine learning approach, Informatics in Medicine Unlocked, 2020.

[8] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, Preeti Nagrath, Heart disease prediction using machine learning algorithms, 2021.

[9] K. Kannan, A. Menaga, Risk Factor Prediction by Naive Bayes Classifier, Logistic Regression Models, Various Classification and Regression Machine Learning Techniques, 2021

[10] Monther Tarawneh, Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques, ACTA SCIENTIFIC NUTRITIONAL HEALTH, 2019.

[11] Mamatha Alex P and Shaicy P Shaji, "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique", International Conference on Communication and Signal Processing, 2019.

[12] Mohammed Khalid Hossen, Heart Disease Prediction Using Machine Learning Techniques, American Journal of Computer Science and Technology, 2022.

[13] Neha Nandal et al, Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis, 2022.

[14] A.Sankari Karthiga et al, Early Prediction of Heart Disease Using Decision Tree Algorithm, International Journal of Advanced Research in Basic Engineering Sciences and Technology, 2017.

[15] Sonam Nikhar et al, Prediction of Heart Disease Using Machine Learning Algorithms, International Journal of Advanced Engineering, Management and Science (IJAEMS), ISSN : 2454-1311, 2016.

[16] Sophie H Bots et al, Sex differences in coronary heart disease and stroke mortality, 2017.

[17] C.Zoccali, Traditional and emerging cardiovascular and renal risk factors: An epidemiologic perspective, International Society of Nephrology, 2006.

[18] https://www.r-project.org/about.html

[19] https://www.healthline.com/health/heart-disease/history

[20] https://www.nhsinform.scot/illnesses-and-conditions/heart-and-blood-vessels /conditions/common-heart-conditions.

MSC PROJECT 22-23 KOWSALYA M

ORIGINALITY REPORT

15%	8%	6%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	2%
2	Submitted to SASTRA University Student Paper	2%
3	K. Kannan, A. Menaga. "Risk Factor Prediction by Naive Bayes Classifier, Logistic Regression Models, Various Classification and Regression Machine Learning Techniques", Proceedings of the National Academy of Sciences, India Section B: Biological Sciences, 2021 Publication	1%
4	Submitted to University of Liverpool Student Paper	1%
5	Submitted to Queen Mary and Westfield College Student Paper	1%
6	rstudio-pubs-static.s3.amazonaws.com Internet Source	1%
7	Submitted to University of Sunderland Student Paper	1%

8	www.hindawi.com Internet Source	1 %
9	www.simplilearn.com Internet Source	<1 %
10	Submitted to University of Wollongong Student Paper	<1 %
11	Submitted to Indian Institute of Technology, Kanpur Student Paper	<1 %
12	f1000research.com Internet Source	<1 %
13	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
14	Submitted to Saveetha Dental College and Hospital, Chennai Student Paper	<1 %
15	Submitted to Liverpool John Moores University Student Paper	<1 %
16	www.researchgate.net Internet Source	<1 %
17	Submitted to Curtin University of Technology Student Paper	<1 %

18	Submitted to The University of Wolverhampton Student Paper	<1%
19	4m6o1.galaxyng.com Internet Source	<1%
20	Submitted to University of Teesside Student Paper	<1%
21	Submitted to Aristotle University of Thessaloniki Student Paper	<1%
22	bscolor.io-bas.bg Internet Source	<1%
23	www.coursehero.com Internet Source	<1%
24	Hafsa Binte Kibria, Abdul Matin. "The severity prediction of the binary and multi-class cardiovascular disease − A machine learning-based fusion approach", Computational Biology and Chemistry, 2022 Publication	<1%
25	Submitted to IIT Delhi Student Paper	<1%
26	uk.teamunify.com Internet Source	<1%
27	www.thefreelibrary.com Internet Source	

		<1 %
28	ijircce.com Internet Source	<1 %
29	www.accentsjournals.org Internet Source	<1 %
30	Submitted to City and Islington College, London Student Paper	<1 %
31	d-nb.info Internet Source	<1 %
32	Fatou NGOM, Ibrahima FALL, Mamadou S CAMARA, Alassane BAH. "A study on predicting and diagnosing non-communicable diseases: case of cardiovascular diseases", 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), 2020 Publication	<1 %
33	www.ijraset.com Internet Source	<1 %
34	"Advances in Artificial Intelligence and Data Engineering", Springer Science and Business Media LLC, 2021 Publication	<1 %
35	Neha Nandal, Lipika Goel, ROHIT TANWAR. "Machine learning-based heart attack	<1 %

prediction: A symptomatic heart attack prediction method and exploratory analysis", F1000Research, 2022
Publication

36 Nandkishor P. Karlekar, N Gomathi. "OW-SVM: Ontology and whale optimization-based support vector machine for privacy-preserved medical data classification in cloud", International Journal of Communication Systems, 2018 <1%
Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		