HEART DISEASE ANALYSIS AND PREDICTION USING MACHINE LEARNING

Project Report submitted to
The Department of Management Studies
in partial fulfillment of the requirements for the award of the degree
Master of Business Administration

THISHONIA PREETHI.S
PRK19MS1107

Under the guidance of

DR. SENITH
ASSOCIATE PROFESSOR

Karunya Institute of Technology and Sciences
[Declared as a Deemed University under sec.3 of the UGC Act, 1956]

April 2021

CERTIFICATE

This is to certify that the project report entitled '**Heart Disease Analysis and Prediction using Machine Learning**' is a bonafide record of work done by **Thishonia Preethi.S (PRK19MS1107)** under my supervision and submitted in partial fulfillment for the award of the degree of Master of Business Administration of Karunya Institute of Technology and Sciences.

Place: Coimbatore
Date: 20-04-2021 Research Supervisor

External Examiner Internal Examiner

Head of the Department

DECLARATION

I, **Thishonia Preethi.S** hereby declare that the project report entitled **'Heart Disease Analysis and Prediction using Machine Learning'** is a bonafide record of the original research work carried out by me in the department of Management studies, Karunya Institute of Technology and Sciences and that it has not been submitted earlier elsewhere for the award of any Degree, Diploma or Fellowship.

- I understand that KITS shall hold the copyrights of all these projects / dissertations submitted to the University.
- I will republish the entire thesis / extracts of the report only with the permission of LITS and I am liable to pay 40% of royalty to KITS.
- If I engage in documenting any research findings with an intention of publishing it for commercial purpose, I shall obtain a NOC from the office of the registrar prior to engaging in such activities.

Signature of the Candidate

Place: Coimbatore

Date: 20-04-2021

ACKNOWLEDGEMENT

I would like to express my sincere thanks to God Almighty, the most gracious and merciful, for his kind blessings to make this internship a successful one. I extend my deep gratitude to DR.C. Joseph Kennady, Dean, SSAMM, Karunya Institute of Technology and Sciences, Coimbatore, for giving me an opportunity and needed facilities to fulfil this project work.

I am greatly obliged to Dr. C. Samuel Joseph, HOD of Management for his wholehearted support and encouragement.

I proudly utilize this privilege to express my heart-felt thanks and sincere gratitude to my inspiring guide Dr. S.Senith Associate Professor, Dept. of Management Studies, Karunya Institute of technology and sciences, Coimbatore. For her kind supervision, valuable guidance and constant encouragement in bringing out this report in time with her untiring involvement and confidence.

It is with great respect; I express sincere gratitude to my beloved parents for their assistance and encouragement. I am also thankful to all my friends who helped me for doing this project. Thank you all for your unwavering support.

Place: Coimbatore

Date: 20-04-2021

Signature of the Candidate

CONTENTS

S. NO	TITLE	PAGE NO.
ACKNOWLEDGEMENT		IV
LIST OF TABLES		2
LIST OF FIGURES		2
I	INTRODUCTION	5
	1.1 INTRODUCTION	6
	1.2 HEART DISEASE	6
	1.3 PROBLEM STATEMENT	9
	1.4 SCOPE OF STUDY	9
	1.5 OBJECTIVE	10
	1.6 OPERATIONAL DEFINITION OF CONCEPTS	10
	1.7 METHODOLOGY AND COLLECTION OF DATA	11
	1.8 PROCESS OF ANALYSIS	11
	1.9 CHAPTER SCHEME	12
II	LITERATURE REVIEW	13
	2.1 LITERATURE REVIEW	14
III	RESEARCH METHODOLOGY	23
	3.1 SUPERVISED MACHINE LEARNING	24
	3.2 MACHINE LEARNING ALGORITHMS	25
	3.2.1 K NEAREST NEIGHBOUR	26
	3.2.2 LOGISTIC REGRESSION	26
	3.2.3 DECISION TREE	27
	3.2.4 RANDOM FOREST	28
	3.2.5 NAÏVE BAYES	29
	3.3 PYTHON FOR DATA ANALYSIS	30
	3.3.1 PANDAS	30
	3.3.2 NUMPY	30
	3.3.3 MATPLOT LIBRARY	31
	3.3.4 SCIKIT LEARN	31
	3.4 TABLEAU FOR PREDICTIVE DASHBOARD	33
IV	DATA ANALYSIS AND INTERPRETATION	35
	4.1 DATA ANALYSIS	36
	4.2 DATA INTERPRETATION	40

	4.2.1 STATISTICS OF DATA	40
	4.2.2 CORRELATION	41
	4.2.3 DISTRIBUTION OF VALUES	41
	4.2.4 ANALYSIS OF ATTRIBUTES	45
	4.2.5 PREDICTIVE DASHBOARD IN TABLEAU	50
	4.3 DISCUSSIONS	51
V	CONCLUSION	54
	5.1 SUMMARY OF FINDINGS	55
	5.2 SUGGESTIONS	55
	5.3 CONCLUSION	56
REFRENCES		57
ANNEXURE		61

LIST OF TABLES

S.NO	TITLE	PAGE NO.
2.1	ACCURACY TABLE	20
4.1	BP READINGS	36
4.2	LDL LEVELS	37
4.3	MAXIMUM HEART RATE ACHIEVED	38
4.4	LDL CATEGORY	39
4.5	MALE FEMALE COUNT	47
4.6	PRESENCE/ABSENCE OF DISEASE	47
4.7	ALGORITHM ACCURACY	49

LIST OF FIGURES

S.NO	TITLE	PAGE NO.
1.1	PROCESS OF ANALYSIS	11
3.1	KNN	26
3.2	LOGISTIC REGRESSION	27
3.3	DECISION TREE	28

3.4	RANDOM FOREST ALGORITHM	29
4.1	STATISTICS	41
4.2	CORRELATION	42
4.3	HISTOGRAM FOR AGE	42
4.4	HISTOGRAM FOR SEX	43
4.5	HISTOGRAM FOR CHEST PAIN	43
4.6	HISTOGRAM FOR RESTING BP	43
4.7	HISTOGRAM FOR CHOLESTROL	43
4.8	HISTOGRAM FOR BLOOD SUGAR	44
4.9	HISTOGRAM FOR MAXIMUM HEART RATE	44
4.10	HISTOGRAM FOR EXERCISE INDUCED ANGIA	44
4.11	TARGET VALUES	45
4.12	SYSTOLIC BLOOD PRESSURE(SBP)	45
4.13	LDL	45
4.14	ADIPOSITY	46
4.15	TYPE A SCORE	46
4.16	FAMILY HISTORY	46
4.17	BAR CHART FOR SEX AND TARGET VALUES	47
4.18	CHOLESTROL LEVELS – GENDERWISE	47
4.19	CORRELATION MAP	47
4.20	TARGET VALUES	48
4.21	CHEST PAIN LEVELS AND TARGET	49
4.22	EFFECT OF CHEST PAIN ON HEART DISEASE	49
4.23	CHEST PAIN AND GENDER	49
4.24	CONFUSION MATRIX OF LOGISTIC REGRESSION	50
4.25	TABLEAU PREDICIVE DASHBOARD	51
4.26	CONFUSION MATRIX	55
4.27	ACCURACY	55

EXECUTIVE SUMMARY

The world has witnessed an explosion of data due to the advancement in technology. The data can be used to bring out insights that help in decision-making. If this data is well utilized, a disease can be detected, predicted, or even cured. The healthcare industry, in particular, incorporates a large quantity of knowledge concerning patients, disease, and treatment procedures. One of the vital organs in the human body is the heart. It circulates oxygen and other vital nutrients through the blood to different parts of the body and helps in metabolic activities. Apart from this it also helps in the removal of metabolic wastes. Thus, even minor problems in the heart can affect the whole organism. Researchers are diverting a lot of data analysis for assisting doctors to predict heart problems. This paper can explain the understanding of predictive analytics and the various machine learning algorithms where the concentration is set upon heart disease prediction to predict who will develop coronary heart disease in the near future. Machine learning techniques are used to compare accuracy among different machine learning algorithms. The machine learning algorithm with the highest accuracy is taken and a predictive model is built using Tableau which enables users to enter in values and get the prediction of the presence/absence probability of heart disease. The findings during this paper embrace heart disease prediction and causes and also the classification of infection among patients' post-surgery. Keywords: Prediction algorithms, Machine learning, big data analytics, Automation, Deep Learning, Coronary Heart Disease, Sepsis, Heart Attack

CHAPTER I

INTRODUCTION & DESIGN OF STUDY

1.1 INTRODUCTION

Cardiovascular diseases are the **leading cause of mortality** globally, claiming the lives of an estimated **17.9 million people each year.** Coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other heart and blood vessel disorders are all categorized as CVDs. Heart attacks and strokes account for four out of every five CVD deaths, with one-third of these deaths occurring before the age of 70.

As the number of adults with congenital heart disease are getting higher, obtained comorbidities play a critical part in dreariness and mortality. India has one of the highest burdens of cardiovascular disease around the world. The yearly number of deaths due to CVD in **India is anticipated to rise from 2.26 million (1990) to 4.77 million (2020).** Coronary heart disease prevalence rates in India have been assessed over the past few decades and have extended from 1.6% to 7.4% in country populations and from 1% to 13.2% in urban populations.

The study is pointed to show that CVD risk components such as abdominal obesity, hypertension, and diabetes influence the onset of heart disease. Predictive analysis plays a major part within the healthcare industry where estimating the onset of disease will diminish the chance that happens to patients. Statistics appear that cardiovascular diseases have expanded the mortality rate. Machine learning which is utilized in developing a predictive model for different domains is nowadays connected within the field of therapeutic diagnostics. In this paper, the analysis of heart disease is done by considering the parameters like age, gender, blood pressure, heart rate, diabetes, and various other risk factors.

1.2 HEART DISEASE

The term "heart disease" refers to a variety of cardiovascular issues. Heart disease encompasses a wide range of disorders and conditions. The following are examples of different types of heart disease:

- Arrhythmia is a condition in which the heart beats irregularly. A heart rhythm abnormality is known as an arrhythmia.
- Atherosclerosis is a disease that affects the arteries. The hardening of the arteries is known as atherosclerosis.
- Cardiomyopathy is a disease that affects the heart. The heart muscles harden or weaken as a result of this disease.
- Heart defects that are present at birth. Heart abnormalities that are present at birth are known as congenital heart defects.
- Coronary artery disease (CAD) is a condition that affects the heart.
- Plaque accumulation in the arteries of the heart causes CAD. Ischemic heart disease is another name for it.

- Infections of the heart. Bacteria, viruses, and parasites may all cause heart infections.

The word "cardiovascular disease" refers to heart diseases that affect the blood vessels primarily.

Heart arrhythmias

Arrhythmias are irregular heartbeats. The symptoms individuals have which vary depending on the type of arrhythmia, fast or slow heartbeats. An arrhythmia can cause the following symptoms:

- Feeling dizzy
- A racing pulse or a fluttering heart
- A sluggish heartbeat
- Spells of fainting
- A feeling of nausea
- Chest discomfort

Atherosclerosis

Atherosclerosis is a disease that affects the arteries. The flow of blood to the extremities is reduced as a result of atherosclerosis. Atherosclerosis signs include

- Chest pain and shortness of breath, as well as:
- Unusual or unexplained pain weakness in the legs and arms
- Coldness, particularly in the limbs
- Numbness, especially in the limbs
- Unusual or unexplained pain

Congenital heart defects

Heart defects that are present at birth. Heart abnormalities that arise as a foetus is developing are known as congenital heart defects. Some heart defects go undetected for years. Others can be discovered as a result of symptoms, such as:

- Swelling of the extremities with a blue tint
- Fatigue and low energy
- Shortness of breath or trouble breathing
- Irregular heart rhythm

Coronary artery disease (CAD)

Plaque accumulation in the arteries that transport oxygen-rich blood through the heart and lungs is known as coronary artery disease (CAD). Symptoms of coronary artery disease include:

- Chest pressure or pain
- Shortness of breath
- Nausea
- A sensation of pressure or squeezing in the chest
- Irritable bowel syndrome (IBS) or gas

Cardiomyopathy

Cardiomyopathy is a disease that affects the heart. Cardiomyopathy is a condition in which the heart muscles enlarge and become stiff, heavy, or frail. The following are some of the signs and symptoms of this condition:

- Exhaustion
- Shortness of breath
- Pounding or rapid pulse
- Bloating swollen legs, particularly ankles and feet

Heart infections

Infections of the heart. Endocarditis and myocarditis are two conditions that can be described as heart infections. A heart infection can cause the following symptoms:

- Chest pain
- Congestion or coughing
- Fever
- Chills
- Rash on the skin

Heart disease is caused by a number of causes. Some things can be regulated, while others cannot. These are some of the risk factors:

- High blood pressure
- High cholesterol
- Low levels of HDL (the “good” cholesterol)
- Obesity due to smoking
- Lack of physical activity

Here, smoking is a preventable risk factor. According to the National Institute of Diabetes and Digestive and Kidney Diseases, smoking doubles the risk of heart disease (NIDDK).

Other heart disease risk factors include:

- Ethnicity
- Sex
- Age
- Family history

While one can't change these risk factors, individuals can keep an eye on how they affect. Diabetes patients may be at an increased risk of heart failure because elevated blood glucose levels raise the risk of:

- Angina
- Coronary artery disease
- Heart attacks
- Stroke

The treatment for heart diseases is mainly

1. Lifestyle changes
2. Medications
3. Surgery

1.3 PROBLEM STATEMENT

Humans in their everyday lives are exposed to a regular and busy schedule, which creates stress and anxiety. Furthermore, the number of obese and cigarette-addicted individuals rises significantly. This leads to the proneness of heart disease. **Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report. 2005** states that “By the year 2030, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs)”. The challenge behind these diseases is their early prediction and treatment at the right time. As the saying goes “Prevention is better than cure”, a priori prediction can help save lives.

1.4 SCOPE OF STUDY

This research would aid in the identification of patients who may be suffering from heart disease. This helps in taking preventive measures and hence try to avoid the possibility of heart disease for the patient. So, when a patient is predicted to have a greater probability for heart disease, then the medical data for the patient can be closely analysed by the doctors. An example would be if the patient has diabetes which may be the cause for heart disease in the future, then the patient can be given treatment to keep diabetes under control which in turn may prevent the onset of heart disease in that patient.

1.5 OBJECTIVE

To identify the various risk factors contributing to heart diseases and compare the accuracy of various supervised machine learning algorithms

To build a predictive model with the algorithm that has the highest accuracy in prediction.

1.6 OPERATIONAL DEFINITION OF CONCEPTS

Attribute Information:

1. Age - Age in years
2. Sex - Male or Female
3. Chest pain type - 4 levels of chest pain
4. Resting BP
5. Serum Cholesterol in mg/dl
6. Fasting blood sugar
7. Resting ECG results
8. Maximum heart rate
9. Maximum heart rate achieves - Maximum heart rate during strenuous activities
10. Exercise-induced angina
11. Oldpeak - ST depression induced by exercise relative to rest
12. The slope of the peak exercise ST segment
13. Number of major vessels - (0-3) colored by fluoroscopy
14. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
15. Sbp (systolic blood pressure)
16. Tobacco (cumulative tobacco (kg))
17. Ldl (low density lipoprotein cholesterol)
18. Adiposity - BMI of a person
19. Famhist - family history of heart disease, a factor with levels "Absent" and "Present"
20. Typea - type-A behaviour intense striving for achievement, competition, easily provoked impatience, time urgency, the abruptness of gesture and speech (explosive voice), hyper-alert posture, overcommitment to vocation or profession
21. Obesity - How obese a person is
22. Alcohol - current alcohol consumption
23. Age - the age at onset
24. Chd - coronary heart disease

1.7 METHODOLOGY AND COLLECTION OF DATA

Secondary data has been used for research. The datasets are taken from Kaggle and UCI Machine Learning Repository.

Dataset 1: https://www.kaggle.com/ronitf/heart-disease-uci

Number of records: 304 observations

Dataset 2: https://www.kaggle.com/yassinehamdaoui1/cardiovascular-disease

Number of records: 462 observations.

The predictive model using Tableau is built on the basis of this dataset.

The Research concepts used in this project are as follows.

This project made use of **quantitative research**. It is a statistical analysis approach that employs numerical data that can be ranked, calculated, or categorized. It aids in the discovery of patterns or relationships, as well as the formulation of generalizations. This form of study is useful for deciding how many, how often, how often, or to what extent.

The primary objective of **predictive research** is to forecast (predict) events, consequences, costs, or effects. This form of research attempts to predict the future by extrapolating from current phenomena, laws, or other institutions. This project aims to predict the incidence of heart disease through predictive analysis.

Predictive analytics is a subset of advanced analytics that is used to forecast uncertain future events. Predictive analytics analyses existing data and make predictions about the future using a variety of techniques including data mining, statistics, simulation, machine learning, and artificial intelligence.

1.8 PROCESS OF ANALYSIS

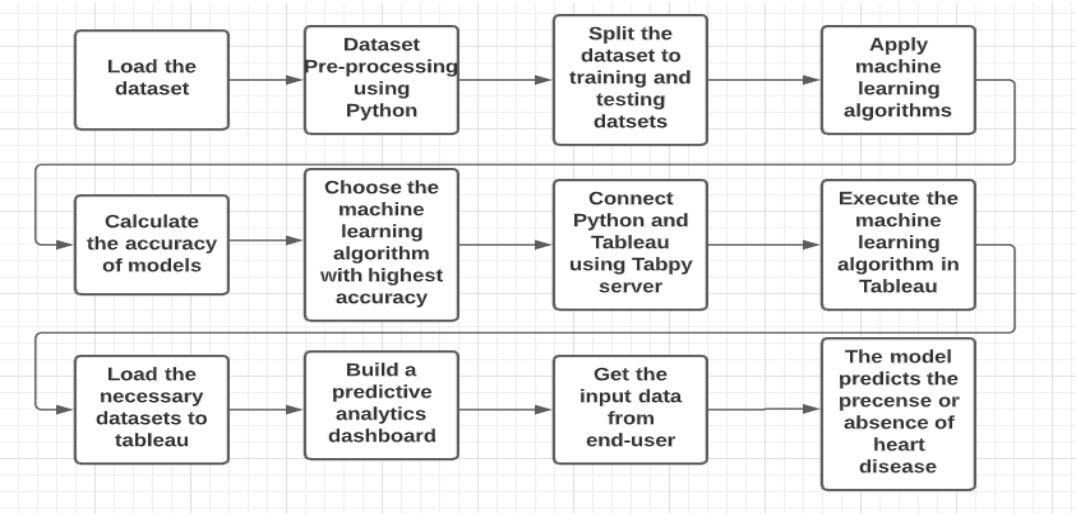


Figure 1.1: Process of Analysis

1.9 CHAPTER SCHEME

The first chapter comprises the Introduction, the problem statement, scope of study, objectives, operational definition of concepts, methodology and collection of data and the analysis process.

The second chapter consists of Literature review and the various accuracies obtained by the researchers.

The third chapter consists of the research methodology, supervised machine learning algorithms, python libraries used and Tableau.

The fourth chapter consists of the data analysis and interpretations, statistical analysis, distribution of data and analysis of various attributes.

The fifth chapter consists of the conclusion, summary of findings and future recommendations.

CHAPTER II

LITERATURE REVIEW

2.1 LITERATURE REVIEW

Monika Gandhi et al. address the different techniques of information abstraction that are being used in today's research for the prediction of heart disease using data mining methods. Using algorithms, the data mining methods Naive Bayes, Neural network, and Decision tree algorithm are evaluated on medical data sets. According to the analysis, neural networks, decision trees, and naive Bayes can all be analyzed in greater depth in order to develop a useful algorithm for healthcare organizations.

J Thomas and R Theresa Princy et al. in view of heart disease prediction, the K nearest neighbor algorithm, neural network, naive Bayes, and decision tree were used. They used data mining methods to accurately detect the risk of heart disease with an accuracy of 80.4 percent.

Sandhya Kumari and Dr. R. Viswanathan's heart disease research paper, prediction seeks to uncover hidden trends in a dataset using Machine Learning techniques and to forecast or know the current value on a scale. Heart disease prediction necessitates a huge amount of data that is too large and complicated to process and analyze using traditional methods. The goal was to find a technique for predicting cardiac disease that was both effective and reliable. The KNN algorithm had an accuracy of 86.30% and the Decision Tree algorithm had an accuracy of 77.58%.

The Heart Disease Prediction System proposed by **AH Chen, SY Huang et al.** is based on a data mining technique with a set of important attributes for heart disease prediction and an artificial neural network for classifying heart disease based on important features. The prediction accuracy rate is nearly 80%.

Using the J48 algorithm, **A. Sheik Abdullah** proposed a data mining model for detecting Coronary Heart Disease and predicting the different events related to each patient record. Using feature selection, this model decreases the number of attributes. This model has chosen 9 attributes and has a 60.74 percent accuracy score.

G Purushottam et al. proposed "An automated system in medical diagnosis would enhance medical care and it can also reduce costs. In this study, we have designed a system that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health." The performances of single models such as Decision tree, artificial neural network, and Naïve Bayes are 85%, 76%, and 69% respectively.

Shanta Kumar, B.Patil, Y.SKumaraswamy's paper suggests that knowing the risk factors for heart disease assists health care practitioners in recognizing patients who are at high risk for heart disease. Healthcare practitioners may use statistical analysis and data processing tools to aid in the diagnosis of heart disease. Coronary heart disease (heart attacks), cerebrovascular disease (stroke), increased blood pressure (hypertension), coronary artery disease, rheumatic heart disease, congenital heart disease, and heart failure have all been reported by statistical analysis. Tobacco use, physical inactivity, an unhealthy diet, and harmful alcohol use are the leading causes of cardiovascular disease.

Vikas Chaurasia and Saurabh Pal's The aim of this study is to find a way to predict the presence of heart disease more accurately with a smaller number of variables. The accuracy was 82% for Naive Bayes and 84% for J48.

Boshra Baharami et al. J48 Decision Tree, k-Nearest Neighbours(k-NN), Naive Bayes(NB), and SMO are some of the classification strategies that have been tested (SMO is widely used for training SVM). With an accuracy of 83.732 percent, J48 is the most accurate.

Sellappan Palaniyappan, Rafiah Awang et al. used Naive Bayes, Decision Trees, and Artificial Neural Networks to build Intelligent Heart Disease Prediction Systems (IHDPS). It presents the findings in both tabular and graphical formats to aid visualization and analysis. The Naive Bayes model tends to outperform the other two, with the largest number of accurate predictions 86.12%, followed by Neural Networks 85.68 %, and Decision Trees 85.68 %.

Himanshu Sharma, M A Rizvi et al. made use of Decision tree, support vector machine, deep learning, K nearest neighbour algorithms. Since the datasets contain noise, they attempted to reduce the noise by cleaning and pre-processing the data, as well as reducing the dataset's dimensionality. They discovered that neural networks can achieve high accuracy.

Animesh Hazra et al. discussed in detail cardiovascular disease and different symptoms of a heart attack. The different types of classification and clustering algorithms and tools were used.

M.A.Nishara Banu and B.Gomathy used the C4.5 algorithm, MAFIA, and K-Means clustering in the year 2014 using 13 attributes in the dataset and achieving 89% accuracy.

Ramandeep Kaur, Er. Prabhakaran Kaur et al. have shown that the heart disease data contains unnecessary, duplicate information. This has to be pre-processed. hey also claim that feature selection on the dataset is needed for better performance. The accuracy of K-means, MAFIA, and C4.5 was 89 percent. KNN has an accuracy of 89.2 percent, while SVM has an accuracy of 85 percent.

G. Parthiban et al. used machine learning approaches to detect heart disease in diabetic patients. WEKA is used to apply Naive Bayes. The researchers used a data collection of 500 patients from the Chennai Research Institute. There are 142 patients who have the disorder and 358 patients who do not. The algorithm provides 74 % accuracy when using the Naive Bayes method.

Vembandasamy et al. diagnosed heart disease using the Naive Bayes algorithm. Bayes' theorem is used in Naive Bayes. Therefore, Naïve Bayes has a powerful principle of independence. The data used are from one of the leading diabetes research institutes in Chennai. The data set consists of 500 patients. WEKA is used as a tool and performs classification using 70% of the Percentage Split. Naive Bayes offers 86.419% accuracy.

Yashvendra K. Singh et al. state that Machine Learning algorithms are becoming more useful in predicting different diseases, according to the authors. Because of the ability of machine learning algorithms to think like humans, this definition is extremely necessary and scalable. The aim of this study is to improve the accuracy of heart disease prediction. The Cleveland heart disease dataset's nonlinear propensity was used to apply Random Forest, which resulted in an accuracy of 85.81 %.

Seyedamin Pouriyeh, Sara Vahid et al. KNN gives an accuracy of 83.16% when the value of k is equal to 9 while using the 10-cross validation technique.

Tan et al. proposed a hybrid method for classification that combines two machine-learning models, such as SVM support vector machine and genetic algorithm. For their experiment, they used four separate data packages from the UCI repository, including iris, diabetic, breast cancer, heart disease, and hepatitis. After applying the hybrid approach to heart disease classification, the accuracy was found to be 84.07 %.

A new approach for detecting and tracking coronary artery disease was proposed by **Ottom et al.** Cleveland heart data was obtained from UCI for this analysis, with 303 cases and 76 attributes in the data package; only 13 attributes are used for detection. They proposed three machine learning algorithms for the detection process, including SVM and Functional Tree. Finally, SVM reaches the highest detection accuracy of 88.3 percent.

Arunpradeep N., G. Niranjana discussed the available machine learning models for prognosis of heart disease with high certitude, precision using KNN, SVM, DT, and RF gave an accuracy of 86%, 75%, 74%, and 84%.

S.Rajathi and Dr.G.Radhamani's paper on Rheumatic Heart Disease efficiently analyses using the KNN with ACO algorithm. Four different algorithms are used to compare the accuracy. The accuracy for the algorithms was 60% for Decision Tree, 65% for SVM, 68% for KNN, and 70% for KNN with ACO.

Ridhi Saini et al. have obtained an efficiency of 87.5% using KNN, which is very good.

Kanika Pahwa and Ravinder Kumar et al. Naive Bayes algorithms have achieved an accuracy of 78% and Random Forest an accuracy of 76%.

Mohammed Abdul Khalee et al. used the tool Weka and classification is executed by using 70% of Percentage Split. The accuracy offered by Naive Bayes is 86.419%. Since the classification task is influenced by the type of dataset and the way the classifier was implemented inside the toolkit, this study concluded that no tool is better than the other when used for a classification task.

M. Marimuthu, S.Deivarani, and Gayathri's paper analyses the various machine learning algorithms such as SVM, Naïve Bayes, decision tree, and k- nearest neighbour. It utilizes data such as blood

pressure, cholesterol, diabetes and then tries to predict the possible coronary heart disease patient in the next 10 years. The algorithm's accuracy was 83% for KNN, 80% for Naive Bayes, 75% for Decision Tree, and 65% for SVM.

Saba Bashir et al. The proposed study's aim is to improve the accuracy of heart disease prediction in patients. The proposed system employs a novel classifier ensemble based on a majority vote to solve problems. Different data mining classifiers should be combined. Heart disease at UCI. The accuracy of the algorithms is 78% for Naive Bayes, 72% for the Decision tree, and 75% for SVM.

Vikas Chaurasia and Saurabh Pal suggested using data mining approaches to detect heart disease. The WEKA data mining tool is used which contains a set of machine learning algorithms for mining purposes. Only 11 attributes are used for prediction. Naive Bayes offers 82.31% accuracy.

A H Chen et al. presented a heart disease prediction method that can help doctors predict heart disease status using patient clinical data. Thirteen essential clinical features were chosen, including age, sex, and form of chest pain. Heart disease was classified using an artificial neural network algorithm. The proposed method for prediction has an accuracy of about 80%.

Asha Rajkumar et al. worked on the diagnosis of heart disease using a classification algorithm based on supervised machine learning. Tanagra tool is used to classify the data, 10 fold cross-validation is used to evaluate the data and the results are compared. Tanagra is a free data mining application that can be used in academic and research areas. Naïve Bayes accuracy was 52.33% and k-NN accuracy was 45.67%.

K.Polaraju and D.Durga Prasad's paper aimed to use the statistical model Multiple Linear Regression Analysis is used to construct a model that reliably predicts the risk of heart disease and helps to diagnose in time to save a person's life. On trained data, Multiple Linear Regression Analysis is used to create a model to which test data is applied. Multiple Linear Regression is proven to be appropriate for estimating the risk of heart disease based on the experimental findings.

Jayamin Patel et al research shows that J48 gives 56.76% which is better than the Logistic Model Tree algorithm with an accuracy of 55.75%.

Ashok Kumar Dwivedi et al paper depicts that Naïve Bayes has 83% accuracy, KNN 80%, Logistic Regression 85%, and Classification Tree 77%.

K. Gomathi et al study results have an accuracy of 78% for Naive Bayes and 77% for J48.

Noura Ajam et al used Artificial Neural Networks that had an accuracy of 88%.

Sairabi H. Mujawar et al. used modified k-means and Naïve Bayes to predict heart disease. It had 89% accuracy in cases where it detected that a patient doesn't have heart disease.

R. Kannan et al compared logistic regression, Random Forest, stochastic gradient boosting, & SVM in R language. Logistic Regression (LR) had the highest accuracy of 87%.

For predicting heart attacks, **Hidayat TAKCI et al** suggested the best feature selection algorithms and machine learning methodology. Various machine learning approaches were used to find the best parameters and feature selection methods. The SVM algorithm with the linear kernel is the best machine learning approach, while the relief technique is the best feature selection algorithm, according to the results.

Rajesh Jangade et al. looked into different classification methods that could be used to predict heart problems. The aim of this project is to find a suitable methodology that will aid future decision-making. The contrast is made between the classifiers to see which one is more effective for the dataset. The accuracy of the decision tree is 75%.

Many elderly doctors assumed that high blood pressure was needed to push blood through the stiffened arteries of the elderly and that it was a natural part of aging. A permissible systolic BP, according to the medical community, was 100 plus the participant's age in millimetres of mercury. **(WB Kannel, JN Mickerson)** For those over the age of 70, the appropriate upper ranges of normal blood pressure were 210 mmHg systolic and 120 mmHg diastolic, according to some experts. **(FI Caird)**

Healthcare analytics is the systematic use of health data and related business insights built through the application of empirical, e.g. applied mathematics, contextual, quantitative, predictive, cognitive, and other models, to drive fact-based higher cognitive processes in care planning, management, measurement, and learning **(Cortada et al. 2012)**. Large data analytics would go beyond increasing income and reducing waste to predict epidemics, cure illnesses, raise the quality of living, and reduce preventable deaths **(Marr 2015).**

(HIMSS. Clinical & business intelligence: Associate in Nursing analytics government review) Healthcare analytics, according to HIMSS, is the "systematic use of information and related clinical and business (C&B) insights built across applied analytical disciplines such as applied mathematics, qualitative, quantitative, predictive, and psychological feature spectrums to drive fact-based higher cognitive processes for idea development, management, evaluation, and learning." Health analytics software is described as "a series of call support technologies for the healthcare industry aimed at enabling information workers such as physicians, nurses, and health officers, as well as health officials and pharmacists, to gain insight and make better and faster health decisions."

"Health analytics is the use of information, data technology, applied mathematics analysis, quantitative strategies, and mathematical computer-based models to assist health care providers gain better insight into these patients and make higher, fact-based decisions," we suggest as another concept.

When using big data, descriptive, predictive, and prescriptive analytics approaches can be used to improve the efficiency of different areas of healthcare.

Medical diagnosis: A data-driven diagnosis can help diagnose diseases early on and reduce care complications. **(Gu et al. 2017; Raghupathi and Raghupathi 2014).**

- **Community healthcare:** Authorities can take measures to prevent chronic disease outbreaks and infectious disease outbreaks in a population **(Lin et al. 2017). (Antoine-Moussiaux et al. 2019).**
- **Hospital monitoring:** Real-time monitoring of hospitals can help government authorities ensure optimal service quality **(Archenaa and Anita 2015).**
- **Patient care:** Customised patient care facilitated by analytics has the potential to provide rapid relief **(Salomi and Balamurugan 2016)** and reduce readmission rates in hospitals **(Gowsalya, Krushitha, and Valliammai 2014).**

(Gopalakrishna Palem) The aim of predictive analytics, according to the author, is to assist businesses in translating data into actionable insights that can help them make better business decisions. Increased global competition and the need for long-term growth are causing a growing number of businesses to pursue analytical methods to gain market insights.

Segments:

- **Critical care intervention:** Clinical surveillance for real-time bedside and remote monitoring solutions offers clinicians proactive warnings for important new values, decreasing the risk of infection, adverse drug events, and other complications for patients.
- **Diagnostic assistance:** Advanced speech and natural language processing methodologies assist clinicians in recommending the appropriate diagnosis routines based on patient symptoms, allowing them to easily get to the root cause of the condition while reducing diagnostic test costs.
- **Clinical decision support:** Intelligent decision support systems assist physicians in assessing the appropriate care plan tailored for each patient by reviewing diagnosis outcomes by cross-referencing patient medical history notes, previous cases, clinical trials, and reference materials in real-time to recommend potential treatment behavior.
- **Disease management:** Clinical incidents and treatment procedures from a variety of health services are reported to central immunity records, which keep track of public health protection in real-time, looking for emerging patterns and possible disease outbreaks.
- **Personal healthcare:** Pervasive and context-aware technologies have long been recognized as potential options for improving the quality of life for both chronic disease patients and their families, as well as lowering long-term health care costs and improving care quality.

Telemedicine, assisted living through supportive sensor aids and safety guides, medication warnings, and wearable body-tracking sensors are just a few of the personal healthcare initiatives that are gaining traction.

- **Readmission prevention:** By collecting patient-specific data, it is possible to minimize readmission rates in advance, equate treatment efficiency to industry-wide results, and concentrate resources on the most successful therapies. Targeting and monitoring hospital discharge and follow-up treatment effectively lowers readmission costs and avoids repeat admissions.

In data mining tasks, classification and clustering have been two major problems. The supervised learning task of identifying common properties among a collection of objects in a database and categorizing them into different classes is known as classification **Chen, M. S et. al.** Since both bring identical objects into the same group, classification and clustering are closely related. In classification, each class's label is a distinct and well-known category, while in clustering problems, the label is an unknown category

Xu, R., and Wunsch say that Clustering was previously thought to be a type of unsupervised classification. The clustering method summarises data patterns from the dataset since there are no current class labels. Traditionally, medical data mining has been viewed as a classification challenge, with the aim of finding the best classifier to identify patients. Data mining is currently being used by researchers. techniques in the diagnosis of a variety of diseases, including diabetes, stroke, cancer, and heart disease, to name a few. **(Jain et al)**

Table 2.1: Accuracy Table

S.NO	AUTHOR	ALGORITHM WITH HIGHEST ACCURACY	PERCENTAGE (%) OF ACCURACY
1	J Thomas and R Theresa Princy et al.	K Nearest Neighbour	80%
2	Sandhya Kumari and Dr. R. Viswanathan	K Nearest Neighbour	86%
3	AH Chen, SY Huang, et al.	Artificial Neural Networks	80%
4	A. Sheik Abdullah	J48	61%
5	G Purushottam et al.	Decision Tree	85%
6	Vikas Chaurasia et al.	Cart	83%

7	Boshra Baharami et al.	J48	84%
8	Sellappan Palaniyappan, Rafiah Awang et al.	Naïve Bayes	86%
9	M.A.Nishara Banu B.Gomathy	C45	89%
10	Ramandeep Kaur, Er. Prabhakaran Kaur et al	C45	89%
11	G. Parthiban et al.	Naïve Bayes	74%
12	Vembandasamy et al.	Naïve Bayes	84%
13	Yashvendra K. Singh et al.	Random Forest	86%
14	Seyedamin Pouriyeh, Sara Vahid et al.	K Nearest Neighbour	83%
15	Tan et al.	Support Vector Machine	84%
16	Ottom et al.	Support Vector Machine	88%
17	Arunpradeep N., G. Niranjana	K Nearest Neighbour	86%
18	S.Rajathi and Dr.G.Radhamani	K Nearest Neighbour	70%
19	Ridhi Saini et al	K Nearest Neighbour	88%
20	Kanika Pahwa and Ravinder Kumar et al.	Naïve Bayes	78%
21	Mohammed Abdul Khalee et al.	Naïve Bayes	86%

22	M. Marimuthu, S. Deivarani and Gayathri	K Nearest Neighbour	83%
23	Saba Bashir et al.	Naïve Bayes	78%
24	Vikas Chaurasia and Saurabh Pal	Naïve Bayes	82%
25	A H Chen et al.	Artificial Neural Networks	80%
26	Asha Rajkumar et al.	Naïve Bayes	52%
27	Jayamin Patel et al	J48	56%
28	Ashok Kumar Dwivedi et al	Logistic Regression	85%
29	K. Gomathi et al.	Naïve Bayes	78%
30	Noura Ajam et al.	Artificial Neural Networks	88%
31	Sairabi H. Mujawar et al.	Naïve Bayes	89%
32	R. Kannan et al.	Logistic Regression	87%
33	Rajesh Jangade et al.	Decision Tree	75%

CHAPTER III

RESEARCH METHODOLOGY

3.1 SUPERVISED MACHINE LEARNING

Machine learning algorithms that are supervised are designed to learn by example. The term "supervised" learning comes from the fact that training such an algorithm is similar to having an instructor oversee the entire operation. The training data for a supervised learning algorithm would consist of inputs that are combined with the right outputs. The algorithm will look for patterns in the data that correspond with the desired outputs during training. Following training, a supervised learning algorithm can take in new unseen inputs and, using prior training data, decide which new inputs will be labelled as. A supervised learning model's goal is to predict the correct label for newly presented data. A supervised learning algorithm can be written as follows in its most basic form:

$$Y=f(x)$$

Where Y is the expected output of a mapping function that assigns a class to an input value x. During preparation, the machine learning model creates the function that connects input features to a projected output. Supervised learning can be split into two subcategories:

1. Classification
2. Regression

Classification

A classification algorithm will be given data points with an assigned category during preparation. A classification algorithm's task is to take an input value and, based on the training data given, assign it to a class, or category, where it belongs. The classification in this study is used to determine whether or not a person is at risk for heart disease. This problem is known as a binary classification problem since there are only two groups to choose from (present or absent). The algorithm will be fed training data containing heart disease data from people with and without the disease. The model will look for features in the data that are associated with either class and use them to construct the mapping function described earlier:

$$Y=f(x)$$

The model will then use this feature to assess whether or not heart disease is present when given new data. A variety of algorithms can be used to solve classification problems, including:

- Linear Classifiers
- Support Vector Machines
- Decision Trees
- K-Nearest Neighbour
- Random Forest
- Naive Bayes

Regression

Regression is a mathematical method that seeks to find a significant association between dependent and independent variables. Regression algorithms come in a variety of shapes and sizes. The three most popular ones are as follows:

- Linear Regression
- Logistic Regression
- Polynomial Regression

Supervised learning is the most basic subcategory of machine learning, and for many machine learning practitioners, it serves as an introduction to the field. The most popular form of machine learning is supervised learning, which has proven to be a useful method in a variety of fields.

3.2 MACHINE LEARNING ALGORITHMS

This research has made use of the following supervised machine learning algorithms.

1. K Nearest Neighbour
2. Logistic Regression
3. Decision Tree
4. Random Forest
5. Naive Bayes

3.2.1 K Nearest Neighbour Algorithm

The KNN algorithm assumes that items that are identical are close together. To put it another way, related things are close together. Both classification and regression predictive problems can be solved with KNN. However, in the market, it is more commonly used in classification problems. We look at three main aspects when evaluating any technique:

1. The performance is simple to understand.
2. Time to calculate
3. Predictive Capacity

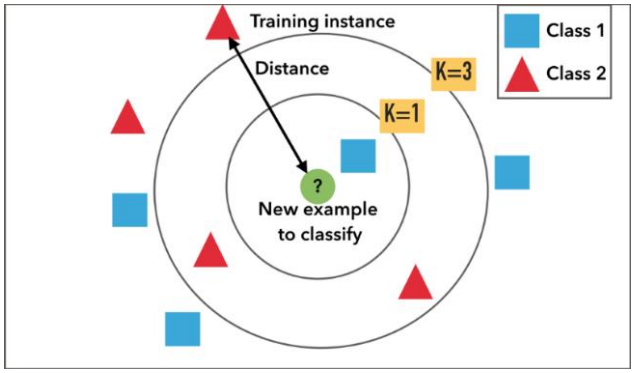


Figure 3.1: KNN

The letter K in KNN refers to the number of nearest neighbours taken into account when assigning a mark to the current point. K is a key parameter, and determining its value is the most difficult problem when using the KNN algorithm. The process of selecting the appropriate value of K is known as parameter tuning, and it is critical for improving accuracy. If K is too small, the model may be overfitted, and if it is too high, the algorithm becomes computationally costly. When there are two groups, most data scientists select an odd number value for K.

Choosing the value of K is a case-by-case decision, and often the best way to choose K is to experiment with various values of K and compare the results. The KNN algorithm can be evaluated for different values of K using cross-validation, and the value of K that results in good accuracy can be considered the optimal value for K.

The key drawback of KNN is that it becomes significantly slower as the amount of data grows, making it an impractical option in situations where rapid predictions are needed. It is called a lazy learner because it does not go through a training process and instead memorises the training dataset. All computations are postponed until the classification is completed. Furthermore, faster algorithms can generate more precise classification and regression results.

3.2.2 Logistic Regression

There are three different Logistic Regression variations that can occur:

- Binary Logistic Regression: There are only two possible outcomes in binary logistic regression (Category).
- Multinomial Logistic Regression: Allows to have more than two categories without having to order them.
- Ordinal Logistic Regression: Ordering allows for more than two categories of ordinal logistic regression.
- In this context, the Binary Logistic Regression method is used.

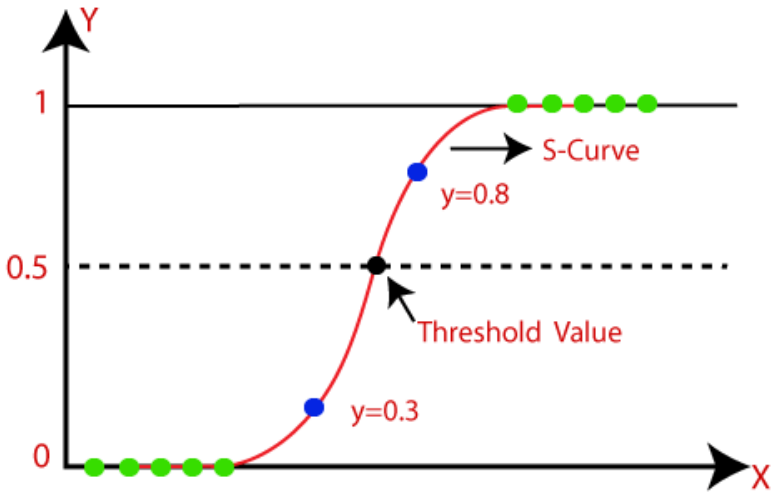


Figure 3.2: Logistic Regression

While logistic regression appears to be a relatively easy algorithm to learn and implement, it has a number of limitations. It can only be used on massive datasets, for example. In order to implement this machine learning algorithm, several assumptions must be made in a dataset.

- In a binary logistic equation, the dependent variable must be binary.
- The dependent variable's factor level 1 should reflect the desired outcome.
- It's possible that using non-meaningful variables will result in errors. Include only the variables that are required and which display a connection.
- The model should have little to no multicollinearity, which means that the independent variables should be completely unrelated to one another.
- The log odds are linearly connected to the independent variables.

One would think that the equation isn't robust enough to be applied to real-world problems because there are too many assumptions to make, but this equation has a lot of uses in the medical sector and is empowering people all over the world with its superpower. The logistic regression equation is widely used when a binary answer is anticipated or inferred. Although it has found the best application in the field of medicine, the possibilities are endless.

Calculating the chances of suffering a heart attack

Medical researchers use data to figure out how the predictor variables interact to predict whether or not a person will have a heart attack. The model's findings show researchers how changes in exercise and weight (predictor variables) influence the likelihood of a heart attack in a given person. In this case, a fitted logistic regression model is used.

3.2.3 Decision Tree Algorithm

Decision Trees are a form of supervised machine learning in which the data is continuously split according to a parameter (describe what the input is and what the corresponding output is in the training data). Two entities, decision nodes and leaves, can be used to illustrate the tree. The decisions or final results are represented by the leaves and the data is divided at the decision nodes. The binary tree can be used to describe a decision tree. Here the aim is to predict whether a person has heart disease or not based on their age, sex, and physical activity, among other factors. Questions like ‘What is his age?', ‘Does he exercise?', and ‘Does he eat a lot of junk food?' are the decision nodes here. And then there are the leaves, which are either ‘presence of heart disease’ or ‘absence of heart disease’' outcomes. A binary classification problem in this case (a yes no type problem).

Trees of classification (Yes/No types)

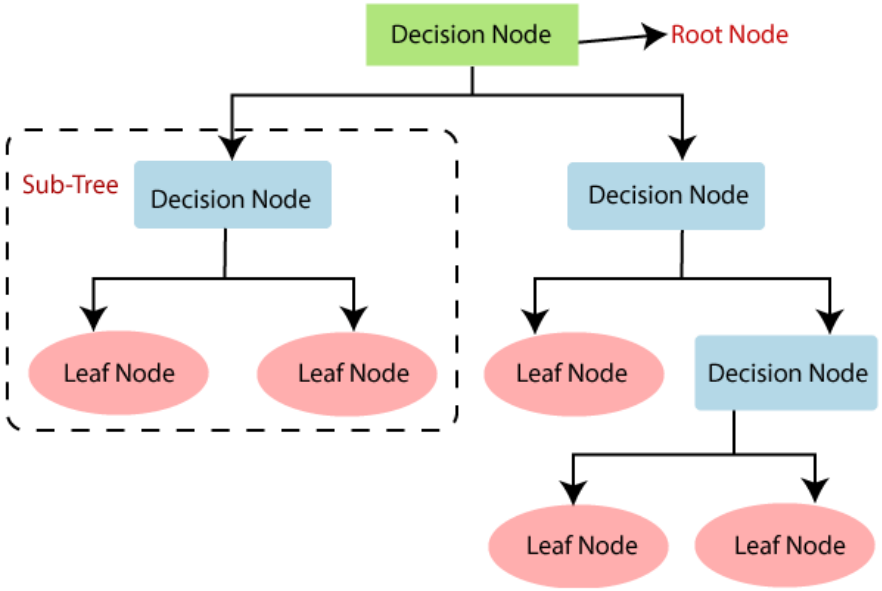


Figure 3.3: Decision Tree

This paper has made use of a classification tree, with the result being a component such as "presence" or "absence" of heart disease. Categorical is the decision variable in this case. At each stage of the tree-building process, information gain is used to determine which function to break on. We want to keep our tree small because simplicity is best. To do so, we should choose the split that produces the purest daughter nodes at each stage. The term "data" refers to a widely used purity metric. The information value for each node of the tree indicates how much information a function provides about the class. The first split will be the one with the highest information gain, and the process will continue until all children nodes are pure, or until the information gain is zero.

Decision Tree's Drawbacks

- Overfitting is a concern.
- Require some kind of evaluation to see how well they're doing.
- Parameter tuning must be done with caution.
- If certain groups dominate, this can result in biased learned trees.

To determine whether to divide a node into two or more sub-nodes, decision trees employ a variety of algorithms. The homogeneity of the resulting sub-nodes improves with the construction of sub-nodes. To put it another way, the purity of the node improves as the goal variable increases. The decision tree splits the nodes into sub-nodes based on all available variables, then chooses the split that produces the most homogeneous sub-nodes. The type of target variables is often taken into account when choosing an algorithm.

3.2.4 Random Forest Algorithm

Random forest is a machine learning algorithm that is supervised. It creates a "forest" out of an ensemble of decision trees, which are normally trained using the "bagging" process. The bagging method's

basic premise is that combining different learning models improves the overall outcome. Random forest has the benefit of being able to solve classification and regression problems, which make up the majority of existing machine learning systems. Since classification is often regarded as a fundamental component of machine learning. A random forest with two trees will look like this:

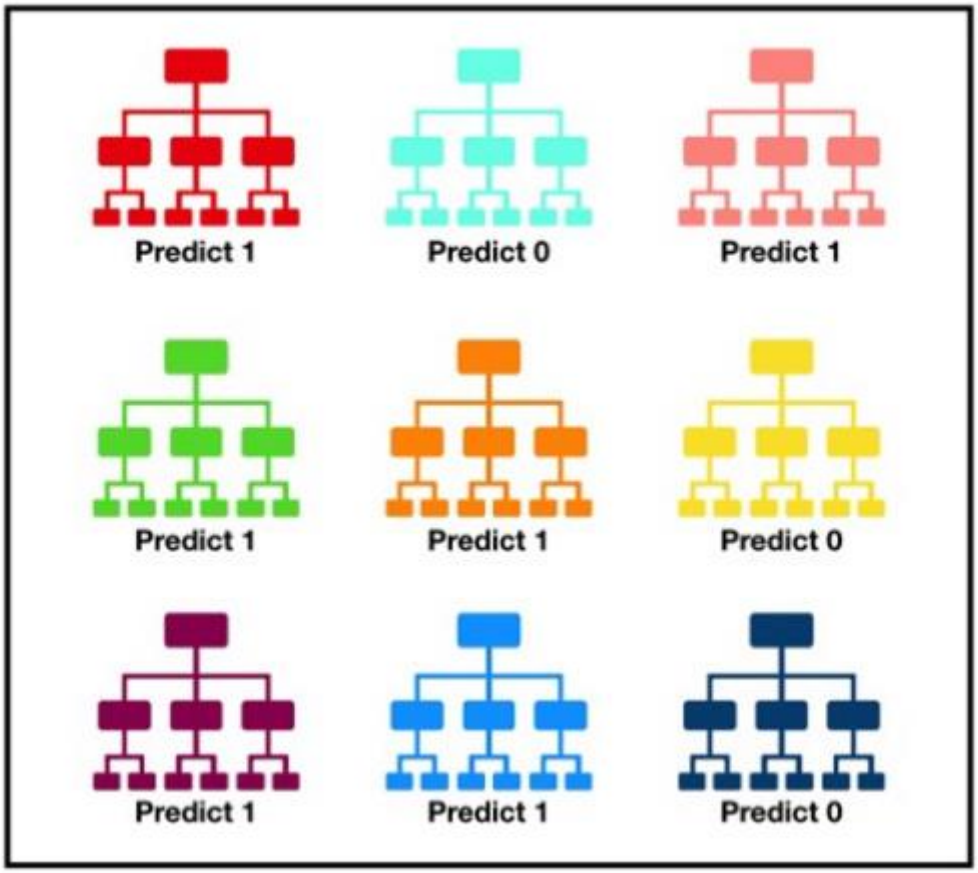


Figure 3.4: Random Forest Algorithm

The hyperparameters of a random forest are very similar to those of a decision tree or a bagging classifier. Fortunately, the classifier-class of random forest can be used instead of combining a decision tree and a bagging classifier. The algorithm's regressor to deal with regression tasks with random forest. While increasing the trees, the random forest adds more randomness to the model. When splitting a node, it looks for the best function among a random subset of features rather than the most appropriate feature. As a consequence, there is a lot of variety, which leads to a better model. As a result, in random forest, the algorithm for splitting a node only considers a random subset of the features. Instead of looking for the best possible thresholds, the trees can be rendered more random by using random thresholds for each element (like a normal decision tree does).

There are some variations between a random forest and a group of decision trees. When a decision is given to a decision tree a training dataset with features and labels, it will generate a collection of rules that will be used to make predictions. Another distinction is that "deep" decision trees can be prone to overfitting. Random forest usually avoids this by generating random subsets of the features and using those subsets to create smaller trees. It then joins the subtrees together. It's important to remember that this doesn't

always work, and it also slows down the calculation depending on how many trees the random forest generates.

3.2.5 Naive Bayes

The Naive Bayes algorithm is a supervised learning algorithm for solving classification problems that is based on the Bayes theorem. It is primarily used in text classification tasks that require a large training dataset. The Naive Bayes Classifier is a simple and effective classification algorithm that aids in the development of fast machine learning models capable of making quick predictions. It's a probabilistic classifier, which means it makes predictions based on an object's likelihood. Spam filtration, sentiment analysis, and article classification are all common uses of the Naive Bayes Algorithm.

The two words Naive and Bayes make up the Naive Bayes algorithm, which can be summarised as follows:

- It's named Naive because it believes that the appearance of one feature is unrelated to the appearance of other features. If the colour, form, and taste of the fruit are used to identify it, a red, spherical, and sweet fruit is known as an apple. As a result, each function contributes to identifying that it is an apple without relying on the others.
- It's named Bayes because it's based on the Bayes' Theorem theory.

Bayes' theorem, also known as Bayes' Rule or Bayes' law, is a mathematical formula for calculating the likelihood of a hypothesis given prior information. It is conditional probability that determines this.
The Bayes theorem's formula is as follows:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

- ➔ Posterior probability (P(A|B)) is the probability of hypothesis A on the observed case B.
- ➔ P(B|A) stands for Potential Probability, which is the probability of the evidence provided that a hypothesis' probability is valid.
- ➔ Prior Probability (P(A)) is the probability of a hypothesis before seeing the proof.
- ➔ P(B) stands for Probability of Evidence Marginal Probability.

Naive Bayes is a fast and simple machine learning algorithm for predicting a class of datasets. It's suitable for both binary and multi-class classifications. In comparison to the other Algorithms, it performs well in Multi-class predictions.
There are three different types of Naive Bayes Models, as described below:

Gaussian: The Gaussian model assumes that features are distributed in a natural manner. If predictors take continuous values rather than discrete values, the model assumes that these values are drawn from a Gaussian distribution.

Multinomial Naive Bayes: When the data is multinomial distributed, the Multinomial Naive Bayes classifier is used. It is mainly used to solve document classification issues, which means determining which group a document belongs to, such as Sports, Politics, or Education. The predictors in the classifier are based on the frequency of terms.

Bernoulli: Like the Multinomial classifier, the Bernoulli classifier uses independent Booleans variables as predictor variables. For example, determining whether or not a specific word appears in a text. This model is also well-known for tasks involving document classification.

In most real-world cases, the conclusions made by Naive Bayes are incorrect. In reality, the presumption of independence is never accurate, but it often works well in practice.

3.3 PYTHON FOR DATA ANALYSIS

Python for Data Analysis is all about manipulating, sorting, cleaning, and crunching data in Python. It's also a realistic, up-to-date introduction to Python scientific computing, with a focus on data-intensive applications. Python is a popular multi-purpose programming language that is popular for its flexibility and large library of libraries, which are useful for analytics and complex calculations. Because of Python's extensibility, thousands of libraries dedicated to analytics exist, including the commonly popular Python Data Analysis Library (also known as Pandas). The NumPy library, which contains hundreds of mathematical equations, operations, and functions, is at least partially derived from data analytics libraries in Python. This research has made use of Python for analysis of the heart disease dataset.

3.3.1 Pandas Library

Pandas is a Python package that provides quick, versatile, and expressive data structures for working with structured (tabular, multidimensional, potentially heterogeneous) and time series data. Its aim is to serve as the foundation for doing realistic, real-world data analysis in Python. Furthermore, it aspires to be the most effective and versatile open-source data analysis and manipulation tool available in any language. It is well on its way to achieving this goal. Pandas' two primary data structures, Series (1-dimensional) and DataFrame (2-dimensional), are capable of handling the vast majority of common use cases in economics, statistics, social science, and many fields of engineering.

3.3.2 Numpy Library

NumPy is a Python library that allows to work with arrays. It also has functions for dealing with matrices, Fourier transforms, and linear algebra. Travis Oliphant invented NumPy in 2005. It is an open-source project that is free to use. Numerical Python is referred to as NumPy. We have lists in Python that serve as arrays, but they are slow to process. NumPy aims to have a 50-fold faster array object than standard Python lists. The array object in NumPy is called ndarray, and it comes with a slew of helper functions to make interacting with it a breeze. In data science, where speed and resources are critical, arrays are

commonly used. NumPy arrays, unlike lists, are stored in a single continuous location in memory, allowing processes to access and manipulate them quickly. In computer science, this is referred to as locality of reference. This is the primary explanation why NumPy outperforms lists. It's also been tweaked to work with the most recent CPU architectures.

3.3.3 Matplot Library

Matplotlib is a visualisation tool that uses a low-level graph plotting library written in Python. John D. Hunter is the creator of Matplotlib. Matplotlib is free and open-source software that we can use. For platform compatibility, Matplotlib is mainly written in Python, with a few segments written in C, Objective-C, and Javascript. Most of the Matplotlib utilities lie under the pyplot submodule, and are usually imported under the plt alias. Matplotlib is a Python library that allows to create static, animated, and interactive visualisations.

3.3.4 Scikit Learn

David Cournapeau started developing Scikit-learn as a Google summer of code project in 2007. Matthieu Brucher later joined the project and began using it as part of his thesis research. INRIA became active in 2010, and the first public update (v0.1 beta) was released in late January. INRIA, Google, Tiny clues, and the Python Software Foundation have all paid sponsorship to the project, which now has over 30 active contributors. Scikit-learn offers a consistent Python framework for a variety of supervised and unsupervised learning algorithms. It is distributed under many Linux distributions and is licenced under a permissive simplified BSD licence, allowing academic and commercial use.

Scikit-learn is based on SciPy (Scientific Python), which must be enabled before one can use it. This stack contains the following items:

- NumPy is a Python package for creating n-dimensional arrays.
- SciPy is a Python-based scientific computing library.
- Matplotlib is a 2D/3D plotting library.
- IPython: A more interactive Python environment
- Symbolic mathematics (Sympy)
- Data structures and analysis with Pandas

Scikit-learn offers a variety of common model classes:

1. Clustering is a technique for organising unlabelled data, such as KMeans.
2. Cross Validation is a technique for estimating the efficiency of supervised models on data that hasn't been seen before.
3. Datasets: for testing and producing datasets with particular properties in order to investigate model behaviour.

4. Principal component analysis, for example, uses dimensionality reduction to reduce the number of attributes in data for summarization, visualisation, and feature selection.
5. Feature extraction is a technique for extracting attributes from image and text data.
6. Feature selection is used to define meaningful attributes from which supervised models can be built.

Seaborn Library

Seaborn is a matplotlib-based Python data visualisation library. It has a high-level GUI for creating visually appealing and insightful statistical graphics. Its plotting functions work with data frames and arrays containing entire datasets, performing the required semantic mapping and statistical aggregation internally to generate informative plots. Its dataset-oriented, declarative API allows to concentrate on the meaning of the plots rather than the mechanics of drawing them.

Train-test-split Procedure

When machine learning algorithms are used to make predictions on data that was not used to train the model, the train-test split method is used to estimate their results. It's a quick and simple procedure that allows to compare the output of different machine learning algorithms for the predictive modelling problem. Although the technique is easy to use and interpret, there are occasions when it should not be used, such as when the dataset is small or when additional setup is required, such as when it is used for classification and the dataset is unbalanced.

Taking a dataset and splitting it into two subsets is the method. The training dataset is the first subset, which is used to match the model. The second subset is not used to train the model; instead, the dataset's input element is given to the model, which then makes predictions and compares them to the expected values. The test dataset is the name given to the second dataset.

- Train Dataset: This is the data set that is used to match the machine learning model.
- Test Dataset: This dataset is used to assess how well a machine learning model suit.

The aim is to estimate the machine learning model's output on new data that was not used to train the model.

Standard Sclaer

When numerical input variables are scaled to a standard range, many machine learning algorithms perform better. This involves algorithms like linear regression that use a weighted sum of the input and algorithms like k-nearest neighbours that use distance steps. Normalization and standardisation are the two most popular methods for scaling numerical data before modelling. Normalization scales each input variable independently to the range 0-1, which is the most precise range for floating-point values. Standardization shifts the distribution to have a mean of zero and a standard deviation of one by subtracting the mean (called centering) and dividing by the standard deviation for each input variable. Differences in scales between input variables can make the problem more difficult to model. Large input values (for

example, a range of hundreds or thousands of units) may lead to a model that learns large weight values. A model with large weight values is more likely to be unstable, which means it will perform poorly during learning and be more sensitive to input values, resulting in higher generalisation error.

Accuracy score

One metric for assessing classification models is accuracy. Informally, accuracy refers to the percentage of correct predictions made by our model. The following is the formal concept of accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

The following formula can be used to measure accuracy in terms of positives and negatives for binary classification:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP stands for True Positives, TN stands for True Negatives, FP stands for False Positives, and FN stands for False Negatives.

A true positive is when the model predicts the positive class correctly. A true negative, on the other hand, is an outcome in which the model correctly predicts the negative class.

A false positive occurs when the model predicts the positive class incorrectly. A false negative is a result in which the model predicts the negative class incorrectly.

Confusion Matrix

A confusion matrix is a method of summarising a classification algorithm's results. If there are an unequal number of observations in each class or if the dataset has more than two classes, classification accuracy alone can be misleading. Calculating a confusion matrix will help see what the classification model is doing correct and where it's going wrong. A confusion matrix is a summary of classification problem prediction outcomes. The number of accurate and incorrect predictions is totaled and broken down by class using count values. The uncertainty matrix's secret is this. The confusion matrix depicts how the classification model becomes perplexed when making predictions. It informs not only about the errors produced by the classifier, but also about the types of errors that are being made. This breakdown overcomes the drawback of relying solely on classification accuracy.

3.4 TABLEAU FOR PREDICTIVE DASHBOARD

In the Business Intelligence Industry, Tableau is a strong and rapidly growing data visualisation tool. It aids in the simplification of raw data into a format that is simple to comprehend. Tableau assists in the creation of data that is understandable for experts at all levels of an enterprise. Non-technical users can

also build personalised dashboards. With Tableau, data analysis is fast, and visualisations are generated in the form of dashboards and worksheets. On Tableau, a dashboard is a series of various types of visualisations or views. We can combine elements from multiple worksheets and display them on a single dashboard. To build a dashboard, we can import and add charts and graphs from worksheets using the dashboard option. We can put related charts and graphs in one view on a dashboard and examine them for better insights.

Tabpy Server

TabPy (the Tableau Python Server) is an external service implementation that extends Tableau's functionality by allowing users to use Tableau's table calculations to execute Python scripts and saved functions. Python is a widely used programming language by analysts and computer scientists in a variety of industries for tasks ranging from data cleaning and manipulation to the implementation of cutting-edge machine learning algorithms.

CHAPTER IV

DATA ANALYSIS AND INTERPRETATION

4.1 DATA ANALYSIS

Heart disease data set has been taken from tableau. The attributes considered for analysis are as follows:

1. Age - Age of the individual in years ranging from 29 years to 77 years
2. Sex - Male or Female

   1 - Male and 0 - Female

3. Chest pain type - 4 levels of chest pain
4. Resting BP

   Resting Blood Pressure on admission to the hospital

   Two numbers are used to calculate blood pressure:

   - The first figure, systolic blood pressure, indicates how much pressure is in the arteries as the heart beats.
   - The second number, diastolic blood pressure, measures the pressure in the arteries between heartbeats.

Table 4.1: BP readings

Normal	systolic: less than 120 mm Hg diastolic: less than 80 mm Hg	Normal	systolic: less than 120 mm Hg diastolic: less than 80 mm Hg
At Risk (prehypertension)	systolic: 120–139 mm Hg diastolic: 80–89 mm Hg	Elevated	systolic: 120–129 mm Hg diastolic: less than 80 mm Hg
High Blood Pressure (hypertension)	systolic: 140 mm Hg or higher diastolic: 90 mm Hg or higher	High blood pressure (hypertension)	systolic: 130 mm Hg or higher diastolic: 80 mm Hg or higher

5. Serum Cholesterol in mg/dl

Cholesterol is often linked to heart disease. Since low-density lipoproteins (LDL) can build up in the arteries and obstruct or block blood flow, this is something one should be aware of. A small amount of cholesterol is still needed by the body for proper digestion, as well as the production of vitamin D and certain hormones. Cholesterol is a fatty substance. It's also known as a lipid. It moves through the bloodstream in the form of tiny molecules encased in proteins. Lipoproteins are the names for these packets. LDL is one of the most common lipoproteins found in the blood. Measuring LDL ("bad") cholesterol levels will assist doctors in determining the likelihood of developing heart disease over the next ten years.

Table 4.2: LDL Levels

Healthy serum cholesterol	less than 200 mg/dL
Healthy LDL cholesterol	less than 130 mg/dL

6. Fasting blood sugar

A blood sugar test is performed after a period of fasting. After an overnight short, a blood sample will be taken. It is common to have a fasting blood sugar level of less than 100 mg/dL (5.6 mmol/L). Prediabetes is described as a fasting blood sugar level of 100 to 125 mg/dL (5.6 to 6.9 mmol/L). I f the blood sugar levels are 126 mg/dL (7 mmol/L) or higher on two different tests then it depicts the presence of diabetes.

7. Resting ECG results

The resting ECG is a straightforward, simple, and painless procedure. The resting ECG can detect heart hypertrophy, ischemia, myocardial infarction, myocardial infarction sequelae, cardiac arrhythmias, and other heart disorders. The test takes about 5 minutes and there is no need to prepare.

8. Maximum heart rate

Subtract age from 220 to get an approximation of the mean age-related heart rate. The estimated maximum age-related heart rate for a 50-year-old male, for example, would be 220 – 50 years = 170 beats per minute (bpm).

9. Maximum heart rate achieves - Maximum heart rate during strenuous activities (71 – 202)

Table 4.3: Maximum Heart rate range

Age	Average Maximum Heart Rate
20	200
30	190
35	185
40	180
45	175
50	170
55	165
60	160
70	155
80	150

10. Exercise-induced angina

1 - Yes, 0 - No

Physical activity is commonly the cause of Angina. The heart needs more blood while climbing stairs, doing exercise, or walking, however narrowed arteries restrict blood flow. Other causes, in addition to physical activity, such as emotional tension, cold temperatures, heavy meals, and smoking, can narrow arteries and cause angina.

11. Sbp (systolic blood pressure)

The top number in 120/80 is systolic blood pressure, which calculates the force the heart exerts on the walls of arteries each time it beats.

12. Tobacco (cumulative tobacco (kg))
13. Ldl (low density lipoprotein cholesterol)

Low-density lipoproteins (LDL) are an acronym for low-density lipoproteins. Since a high LDL level causes an accumulation of cholesterol in the arteries, it is often referred to as "poor" cholesterol.

Table 4.4: LDL Category

LDL (Bad) Cholesterol Level	LDL Cholesterol Category
<100 mg/dL	Optimal
100-129mg/dL	Near optimal/above optimal
130-159 mg/dL	Borderline high
160-189 mg/dL	High

14. Adiposity - BMI of a person

The BMI is a straightforward measure based on a person's height and weight. BMI = kg/m2, where kg represents a person's weight in kilogrammes and m2 represents their height in metres squared. Overweight is described as a BMI of 25.0 or higher, while the healthy range is 18.5 to 24.9. Most adults between the ages of 18 and 65 have a BMI.

15. Famhist - Family history of heart disease, whether a family member has suffered from a heart disease or not.

A factor with levels "0 - Absent" and "1 - Present"

16. Typea

Type-A behaviour intense striving for achievement, competition, easily provoked impatience, time urgency, the abruptness of gesture and speech (explosive voice), hyper-alert posture, overcommitment to vocation or profession

17. Alcohol - current alcohol consumption
18. Chd - coronary heart disease

0 - Absent, 1 - Present

4.2 DATA INTERPRETATION

4.2.1 Statistics of the data

	age	sex	cp	trestbps	chol
count	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026
std	9.082101	0.466011	1.032052	17.538143	51.830751
min	29.000000	0.000000	0.000000	94.000000	126.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000
max	77.000000	1.000000	3.000000	200.000000	564.000000

fbs	restecg	thalach	exang
303.000000	303.000000	303.000000	303.000000
0.148515	0.528053	149.646865	0.326733
0.356198	0.525860	22.905161	0.469794
0.000000	0.000000	71.000000	0.000000
0.000000	0.000000	133.500000	0.000000
0.000000	1.000000	153.000000	0.000000
0.000000	1.000000	166.000000	1.000000
1.000000	2.000000	202.000000	1.000000

Figure 4.1: Statistics

Statistical Details Describe provides us with statistical information in the numerical format. It can be inferred that in the AGE column the minimum age is 29yrs and maximum is 77yrs mean of age is 54yrs. The quartile details are given in the form of 25%, 50% and 75%. The data is divided into 3 quartiles or 4 equal parts. so, 25% values lie in each group. standard deviation and mean are statistical measures which give us an idea of the central tendency of the data set.

The data set is checked to see if there are any null or missing values, the data used in this research does not have any null values.

4.2.2 Correlation

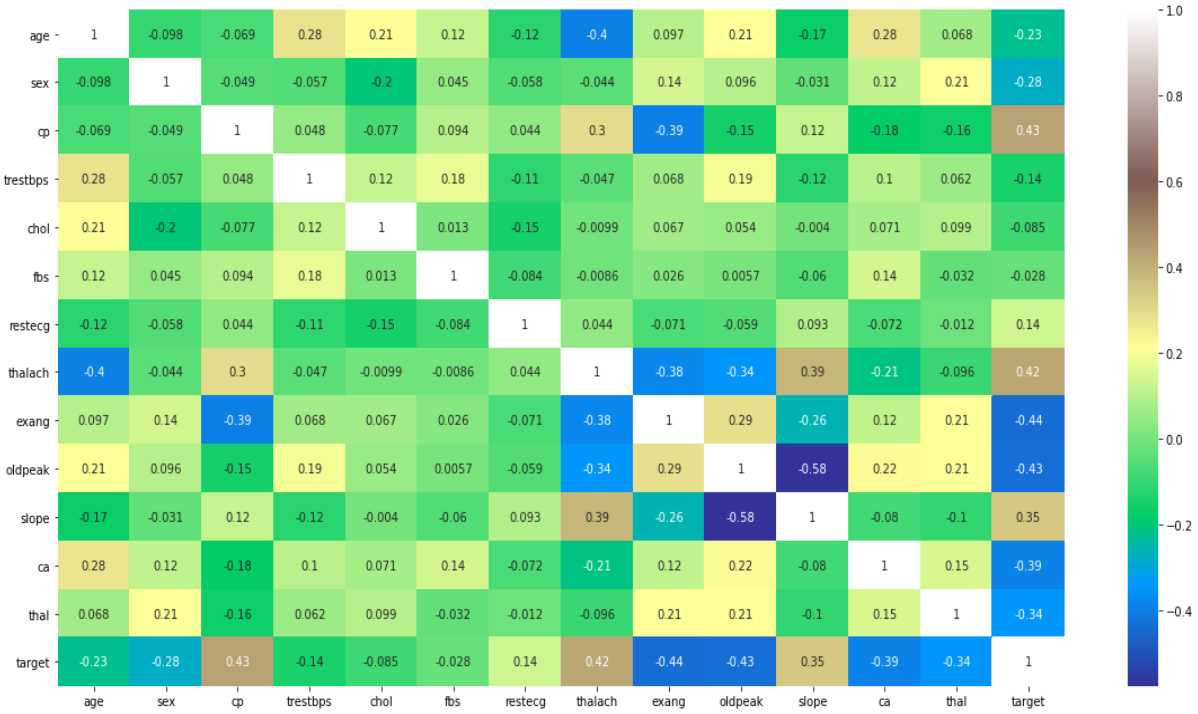


Figure 4.2: Correlation

The degree to which two variables change in relation to each other is measured by correlation. Values closer to +1 means there is a positive correlation and values closer to -1 depict a negative correlation.

From the correlation diagram the following can be depicted:

- As age increases individuals are prone to high blood pressure.
- If an individual has chest pain the chances of heart disease are high.
- The sex of a person does not influence the onset of a heart disease to a great extent.

4.2.3 Distribution of values

A histogram has been used to show the distribution of data for each attribute.

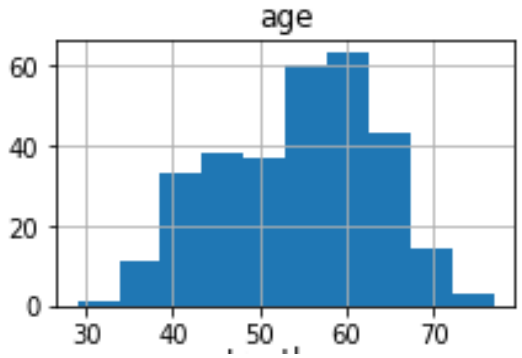


Figure 4.3: Histogram for age

Age is spread across 30 - 77 years.

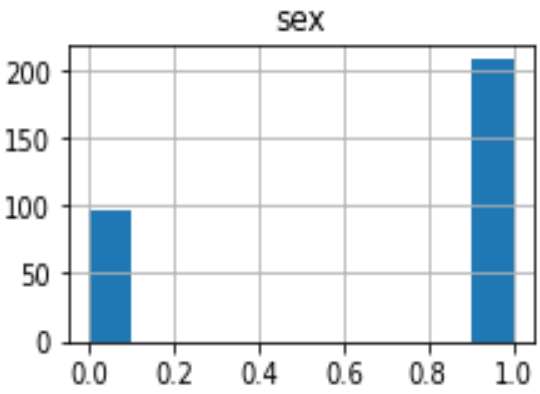


Figure 4.4: Histogram for Sex

There are more males in this data than females. 1 - Male and 0 - Female.

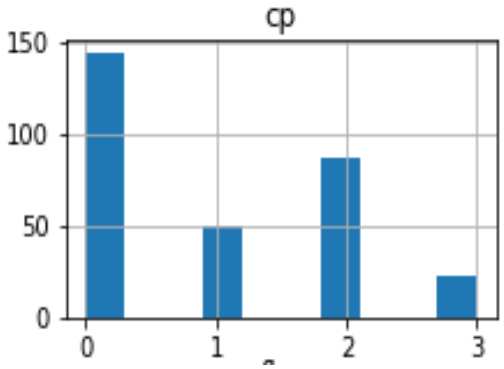


Figure 4.5: Histogram for chest pain

Individuals with chest pain of level 0 are high followed by level 2.

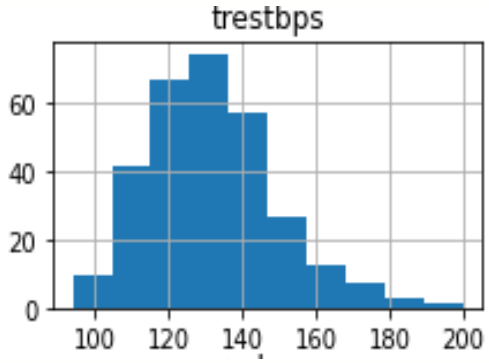


Figure 4.6: Histogram for Resting BP

The blood pressure values range from 80 to 200.

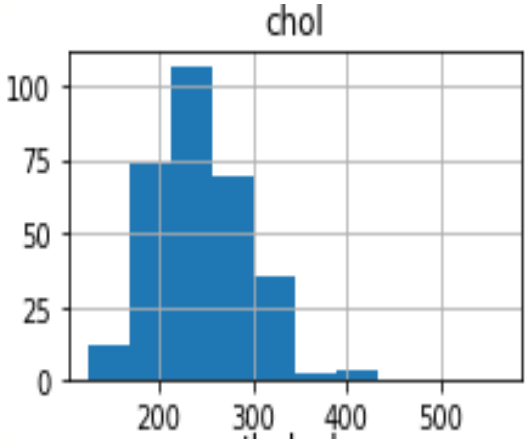


Figure 4.7: Histogram for Cholesterol

Cholesterol levels range from 100 to 450.

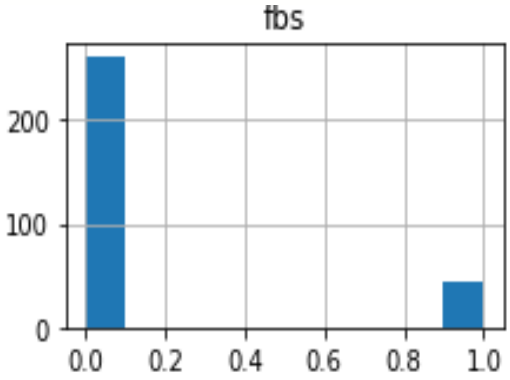


Figure 4.8: Histogram for Blood Sugar

1 - High fasting blood sugar, 0 - Normal Fasting Blood sugar.

This data set has more data of individuals without diabetes.

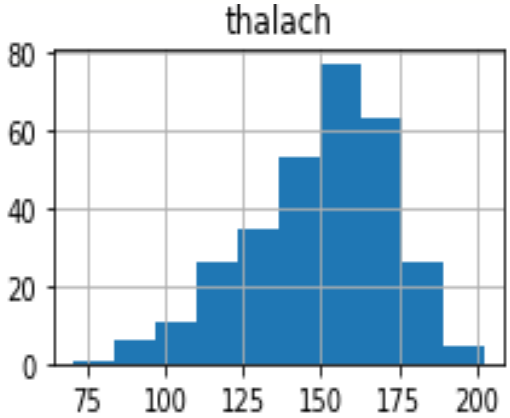


Figure 4.9: Histogram for Maximum Heart Rate

The maximum heart rate achieved ranges from 70 to 210.

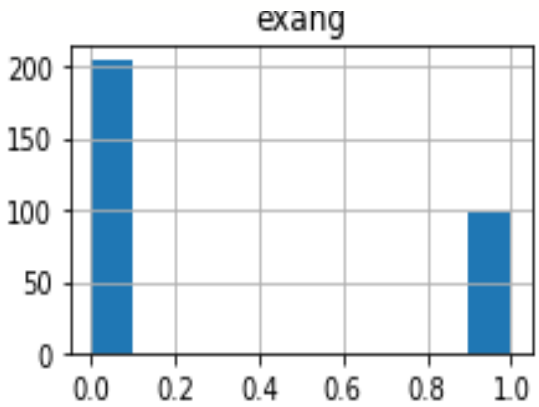


Figure 4.10: Histogram for Exercise induced angia

Exercise induced angina is present in 50% of the cases.

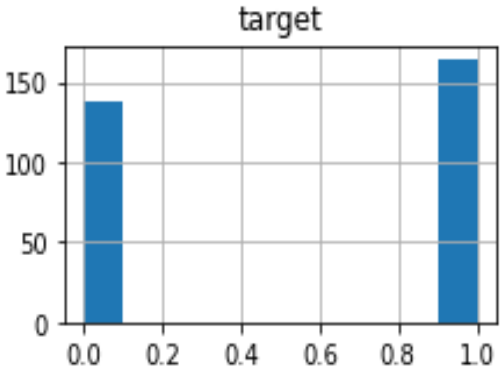


Figure 4.11: Histogram for Target values

Most cases in this dataset suffer from heart disease.

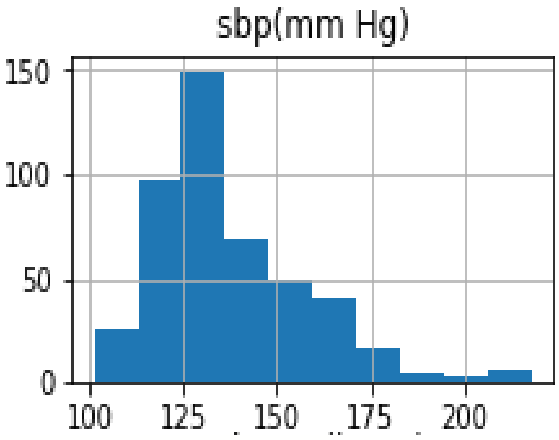


Figure 4.12: Histogram for Sbp

Blood pressure ranges from 100-280

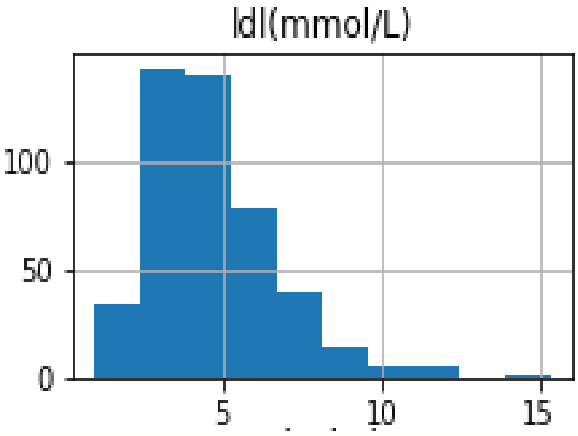


Figure 4.13: Histogram for LDL

LDL levels range from 1 -15

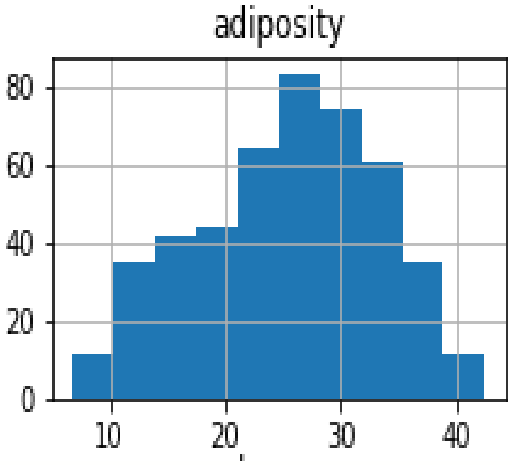


Figure 4.14: Histogram for Adiposity

BMI values range from 10 - 45

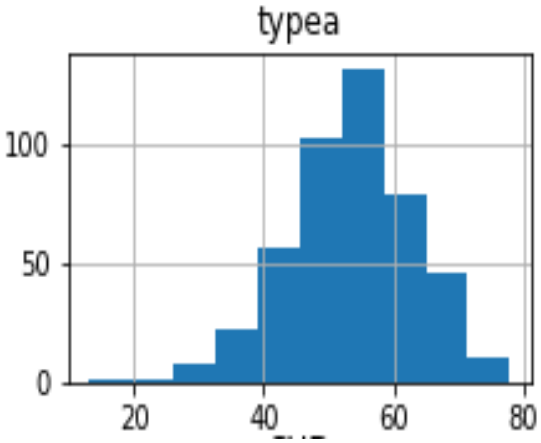


Figure 4.15: Histogram for Type A Scores

Type A scores range from 10-80

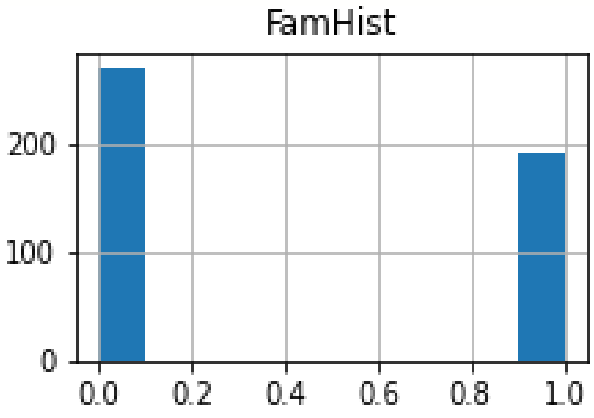


Figure 4.16: Histogram for Family history

Family History is not there in most of the cases.

4.2.4 Analysis of attributes

Here 1 means male and 0 denotes female. It is observed females having heart disease are comparatively less when compared to males Males have low heart diseases as compared to females in the given dataset.

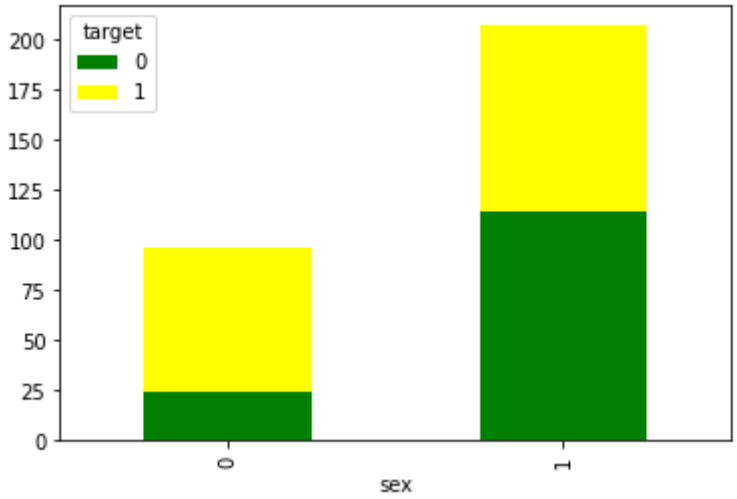


Figure 4.17: Bar chart for sex and target values

Cholesterol contributes to the onset of heart disease.

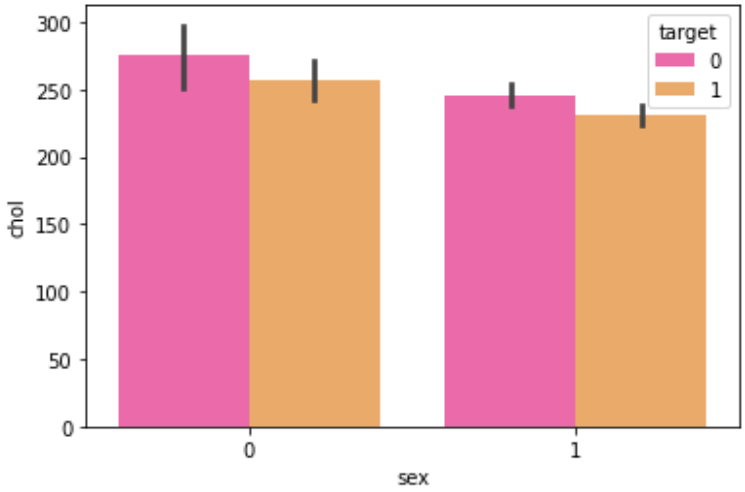


Figure 4.18: Cholesterol levels among male and female

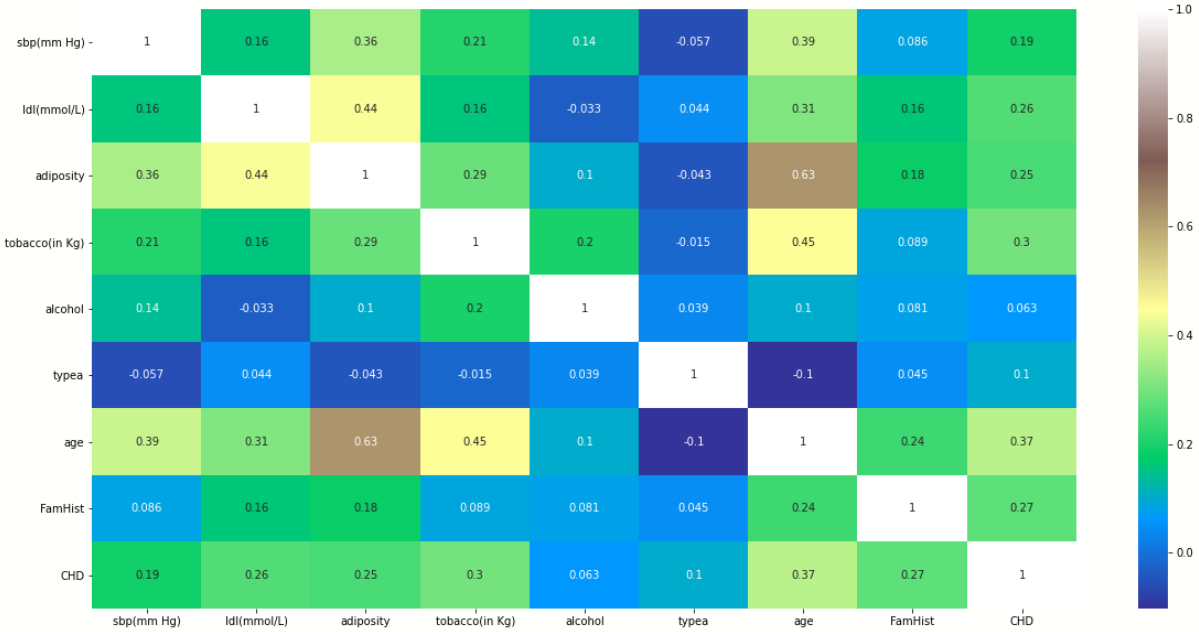


Figure 4.19: Correlation map

The observation of the coefficients of the correlation matrix above suggests that:

- Age is strongly correlated with adiposity, tobacco consumption/smoking, systolic blood pressure and elevated LDL cholesterol.
- Adiposity is strongly correlated with obesity.
- LDL is strongly correlated with Obesity and Adiposity both.
- To conclude Obesity (high BMI), Adiposity (excessive fats), Smoking (tobacco)Blood pressure and old age are all correlated with heart diseases

The male female count is as follows:

Table 4.5: Male and Female Count

Male	206
Female	96

Presence and absence of heart disease

Table 4.6: Presence or Absence of Disease Count

Present	55%
Absent	46%

It is observed the count for not having heart disease and having heart disease are almost balanced, not having frequency count is 140 and those having heart disease the count is 160.

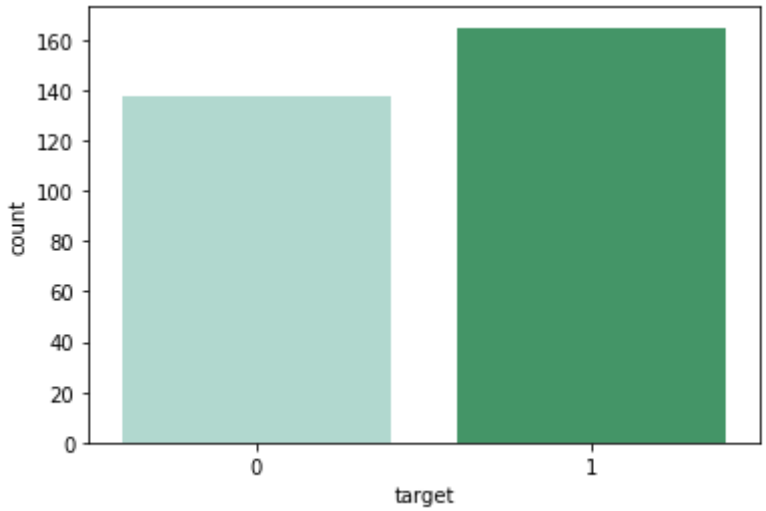


Figure 4.20: Target values

Individuals with chest pain of level 2 are more prone to heart disease.

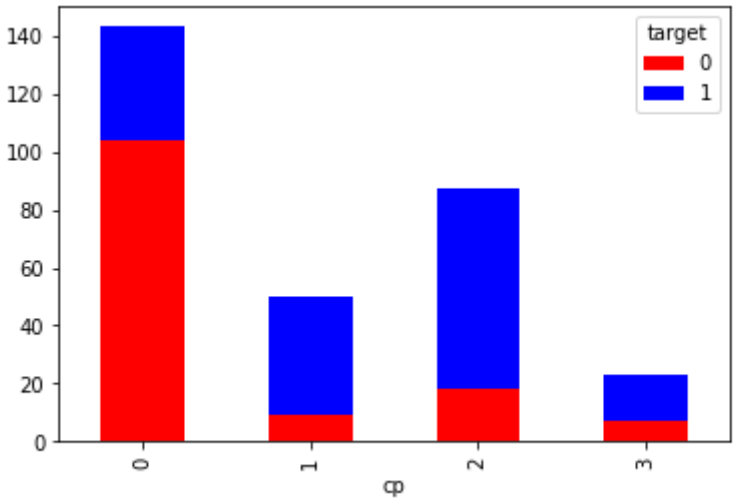


Figure 4.21: Chest pain levels and target

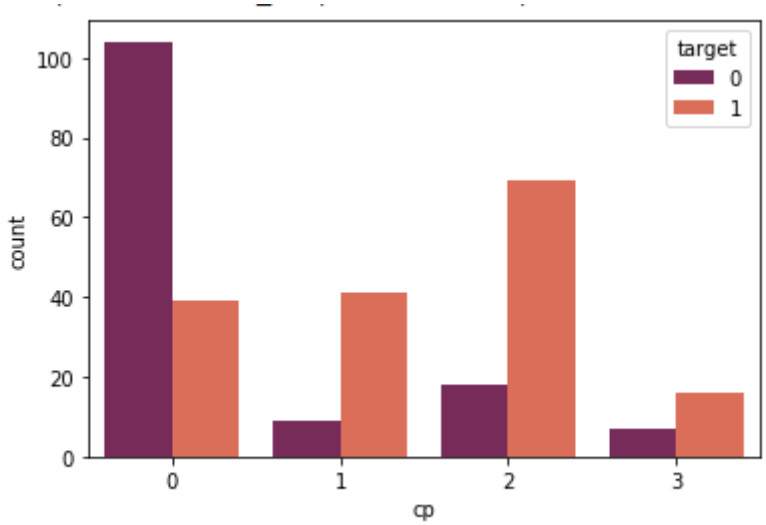


Figure 4.22: Effect of chest pain on heart disease

Males suffer from chest pain more than females

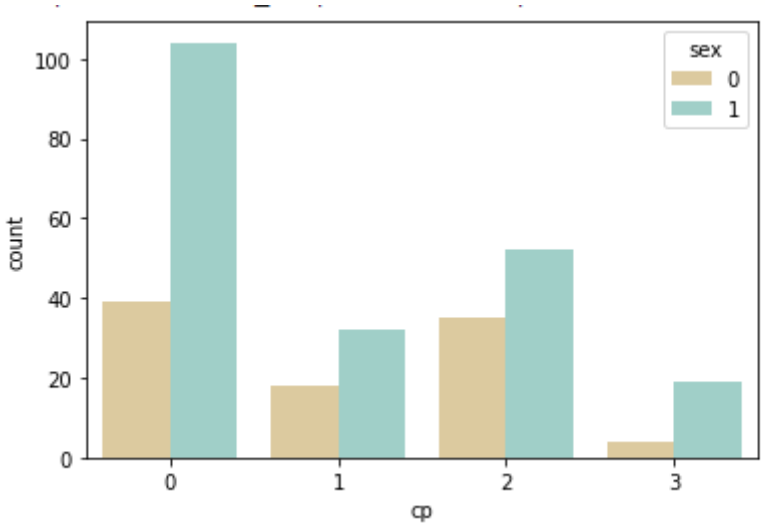


Figure 4.23: Chest pain and gender

4.2.4 Finding the accuracy

- The data is split into training and testing data.
- The accuracy is then determined.
- The columns are fitted to scale to achieve a better accuracy.
- 70% data is taken for training and 30% data is taken for testing.
- The machine learning algorithms are them compared for accuracy.

Table 4.7: Algorithm Accuracy

Algorithm	Accuracy %
LOGISTIC REGRESSION	92%
NAIVE BAYE	89%
RANDOM FOREST	86%
K NEAREST NEIGHBOUR	86%
DECISION TREE	76%

Logistic regression has the highest accuracy of 92% and this algorithm is taken to build a predictive model in tableau.

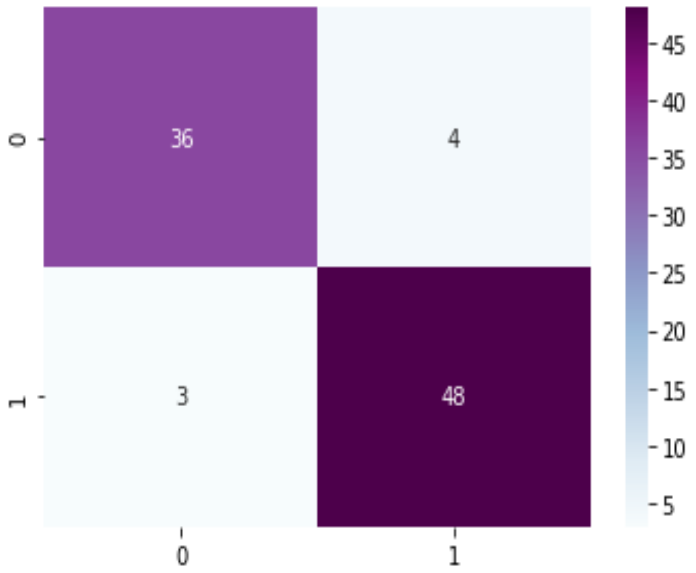
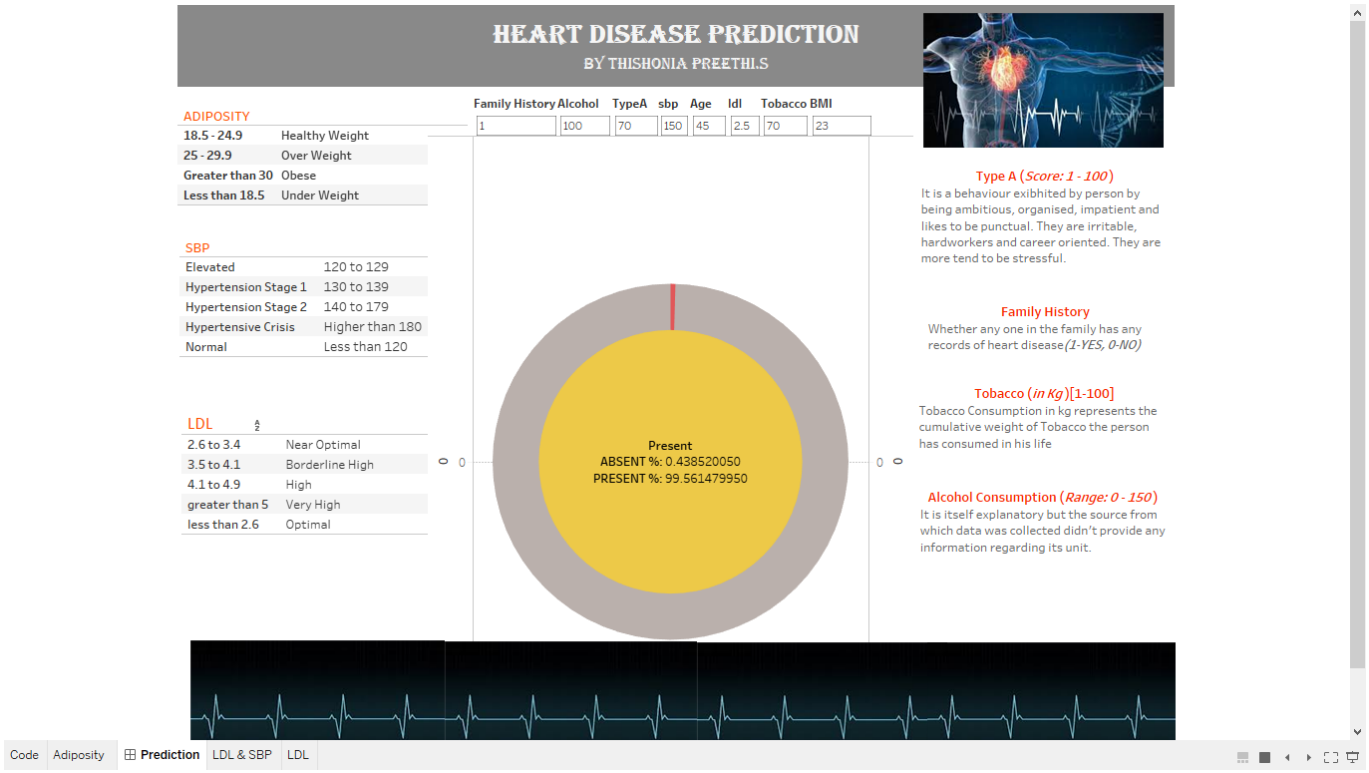


Figure 4.24: Confusion matrix of Logistic Regression

4.2.5 Predictive dashboard in Tableau

The python code is executed in Tableau using the Tabpy server. A predictive dashboard is built where the user/medical practitioners can enter the values. The model will then predict the presence or absence of heart disease. This will help in timely treatment and save lives of patients.

Figure 4.25: Tableau Predictive Dashboard



4.3 DISCUSSIONS

The key findings of this research are that feature selection and scaling play an important role in increasing the accuracy of algorithms. The features have been selected based on their impact on heart disease and the best algorithms for classification problems have been used. KNN, Naïve Bayes, Random Forest, Logistic Regression and Decision Tree are all proven to well suit classification problems. The classification problems are the ones which have 0 and 1 as the end result, in this case presence or absence of heart disease. Logistic Regression Algorithm has proven to help in decision making in the field of medicine. Ashok Kumar et al had an accuracy of 85% and Kannan et al and accuracy of 87% using Logistic regression. The accuracy at the end of this research is 92% which is pretty good in comparison to the previous studies.

Accuracy is a very important factor when it comes to the field of medicine as we are dealing with human lives here and a low accuracy cannot be admissible. The accuracy has been determined using the following formula.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Pos} + \text{False Pos} + \text{True Neg} + \text{False Neg}}$$

A confusion matrix has been formulated that depicts the true and false positives and negatives.

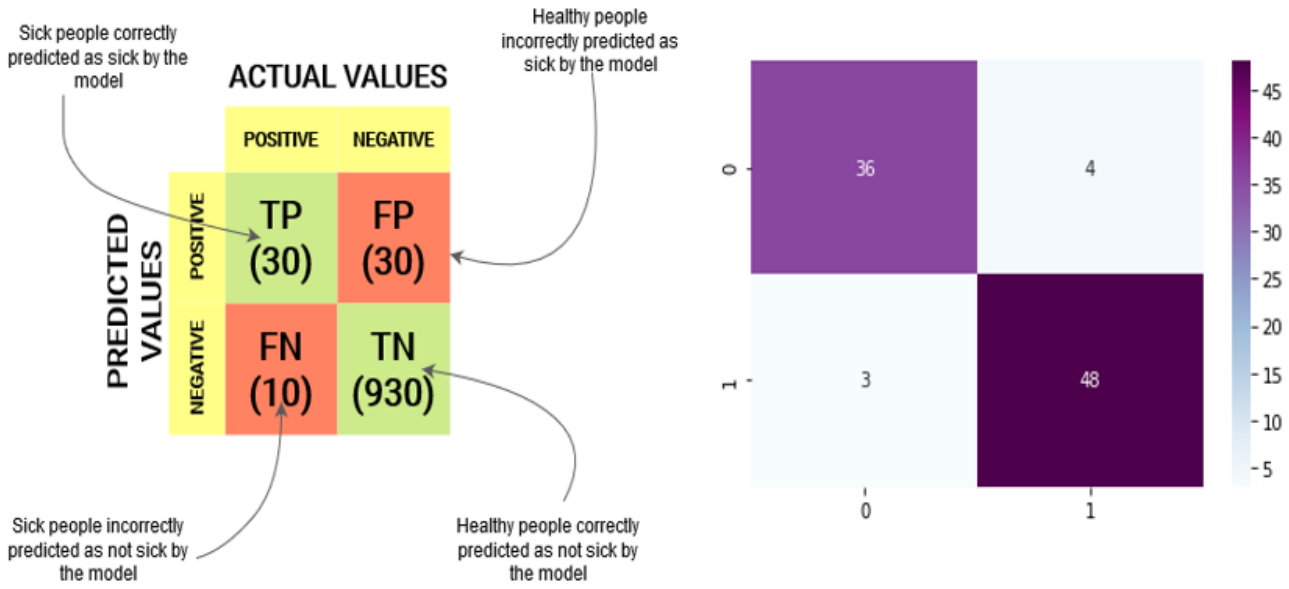


Figure 4.26: Confusion matrix

This clearly shows that the model is good and can be used for prediction. However, there is room for more improvement.

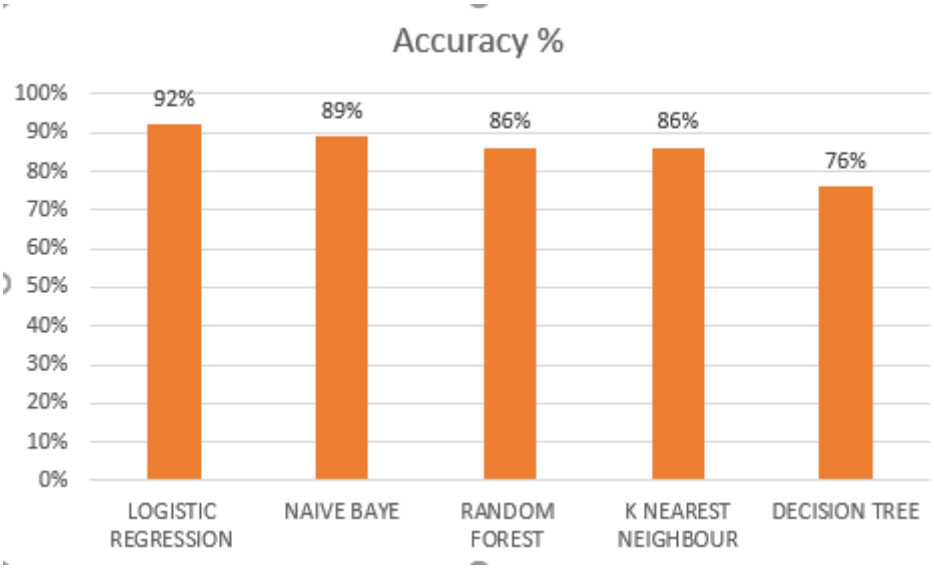


Figure 4.27: Accuracy

The data has been fed into Tableau and the Python code has been modified accordingly. The linkage between Tableau and Python has been done using the Tabpy server. The predictive dashboard built can

be used by normal people or medical practitioners to predict the onset of heart disease in the near future. The dashboard has a clear picture of the range of values to give the presence or absence probability of the heart disease. This research paper has gone a step ahead to build a predictive dashboard in Tableau without stopping at finding the highest accuracy.

The results tie well with previous studies where the researchers have made an analysis comparing different machine learning algorithms and finding the best fit. This paper will serve as a starting point for learning how to use automated learning to diagnose heart disease, and it can be expanded for future study. This research has many limitations, including, the instruments used in the study, such as the computer's computing capacity, and the time limit available for the study. This form of research precludes cutting-edge resources and experience in the relevant fields.

CHAPTER V

CONCLUSION

5.1 SUMMARY OF FINDINGS

- Family history has an impact on the onset of heart disease
- Alcohol consumption has an impact on the onset of heart disease
- Type A behaviour with a score above than 50 has an impact on the onset of the disease
- Tobacco consumption has an impact on the onset of heart disease
- Age greater than 50 has an impact on the onset of heart disease
- High blood pressure is associated with the onset of heart disease
- Males are more prone to heart diseases than female
- Chest pain type 2 has a major role in the onset of heart disease
- An obese person is more prone to heart disease
- Diabetic patients are more prone to heart disease
- As age increases the BMI increases
- As people age few indulge in tobacco and smoking
- Cholesterol is one of the reasons for obesity

5.2 SUGGESTIONS

- Maintaining a healthy blood pressure level. Heart disease is caused by high blood pressure, which is a significant risk factor. It's important to have blood pressure tested on a regular basis - at least once a year for most adults, and even more often if a person has high blood pressure. Taking action to avoid or regulate high blood pressure, including making lifestyle changes.
- Maintaining a healthy cholesterol and triglyceride level. Cholesterol levels that are too high can clog arteries, increasing the risk of coronary artery disease and heart attack. Cholesterol can be reduced by a combination of lifestyle changes and medications.
- Maintaining a balanced body weight. Obesity or being overweight can increase the risk of heart disease. This is primarily due to their association with other heart disease risk factors such as elevated blood cholesterol and triglyceride levels, high blood pressure, and diabetes. These dangers can be reduced if one maintains a healthy weight.
- Eating a well-balanced diet. Saturated fats, high-sodium foods, and added sugars can all be avoided. Fruits, herbs, and whole grains can all be consumed in moderation. This kind of a meal plan that can help to lower blood pressure and cholesterol levels, two factors that can lower your risk of heart disease.
- Getting any exercise on a daily basis. Exercise has many advantages, including strengthening the heart and improving circulation. It may also aid in the maintenance of a healthy weight as well as the reduction of cholesterol and blood pressure. Both of these things will help to avoid heart disease.

- Alcohol should be consumed in moderation. Too much alcohol will cause the blood pressure to rise. It also contributes more calories, potentially leading to weight gain. Any of these factors increase the chances of developing heart disease.
- Avoid smoking. Cigarette smoking raises blood pressure and increases the risk of heart disease and stroke.
- Taking care of stress which is associated with Type A behaviour. In several ways, stress is related to heart disease. It has the potential to increase blood pressure. A heart attack may be triggered by extreme stress. Furthermore, some common stress-relieving behaviours, such as overeating, heavy drinking, and smoking, are harmful to the heart. Exercising, listening to music, concentrating on something calm or happy, and meditating are all good ways to relieve stress.
- Taking care of diabetes. Diabetic heart disease is two times more likely. Diabetes causes high blood sugar, which damages the blood vessels and the nerves that regulate heart and blood vessels over time. So, it is important to get tested for diabetes, and keep it under control.

5.3 CONCLUSION

The data set was subjected to a variety of machine learning algorithms, including Logistic Regression, Nave Bayes, Decision Tree Random Forest, and K - nearest neighbour. It uses data such as blood pressure, cholesterol, and diabetes to predict who will develop coronary heart disease in the next ten years. As previously stated, a family history of heart disease may also be a factor in developing heart disease. As a result, the patient's data can also be used to improve the model's accuracy. This research would be helpful in finding potential heart attack patients in the near future. This can aid in the implementation of preventative measures, thereby reducing the risk of heart disease in the patient. When a patient's medical data is predicted to be positive for heart disease, physicians will closely examine the patient's medical records. For example, if a patient has diabetes, which may lead to heart disease in the future, the patient may be treated to keep the diabetes under control, potentially preventing heart disease. Other machine learning algorithms can be used to predict heart disease. In this case of binary classification problems, logistic regression has proven to be more accurate.

REFERENCES

1. Thapliyal M, Gochhait S. Predictive analytics in healthcare. Eur J Mol Clin Med. 2020;7(6):2558-2576.

2. Engelings CC, Helm PC, Abdul-Khaliq H, et al. Cause of death in adults with congenital heart disease - An analysis of the German National Register for Congenital Heart Defects. Int J Cardiol. 2016;211:31-36. doi:10.1016/j.ijcard.2016.02.133

3. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. Proc - IEEE Symp Comput Commun. 2017;(July):204-207. doi:10.1109/ISCC.2017.8024530

4. Saini R, Bindal N, Bansal P. Classification of heart diseases from ECG signals using wavelet transform and kNN classifier. Int Conf Comput Commun Autom ICCCA 2015. Published online 2015:1208-1215. doi:10.1109/CCAA.2015.7148561

5. Arun Pradeep N, Niranjana G. Different Machine Learning Models Based Heart Disease Prediction. Int J Recent Technol Eng. 2020;8(6):544-548. doi:10.35940/ijrte.f7310.038620

6. Rajathi S, Radhamani G. Prediction and analysis of Rheumatic heart disease using kNN classification with ACO. Proc 2016 Int Conf Data Min Adv Comput SAPIENCE 2016. Published online 2016:68-73. doi:10.1109/SAPIENCE.2016.7684132

7. Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. 2017 4th IEEE Uttar Pradesh Sect Int Conf Electr Comput Electron UPCON 2017. 2017;2018-January:500-504. doi:10.1109/UPCON.2017.8251100

8. Marimuthu M, Abinaya M, S. K, Madhankumar K, Pavithra V. A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach. Int J Comput Appl. 2018;181(18):20-25. doi:10.5120/ijca2018917863

9. Solanki Y. A Survey on Risk Assessments of Heart Attack Using Data Mining Approaches. Int J Inf Eng Electron Bus. 2019;11(4):43-51. doi:10.5815/ijieeb.2019.04.05

10. Polaraju K, Durga Prasad D, Tech Scholar M. Prediction of Heart Disease using Multiple Linear Regression Model. Int J Eng Dev Res. 2017;5(4):2321-9939. www.ijedr.org

11. Vikas C, Saurabh P. Data Mining Approach to Detect Heart Diseases. Int J Adv Comput Sci Inf Technol. 2013;2(4):56-66.

12. Parthiban G, S.K.Srivatsa A, Rajesh A. Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method. Int J Comput Appl. 2011;24(3):7-11. doi:10.5120/2933-3887

13. Animesh H, Subrata KM, Amit G, Arkomita M, Mukherjee A. Heart Disease Diagnosis and Prediction Using Machine LearAnimesh, Hazra et al. 2017. “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review.” Advances in Computational Sciences and Technology 10(7): 2137–59. ht. Adv Comput Sci Technol. 2017;10(7):2137-2159. http://www.ripublication.com

14. Bashir S, Qamar U, Javed MY. An ensemble based decision support framework for intelligent heart disease diagnosis. Int Conf Inf Soc i-Society 2014. Published online 2015:259-264. doi:10.1109/i-Society.2014.7009056

15. Priya NH, Gopika Rani N, Gowri SS. Analysis of Heart Disease Prediction Using 5achine Learning Techniques. Handb Artif Intell Biomed Eng. Published online 2020:173-194. doi:10.1201/9781003045564-8

16. Computing M, Kaur R. A Review - Heart Disease Forecasting Pattern using Various Data Mining Techniques. Int J Comput Sci Mob Comput. 2016;5(6):350-354.

17. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. AICCSA 08 - 6th IEEE/ACS Int Conf Comput Syst Appl. Published online 2008:108-115. doi:10.1109/AICCSA.2008.4493524

18. Animesh H, Subrata KM, Amit G, Arkomita M, Mukherjee A. Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Adv Comput Sci Technol. 2017;10(7):2137-2159. http://www.ripublication.com

19. Purushottam, Saxena K, Sharma R. Efficient Heart Disease Prediction System. Procedia Comput Sci. 2016;85:962-969. doi:10.1016/j.procs.2016.05.288

20. Kumari S, Viswanathan R. Heart Disease Prediction System. 2019;9(7):2019. doi:10.21275/SR20530171710

21. Noorul Islam Centre for Higher Education. Department of Electrical and Electronics Engineering, IEEE Electron Devices Society. India Chapter, Institute of Electrical and Electronics Engineers. Proceedings of IEEE International Conference on Circuits, Power and Computing Technologies : ICCPCT-2016 on 18th & 19th March 2016. Published online 2016.

22. Bharti S, Singh SN. Analytical study of heart disease prediction compared with different algorithms. Int Conf Comput Commun Autom ICCCA 2015. Published online 2015:78-82. doi:10.1109/CCAA.2015.7148347

23. Gandhi M, Singh SN. Predictions in heart disease using techniques of data mining. 2015 1st Int Conf Futur Trends Comput Anal Knowl Manag ABLAZE 2015. Published online 2015:520-525. doi:10.1109/ABLAZE.2015.7154917

24. Bahrami B, Shirvani MH. Prediction and Diagnosis of Heart Disease by Data Mining Techniques. J Multidiscip Eng Sci Technol. 2015;2(2):3159-3199. www.jmest.org

25. Abdullah H.Wahbeh, Qasem A. Al Radaideh, Mohammed N. Al Kabi, and Emad M. Al Shawakfa. A Comparison Study between Data Mining Tools over some Classification Methods. Int J Adv Comput Sci Appl. 2011;1(3). doi:10.14569/specialissue.2011.010304

26. Monika Gandhi and Dr. Shailendra Narayan Singh (2015),Predictions in Heart Disease UsingTechniques of Data Mining, International Conference on Futuristic trend in Computational Analysis and Knowledge Management

27. Thomas, J., & Princy, R. T. (2016). Human heart disease prediction system using data mining techniques. 2016 International Conference on Circuit, Power and Computing Technologies

28. Sandhya Kumari and Dr. R. Viswanathan (2018), Heart Disease Prediction System, International Journal of Science and Research (IJSR)

29. AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin, "HDPS:Heart Disease Prediction System"10 Department of Medical Informatics, Tzu Chi University, Hualien City, Taiwan.

30. Sheik Abdullah, " A Data Mining Model to Predict and Analyze the Events Related to Coronary Heart Disease using Decision Trees with Particle Swarm Optimization for Feature Selection", International Journal of Computer Applications Volume 55– No.8, October 2012.

31. Shanta kumar, B.Patil,Y.S.Kumaraswamy, "Predictive data mining for medical diagnosis of heart disease prediction" IJCSE Vol .17, 2011

32. Boshra Bahrami, Mirsaeid Hosseini Shirvan (2015) Prediction and Diagnosis of Heart Disease by Data Mining Techniques,Journal of Multidisciplinary Engineering Science and Technology (JMEST),Vol. 2 Issue 2

33. Himanshu Sharma,M A Rizvi (2017) International Journal on Recent and Innovation Trends in Computing and Communication,Volume:5

34. Ramandeep Kaur, Er. Prabhsharn Kaur (2016), A Review - Heart Disease Forecasting Pattern using Various International Journal of Computer Science and Mobile Computing Data Mining Techniques, Vol.5 Issue.6, June- 2016, pg. 350-354

35. M.A.Nishara Banu and B. Gomathy "Disease Forecasting System using Data Mining Systems", International Conference on Intelligent Computing Systems, 2014

36. K. Vembandasamy, R. Sasipriya, and E. Deepa,"Heart Diseases Detection Using Naive Bayes Algorithm", IJISET-International Journal of Innovative Science, Engineering & Technology, Vol.2, pp.441-444, 2015.

37. Seyedamin Pouriyeh, Sara Vahid, Hamid Reza Arabnia (2021) A Comprehensive Investigation on Comparison of Machine Learning Techniques On Heart Disease

38. K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, (2009) A hybrid evolutionary algorithm for attribute selection in data mining, Expert Systems with Applications

39. A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour,Effective diagnosis and monitoring of heart disease, International Journal of Software Engineering and Its Applications, 2015.

40. Vikas Chaurasia and Saurabh Pal,Data Mining Approach to Detect Heart Diseases, International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, Month Year, Page: 56-66

41. Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report. 2005

42. Boshra Bahrami, Mirsaeid Hosseini Shirvan (2015) Prediction and Diagnosis of Heart Disease by Data Mining Techniques,Journal of Multidisciplinary Engineering Science and Technology (JMEST)

43. Chen, M. S., Han, J., and Yu, P. S." Data mining: An overview from a database perspective," IEEE Transactions on Knowledge and Data Engineering,8, 866–883,1996.

44. Xu, R. and Wunsch, D. "Survey of clustering algorithms,"IEEE Transactions on Neural Networks, 2005.

45. Jain, A. K., Murty, M. N., and Flynn, P. J. Data Clustering: a review. ACM computing surveys (CSUR), 264-323,1999.

46. Asha Rajkumar and Mrs G Sophia ReenaDiagnosis Of Heart Disease Using Data Mining Algorithm, ResearchGate November 2009

47. Ashok kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation", Springer,17 September 2016.

48. Jaymin Patel, Prof. Tejal Upadhyay, Dr.Samir Patel, "Heart Disease Prediction using Machine Learning and Data Mining Technique", International Journal of Computer Science and Communication, September 2015-March 2016, pp.129-137.

49. K.Gomathi, Dr.D.Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", International Journal of System and Software Engineering, December 2016, pp.12-14.

50. Noura Ajam, "Heart Disease Diagnosis using Artificial Neural Network", The International Institute of Science, Technology and Education, vol.5, No.4, 2015, pp.7-11.

51. Sairabi H. Mujawar1, P. R. Devale, "Prediction of Heart Disease using Modified K-means and by using Naive Bayes", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297:2007 Certified Organization) Vol. 3, Issue 10, October 2015

52. R. Kannan, V. Vasanthi, "Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease", Springer Briefs in Forensic and Medical Bioinformatics, 14 June 2018,

53. Hidayat TAKCI," Improvement of heart attack prediction by the feature selection methods", Turkish Journal of Electrical Engineering & Computer Sciences 2018,

54. Rajesh Jangade, Ritu Chauhan, Ruchita Rekapally," Classification Model for Prediction of Heart Disease", Springer Nature Singapore Pte Ltd. 2018

55. Kannel WB. Fifty years of Framingham study contributions to understanding hypertension. J Hum Hypertens. 2000

56. Mickerson JN. Heart failure in hypertensive patients. Am Heart J. 1963

57. Caird FI. Heart disease in old age. Postgrad Med J. 1963

58. HIMSS. Clinical & business intelligence: Associate in Nursing analytics government review; 2013,. Sulkers P. care analytics, 2011

59. The Practice of Predictive Analytics in Healthcare April 2013 Gopalakrishna Palem

ANNEXURE:

Source code (Python):

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

#Load the data

from google.colab import files

uploaded=files.upload()

df = pd.read_csv('heart.csv')

df.head(10)

#number of rows and columns

df.shape

#column names

df.columns

#statistical analysis

df.describe()

df.isnull().sum()

#data type and null value check

print(df.info())

#annot-print values inside

plt.figure(figsize=(20,10))

sns.heatmap(df.corr(), annot=True, cmap='terrain')

#distribution of data (5 rows-3col)

df.hist(figsize=(12,12), layout=(5,3));

# box and whiskers plot-outliers
```

```
df.plot(kind='box', subplots=True, layout=(5,3), figsize=(12,12))

plt.show()

sns.catplot(data=df, x='sex', y='age',  hue='target', palette='husl')

#Female - cholestrol level

sns.barplot(data=df, x='sex', y='chol', hue='target', palette='spring')

#How many male and femlae

df['sex'].value_counts()

#presence and absence of heart disease

df['target'].value_counts()

#male and female having heart disease is plotted

sns.countplot(x='sex', data=df, palette='husl', hue='target')

#plotting having and not having disease

sns.countplot(x='target',palette='BuGn', data=df)

#xhest pain

df['cp'].value_counts()  # chest pain type

sns.countplot(x='cp' ,hue='target', data=df, palette='rocket')

#chest pain male and female

sns.countplot(x='cp', hue='sex',data=df, palette='BrBG')

sns.barplot(x='sex', y='cp', hue='target',data=df, palette='cividis')

gen = pd.crosstab(df['sex'], df['target'])

print(gen)

gen.plot(kind='bar', stacked=True, color=['green','yellow'], grid=False)

chest_pain = pd.crosstab(df['cp'], df['target'])

chest_pain

chest_pain.plot(kind='bar', stacked=True, color=['red','blue'], grid=False)

#all values are distributed - so we scale

from sklearn.model_selection import train_test_split
```

```
from sklearn.preprocessing import StandardScaler
StandardScaler = StandardScaler()
columns_to_scale = ['age','trestbps','chol','thalach']
df[columns_to_scale] = StandardScaler.fit_transform(df[columns_to_scale])
X= df.drop(['target'], axis=1)
y= df['target']
X_train, X_test,y_train, y_test=train_test_split(X,y,test_size=0.3,random_state=40)
from sklearn.linear_model import LogisticRegression
lr=LogisticRegression()
model1=lr.fit(X_train,y_train)
prediction1=model1.predict(X_test)
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,prediction1)
cm
sns.heatmap(cm, annot=True,cmap='BuPu')
TP=cm[0][0]
TN=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
print('Testing Accuracy:',(TP+TN)/(TP+TN+FN+FP))
from sklearn.metrics import accuracy_score
accuracy_score(y_test,prediction1)
from sklearn.metrics import classification_report
print(classification_report(y_test,prediction1))
from sklearn.tree import DecisionTreeClassifier
dtc=DecisionTreeClassifier()
model2=dtc.fit(X_train,y_train)
```

```
prediction2=model2.predict(X_test)

cm2= confusion_matrix(y_test,prediction2)

accuracy_score(y_test,prediction2)

from sklearn.ensemble import RandomForestClassifier

rfc=RandomForestClassifier()

model3 = rfc.fit(X_train, y_train)

prediction3 = model3.predict(X_test)

confusion_matrix(y_test, prediction3)

accuracy_score(y_test, prediction3)

from sklearn.naive_bayes import GaussianNB

NB = GaussianNB()

model5 = NB.fit(X_train, y_train)

prediction5 = model5.predict(X_test)

cm5= confusion_matrix(y_test, prediction5)

accuracy_score(y_test, prediction5)

from sklearn.neighbors import KNeighborsClassifier

KNN = KNeighborsClassifier()

model6 = KNN.fit(X_train, y_train)

prediction6 = model6.predict(X_test)

cm6= confusion_matrix(y_test, prediction5)

cm6

accuracy_score(y_test,prediction6)

print('KNN :', round((accuracy_score(y_test, prediction6))*100))

print('LOGISTIC REGRESSION :', round((accuracy_score(y_test, prediction1))*100))

print('DECISION TREE :', round((accuracy_score(y_test, prediction2))*100))

print('RANDOM FOREST :', round((accuracy_score(y_test, prediction3))*100))

print('NNAIVE BAYES: ', round((accuracy_score(y_test, prediction4))*100)
```

Tableau code:

```
SCRIPT_STR(
"
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from ast import literal_eval
d={
'ad':literal_eval(_arg1[0]),
'ag':literal_eval(_arg2[0]),
'al':literal_eval(_arg3[0]),
'fam':literal_eval(_arg4[0]),
'ldl':literal_eval(_arg5[0]),
'sbp':literal_eval(_arg6[0]),
'tob':literal_eval(_arg7[0]),
'typea':literal_eval(_arg8[0]),
'chd':literal_eval(_arg9[0]),
}
df=pd.DataFrame(d)
x=df.iloc[:,:8]
y=df.iloc[:,8]
sc=StandardScaler()
x_t=sc.fit_transform(x)
lr=LogisticRegression()
lr.fit(x_t,y)
input_list=[_arg10[0],_arg11[0],_arg12[0],_arg13[0],_arg14[0],
```

```
_arg15[0],_arg16[0],_arg17[0]]

inp=np.array(input_list).reshape(1,-1)

inp=sc.transform(inp)

pred=lr.predict(inp)

prob=lr.predict_proba(inp)

return str(pred[0])

return str(lr.score(x_t,y))

",

ATTR([Adiposity]),ATTR([Age]),

ATTR([Alcohol]),ATTR([Fam Hist]),

ATTR([ldl(mmol/L)]),ATTR([sbp(mm Hg)]),

ATTR([tobacco(in Kg)]),ATTR([Typea]),ATTR([CHD]),

[Parameters].[Adiposity],[Parameters].[Age],[Parameters].[Alcohol],

[Family History],[ldl],[sbp],[tobacco],[TypeA]

)

inp=np.array(input_list).reshape(1,-1)

inp=sc.transform(inp)

pred=lr.predict(inp)

prob=lr.predict_proba(inp)

return str((prob[0][0]*100))

return str(lr.score(x_t,y))
```