A COMPARATIVE STUDY ON HEART DISEASE PREDICTION USING DIFFERENT MACHINE LEARNING CLASSIFICATION ALGORITHMS

MD. IMAM HOSSAIN
MEHADI HASAN MARUF

BACHELOR OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

A COMPARATIVE STUDY ON HEART DISEASE PREDICTION USING DIFFERENT MACHINE LEARNING CLASSIFICATION ALGORITHMS

MD. IMAM HOSSAIN
MEHADI HASAN MARUF

Thesis submitted in fulfillment of the requirements for the award of the degree of Bachelor of Science in Information and Communication Engineering

Department of Information and Communication Engineering
NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

JANUARY 2020

DECLARATION

This thesis is submitted to the Department of Information and Communication Engineering of Noakhali Science & Technology University in partial fulfillment of the requirements for the award of the degree of Bachelor of Science (B.Sc.) Engineering in Information and Communication Engineering (ICE). So, we hereby declare that this thesis report has not been submitted elsewhere for the requirement of any kind of degree, diploma or publication.

Md. Imam Hossain
Student of Information and Communication Engineering, NSTU
Roll No. ASH1611008M
Session 2015-16

Mehadi Hasan Maruf
Student of Information and Communication Engineering, NSTU
Roll No. ASH1611041M
Session 2015-16

ACCEPTANCE

I hereby declare that I have checked this thesis, and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science (B.Sc.) degree in Information & Communication Engineering.

Professor Dr. Md Ashikur Rahman Khan
Chairman
Department of Information and Communication Engineering
Noakhali Science & Technology University
Noakhali - 3814, Bangladesh

ACKNOWLEDGMENT

We are grateful and would like to express our sincere gratitude to our supervisor Professor Dr. Md Ashikur Rahman Khan for his germinal ideas, invaluable guidance, continuous encouragement and constant support in making this research possible. He has always impressed us with his outstanding professional conduct, his strong conviction for science. We appreciate his consistent support from the first day We applied to graduate program to these concluding moments. We are truly grateful for his progressive vision about our training in science, his tolerance of our naive mistakes, and his commitment to our future career.

Our sincere thanks to all our lab mates and members of the staff of the Information and Communication Engineering Department, NSTU, who helped us in many ways. And, thanks to the director of those hospitals who approved us to collect data.

We acknowledge our sincere indebtedness and gratitude to our parents for their love, dreams and sacrifice throughout our life, Special thanks should be given to our committee members and Saiful Islam Mamun who helped us to collect data and with his medical science knowledge. I would like to acknowledge their comments and suggestions, which were crucial for the successful completion of this study.

ABSTRACT

Heart disease is the leading cause of death over the past 10 years all over the world. The risk factors of heart disease may hold numerous measures and considering all the measures increases the time of medical practitioners for the decision-making process. The main aim of this work is to pave the way to maintain the huge datasets with the appropriate measures and choose the right approach which would provide a faster and more accurate prediction of heart disease depending on the accuracy rate of the measures. In this thesis, we examine and compare the performance metrics of different machine learning classification algorithms to predict heart disease. This comparison shows the different accuracy, precision, and recall rates of different techniques and the reasons behind their variations. We used collected data that we collect from the different hospitals in Bangladesh. We preprocess our collected data and then divide it into two sections named training and testing datasets. The Logistic Regression, Naive Bayes Classifier, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, and Multi-Layer Perceptron Neural Networks techniques have been investigated in this research. By the end of the implementation part, we have found Random Forest is giving the maximum score in all sectors like accuracy, precision, recall, and f1 score in our dataset, and Multi-Layer Perceptron is performing very poorly. Random Forest gives a maximum of 90 percent accuracy and KNN gives the second-best 85 percent accuracy. Other algorithms like KNN, SVM, and Decision Tree also show overall good performances. The reasons for variations of these different techniques by analyzing their characteristics and behavior with respect to the dataset have been understood by the study conducted for this thesis. The research outcome is the implementation of machine learning on healthcare data and finding a better technique for prediction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

According to the World Health Organization (WHO), 17.9 million people die each year from CVD (Cardiovascular Disease), an estimated 31% of all deaths worldwide and 85% are due to heart attack and stroke [1]. In 2018 WHO has marked Heart Disease as one of the biggest causes of death all around the World. Populations most affected are from low and middle-income countries like Bangladesh, where 80% of these deaths occur [2]. There are also a lot of people affected at heart diseases in Bangladesh, and the number is increasing day by day. A study from rural Bangladesh demonstrated a rapid increase in CVD mortality rates from 16 deaths per 100,000 to 483 deaths per 100,000 among males and from 7 deaths per 100,000 to 330 deaths per 100,000 in females [3].

There are several risk factors for heart disease. Some factors can be controlled, others can't. Ones that can't be controlled include:

- Gender (males are at greater risk)
- Age (the older you get, the higher your risk)
- A family history of heart disease
- Being post-menopausal

Still, making some changes in our lifestyle we can reduce the chance of having heart disease. Controllable risk factors include:

- Smoking
- High LDL, or "bad" cholesterol, and low HDL, or "good" cholesterol
- Uncontrolled hypertension (high blood pressure)
- Physical inactivity
- Obesity
- Uncontrolled diabetes
- Uncontrolled stress and anger

Heart disease prediction is one of the fields where machine learning can be implemented. Machine learning is a subfield of computer science and a rapidly up surging topic in today's context and is expected to rise more in the coming days. Our world is flooded with data and data is being created rapidly every day all around the world. Efficient machine learning algorithms can predict heart disease more accurately with patient's diagnosis test data. Machine learning helps to gain insight from a massive amount of data which is very cumbersome and impossible to humans.

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. Classification algorithms and regression algorithms are types of supervised learning. In unsupervised learning, the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points.

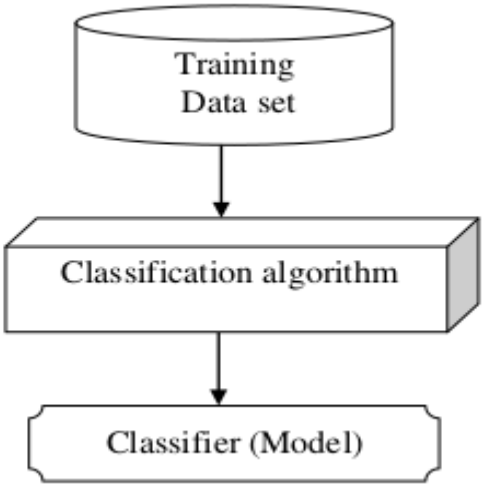

Figure 1.1: Machine Learning Model Construction Steps

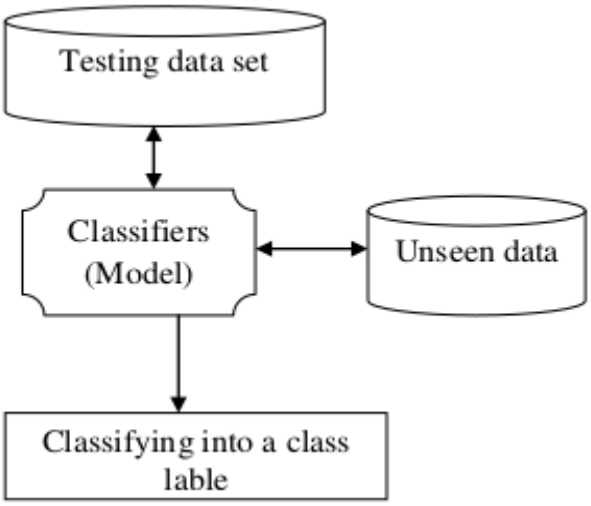

Figure 1.2: Use of Machine Learning Model

1.2 MOTIVATIONS

Nowadays, the healthcare industry contains a huge amount of healthcare data, which contains hidden information. This hidden information is useful for making effective decisions. Our study's objective is to prioritize the diagnosis test and see some of the daily habits which contribute to heart disease. We will make heart disease

prediction model using supervised machine learning classification algorithms such as Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, K-Nearest Neighbor, Naive Bayes Classifier and Multi-Layer Perceptron Neural Networks. Furthermore, we will compare different machine learning algorithms according to different performance metrics like accuracy, precision, recall, f1-score, roc auc score etc.

Many different factors trigger heart disease in young people. Doctors depend on certain factors like age, cholesterol level, and blood pressure. Unfortunately, that is often insufficient. Also, the information provided by the patients may include redundant and interrelated symptoms and signs especially when the patients suffer from more than one type of disease of the same category. Thus physicians may not able to diagnose it correctly. With this concern in recent times, computer technology and machine learning techniques are being used to develop disease prediction system to assist doctors in making the decision of heart disease in the preliminary stage. But the accuracy and efficiency of the prediction system depend on better algorithms.

1.3 RESEARCH OBJECTIVES

The main objectives of this research is to analyze various machine learning classification algorithms on heart disease prediction and to conclude which techniques are more effective and accurate.

The other purposes of this study are listed below:

i. To compare different algorithms performance to find the more efficient and accurate technique to build efficient heart disease prediction model which can classify the heart disease correctly based on different performance metrics.

ii. To analyze various research works done on heart disease prediction and classification using various machine learning and deep learning techniques.

iii. To better understanding and application of machine learning in the medical domain.

iv. To obtain a clear idea of machine learning classification algorithms and see how each of them perform.

1.4 THESIS OUTLINE

Remaining part of this thesis report has been organized as follows:

Chapter 2 briefly reviews some of the machine learning algorithms and related concepts from the literature. Chapter 3 describes some related works our thesis. Chapter 4 discusses about the materials and methods that we use in our thesis and we show exploratory data analysis of our collected data we also discuss about our data preprocessing techniques. In chapter 5 we represent our result that we got from our data analysis and model prediction, we also discuss about model performance based on their algorithm characteristics. In chapter 6 we conclude the report with a summary of our work and our future plans.

CHAPTER 2

THEORETICAL BACKGROUND

2.1 INTRODUCTION TO MACHINE LEARNING

Machine learning is a branch of artificial intelligence, a science that researches machines to acquire new knowledge and new skills and to identify existing knowledge. Machine learning has been widely used in data mining, computer vision, natural language processing, biometrics, search engines, medical diagnostics, detection of credit card fraud, securities market analysis, DNA sequence sequencing, speech and handwriting recognition, strategy games and robotics. The precise definition of machine learning is: It's a computer program learning from experience E with respect to some task T and some performance measure P, if its performance on T as measured by P, improves with E [4]. It is necessary to understand data and gain insights for better understanding of a human world. The data amount is so huge today that traditional methods cannot be used. Analyzing data or building predictive models manually is almost impossible in some scenarios and also time consuming and less productive. Machine learning, on the other hand, produces reliable, repeatable results and learns from earlier computation. It is often also referred to as predictive analytics, or predictive modeling. Its goal and usage is to build new and/or leverage existing algorithms to learn from data, in order to build generalizable models that give accurate predictions, or to find patterns, particularly with new and unseen similar data.

Imagine a dataset as a table, where the rows are each observation (aka measurement, data point, etc.), and the columns for each observation represent the features of that observation and their values. At the outset of a machine learning project, a dataset is usually split into two or three subsets. The minimum subsets are

the training and test datasets, and often an optional third validation dataset is created as well. Once these data subsets are created from the primary dataset, a predictive model or classifier is trained using the training data, and then the model's predictive accuracy is determined using the test data. As mentioned, machine learning leverages algorithms to automatically model and find patterns in data, usually with the goal of predicting some target output or response. These algorithms are heavily based on statistics and mathematical optimization.

2.1.1 Types of Machine Learning

Supervised learning: When an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of Supervised learning. This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples.

Unsupervised learning: When an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms. As a kind of learning, it resembles the methods humans use to figure out that certain objects or events are from the same class, such as by observing the degree of similarity between objects. Some recommendation systems that you find on the web in the form of marketing automation are based on this type of learning.

Reinforcement learning: Input data as feedback to the model, emphasizing how to act based on the environment to maximize the expected benefits. The difference between supervised learning is that it does not require the correct input/output pairs and does not require precise correction of sub-optimal behavior. Reinforcement learning is more

focused on online planning and requires a balance between exploration (in the unknown) and compliance (existing knowledge).

2.2 MACHINE LEARNING ALGORITHMS

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). Classification belongs to the category of supervised learning where the targets also provided with the input data. Next we are going to discuss about some of the supervised classification algorithms.

Logistic Regression: Linear regression predictions are continuous values (i.e., rainfall in cm), logistic regression predictions are discrete values (i.e., whether a student passed/failed) after applying a transformation function. Logistic regression is named after the transformation function it uses, which is called the logistic function h(x)= 1/ (1 + ex). This forms an S-shaped curve.

In logistic regression, the output takes the form of probabilities of the default class (unlike linear regression, where the output is directly produced). As it is a probability, the output lies in the range of 0-1.
This output (y-value) is generated by log transforming the x-value, using the logistic function h(x)= 1/ (1 + e^ -x). A threshold is then applied to force this probability into a binary classification [5].

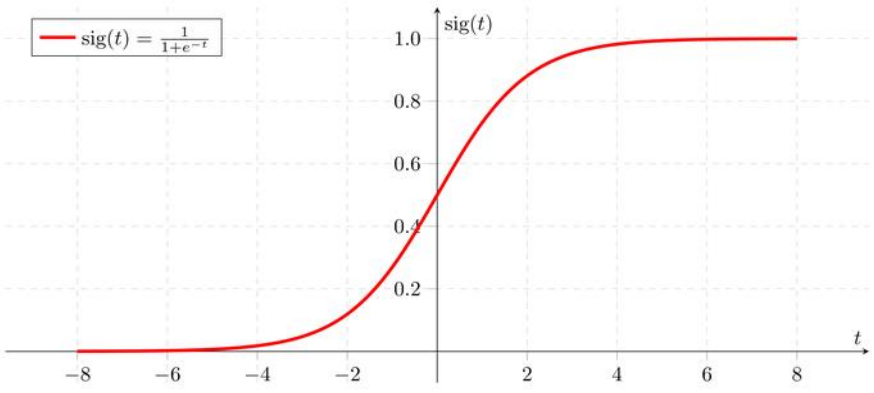

Figure 2.1: Logistic Regression Algorithms Graph

Naive Bayes: It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple. Naive Bayesian model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

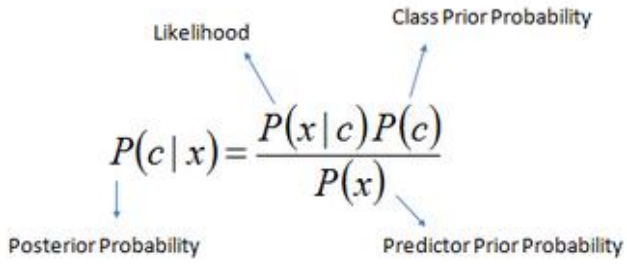

Figure 2.2: Equation of Naive Bayes

KNN (k- Nearest Neighbors): It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. *K nearest neighbors* is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function. These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and the fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the

class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modeling [5].

Figure 2.3: k- Nearest Neighbors visualization

Support Vector Machine (SVM): It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate.

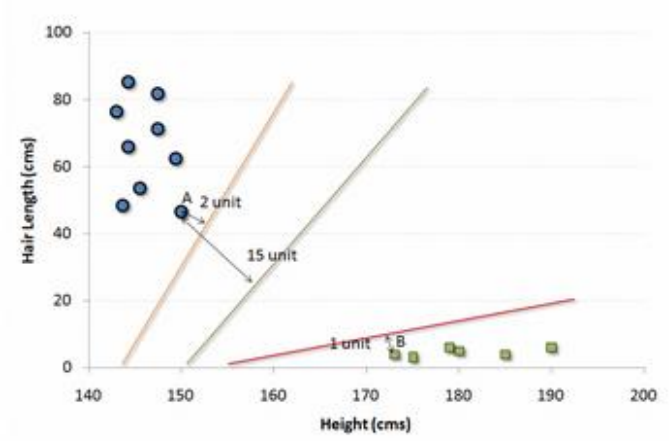

Figure 2.4: Support Vector Machine Algorithms Visualization

Decision Tree: A decision tree is a classifier and used recursive partition of the instance space. This model consists of nodes and a root. Nodes other than root have exactly one incoming edge. Intermediate node is test nodes after performing a test they generate outgoing edge. Nodes without outgoing are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces a certain discrete function of the input attributes values.

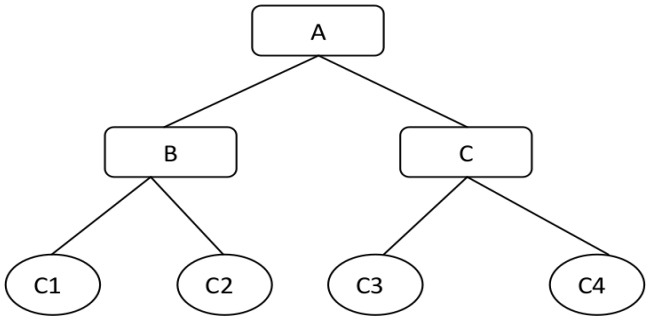

Figure 2.5: Decision Tree Algorithm Visualization

Random Forest: It is a supervised classification algorithm. Multiple number of decision trees taken together forms a random forest algorithm i.e. the collection of many classification trees. Each decision tree includes some rule based system. For the given training dataset with targets and features, the decision tree algorithm will have a set of rules. In random forest unlike decision trees there is no need to calculate information gain to find root node. It uses the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome. Further it calculates the vote for each predicted target. Thus high voted prediction is considered as the final prediction from the random forest algorithm.

Multi-Layer Perceptron: In the Multilayer perceptron, there can be more than one linear layer (combinations of neurons). If we take the simple example the three-layer network, first layer will be the *input layer* and last will be *output layer* and middle layer will be called *hidden layer*. We feed our input data into the input layer and take the output from the output layer. We can increase the number of the hidden layer as much as we want, to make the model more complex according to our task.

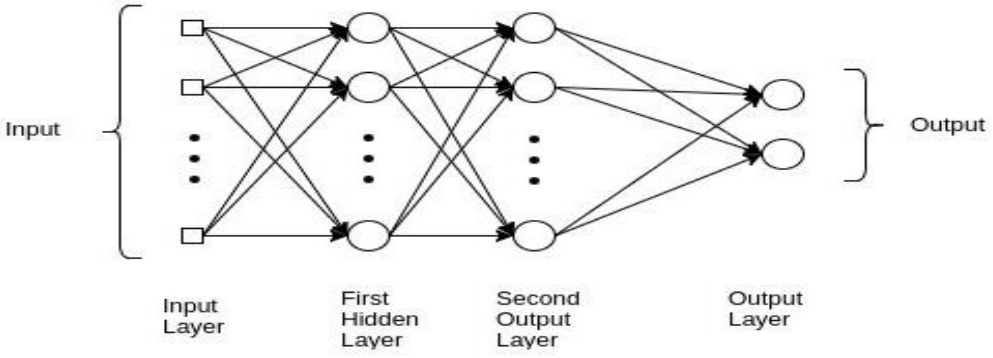

Figure 2.6: Multi-Layer Perceptron Neural Networks

2.3 PERFORMANCE METRICS OF ALGORITHMS

In a machine learning context, performance metrics is the measurement of algorithm on how well algorithm performs based on different criteria such as accuracy, precision, recall etc. Different performance metrics are discussed below.

Confusion Matrix: Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model. It is the easiest way to measure the performance of a classification problem where the output can be of two or more types of classes. A confusion matrix is nothing but a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, “False Negatives (FN)”.

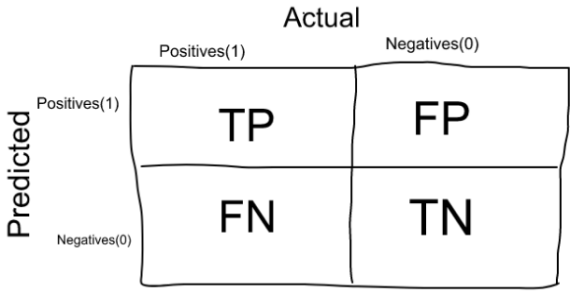

Figure 2.7: Confusion Matrix Structure

Explanation of the terms associated with confusion matrix are as follows –

- True Positives (TP) – It is the case when both actual class & predicted class of data point is 1.
- True Negatives (TN) – It is the case when both actual class & predicted class of data point is 0.
- False Positives (FP) – It is the case when actual class of data point is 0 & predicted class of data point is 1.
- False Negatives (FN) – It is the case when actual class of data point is 1 & predicted class of data point is 0.

Accuracy: Accuracy is what we usually mean when we use the term accuracy. It is the ratio of the number of correct predictions to the total number of input samples. It is most common performance metric for classification algorithms. It may be defined as the number of correct predictions made as a ratio of all predictions made.

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

It works well only if there are an equal number of samples belonging to each class.

Precision: Precision is defined as the proportion of true positive instances which are classified as positive. It shows how close predicted values are to each other. Formula for calculating precision is given below.

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

Precision is a measure of how close or near the predicted values are to each other. For example, if the actual value of a person's height is 6 feet and measured value is or predicted value is 5.5 and every time measurement is taken height is 5.4 or 5.6 then prediction is quite precise but not accurate.

Recall or Sensitivity: Recall is defined as the proportion of positive instances are correctly classified as positive. Recall is also often called sensitivity. Formula for calculating recall is given below.

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

Recall is simply the measure of how many true samples are predicted from the all the samples.

F1 Score: F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as :

$$F1 = 2 * (precision * recall)/(precision + recall)$$

F1 Score tries to find the balance between precision and recall.

AUC (Area Under the ROC curve): AUC (Area Under Curve)-ROC (Receiver Operating Characteristic) is a performance metric, based on varying threshold values, for classification problems. As the name suggests, ROC is a probability curve and AUC measure the reparability. In simple words, AUC-ROC metric will tell us about the capability of model in distinguishing the classes. Higher the AUC, better the model. Mathematically, it can be created by plotting TPR (True Positive Rate) i.e. Sensitivity or recall vs FPR (False Positive Rate) i.e. 1-Specificity, at various threshold values.

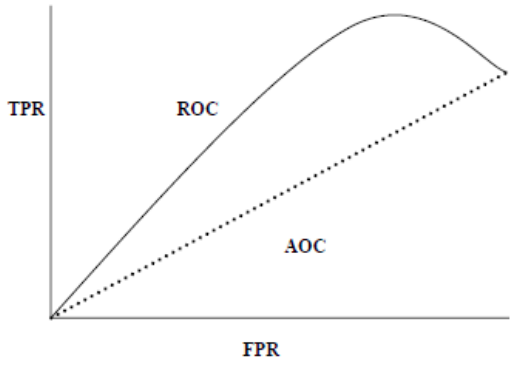

Figure 2.8: Area Under the ROC curve

2.4 DATA PREPROCESSING

In the modern world, there is a large amount of data collected through various sources such as the internet, surveys and experiments etc. But most of the time the data to be used are full of missing values, noises and distortions. Since, the majority of data collected are not taken from controlled environment, the data can contain garbage in it. “Garbage in – Garbage out” principles work here also. No matter how well data is modelled, if data contains garbage then output obtained is always garbage. Therefore, data pre- processing is a fundamental step to data analysis and machine learning. Data pre- processing or often called data wrangling is a set of procedures which transforms raw data into understandable format.

Data pre-processing includes detecting outliers, making decisions what to do with the outliers, finding and filling appropriate missing values and searching for inconsistency in the data. Data needs to be normalized or standardized and reduced before applying to machine learning algorithms. Normalization of data is done when the features of dataset have different measuring units. For example, Fahrenheit and Celsius, though both are units of temperature their measuring units are different. Standardization is used to scale the data and it gives the information on how many standard deviations the data is from its mean value. Standardization rescales the data to have the mean (μ) of 0 and the standard deviation (σ) of 1.

Steps Involved in Data Preprocessing:

1. Data Cleaning: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data, etc.

2. Data Transformation: This step is taken in order to transform the data in appropriate forms suitable for mining process. It involves Normalization, Attribute Selection, Discretization, etc.

3. Data Reduction: Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

CHAPTER 3

LITERATURE REVIEW

In this chapter, we discussed about various machine learning classifiers and previous work on the heart disease prediction using different classifiers.

R. Rajalakshmi has discussed why we need to apply data mining techniques on the CVD related data and applied Multilayer Perceptron, Simple Naive Bayes, J48 Decision Classifier, Decision Table Classifier algorithms to diagnose the risk factors of CVD. They work with a data set that contains 10 attributes. It is concluded that Multilayer Perceptron outperforms others with 91% accuracy. J48 and Decision Table shows 85% and 84% accuracy. Simple Naive Bayes has the lowest accuracy of 77% [6].

Koushik Chandra, Md. Shahriare Satu and Arup Mazumder have discussed the alarming death rate from CHD all over the world as well as in Bangladesh and the need for data mining in this sector. They applied 7 data mining algorithms and used 34 attributes to predict Coronary Heart Disease. The result of the research concluded that there is not much difference in the accuracy of the used algorithms in predicting the occurrence of heart diseases overall Naive Bayes and Kstar have the best accuracy among all [7].

Jyothi Rohilla has discussed the advantages of applying data mining techniques to maintain huge volumes of data. It is also discussed that the proportion of deaths caused by heart diseases is greater than other diseases. Performance analysis of various classification algorithms are also conducted and it is concluded that ID3 and decision tree outperforms than other techniques [8].

Mr. Sudhir M. Gorade, Prof. Ankit Deo studying various data mining classification techniques like Decision Tree, K-Nearest Neighbor, Support Vector Machines, Naive Bayesian Classifiers, and Neural Networks they make a decision about the advantages and disadvantages of these methods like K-Nearest Neighbor is effective if training data is large, Support Vector Machines is useful for non-linearly separable data, Neural Networks extracting the knowledge (weights in ANN) is very difficult [9].

Saxena, K., Sharma design a system that could efficiently discover the rules to predict the risk level of patients based on the parameter about patient's health. The rules could be prioritized according to the user's specifications. The performance of the system was tested and marked in terms of classification precision and results pointed out that the system had great potential in predicting heart disease risk level [10].

Priyanka P. Shinde, Kavita S. Oza, Rajanish research for the better solution of diagnosis, they used medical data related to heart disease and studied many algorithms like neural networks, decision tree, etc., and used analytical tools like R, Weka, etc. for the prediction of heart disease. Their Experimental results offer 88.5% accuracy when they use the R Analytical Tool and decision tree algorithm [11].

Here used feed-forward backpropagation network and they took eight input variables, four hidden variables, and two output variables achieve results comparable to physicians. However, but they don't get acceptable results [12].

V. Manikandan proposed that association rule mining is used to extract the item set relations. The data classification was based on MAFIA algorithms which resulted in better accuracy. The data was evaluated using entropy based cross validation and partition techniques and the results were compared. MAFIA (Maximal Frequent Item set Algorithm) used a dataset with 19 attributes and the goal of the research work was to have highly accurate recall metrics with higher levels of precision. They get 74% accuracy, precision 0.78, recall 0.67 when they use k-mean based on MAFIA and get 92% accuracy, precision 0.82, recall 0.92 when they use k-mean based on MAFIA with ID3 and C4.5 [13].

Aditya Methaila research using different algorithms and combinations of several targets attributes for effective heart attack prediction. Decision Tree has outperformed with 89% accuracy by using 15 attributes, and using Naïve Bayes they got 86.53%. The accuracy of the Decision Tree and Bayesian Classification further improves after applying the genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction [14].

CHAPTER 4

MATERIALS AND METHODS

This section describes the different methods and materials used for this study i.e., research design, data collection and description, exploratory data analysis, tools and data scaling used for this study.

4.1 RESEARCH DESIGN

This section illustrates the research design; it describes the actual flow of the entire research. The proposed model of our research work is divided into three main modules. By considering the result of each module we moved forward to the next step, and those main modules was also divided into some other submodule.

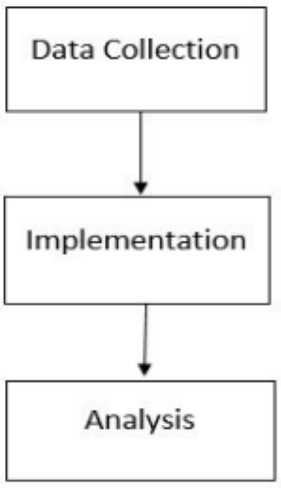

Figure 4.1: Workflow of Our Research

4.1.1 Data Collection

We collect data of heart disease risk factor from patients, hospitals, diagnostic center, clinic center. We collect data from different places of Bangladesh. Basically we collect maximum data from hospital admitted patients. We ask questions to patients and we analyze their test report and collect important attributes result. But there was little bit missing value of different attributes. The proposed model efficiency largely depends upon the quantity and quality of the data that we collect. The more data we collect; our model will be more efficient. But we got a total 59 patients test result with their answer of different question. It was a tough challenge for us to collect data from different sources. Our data collection process was iterative process we repeatedly visited different hospital after a time period. And finally we got data from different places of Bangladesh. Our data set consists of 19 attributes.

4.1.2 Implementation

This module consists of different process of building our different model and measuring their performance. We Preprocess collected data and handle missing value after that we map all the attributes to numerical value so that our model can perform better. We also scale our dataset and normalize value. Implementation module consists different submodule like data preprocessing, feature extraction, training model, testing model.

Data Preprocessing: This is one of the most important states of the implementation module. We collect a large amount of data from several sources but all of them may not be valid. There are many unclear, duplicate and missing values. Considering these we follow the following steps for data preprocessing.

a. Remove Unclear Value: There was many unclear value. Those value actually does not mean anything. So we remove those unclear value. As a result, out datasets attributes get reduced. But we get better result for this process.

b. Remove Duplicate Value: There were many duplicate values in our collected data. So we find those duplicate values and remove them manually as a result our dataset items get reduced.

c. Replace Missing Value: It is not possible to get hundred percent data form collection. There were few missing values in our dataset. Especially different test result was missing, and we take different steps to handle missing values. Like filling missing values with median, mode. Or we make different model to predict missing values.

Feature Extraction: The Proposed model will use CBFSST (Correlation Based Feature Subset Selection Technique) with BFS (Best First Search) in this state. In our collected data all attributes are not as important. We find best subset of attributes using Correlation with target value and different attributes values.

Training the model: We split our processed data into two categories

a. Test Data Set: This part was used to test our model performance.
b. Train Data Set: This part was used to train our model.

We train our model with test dataset using seven machine learning classification algorithms.

Test The Model: After Training the model we evaluate our model performance for those classification algorithms. We record these performance results for the analysis.

4.1.3 Analysis

The analysis module consists of two state

a. Compare the performance result: For each classification algorithm we calculate the precision, accuracy, recall, F1-score, and ROC curve then compare it with others.

b. Identify the conclusion: Finally calculating all the selected machine learning algorithms result, we find out a more efficient algorithm that outperforms all the others performance.

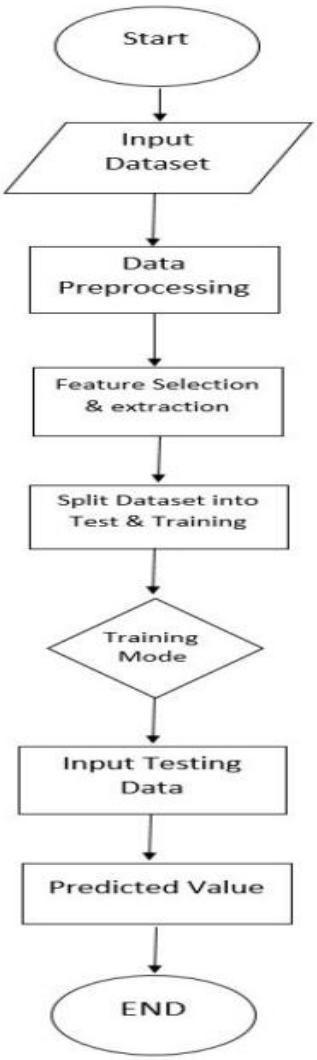

Figure 4.2: Flow Chart of Model Building Process

4.2 DATASET DESCRIPTION

Data collection was a long journey for us. On the time frame of 3 months we collect data from different hospitals, clinic center. Initially we select 30 attributes and we started to collect data on those 30 attributes, but we didn't get all attributes from maximum patients. So we try to collect as much as possible data from patients. We get total 59 sets of data with different number of attributes and we select those attributes that are important and also common on maximum patients. We select 19 attributes, here 18 attributes are independent and 1 attributes that is target value (is the patient has heart disease) is dependent. So our dataset contained 19 attributes and 59 rows although the numbers are quite small, we got a satisfactory performance. Complete attribute description of the dataset is given below.

1. Age: Age of the patient. Age can be numeric value like 10 - 70.
2. Sex: Sex of the patient can be male or female.
3. Occupation: What is patient occupation? It can be nominal value like businessman, housewife, teacher etc.
4. Family History: Is any family member of patients was affected by heart disease. The relationship between heart disease affected family member and patient. It can be nominal like father, mother, brother, etc.
5. Smoking: Does the patient have a smoking habit? Heart attacks are more common in smokers than in nonsmokers. It damages the lining of arteries and builds up a fatty material called atheroma which narrows the arteries causing heart attacks. It can be nominal value like yes or no.
6. Obesity: Does the patient have obesity? Obesity is an abnormal accumulation of body fat, usually 20% or more over an individual's ideal body weight. It can be nominal value like yes or no.
7. Diet: In nutrition, diet is the sum of food consumed by a person. An unhealthy diet is one of the major risk factors for a range of chronic diseases, including cardiovascular diseases, cancer, diabetes and other conditions linked to obesity. It is also advisable to choose unsaturated fats instead of saturated fats and towards the elimination of trans-fatty acids. It can be nominal value like normal means healthy or abnormal means unhealthy.

8. Physical Activity: Physical activity is defined as any bodily movement produced by skeletal muscles that require energy expenditure. Physical inactivity (lack of physical activity) has been identified as the fourth leading risk factor for global mortality (6% of deaths globally) [1]. It can be nominal value like yes or no.

9. Stress: Stress is the body's reaction to harmful situations -- whether they're real or perceived. When men feel threatened, a chemical reaction occurs in the body that allows acting in a way to prevent injury. It can be nominal value like normal levels or high levels.

10. Chest Pain Type (angina): What is the patient chest pain type. There are four types of chest pain such as typical, atypical, asymptomatic, non-cardiac chest pain. Angina is chest pain or discomfort caused when the heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in the chest. This usually happens because one or more of the coronary arteries is narrowed or blocked. It can be nominal value like typical, atypical, asymptomatic, non.

11. Previous Chest Pain: If any kinds of chest pain occur before. Chest pain history, it can be nominal value like yes or no.

12. Edema: Edema is swelling that occurs when too much fluid becomes trapped in the tissues of the body, particularly the skin. Edema can be the result of medication, pregnancy or an underlying disease often congestive heart failure, kidney disease or cirrhosis of the liver. When heart weakens and pumps blood less effectively than the fluid can build up and create edema. It can be nominal value which represents the position of the edema like leg or no.

13. Blood Pressure Systolic: Systolic blood pressure (the first number) indicates how much pressure blood is exerting against artery walls when the heartbeats. When heartbeats, it squeezes and pushes blood through arteries to the rest of the body. This force creates pressure on those blood vessels, and that's systolic blood pressure. Normal systolic pressure is below 120. A reading of 120-129 is elevated. It can be a numeric value.

14. Blood Pressure Diastolic: Diastolic blood pressure (the second number) indicates how much pressure blood is exerting against artery walls while the heart is resting between beats. Normal diastolic blood pressure is lower than 80. It can be a numeric value.

15. Heart Rate: Heart rate, also known as pulse, is the number of times a person's heart beats per minute. Normal heart rate varies from person to person, but a normal range for adults is 60 to 100 beats per minute. It can be a numeric value.

16. Diabetes: Diabetes is a disease in which blood glucose, or blood sugar, levels are too high. Glucose comes from the foods. Insulin is a hormone that helps the glucose get into cells to give them energy. Diabetes can also cause heart disease, stroke and even the need to remove a limb. It can be nominal value like yes or no.

17. Troponin: Troponins are proteins found in the cardiac and skeletal muscles. When the heart is damaged, it releases troponin into the bloodstream. Doctors measure your troponin levels to detect whether or not you’re experiencing a heart attack. This test can also help doctors find the best treatment sooner. It can be nominal value like positive or negative.

18. ECG: An ECG is a test that measures the electrical activity of the heart through small electrode patches attached to the skin of the chest, arms, and legs. It may be used as a test for heart disease. By examining changes from normal on the ECG, clinicians can identify a multitude of cardiac disease processes. With each beat, an electrical impulse (or “wave”) travels through the heart. This wave causes the muscle to squeeze and pump blood from the heart. A normal heartbeat on ECG will show the timing of the top and lower chambers. It can be nominal value like normal, abnormal or undefined.

19. Target: This value indicates whether the patient has heart disease or not. Diagnosis of heart disease (0: not heart disease, 1: heart disease).

4.3 DATA CLEANING AND PROCESSING

Our Collected data has 59 instances so we create dataset with this 59 instances as result out dataset contain 59 rows; in our dataset we take total 19 attributes along with target value. In this 19 attributes there has five numerical attributes, and 14 categorical attributes. The data set has some missing values. Since the missing values are from categorical attributes it was replaced with a mode value of the respective columns. Especially this procedure takes on troponin column there was 8 missing value we fill missing troponin values with the mode. And there were some more missing values in bp_systolic, bp_diastolic and heart_rate attributes we fill all those attributes missing value with corresponding column median value. Data falling outside the 3 standard deviation are considered outliers. Patients who are diagnosed with heart disease have these outliers, meaning some data is above the normal scale. So, remove those outliers

We mapped our categorical attribute to numeric values, this process was selected manually but done with programming. After mapping values, we convert categorical values to dummies value. Dummy variables (or binary variables) are commonly used in statistical analyses and in more simple descriptive statistics. A dummy column is one which has a value of one when a categorical event occurs and a zero when it doesn’t occur. We use one hot encoding technique to create dummy variables of categorical data. A one hot encoding allows the representation of categorical data to be more expressive.

We use feature scaling technique to rescale the data. That helped the machine learning models to perform better. A generic dataset is made up of different values which can be drawn from different distributions, having different scales and, sometimes, there are also outliers. A machine learning algorithm isn't naturally able to distinguish among these various situations, and therefore, it's always preferable to standardize datasets before processing them. A very common problem derives from having a non-zero mean and a variance greater than one [15].

We use standard scaler of scikit learn library to rescale our numerical data between 0 and 1. Standard scaler used on age, bp_systolic, bp_diastolic and heart_rate attribute. Scaling data performs better on KNN, SVM like algorithms.

4.4 TOOLS USED

Different tools are used for this study. All of them are free and open source.

1. Python: Python is a high level general programming language and is very widely used in all types of disciplines such as general programming, web development, software development, data analysis, machine learning etc. Python is used for this project because it is very flexible and easy to use and also documentation and community support is very Large [16].

2. NumPy: NumPy is very powerful package which enables us for scientific computing. It comes with sophisticated functions and is able to perform N-dimensional array, algebra, Fourier transform etc. NumPy is used very where in data analysis, image processing and also different other libraries are built above NumPy and NumPy acts as a base stack for those libraries [17].

3. Matplotlib: Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits [18].

4. Pandas: Pandas is open source BSD licensed software specially written for python programming language. It provides a complete set of data analysis tools for python and is best competitor for R programming language. Operations like reading data-frame, reading csv and excel files, slicing, indexing, merging, handling missing data etc can be easily performed with Pandas. Most important feature of Pandas is, it can perform time series analysis [19].

5. Seaborn: Seaborn is library used for data visualization and is created by using the Python programming language. It is high level library stacked on top of matplotlib. Seaborn is more attractive and informative than matplotlib and very easy to use and is tightly integrated with NumPy and Pandas. Seaborn and matplotlib can be used essentially side by side to derive conclusions from the datasets [20].

6. SciPy, Scikit-learn: SciPy is a collection of fundamental mathematical functions and is built on the top of Numpy while Scikit-learn is widely used popular library for machine learning, it is third party extension to SciPy. Scikit-learn includes all the tools and algorithms needed for most of machine learning tasks. Scikit-learn supports regression, classification, clustering, dimensionality reduction and data pre-processing. For this study, scikit-learn is used because it is based on python and can interoperate to NumPy library. It is also very easy to use [21].

4.5 SUMMARY

In our material and methods section we discussed about our data collection process, data preprocessing technique. We plot our data into different graphs to get the insight our dataset. We select two subsets of our dataset one consisting all the features and another one contains only important features. We consider important features by checking the correlation with the target value and features. All the processes and techniques used some tools and materials, all those tools helps us to find the best technique to do the task.

CHAPTER 5

RESULT AND DISCUSSION

In our previous chapters we have discussed about different algorithms, previous works in this field and the dataset we used for our experiments. All those were the foundation for this chapter. In this chapter we discussed about results that we found after implementing the algorithms and analyzed them.

5.1 INTRODUCTION

Comparative analysis of different machine learning algorithms like Logistic Regression, Naive Bayes Classifier, K Nearest Neighbor, Support Vector Machines, Decision Trees, Random Forest, Multi-Layer Perceptron Neural Networks gives us the insight about which algorithms performs better for disease prediction. For the result we can ensure how to get better performance of an algorithms so that we can build a perfect model that can predict any outcome with the largest number of accuracy. In our thesis we collect data, preprocess data, and then build model by training model with those data. After all we get different performance metrics like accuracy, precision, recall, f1 score and roc auc score. By comparing those score we can easily find out which algorithm is better than other algorithms. We take our dataset two format one consisting all the 19 attributes and another one consisting only 14 attributes we reduced some attribute based on their correlation with target and their impact to predict heart disease. So in this thesis we will compare the performance of different algorithms using one dataset with two variations of attributes set.

5.2 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) from statistics to linear algebra by using simple plotting tools, to understand what our dataset is, before going to do actual machine learning. Visualizing attributes is perhaps the most important/interesting part of EDA. Anything in this world could be better understood if we have an image/visualization of it.

In short, EDA is “A first look at the data”. It is a critical step in analyzing the data from an experiment. It is used to understand and summarize the content of the dataset to ensure that the features which we feed to our machine learning algorithms are refined and we get valid, correctly interpreted results.

5.2.1 Data Visualization

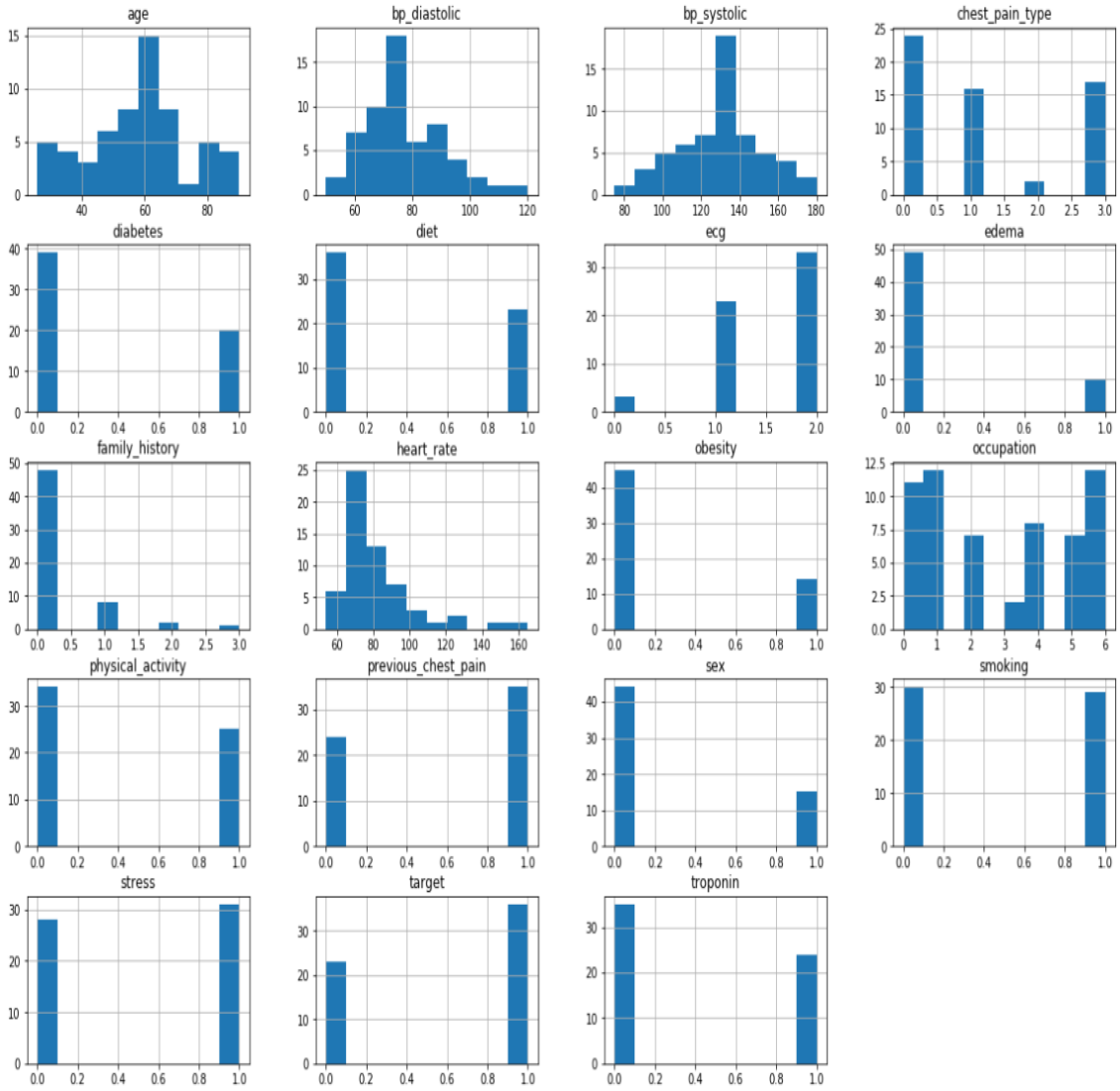

Figure 5.1: Frequency Distribution of Features for All Participants

Here we visualize the dataset for all patients with all 19 attributes. It is now easy to see the frequency distribution of dataset. So we can say that age, bp_systolic, bp_diastolic looks like normal distributed.

5.2.2 Descriptive statistics of Dataset

Out of 19 attributes, the data contains only five numeric attributes.

Table 5.1: Descriptive Statistics of Numeric Value

	Age	bp_systolic	bp_diastolic	heart_rate	Target
count	59.0	46.00	46.00	47.00	59.00
Mean	15.76	128.26	77.86	83.95	0.61
Std	15.87	24.06	15.44	22.70	0.49
Min	26.00	75.00	50.00	54.00	0.00
25%	49.00	110.00	66.25	70.50	0.00
50%	60.00	130.00	75.00	76.00	1.00
75%	65.00	143.75	90.00	90.00	1.00
Max	90.00	180.00	120.00	165.00	1.00

From this table we see that there has 59 values of age, 46 values of bp_systolic, 46 values of bp_diastolic, 47 values of heart_rate, and 59 values of target. Actually there has 59 patient's data and there has some missing values in bp_systolic, bp_diastolic and heart_rate attributes. So before model building we fill up those missing values with different techniques. We also see that mean of age if 57 so we can say that most of the patients was old. Minimum values of age were 26 years and maximum value was 90 years. Age distribution of data is left skewed which shows population with low age are absent. Standard deviation is 15.87 which shows the sparseness of the age, ranging from 57 years to 73 years, this age range visits the doctor most. Similarly, bp_systolic mean was 128 which show that most patients bp systolic was normal. Minimum value of bp_systolic was 75Hgmm and maximum was 180Hgmm, and bp_diastolic values mean was 77, minimum value was 50Hgmm, and maximum value was 120Hgmm. There were 47 heart_rate so we can say some value

was missing we use median to fill those missing value. Mean of heart rate was 83bpm and max heart rate found was 165bpm (bit per minute).

Correlation between all the numeric data is given below.

Table 5.2: Correlation Among Numeric Value

	Age	bp_systolic	bp_diastolic	heart_rate	Target
Age	1.000000	0.0199994	0.069676	-0.030911	0.288249
bp_systolic	0.019994	1.000000	0.805017	0.150733	0.132768
bp_diastolic	0.069676	0.805017	1.000000	0.208148	0.205186
heart_rate	-0.030911	0.150733	0.208148	1.000000	0.145232
target	0.288249	0.132768	0.205186	0.145232	1.000000

From the above table we see that there is a close correlation between bp_systolci and bp_diastolic. If we see the heat of of this correlation table we can understand better. Heatmap of correlation between numeric value is given below.

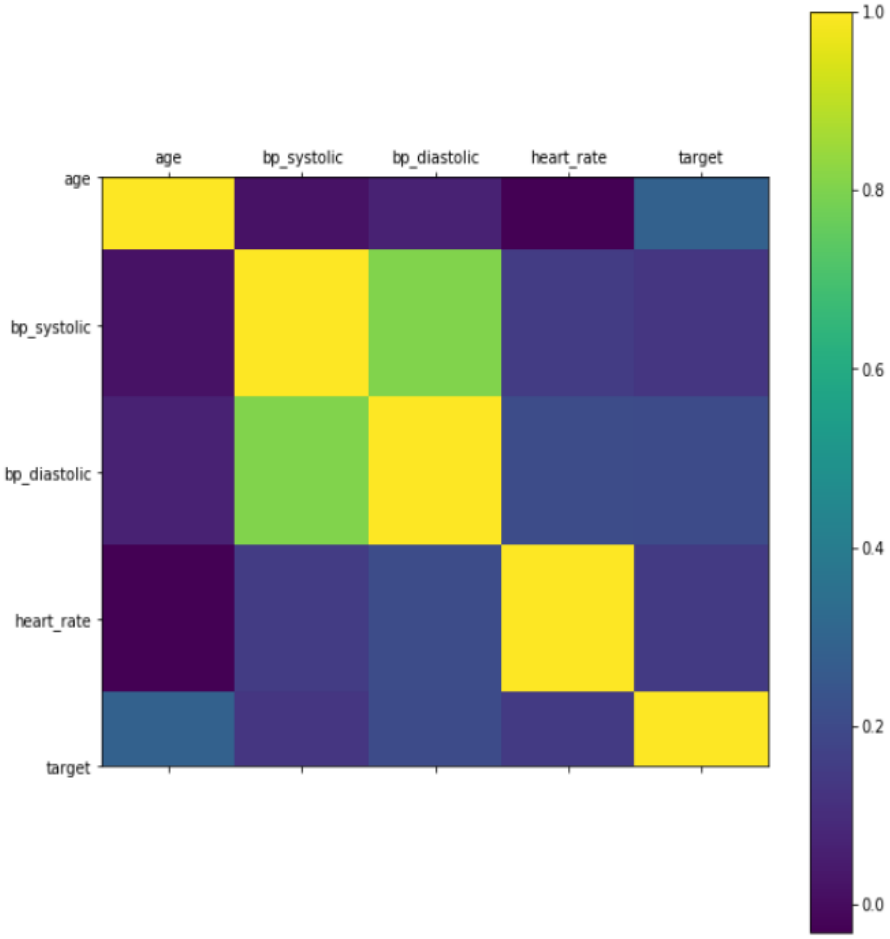

Figure 5.2: Correlation Numeric Value Visualizing using Heat map

So we should depreciate bp_systolic or bp_diastolic. But from the correlation with target and all other attributes we understand that bp_diastolic has more correlation with target. So we deprecate bp_systolic for our selected dataset.

From the exploratory analysis of the data, the following insights are gained.

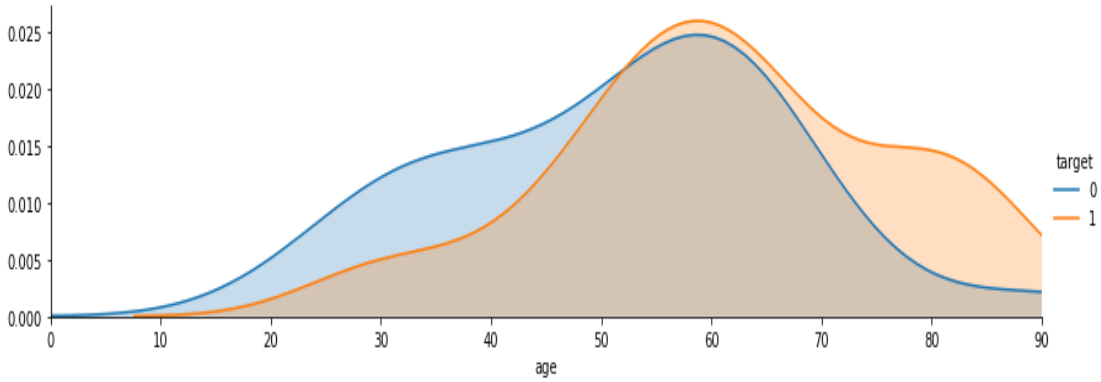

Figure 5.3: Distribution of People's Age Corresponding to Target

From the above graph we see that in our dataset people who are older have more possibility to be a heart disease patient that younger people. Especially people with age more than 50 years have the possibility to be a patient of heart disease. And we also see people who are younger that means people within the age up to 50 years have less heart disease. And the heart disease patients are decreasing with the decrease of age. We take age as an important factor to predict heart disease.

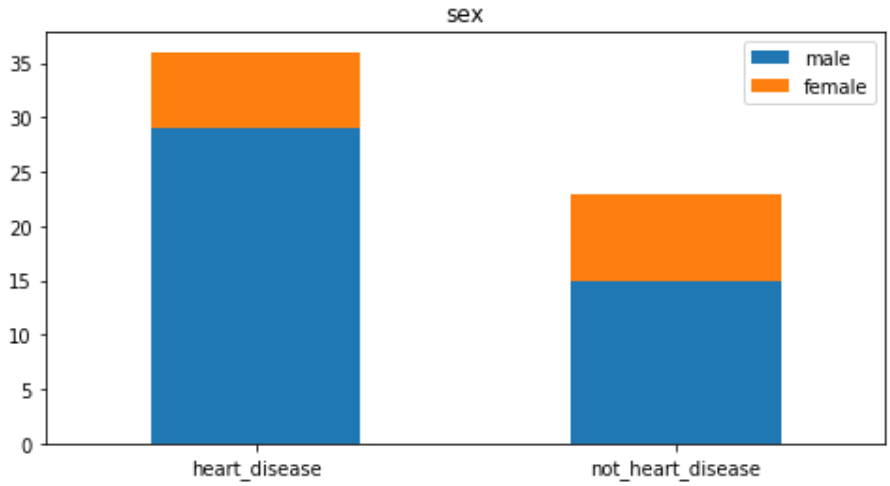

Figure 5.4: Sex vs Patients

In our data set we found that male have more heart disease than females. From the above Bar chart we can say that in country male has more risk to heart disease than female. So take age as an important risk factor for heart disease prediction. Our dataset contains more male patient’s data than female patient’s data.

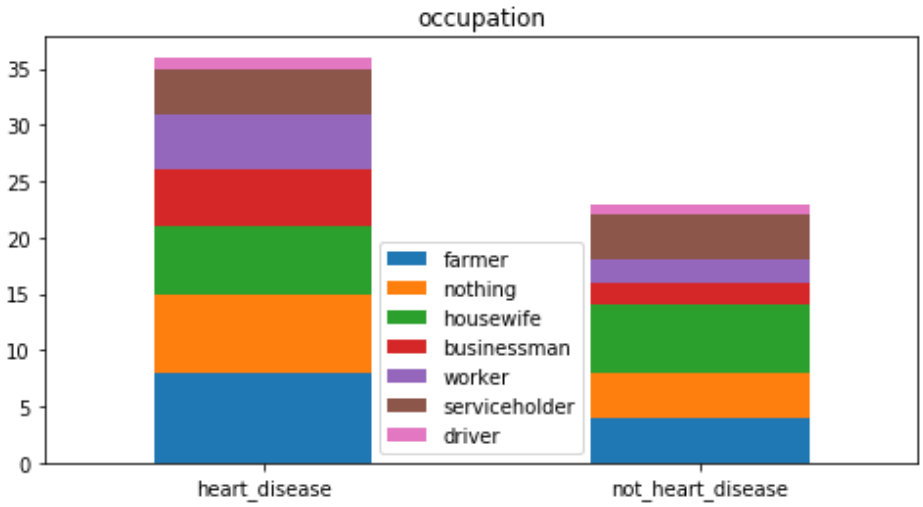

Figure 5.5: Occupation vs Patients

In our dataset, we see that all patients are distributed to different kinds of occupation. From her we see there is not very much impact of occupation to heart disease. We conclude that occupation is not an important risk factor.

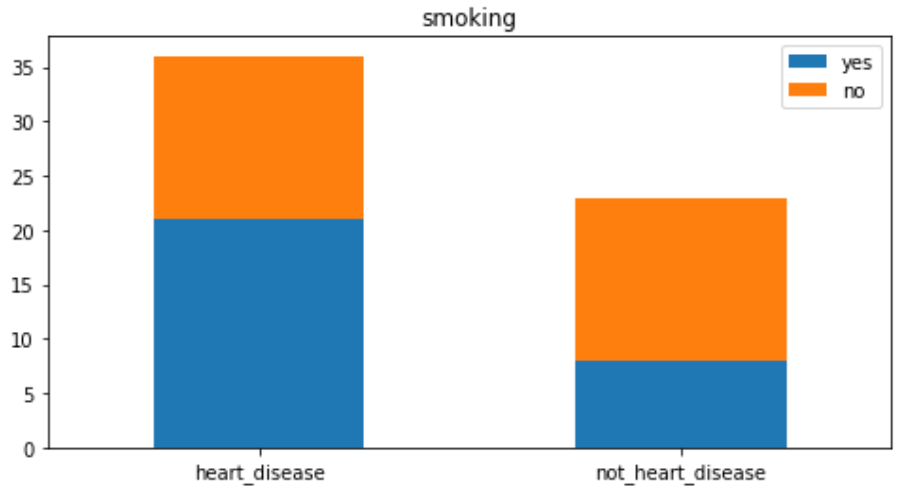

Figure 5.6: Smoking Habit Vs Patients

From the above bar chart we see smoking habit has a great impact on heart disease. Here we see that the patients have smoking habit has more possibility to be a heart disease patient. So we take smoking as an important risk factor of heart disease.

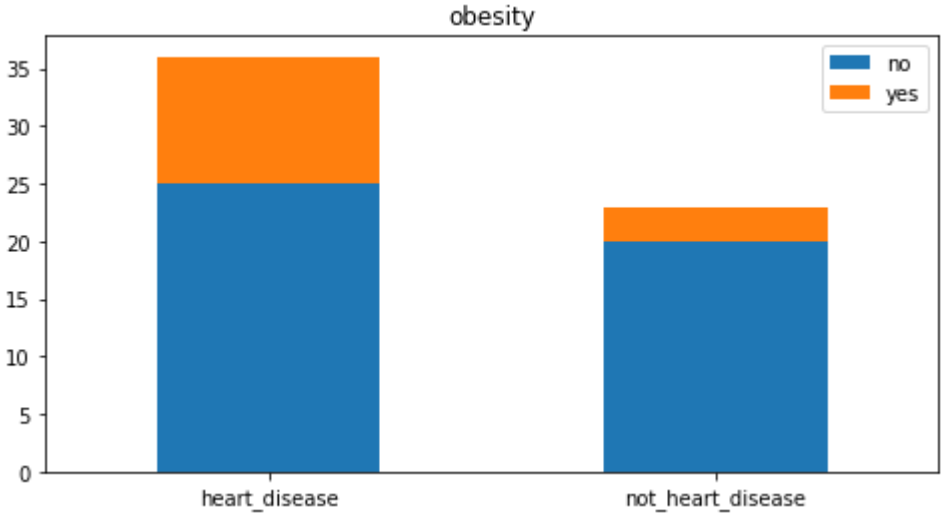

Figure 5.7: Obesity vs Patients

Obesity also has some impact on heart disease. From our dataset we see overweight people has the possibility to be a heart disease patient. Cause obesity make block of coronary artery. So we take it as an important risk factor.

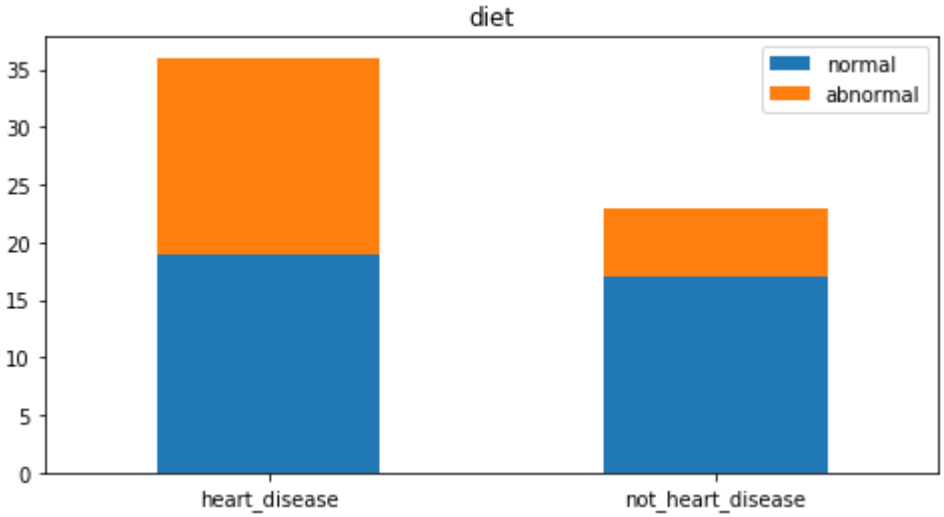

Figure 5.8: Diet vs Patients

From the above bar chart we see people who do not control their diet has the possibility to be a heart disease patient. Because diet is an important factor to control different disease. So we take diet as an important risk factor to predict heart disease.

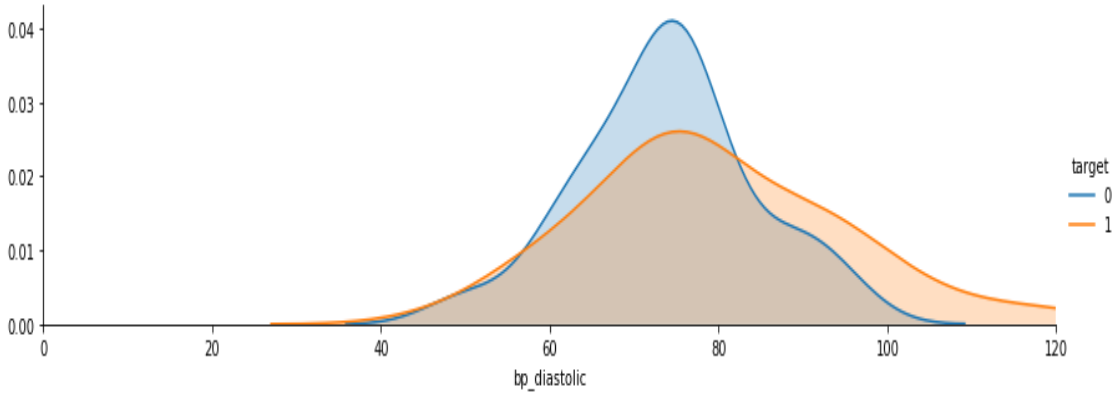

Figure 5.9: Distribution of People's Diastolic Blood Pressure Corresponding to Target

We see that people with normal blood pressure has more possibility to be not heart disease patients. When there has been some abnormal behavior of blood pressure like increasing blood pressure decreasing blood pressure than there has the possibility to be a patient of heart disease. So we consider bp_diastolic as an important factor for heart disease prediction.

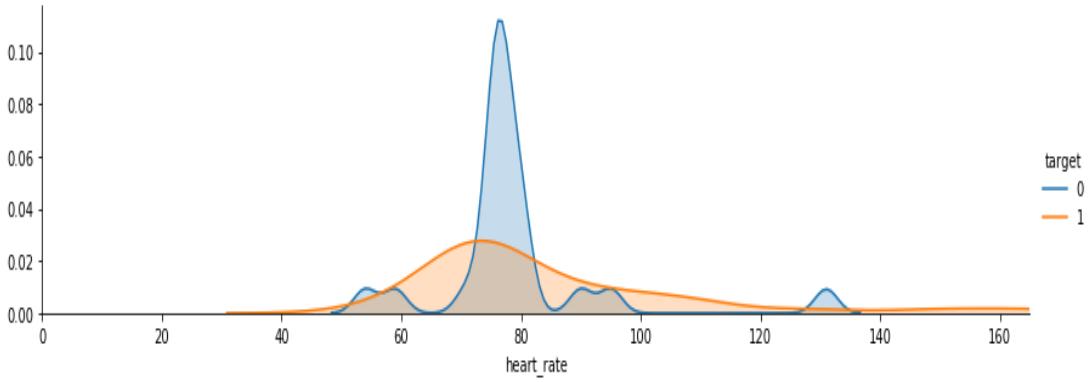

Figure 5.10: Distribution of People's Heart Rate Corresponding to Target

People with normal heart rate has more possibility to be a heart disease patient. When heart rate decrease or increase possibility of heart disease also increase. And we take heart rate as an important factor.

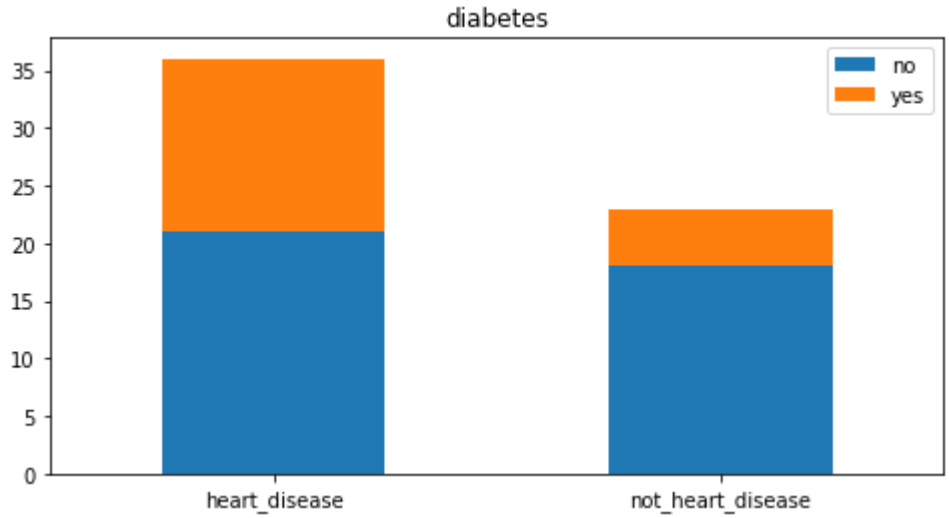

Figure 5.11: Diabetes vs Patients

Diabetes patients have more risk of heart disease. Cause diabetes can create other kinds of disease along with heart problems. We see that in our dataset heart disease patients have diabetes more than not heart disease patients. So we take diabetes as an important factor.

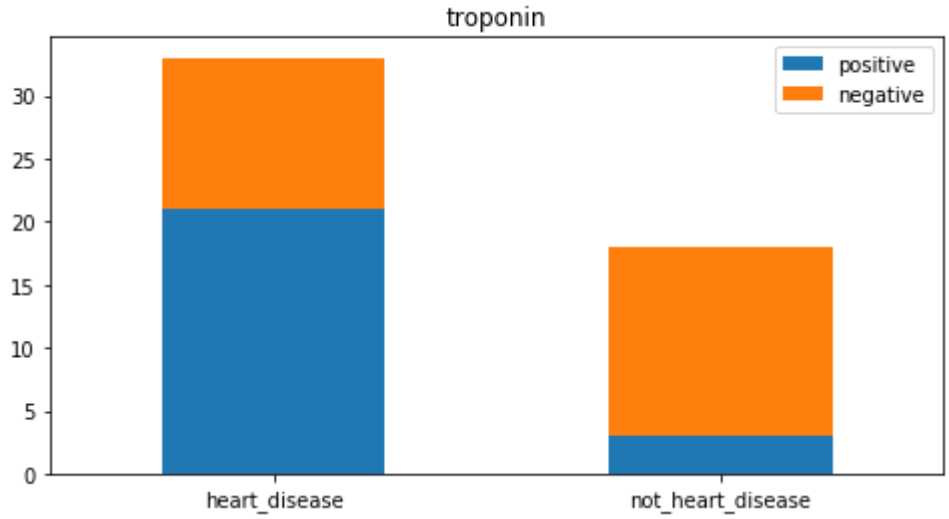

Figure 5.12: Troponin vs Patients

Troponin, or the troponin complex, is a complex of three regulatory proteins that are integral to muscle contraction in skeletal muscle and cardiac muscle, but not smooth

muscle. Heart disease patients has more troponin positive compared to not heart disease patients. So we take it as an important factor to predict heart disease.

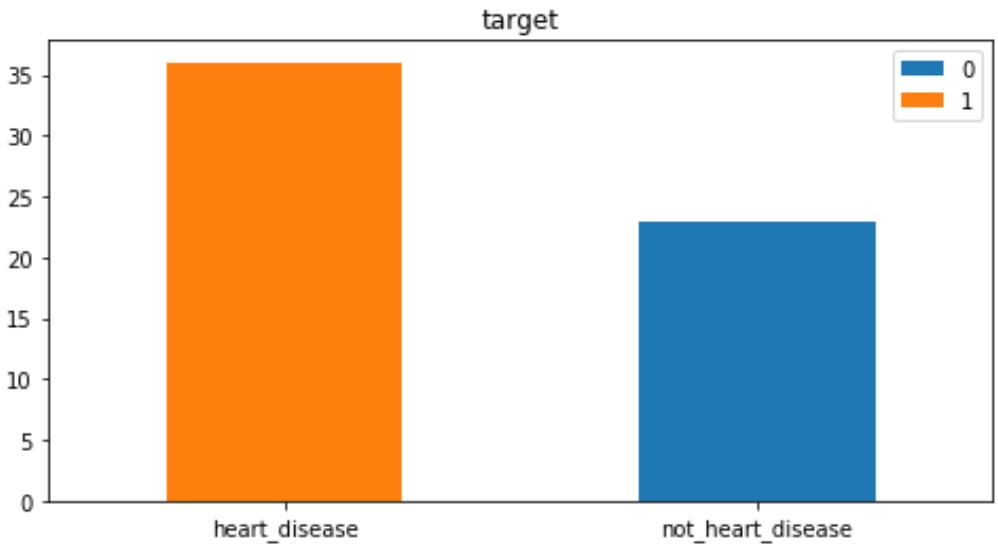

Figure 5.13: Heart Disease Vs Not Heart Disease Patients

At last we see the count number of heart disease patients and not heart disease patients. In our dataset we have more heart disease patients than not heart disease patients. We have almost 35 instances of heart disease and the rest are not heart disease patients.

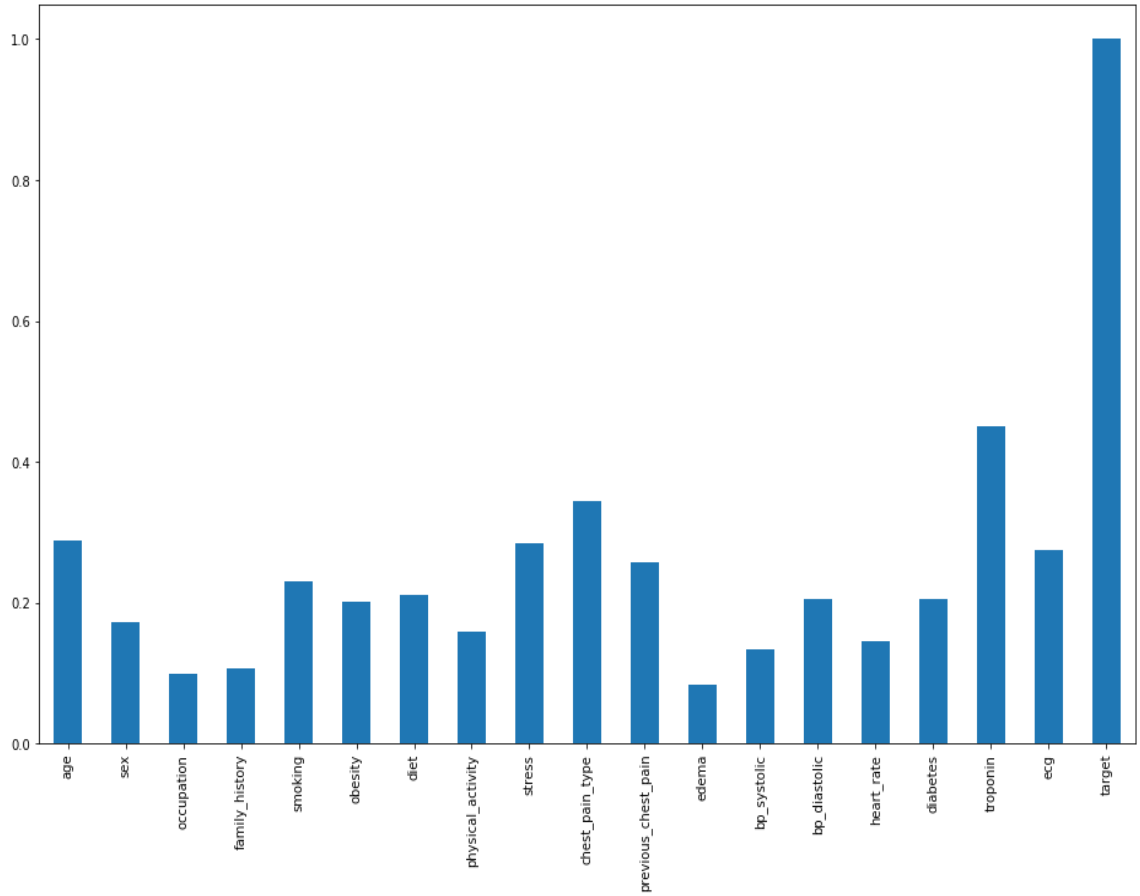

Figure 5.14: Correlation of All Attributes with Target Value

From the above bar chart we see different attribute has different correlation with target. We make a subset of attribute form correlation with target, and this subset contain age, sex, smoking, obesity, diet, physical_activity, stress, chest_pain_type, previous_chest_pain, bp_diastolic, diabetes, troponin, egg and target.

5.3 EXPERIMENTS

Different machine learning models were experimented using our collected data. In this study initially data set was modelled without feature selection and the results were obtained and in the second phase, data set was modelled only with features obtained from correlation based feature selection technique. All experiments included methods accuracy, precision, recall, f1 score, auc roc score. K-fold cross validation is a technique used in a dataset to avoid overfitting and under-fitting of the model and

parameter tuning is a technique which assists to find the best parameters for the model being used.

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. Training set is the portion of data in which the model is trained. In this study, 67 percent of data was used for training. So from our 59 instances of dataset 39 instances used for training set. In general, in machine learning communities, it is normal to use 60 to 70 percent of data for training but it varies diversely according to the need and purpose of the experiment. In data training, often the accuracy of training is high, meaning the model shows high level of accuracy performance in the training set but when tested against the test set, the performance is poor. So to avoid performance error, k-fold cross validation was used. In k-fold cross validation, for example 10-fold cross validation, training set is split into 10 parts and and from each 10 part, training and test set is defined and model is employed and the result of all the 10 parts are averaged, this helps to minimize the overfitting and under fitting of the data.

The test set is the portion of data where the model is tested, it is often the dependent variable of the data. In this study, 33 percent of the data was used for testing. From our data set 20 instances used for testing dataset. When cross validated data is tested, it will perform better or worse depending on the model used. So, to ensure every model is functioning in its optimum, a technique named parameter tuning was used. We use sklearn library to split our data set as train and test set. Sklearn model selection train test split library component split the dataset randomly with specified portion and we get random train and test part from full dataset.

5.4 RESULTS

The aim of the entire project was to test which algorithm classifies diseases the best. This section includes all the results obtained from the study and introduces the best performer according to various performance metrics. First, performance was obtained by using the dataset that contain all 19 attributes. Secondly, performance was obtained just by considering 14 attributes that has strong correlation with the target value. Then at last we compare different performance metrics of different algorithms.

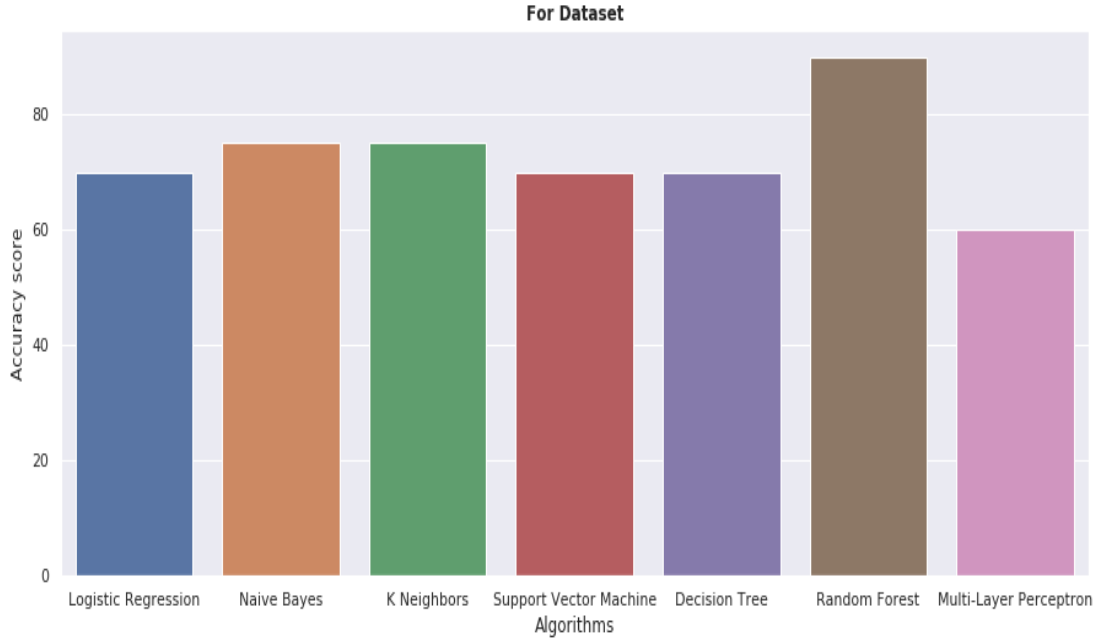

Figure 5.15: Accuracy of Different Models for All Attributes

From the above bar chart we see that using Random Forest algorithms we got highest accuracy than any other algorithms, after Random Forest we see that Naive Bayes and K Neighbors gives better accuracy. Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data. But all the time accuracy does not give good performance metrics to compare algorithms so take other metrics like precision, recall, f1 score and auc roc score.

Different Performance Metrics of Models with All Features

Table 5.3: Evaluation of algorithms in test dataset with all features

	Accuracy	Precision	Recall	F1-score	ROC AUC Score
Logistic Regression	70.0	69.23	81.82	75.00	68.69
Naive Bayes	75.0	87.50	63.64	73.68	76.26
K Neighbors	75.0	75.00	75.00	75.00	75.00
SVM	70.0	72.73	72.73	72.73	69.70
Decision Tree	70.0	69.23	81.82	75.00	68.69
Random Forest	90.0	84.62	100.00	91.67	88.89
Multilayer Perceptron	60.0	60.00	81.82	69.63	57.58

Referring to the above table we see Random Forest perform better in all the performance metrics like accuracy, precision, recall, f1 score, roc auc score. We go highest 90 percent accuracy with highest precision 84 percent, highest recall 100 percent, highest f1 score 91 percent, and highest roc auc score 88 percent. So we can say for our dataset with 19 attributes Random Forest is better than any other six classification algorithms. After that K Neighbors and Naive Bayes performs good but Naive Bayes has max score in precision than K Neighbors. But recall and f1 score is less than K Neighbors although ROC AUC Score is better than K Neighbors for Naive Bayes we can say K Neighbors is better than Naive Bayes considering other metrics. Logistic Regression give 70 percent accuracy with 75 percent F1 Score. SVM gives 70 percent accuracy with 72 percent F1 Score. Decision Tree gives 70 percent accuracy with 75 percent F1 Score So Logistic Regression and Decision Tree are the same position on the race. But Multi-Layer Perceptron did not show good performance compared to other algorithms.

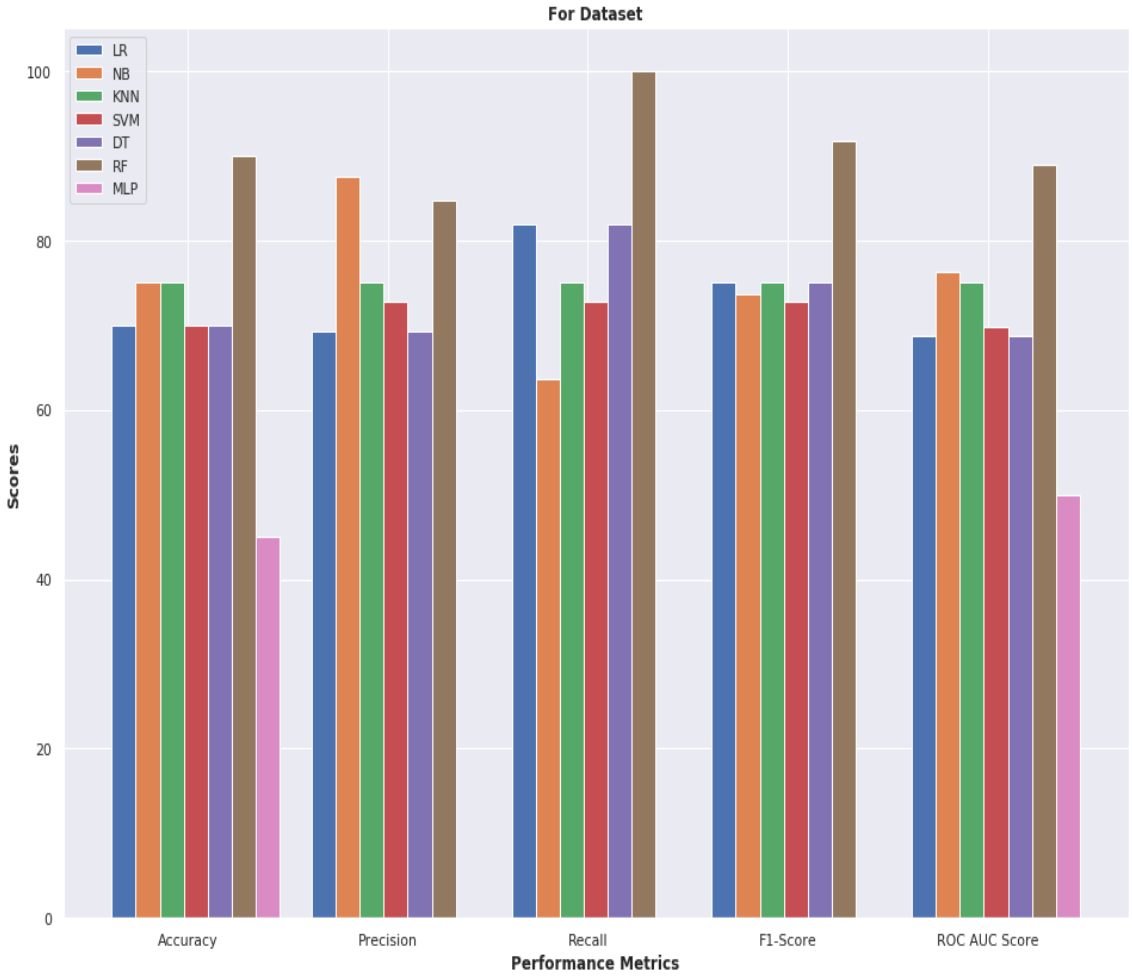

Figure 5.16: Performance metrics of different algorithms visualized using bar chart.

So we can conclude that Random Forest is best classification algorithms for out dataset and Multi-Layer Perceptron is not so good compared to other.

When we narrow down the attributes set to 14 attributes than we test another performance metrics to all algorithms. Let's see the accuracy of models with selected subset of attributes.

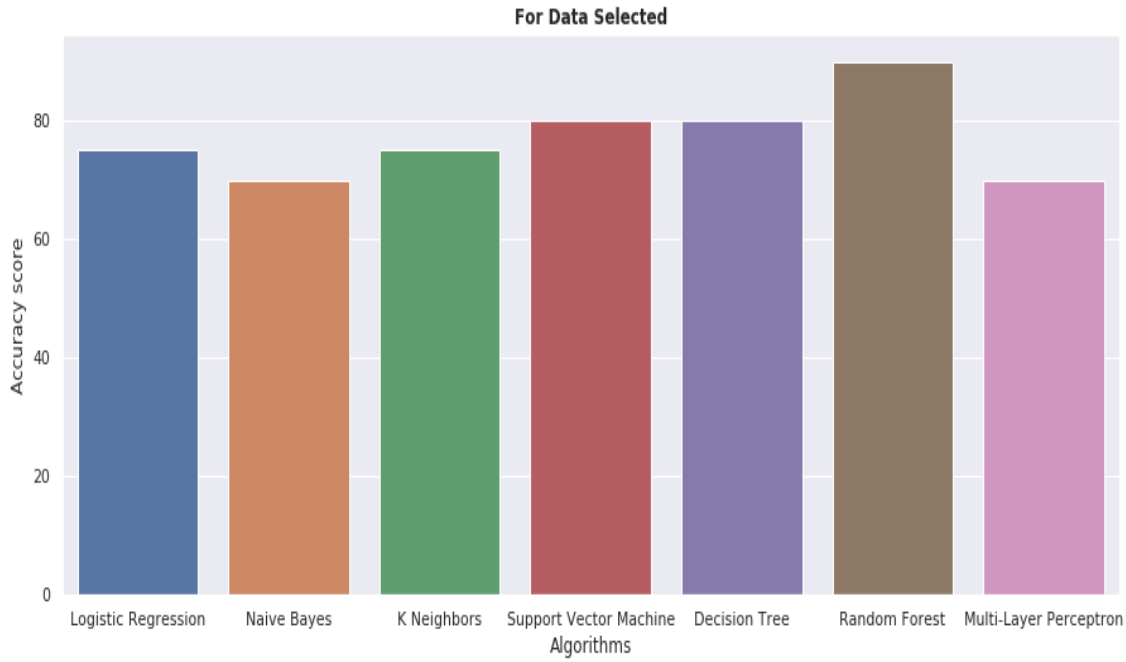

Figure 5.17: Accuracy of Different Models for Selected Attributes

From the above bar chart we see we got slightly different results than previous results. This time also Random Forest gives best accuracy about 90 percent. K Neighbors gives 85 percent which is second best accuracy this dataset. And Decision Tree and SVM increase their accuracy percent about 80 percent. Logistic Regression increase accuracy to 75 percent. Also Multi-Layer Perceptron gives good accuracy than previous test. But Naive Bayes decrease its accuracy for this dataset. So we conclude that by considering only important feature we can increase accuracy of different classification algorithms.

Different Performance Metrics of Models with Subset of Features

Table 5.4: Evaluation of algorithms in test dataset with selected features

	Accuracy	Precision	Recall	F1-score	ROC AUC Score
Logistic Regression	75.0	75.00	81.82	78.26	74.24
Naive Bayes	70.0	77.78	63.64	70.00	70.71
K Neighbors	85.0	85.00	85.00	85.00	85.00
SVM	80.0	76.92	83.33	83.33	78.79
Decision Tree	80.0	81.82	81.82	81.82	79.80
Random Forest	90.0	90.91	90.91	90.91	89.90
Multilayer Perceptron	70.0	66.67	76.92	76.92	67.68

Referring to the above table we see there are many differences in the performance metrics with the previous test result. Precision increase in Logistic Regression, K Neighbors, SVM, Decision Tree, Multi-Layer Perceptron. But precision decrease for Naive Bayes like accuracy decrease. Random Forest has a max precision and Multi-Layer Perceptron has less precision. Recall increase for K Neighbors, SVM, Decision Tree and Multi-Layer Perceptron. But this time there is no decrease in the recall for any algorithms. F1 Score gives the full overview of precision and recall simultaneously so we see F1 Score increase for Logistic Regression, K Neighbors, SVM, Decision Tree and Multi-Layer Perceptron. Naive Bayes decrease it's F1 Score. ROC AUC score is increase for all algorithms but not for Naive Bayes. ROC (*Receiver Operating Characteristic*) Curve tells us about how good the model can distinguish between two things (*e.g. If a patient has a disease or no*). Better models can accurately distinguish between the two. Whereas, a poor model will have difficulties in distinguishing between the two. So from the above table we conclude that Random Forest is here giving the best performance. After that K Neighbor gives good performance. Here SVM and Decision Tree are almost equal to the race, but Decision Tree perform slightly better than SVM. and Naive Bayes decrease its performance in all sectors. Multi-Layer Perceptron also increase its performance relatively previous test.

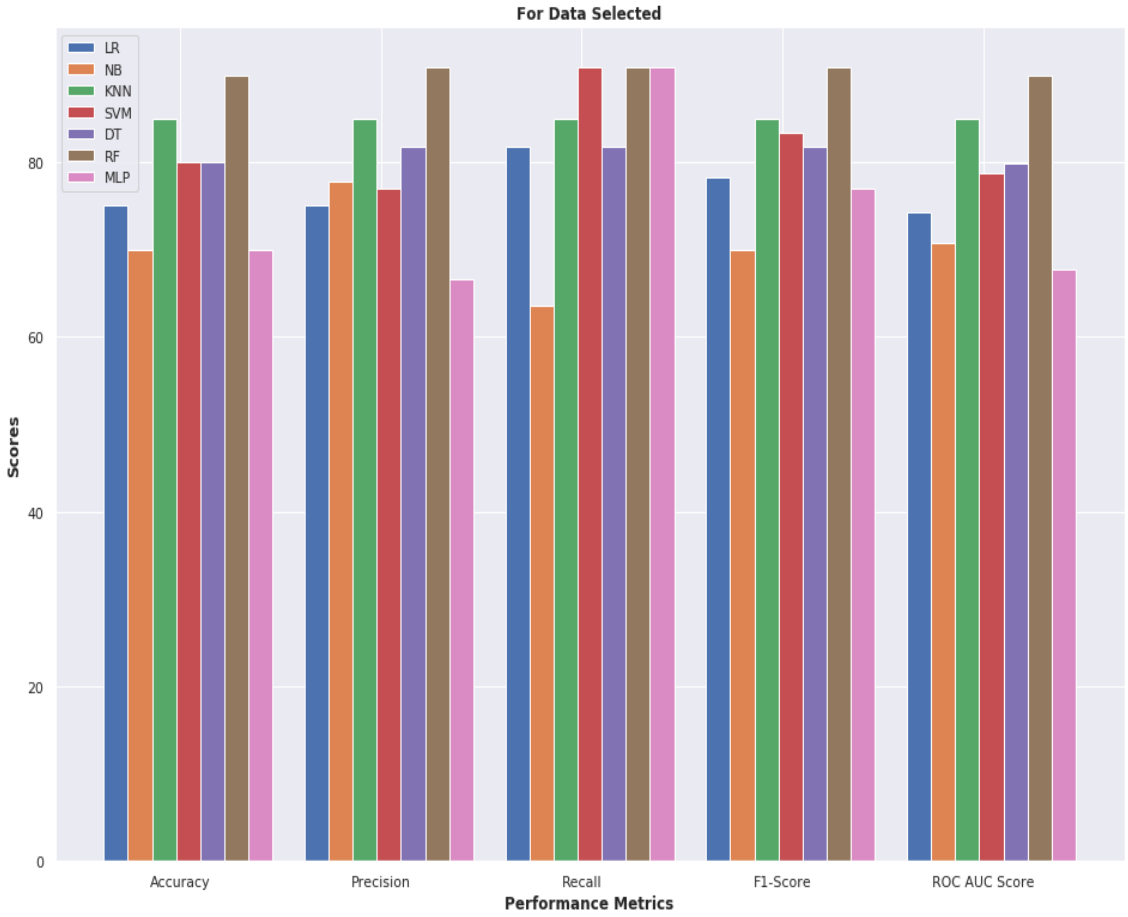

Figure 5.18: Performance metrics of different algorithms visualized using bar chart.

5.5 DISCUSSION

From all the tables above, we see that different algorithms performed better depending upon the situation whether subset of feature is selected or all the features is used. Every algorithm has its intrinsic capacity to outperform other algorithms depending upon the situation. For example, Random Forest performs better than any other algorithms in our dataset. But we know SVM (Support Vector Machine) performs better for small size of data, and for large set of data ensemble type classifier like random forest perform better. In the case of decision tree missing values play an important role. Even after imputing it can't give the result which it can with a perfect dataset.

Gaussian naive bayes is one of the good classifiers. But our dataset it did not perform well. The reason of its pre-assumption that all the attributes are independent. If there was a dependency between the attributes in the dataset it Results and Analysis would have given less accuracy. K-nearest neighbor's accuracy increases with the number of 'k' we pick. It assures the similarity between the given point and the dataset. Multi-Layer Perceptron works better when the dataset has a large number of instances. It helps Multi-Layer Perceptron to understand the underneath patterns in the dataset and to understand it's behavior.

Performance of algorithms increased after reducing the attribute in the dataset. When there are many attributes classifier algorithms get complex and their prediction results also varies. For the comparison of dataset, performance metrics after feature selection, data standardization, data scaling, and converting dummies value is used because this is the standard process of evaluating algorithms.

CHAPTER 6

CONCLUSION

6.1 INTRODUCTION

There are numerous medical diagnosis procedures, it has become a much more complicated job in evaluating the risk factors of any disease. Because the quantity of data has been increased, it has become a critical task to maintain, store and evaluate the data in the medical diagnosis. The risk factors may hold numerous measures in the dataset and considering all the measures increases the time of medical practitioners for the decision-making process. So we try to use these machine learning algorithms to analyze these large amounts of data. Our experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly. The model from the classification will be able to answer more complex queries in the prediction of heart attack diseases. To be able to predict the heart disease accurately using machine learning algorithms implemented model has many significances.

6.2 CONCLUSION

Through our dataset is relatively small for data analysis and prediction model building but try to find out which classification algorithms performs better in terms of disease prediction especially heart disease. After processing collected data we conclude that Random Forest performs better than any other classification algorithms for out data set, and we find out the cause of its performance that is an ensemble technique which create it’s individuals tree and combine all results to give us the final

result. Random Forest gives accuracy score 90 percent for when we consider all features. KNN gives second best accuracy score 85 percent when we narrow down features set. And Decision Tree and SVM gives accuracy score 80 percent. So after considering all the performance metrics we conclude that Random Forest, KNN, Decision Tree and SVM are preferable classification algorithms for heart disease prediction.

6.3 FUTURE SCOPE

In future by collecting more number of data instances we can build a more accurate model which can be used to manufacture different devices and which will help to monitor the heart-related activities and diagnose the disease. These devices will be more helpful when they can predict heart disease more accurately and efficiently because heart disease experts are not available everywhere in our country. The study in this paper should carried to find the best feature selection model that can provide higher accuracy.

REFERENCES

[1] WHO (World Health Organization). World Heart Day (online). https://www.who.int/cardiovascular_diseases/world-heart-day/en (2 January 2020).

[2] Laslett LJ, Alagona P, Clark BA; 2012. The worldwide environment of cardiovascular disease: prevalence, diagnosis, therapy, and policy issues: a report from the American college of cardiology. *J Am Coll Cardiol*.

[3] AhsanKarar Z., Alam N., Kim Streatfield P. 2009. Epidemiological transition in rural Bangladesh, 1986–2006. *Glob Health Action*.

[4] Tom Mitchell, McGraw Hill. 1997. *Machine Learning.* ISBN 0070428077.

[5] Reena Shaw. 2019. The 10 Best Machine Learning Algorithms for Data Science Beginners(online). https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners (5 January 2020).

[6] R. Rajalakshmi, Dr.R.Lath. 2016. Performance Evaluation of Classification Techniques in Diagnosing the Risk Factors of CVD. *International Journal of Innovative Science and Engineering Research(IJFISER)*, **Vol. 2**(2).

[7] Koushik Chandra Howlader, Md. Shahriare Satu, Arup Mazumder. 2017. Performance Analysis of Different Classification Algorithms that Predict Heart Disease Severity in Bangladesh. *International Journal of Computer Science and Information Security (IJCSIS)* **Vol. 15**.

[8] Jyothi Rohilla, Preeti Gulia. 2015. Analysis of Data Mining Techniques for Diagnosing Heart Disease. *International Journal of Advanced Research in Computer Science and Software Engineering*. **Vol. 5**(7).

[9] Mr. Sudhir M. Gorade, Prof. Ankit Deo, Prof.Pritesh Purohit. 2017. A Study of Some Data Mining Classification Techniques. *International Research Journal of Engineering and Technology (IRJET)*. **Vol. 4**(4).

[10] Saxena, K., Sharma. 2015. Efficient heart disease prediction system using the decision tree. *In Computing, Communication & Automation (ICCCA)*. **Vol. 5**(3).

[11] Priyanka P. Shinde, Kavita S. Oza, Rajanish. 2018. Analytical Perception for Medical Data. *International Journal of Pure and Applied Mathematics* **Vol. 119.**

[12] Anjali Burande Pankaj Srivastava, Amit Srivastava, and Amit Khandelwal. 2013. A note on hypertension classification scheme and soft computing decision-making system. *ISRN Biomathematics*. **Vol. 11.**

[13] V. Manikandan and S. Latha. 2013. Predicting the Analysis of Heart Disease Symptoms Using Medical Data Mining Methods. *International Journal of Advanced Computer Theory and Engineering*. **Vol. 2**(2).

[14] Aditya Methaila.2014. Early Heart Disease Prediction Using Data Mining Techniques. *CCSEIT, DMDB, ICBB, MoWiN, AIAP*.

[15] Giuseppe Bonaccorso. 2017. Machine Learning Algorithms.

[16] About Python (online). https://www.python.org/about (7 January 2020).

[17] About Numpy (online). https://numpy.org/ (7 January 2020).

[18] About Matplotlib (online). https://matplotlib.org/ (7January 2020).

[19] About Pandas (online). http://pandas.pydata.org/ (7 January 2020).

[20] An introduction to seaborn (online). http://seaborn.pydata.org/introduction.html (7 January 2020).

[21] Scikit-learn (online). https://scikit-learn.org (7 January 2020).

APPENDICES

Appendix 1: Sample of Dataset after mapping with numeric value

	age	sex	smoking	obesity	diet	physical_activity	stress	chest_pain_type	previous_chest_pain	bp_diastolic	diabetes	troponin	ecg	target
0	60	0	1	0	0	1	0	0	0	75.0	1	0.0	1	0
1	54	0	1	0	0	0	0	1	0	90.0	0	0.0	2	1
2	45	0	1	0	0	1	1	0	1	60.0	0	1.0	1	1
3	60	0	0	0	0	0	0	3	1	90.0	0	0.0	2	0
4	52	0	1	0	0	0	1	0	0	60.0	0	1.0	1	1
5	60	0	0	0	0	0	0	3	0	65.0	0	0.0	2	0
6	84	0	1	0	1	0	1	1	0	90.0	0	1.0	2	1
7	65	0	1	1	0	0	0	0	0	95.0	0	1.0	2	1
8	42	0	1	0	1	1	1	1	0	80.0	0	1.0	2	1
9	55	0	0	0	1	1	1	0	0	92.0	1	0.0	2	1
10	63	0	0	1	1	0	1	1	1	75.0	0	0.0	2	1
11	35	1	0	0	0	0	1	3	0	70.0	0	0.0	1	0
12	80	0	1	1	1	0	1	0	1	75.0	0	1.0	0	1
13	88	1	0	1	0	0	1	1	1	60.0	0	1.0	2	1
14	90	1	0	0	0	1	0	3	0	80.0	0	0.0	1	0
15	80	1	0	0	0	0	1	3	1	70.0	1	0.0	0	1
16	70	1	0	1	1	0	1	1	1	80.0	0	1.0	0	1
17	60	0	0	0	1	0	0	3	1	75.0	0	0.0	1	0

Appendix 2: Sample of Code

```
# importing necessary module
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.metrics import accuracy_score

# reading dataset to program
dataset = pd.read_csv("heart_disease_data.csv")

# mapping dataset to numeric value
sex_mapping = {"male": 0, "female": 1}
dataset['sex'] = dataset['sex'].map(sex_mapping)
occupation_mapping = {"nothing": 0, "farmer": 1, "businessman": 2, "driver": 3,
"serviceholder": 4,
```

```
        "worker": 5, "housewife": 6, "military": 7, "doctor": 8}
dataset['occupation'] = dataset['occupation'].map(occupation_mapping)

family_history_mapping = {"no": 0, "father": 1, "brother": 2, "son": 3}
dataset['family_history'] = dataset['family_history'].map(family_history_mapping)

smoking_mapping = {"no": 0, "yes": 1}
dataset['smoking'] = dataset['smoking'].map(smoking_mapping)

obesity_mapping = {"no": 0, "yes": 1}
dataset['obesity'] = dataset['obesity'].map(obesity_mapping)

diet_mapping = {"normal": 0, "abnormal": 1}
dataset['diet'] = dataset['diet'].map(diet_mapping)

physical_activity_mapping = {"no": 0, "yes": 1}
dataset['physical_activity'] =
dataset['physical_activity'].map(physical_activity_mapping)

stress_mapping = {"normal": 0, "high": 1}
dataset['stress'] = dataset['stress'].map(stress_mapping)

chest_pain_type_mapping = {"typical": 0, "atypical": 1, "asymptomatic": 2, "non": 3}
dataset['chest_pain_type'] = dataset['chest_pain_type'].map(chest_pain_type_mapping)

previous_chest_pain_mapping = {"no": 0, "yes": 1}
dataset['previous_chest_pain'] =
dataset['previous_chest_pain'].map(previous_chest_pain_mapping)

edema_mapping = {"no": 0, "yes": 1}
dataset['edema'] = dataset['edema'].map(edema_mapping)

diabetes_mapping = {"no": 0, "yes": 1}
dataset['diabetes'] = dataset['diabetes'].map(diabetes_mapping)

troponin_mapping = {"negative": 0, "positive": 1}
dataset['troponin'] = dataset['troponin'].map(troponin_mapping)

ecg_mapping = {"undefined": 0, "normal": 1, "abnormal": 2}
dataset['ecg'] = dataset['ecg'].map(ecg_mapping)

# filling missing value using median
dataset["bp_systolic"].fillna(dataset.groupby("sex")["bp_systolic"].transform("median
"), inplace=True)
dataset["bp_diastolic"].fillna(dataset.groupby("sex")["bp_diastolic"].transform("media
n"), inplace=True)
dataset["heart_rate"].fillna(dataset.groupby("sex")["heart_rate"].transform("median"),
inplace=True)
dataset.troponin.fillna(dataset.troponin.mode()[0], inplace=True)
```

```
# spliting dataset to train and test set
predictors = dataset.drop("target",axis=1)
target = dataset["target"]

X_train,X_test,y_train,y_test =
train_test_split(predictors,target,test_size=0.33,random_state=0)

lr = LogisticRegression()

# training the model with training set
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)

accuracy_lr = round(accuracy_score(y_test, y_pred_lr)*100, 2)
precision_lr = round(precision_score(y_test, y_pred_lr)*100, 2)
recall_lr = round(recall_score(y_test, y_pred_lr)*100, 2)
f1_score_lr = round(f1_score(y_test, y_pred_lr)*100, 2)
roc_auc_lr = round(roc_auc_score(y_test, y_pred_lr)*100, 2)

print("Accuracy score of Logistic Regression: "+str(accuracy_lr)+" %")
print("Precision score of Logistic Regression: "+str(precision_lr)+" %")
print("Recall score of Logistic Regression: "+str(recall_lr)+" %")
print("F1 score of Logistic Regression: "+str(f1_score_lr)+" %")
print("ROC AUC score of Logistic Regression: "+str(f1_score_lr)+" %")

#printing classification report
print(classification_report(y_test,y_pred_lr))
```