A Thesis On

Efficient Data Mining Techniques for Heart Disease Prediction and Comparative Analysis of Classification Algorithms

Submitted By

Masudur Rahman
Roll: ASH1711M.Sc124M
Department of Information & Communication Engineering
Noakhali Science & Technology University

Thesis Supervisor

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

Thesis Co-Supervisor

Zayed- Us- Salehin
Assistant Professor
Department of Information and Communication Engineering
Noakhali Science and Technology University,
Noakhali- 3814, Bangladesh.

December 2020

SUPERVISOR'S DECLARATION

I hereby declare that Masudur Rahman, ID: ASH1711MSc124M, is a Masters student and I am his supervisor. 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 Master of Science in Information and Communication Engineering.

_____

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

CO-SUPERVISOR'S DECLARATION

I hereby declare that, Masudur Rahman ,ID:ASH1711MSc124M, is a Masters students and I am his co-supervisor. 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 Master of Science in Information and Communication Engineering.

_____

Zayed- Us- Salehin
Assistant Professor
Department of Information and Communication Engineering
Noakhali Science and Technology University,
Noakhali- 3814, Bangladesh.

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged. The thesis has not been accepted for any degree and is not concurrently submitted for award of other degree.

_____

Masudur Rahman

Department of ICE, NSTU

Roll No: ASH1711MSc124M

ACKNOWLEDGEMENTS

I am grateful and would like to express my sincere gratitude to my Professor Dr. Md. Ashikur Rahman Khan for his ideas, invaluable guidance, continuous encouragement and constant support in making this research possible. He has always impressed me with his outstanding professional conduct. I appreciate his consistent support from the first day I applied to this program to these concluding moments. I am truly grateful for his progressive vision about my training in science, his tolerance of my naïve mistakes, and his commitment to my future career. I also would like to express very special thanks for his suggestions and for the time spent proofreading and correcting my many mistakes. I also sincerely thanks my co-supervisor Zayed- Us- Salehin for his contributions and his inspirations.

My sincere thanks to all my lab mates and members of the staff of the Information and Communication Engineering Department, NSTU, who helped me in many ways and made my stay at NSTU pleasant and unforgettable. Many special thanks for the hospital members who help me for collecting data.

I acknowledge my sincere indebtedness and gratitude to my parents for their love, dream and sacrifice throughout my life. Special thanks to other members who helped me to complete this research. I would like to acknowledge their comments and suggestions, which was crucial for the successful completion of this study

ABSTRACT

Data mining techniques are used to extract interesting patterns and discover meaningful knowledge from huge amount of data. There has been increasing in usage of data mining techniques on medical data for determining useful trends and patterns that are used in analysis and decision making. About eighty percent of human deaths occurred in low and middle-income countries due to heart diseases. The healthcare industry generates large amount of heart disease data which are not organized. These data make the prediction process more complicated and voluminous. Data mining provides the techniques for fast and accurate transformation of data into useful information for heart diseases prediction. The main objectives of this research is to predict heart diseases more accurately using Naïve Bayes, Decision Tree, Neural Network, Random Forest classification algorithms and compare the performance of classifiers. The research uses raw dataset for performance analysis and the analysis is based on Jupyter. This research also shows better classification technique from them which is Random Forest on the basis of accuracy.

Keywords: KDD, Jupyter, Naïve Bayes, Decision Tree, Multilayer Perceptron, CSV.

.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Data mining is becoming one of the most dynamic and popular technique in different sectors of our daily life. Data mining techniques means extracting or mining knowledge from huge amount of data. It is also a process of discovering interesting patterns or information and knowledge from huge collection of data. Data mining is a mixture of algorithmic techniques for extraction of informative patterns from raw data. Datamining technique used in healthcare for improving care and reducing costs. On the other hand, Data mining has successfully been used in different fields of life including marketing, banking, businesses etc.

An algorithm in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyses the provided data, looking for specific types of patterns or trends. The algorithm uses the results of this analysis over many iterations to find the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics. Choosing the best algorithm to use for a specific analytical task can be a challenge.
Data mining used different techniques for prediction of diseases. It includes classification, clustering, association etc. Classification based on categorical (i.e. discrete, unordered) data. This technique based on the supervised learning (i.e. desired output for a given input is known) [1]. It divides data samples into target classes. The classification technique predicts the target class for each data points. In order to process the prediction, it mainly analyses the input. Training data set consists all the information regarding patient which were recorded previously, whether a patient had a Heart disease or not is the prediction attribute there [2]. Classification aims to categorize unseen input data records into known classes. The goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. Clustering partitioned the data points based on the similarity measure [3]. Association analysis discovers relationships between records within the same data set. Association also has great impact in the health care industry to discover the relationships between diseases, state of human health and the symptoms of disease.

The number of diseases in the world increasing day by day as a result the amount of medical data in the health sector also increasing and it becomes one of the challenges in medical science. At present, many people have been died due to heart diseases. The World Health Organization (WHO) has calculated that about 12 million deaths occur worldwide, every year due to the Heart diseases. About eighty percent of deaths in world are because of Heart disease. World health organization has estimated that about 17.5 million people in the world died from cardio vascular diseases in 2012, almost thirty one percent of all global deaths. Out of these reports, an estimated prepared that almost 7.4 million were due to coronary heart disease and 6.7 million were due to stroke [4]. According to the WHO country profile for 2018, cardiovascular disease alone kills 2.56 lakh people in Bangladesh accounting for 30 per cent of deaths caused by Non-Communicable Diseases. Almost 23.6 million people will die due to Heart disease by 2030, as written in the report of WHO. However, ninety percent of those deaths were estimated to be preventable if patients have accurately been diagnosed early and they improved their characteristics such as: healthy eating, exercise, and alike [4].

Today's Healthcare industry produces large amount of multifarious data about hospitals, resources, disease diagnosis, electronic patient records etc. Many hospitals keep their present data in an electronic form by using their hospital database management system. These systems generate huge amount of data on daily basis. This data may be in form of free text, structured as in databases or in form of images [5]. The large amount of data is determined for processing and anatomized for knowledge extraction that allows support for understanding the prevailing circumstances in healthcare industry. Large number of workers are involved in the diagnosis of heart disease, researchers have been examining the use of data mining techniques for helping the professionals. Data mining has recently become one of the most advanced and encouraging fields for the extraction and manipulation of data to produce useful information. It uses a series of pattern recognition technologies and statistical and mathematical techniques to invent the feasible rules or relationships which control the data in the medical database system. These techniques include framing a suspicion, data collection, pre-processing, designing the model, and understanding the model and finally draw the conclusions [6].

Now a days, the researcher all over the world use Knowledge Discovery Diagram process for finding meaningful patterns from large datasets. These patterns can be further analyzed and the result can be used for further valuable decision making and analysis. It has helped to confirm the best prediction technique in terms of its accuracy and error rate on the specific dataset.

Knowledge Discovery (KD) uses data mining generally consists of some phases: Data Selection, Data preprocessing, Data transformation, Data mining, Pattern evaluation, Knowledge presentation. So, data mining process is to extract information from a data set and transform it into an understandable structure for further use.

this paper is to the Different classification techniques of data mining used for prediction of heart disease by using python. Life is dependent on efficient working of heart because heart is essential part of our body. Here different risk factors which increases the risk of heart disease were analyzed. In this paper, various classification algorithms are applied on raw data for experiment and compared on the basis of their performance.

1.2 PROBLEM STATEMENTS

Today, Healthcare organizations produces huge amounts of multifarious data about hospitals, resources, disease diagnosis, electronic patient records, etc. This increase in data volume automatically requires the data to be retrieved when needed. The large amount of data is critical to be processed and analyzed for knowledge extraction that empowers support for understanding the prevailing circumstances in healthcare industry. With the use of data mining techniques is possible to extract the knowledge and determine interesting and useful patterns. The knowledge gained in this way can be used in the proper in order to improve work efficiency and enhance the quality of decision making.

A major problem in health science or bioinformatics exploration is in managing the correct diagnosis of certain important information. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications. Detection of the disease in the early stage become difficult for large dataset. The accuracy of heart diseases prediction become lower with higher runtime.

Most of the algorithms (like ID3) require that the target attributes have only discrete values because decision trees use the divide and conquer method. If there are more complex interactions among attributes exist then performance of decision trees is low. Their performance is better only when there exist a few highly relevant attributes. On the other hand, training or learning process in large dataset is very slow and computationally very expensive.

Data sharing is another major problem neither patients nor healthcare organizations are interested in sharing of their private data.

The researchers all over the world are working on this issue. They are trying to develop better algorithm comparatively lower execution time and better accuracy for easily detect this disease than the traditional techniques. We use different classification methods including Decision tree, Bayesian algorithms (Naive Bayes), Neural Network (Multilayer Perceptron) and others classification methods to predict heart diseases easily with lower execution time and higher accuracy. Hence it will provide revolutionary changes in medical sector.

1.3 RESEARCH OBJECTIVES

This research paper works with classifier algorithms and provides better result that will directly help to the physicians for accurate diagnosis. The main purposes of our research are as follows:

i. Applying different classification algorithms for better accuracy
ii. Reducing the complexity of disease prediction
iii. Comparative analysis of classifier algorithms and find out the best one.

1.4 THESIS OUTLINE

The thesis comprises of five chapters including: introduction, literature review, methodology, experimental result discussion, conclusion, list of reference.

Chapter 1 - Introduction: The introductory chapter discuses about the general structure and framework of the thesis. Shortly explains of purpose of thesis, thesis topic, problem statements.

Chapter 2 - Literature Review: This chapter review, analyze and summarize of all article those are used in this thesis. Describes previous works regarding thesis paper.

Chapter 3 – Methodology: In this chapter, show that the represent of how to the data collection, data analysis, model and formula to use this paper. This describes the proposed model of this research. All chapters are shown in Figure.1.1.

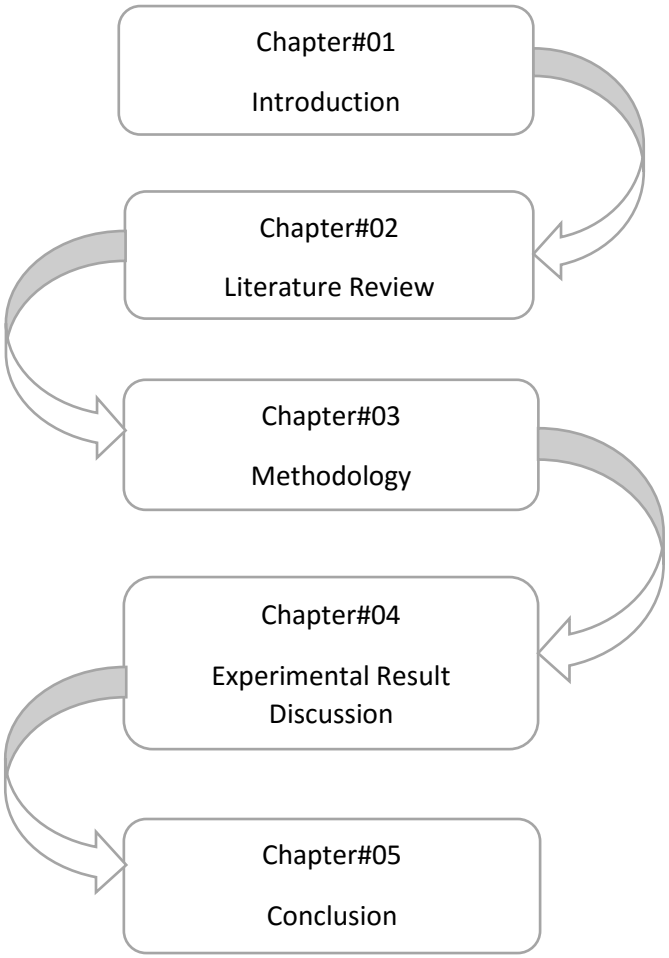


Figure 1.1: Diagram of thesis outline

Chapter 4- Experimental Result Discussion: In this chapter all of the data are process by using classifiers, experimental analysis of prediction results, graphical analysis, statistical analysis, comparison of classifiers results, outcome of the analysis, and declaration of the best one between those classifiers.

Chapter 5- Conclusion: In this chapter conclusion shows that the best performance among all classifiers and some suggestion about the future works and limitations of this thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

To early diagnosis of heart disease accurate prediction is highly required. Many researchers used different techniques for this purpose. A heart disease predictor should meet the following specification; efficient modelling, applicability, accuracy and be trusted. It should be compatible with various diagnostic techniques Here are the discussion for some of the related works.

2.2 RELATED WORKS

A lot of work has been done related to heart diseases using data mining techniques. Many experiments have been carried out for diagnosing heart diseases. In this context, researchers have been applied different data mining techniques such as: Association Rules Technique, Clustering, and Classification Algorithms to extract significant patterns for prediction of heart diseases with good accuracy.

M. Anbarasi et al. Proposed a Genetic Algorithm to improve Prediction of Heart Disease. This paper used Weka data mining tool with 435 instances for experiments. The Genetic Algorithm is an optimization technique inspired by natural selection and natural genetics. This paper also used classifiers Decision tree Classification with clustering and Naive Bayes were used for diagnosis of patients with heart disease. A Heart disease system was developed to predict accurately the presence of heart disease with reduced number of attributes [19].

AH Chen et al. introduced HDPS: Heart Disease Prediction System in which they used for classifying the heart disease based on 13 different attributes. Only one data mining algorithm is used that is artificial neural network (ANN) The data set is used in this system is taken from UCI machine learning repository having 303 instances [10].

Chaitrali S. Dangare et al. used data mining neural network approach for prediction about heart disease [11]. This paper used total 15 attributes to increase the accuracy of the prediction. WEKA data mining tool used for the experiments and the data set for this contains 573 records

from UCI repository which are divided into two parts training and testing. They have taken Multilayer Perceptron Neural Network (MLPNN) with Back propagation algorithm (BP) for developing their system. This paper shows batter accuracy about 100% accuracy using neural network approach for prediction about heart disease. [11].

Jyoti Soni et al. developed a data mining technique for heart disease prediction system using Weighted Associative Classifiers (WAC), where different weights are assigned to the attributes according to their capability about predicting. They implemented a system by using JAVA platform and benchmark data from online available UCI repository. The data set contain 303 records and 14 different attributes. They got the results of experiments about 80% accuracy of using a new techniques WAC [12].

Vikas Chaurasia et al. used different data mining approaches to predict heart diseases. The data set was contained only 11 attributes for experiments. This paper used different data mining approaches to predict heart diseases. The results of these experiments were compared with each other using 10 cross validations with and without bagging. Bagging means Bootstrap aggregation which is used to increase the accuracy of the classification. Three techniques were used and compare in this research i.e., Naïve Bayes, J48 Decision Tree and Bagging. The experiments also show that the bagging has the highest accuracy i.e., 85.03 % and J48 decision tree and naïve Bayes have 84.35% and 82.31% accuracy respectively [13].

Shadab et al. used Naive Bayes technique in the year 2012 for the heart diagnosis in heart prediction system. They used 12 attributes for predection purpose and performed analysis used python[14].

Rashedur et al. in the year 2013 used Neural network technique and to compare various classification techniques, he used another technique fuzzy logic using TANGRA data mining tool. This paper worked with 330 instances. They achieved 79.19% accuracy using neural network, he used another technique fuzzy and achieved 83.85% accuracy [15].

My Chau Tu's used bagging algorithm to identify the heart disease [16]. This paper worked with 350 instances.

Aqueel Ahmed et al. found that decision tree and SVM perform classification more accurately than the other methods and was able to achieve 91% accuracy [17].

Moreno et al. applied an association algorithm successfully for prediction in health domain, they discovered huge number of rules (some of them are not interested), in addition generally the performance of association algorithms is low [18].

Chitrali S. et al. used 15 attributes for heart disease prediction system. The research work included two extra attributes obesity and smoking for efficient diagnosis of heart disease in developing effective heart disease prediction system. They showed that Artificial Neural Network outperforms other data mining techniques such as Decision Tree and Naïve Bayes [19].

Kuma ravel et al. have proposed automatic diagnosis system for heart diseases using 38 input parameters. They found better accuracy in neural network with accuracy of 63.6 - 82.9% [20].

B. Venkata Lakshmi et al. created an analysis on heart disease diagnosis using data mining techniques with data set of 13 attributes was used. They performed an analysis on heart disease diagnosis. Data set of contained 294 records. They performed an analysis on heart disease diagnosis using Naïve Bayes and Decision Tree techniques. Different sessions of experiments were conducted with the same datasets [21].

Aditya Methaila et.al. In their research work made combinations of several target attributes for effective heart attack prediction using data mining. This paper also used 15 attributes. They found that Decision Tree has outperformed with 99.62% accuracy. Also, the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction [22].

Mamta Sharma et al. in the year 2017 used different classifiers for heart disease prediction using datamining techniques. They used 15 attributes and Weka for prediction purpose. They used Decision tree, Neural network and Naïve Bayes. They show that among all the classifiers such as Decision tree, Neural network and Naïve Bayes, Naïve Bayes algorithm gives a better average prediction with 90% accuracy [23].

Ahmed Zriqat, Ahmad Mousa Altamimi et al. performed a Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods like decision tree and Random forest. Data set of 303 records with 14 attributes were used. This research shows the results revealed that decision tree outperforms other classifiers with an accuracy rate of 99.0% followed by Random forest [24].

Saravana Kumar et al.in March 2014, Frequent Feature Selection method was used for predicting about the heart diseases fuzzy measure and the relevant nonlinear integral are use with this method to give good performance. There are different stages of mining using feature subset selection method, first is Weighted Page 13 Support in which each record contains relative weight of all transaction and the second is Maximal Frequent Itemset Algorithm (MAFIA) which combines old and new algorithmic ideas which mine the frequent item set and at the end significance weight age of each pattern is calculated [25].

Sultana highlighted, performance measures for different classification models are as follows on the basis of performance accuracy Bayes Net (81.11%), K star (75.182%) and SMO (84.074%). And from their study it is showed that the performance accuracy of support vector machine is better than other algorithm which they used in their paper [26].

H. Benjamin Fredrick David et al. performed an analysis on heart disease prediction using data mining techniques. They used three data mining classification algorithms like Random Forest, Decision Tree and Naïve Bayes are addressed and used to develop a prediction system in order to analyses and predict the possibility of heart disease. The experimental setup has been made for the evaluation of the performance of algorithms. They used Weka tool for research purpose where the dataset retrieved from UCI machine learning repository. They found that Random Forest algorithm performs best with 81% precision when compared to Decision Tree and Naïve Bayes for heart disease prediction [27].

According to the previous study of literature review, it has seen that all the researchers used different classifiers, various kinds of attributes of heart diseases from different sources in order to increase the accuracy. Most of them used online data for prediction purpose. In our research raw data are collected from hospitals and accuracy of algorithms tried to increase according to the heart disease situation of Bangladesh.

2.3 TABLE OF PREVIOUS WORK

In this section the previous works related to this research is summarized by a table 2.1

Table 2.1 Table of related work

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
1	Disease Forecasting System Using Data Mining Methods	Nishara Banu	2017	Sex, Age, Ca, Cp, Fbs, ECG, Bp, slope	UCI repository	K-means algorithm	92%
2	Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection	M.Akhil Jabbar	2013	Oldpeak, Slope, Thal, ECG, Ca	UCI repository	Decision trees	88%
3	Diagnosis of heart disease using genetic algorithm based trained	Sagir, Sathasivam	2017	Chol, Trestbps, Fbs,	Kaggle	Adaptive Neuro Fuzzy model	88%
4	Prediction of heart disease using multilayer perceptron neural network	Jayshril S. Sonawane and Patil	2014	Oldpeak, Slope, Thal, ECG, Ca	Kaggle	Multilayer perceptron	98%
5	Prediction of Heart Disease Using Machine Learning	Aditi Gavhane	2018	Chol, Trestbps, Fbs	Kaggle	Support Vector Machine	82%
6	Analysis of heart disease prediction using neural network	Tulay karawulanet		Age, Sex,Chol	UCI repository	Backpropagation algorithm	95%.

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
7	Heart disease diagnosis using data mining technique	Sarath Babu	2017	Oldpeak, Slope, Thal, ECG, Ca	UCI repository dataset	K-means algorithm, MAFIA algorithm and Decision tree	81%
8	Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction	Jyoti Soni	2017	Oldpeak, Slope, Thal, ECG, Ca	Kaggle	Naive Bayes, Decision List and KNN	88%
9	Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules	P.K Anooj	2019	Trestbps, Fbs, Thalach, Ca	Kaggle	Weighted fuzzy rule-based system	85%
10	An Analysis of Heart Disease Prediction using Different Data Mining Techniques	Nidhi Bhatla	2012	Age, Sex, Ca,Fbs	UCI repository dataset	Descision tree	87%
11	Early Heart Disease Prediction Using Data Mining Techniques	Aditya Methaila	2014	Sex, Age, Ca, Bp	UCI repository dataset	Decision Trees, Naive Bayes	91%
12	Heart disease diagnosis using data mining technique	Shimpy goyal	2019	Age, Sex, Cp, Ca, ECG	Kaggle	K-means and apriori algorithm	88%
13	Heart disease analysis using genetic algorithm	M. Akhil jabbar	2018	Age, Sex, Cp, Ca, ECG	Kaggle	Genetic algorithm	

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
14	Heart disease detection using neural network	R Sethukkarasi	2018	Age, Sex, Ca,Fbs	Kaggle	Neuro fuzzy techniques	
15	Association Rule Mining based on a Modified Apriori Algorithm in Heart Disease Prediction	Mohammed Abdul Khaleel	2018	Age, Sex, Ca,Fbs	UCI repository	Apriori data mining technique	
16	A Literature Survey on Applications of Data Mining Techniques to Predict Heart Diseases	Boshra Bahrami	2015	Age, Sex, Ca,Fbs, ECG	UCI repository	Decision Tree, KNN, and Naive Bayes	
17	A Literature Review on Heart Disease Prediction Based on Data Mining Algorithms	Deepika N	2018	Age, Sex, Ca,Fbs	UCI repository	Pruning Classification Association Rule (PCAR)	
18	Review of heart disease analysis using neural network	Sonam Nikhar and A.M. Karandikar	2019	Age, Sex, Bp, Bs,ECG, Ca	UCI repository	Naive Bayes and decision tree classifier	
20	Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm	Zeinab Arabasadi	2019	Sex, Cp, Chol, Trestbps, Fbs	Z-Alizadeh Sani dataset	Genetic algorithm	92%

Sl No.	Title	Author	Year	Considered Parameters	Source of Data	Implemented Methods/ Techniques	Result
21	Neural Network Based Intelligent System for Predicting Heart Disease	V. B. K. Subhadra	2019	Sex, Age, Ca, Cp, Fbs, Bp	Kaggle	Multilayer perceptron	95%
22	Heart Disease Classification Using Neural Network and Feature Selection	V. B. Anchana Khemphila	2011	Age, Sex, Cp, Ca, Fbs	Kaggle	Genetic algorithm	85%

From the above discussion, it is shown that authors mainly used the dataset from the UCI repository. They had found a good accuracy though but this research uses the real dataset collected from different hospitals.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

In this chapter, the heart disease dataset that had been studied was discussed in detail. We elaborated briefly how the current data had been collected and then moved on to how the datasets were built and formatted to train the classifiers. The current research intends to improve the diagnosing of heart diseases by examining the patient's symptoms using data mining classification methods. To achieve this goal, a literature review was carried out to review the data mining works related to diagnose heart diseases. For this purpose, we need follow the following process carefully.

3.2 ATTRIBUTES SELECTION AND DESCRIPTION

There are many attributes used for heart disease diagnosis. However, most of the researchers commonly refer to 13 of them. This research used different types of input attributes for the heart disease are age, sex, chest pain type, blood pressure, cholesterol, blood sugar ECG result, old peak, slop etc. Here is given 13 attributes where some of attributes are nominal and others attributes are numeric and one attribute is class variable. The attributes of the dataset and their description is given in Table 3.1. The dataset is typically used for various types of heart diseases such as typical angina, atypical angina, and non-anginal pain and asymptomatic. This research work is aimed at predicting the heart disease irrelevant of the disease types. The attribute is a numeric data type that represents the age of the patient and ranges from 29 to 65 years. The Cp is an attribute for determining the pain type, represented from the range1 to 4. The trestbpd is a resting blood pressure that lies between 92 and 100; the fbs is fasting blood sugar level that is either a 1 or 0 representing Boolean values true or false. The restecg is the resting electro cardio graphic result represented as three cases from 0 to 2. The thalach is the maximum heart rate achieved ranging from 82 to 185. The exang is the exercise induced angina that is a Boolean value. The disease is the target class of the dataset denoting the heart disease presence with a yes or a no.

Table 3.1: Attributes and their descriptions

Attribute Name	Description	Value
Age	patient's age in years	
Gender	sex of patient	Male,Female
Chest pain	chest pain	Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
Blood_Pressure	resting blood pressure (in mm Hg)	
Chol	serum cholesterol in mg/dl	
Fasting_Blood_sugar	fasting blood sugar > 120 mg/dl	1 = true 0 = false
Resting_ECG_results	resting electrocardiographic results	0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: Ventricular hypertrophy by Estes' criteria
Maxi_heart_rate	maximum heart rate achieved	
Exercise	Exercise include angina	1. yes & 2.no
Old peak	ST depression induced by exercise relative to rest	
Slope	the slope of the peak exercise ST segment	1: upsloping 2: flat 3: down sloping
Number of major vessels	number of major vessels colored by fluoroscopy	attributes values can be 0 to 3

Defect type	type of defect	3 = normal 6 = fixed defect 7 = reversible defect
Disease	diagnosis of heart disease	1=Present 0=Absent

3.3 DATA COLLECTION

Data are collected from different hospitals of heart patients for the research purpose. Raw data was collected from heart disease patients for making heart disease predictions. Data collected from National Heart Foundation of Bangladesh, Ad-Din medical college and hospital, Bangabandhu Sheikh Mujib Medical University and other patients. These datasets are integrated into a single dataset. It includes 13 attributes and one class value. The heart disease dataset included in this research work consists of total 305 instances with no missing values. In which 155 are heart patients records and others are non-heart patients' records. All attributes values are numeric. If a person is affected by heart disease, then he in Class Present otherwise in Absent.

3.4 DATA PRE-PROCESSING

In real world, data are generally incomplete, noisy and inconsistent. For predicting data mining, data in raw form are not best for analysis. Missing values are extremely common in medical records. However, these values must be treated before being used as they may lead to failure classification or incorrect disease prediction [28]. Knowledge Discovery (KDD) is a process that allows automatic scanning of high-volume data in order to find useful patterns that can be considered knowledge about the data. Data preprocessing is a first step of the Knowledge discovery in databases (KDD) process. The data must be preprocessed and transformed to get the best mineable form. There are number of different tools and methods available for data preprocessing. The tasks in the data preprocessing are: Data Cleaning, Data Integration, Data Transformation, Data Reduction, and Data evaluation. This paper used Jupyter notebook tool for that purpose. We used python programming language for data preprocessing and prediction. To do so, we created CSV (Comma Separated Value) file format of the dataset by using Microsoft Excel Program. Dataset is loaded into the jupyter notebook. Below a small amount of data are shown.

	AGE	SEX	CHEST PAIN	BLOOD PRESSURE	CHOLESTORAL	BLOOD SUGAR	ECG	HEART RATE	ANGINA	OLDPEAK	SLOPE	VESSELS	THALASSEMIA	CLASS
0	69	1	4	130	328	1	2	105	1	2.2	3	3	6	1
1	66	1	3	120	445	0	1	156	0	1.4	2	0	7	0
2	58	0	3	125	260	0	0	145	0	0.4	1	2	6	1
3	65	1	4	130	263	0	1	105	0	0.2	3	1	7	0
4	73	0	2	120	270	1	2	124	1	0.2	1	0	3	0
5	65	1	3	115	177	0	0	142	0	0.6	1	0	7	0
6	55	0	4	126	253	1	1	148	0	0.6	2	1	6	1
7	57	1	3	110	239	0	2	145	1	1.4	2	0	7	1
8	61	1	4	138	290	0	2	170	0	1.3	2	2	6	1
9	62	1	4	145	407	0	0	155	0	3.0	3	3	7	1

Figure 3.1: Dataset with attributes value

3.5 DATA ANALYSIS AND VISUALIZATION

This step will represent “PRESENT” by value “1” and “ABSENT” by “0” of heart disease among the instances. This will show that among all instances, some instances are classified as present of heart disease and others instances are classified as absent of heart disease. Here, this paper visualized all attributes using Jupyter notebook web-based software. We checked if there is there any null values of the attributes. There are no null value attributes. These figures show the distribution of the attribute values with respect to number of patients. In this picture x-axis representing the value of each attributes and y-axis representing the number of instances.

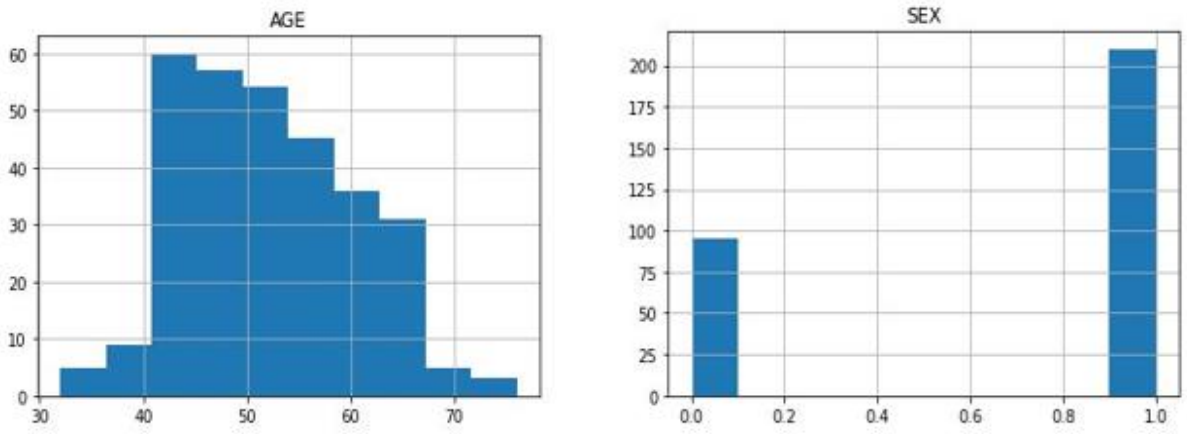


Figure 3.2: Attribute value visualizations with respect to number of instances

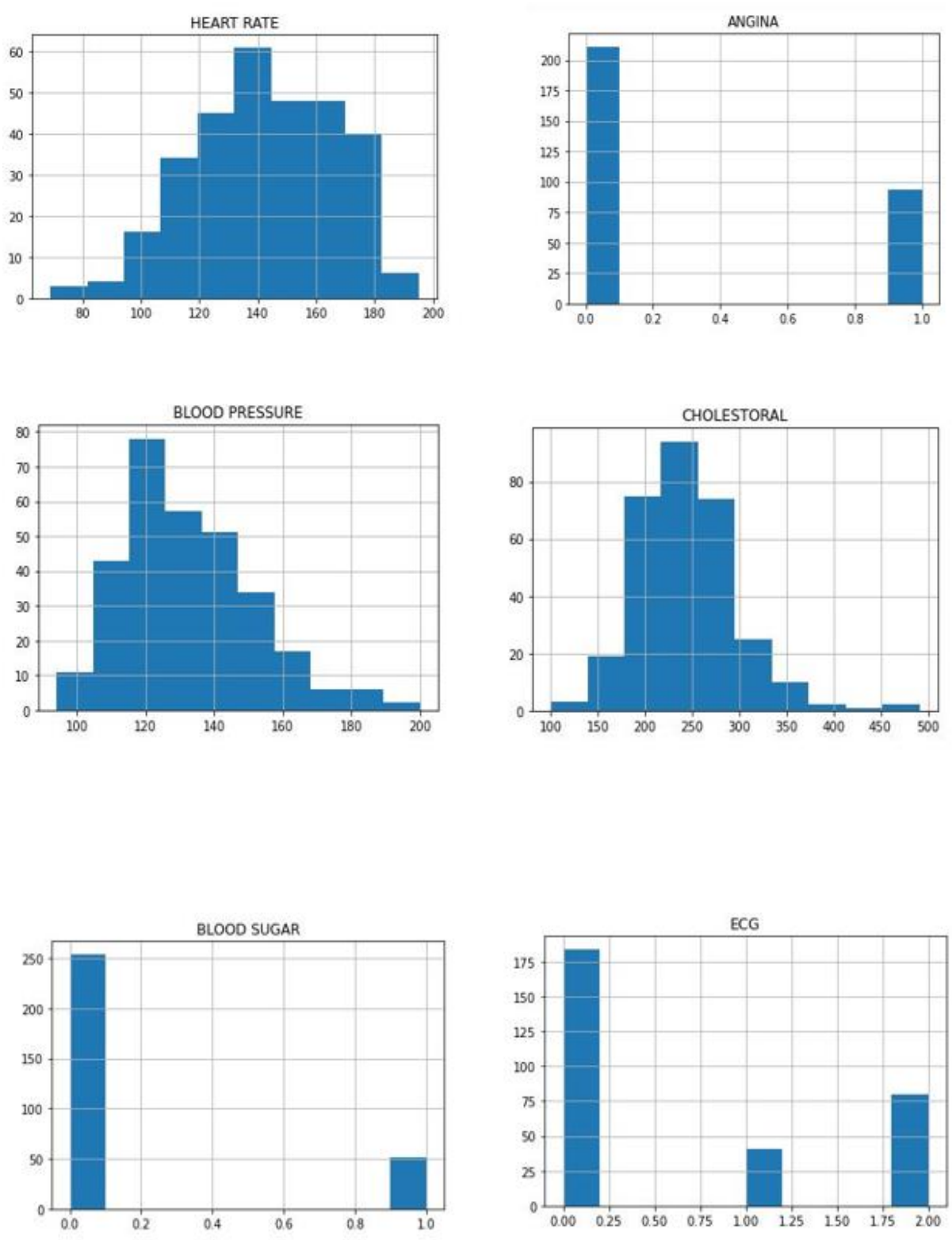


Figure 3.3: Attribute value visualizations with respect to number of instances

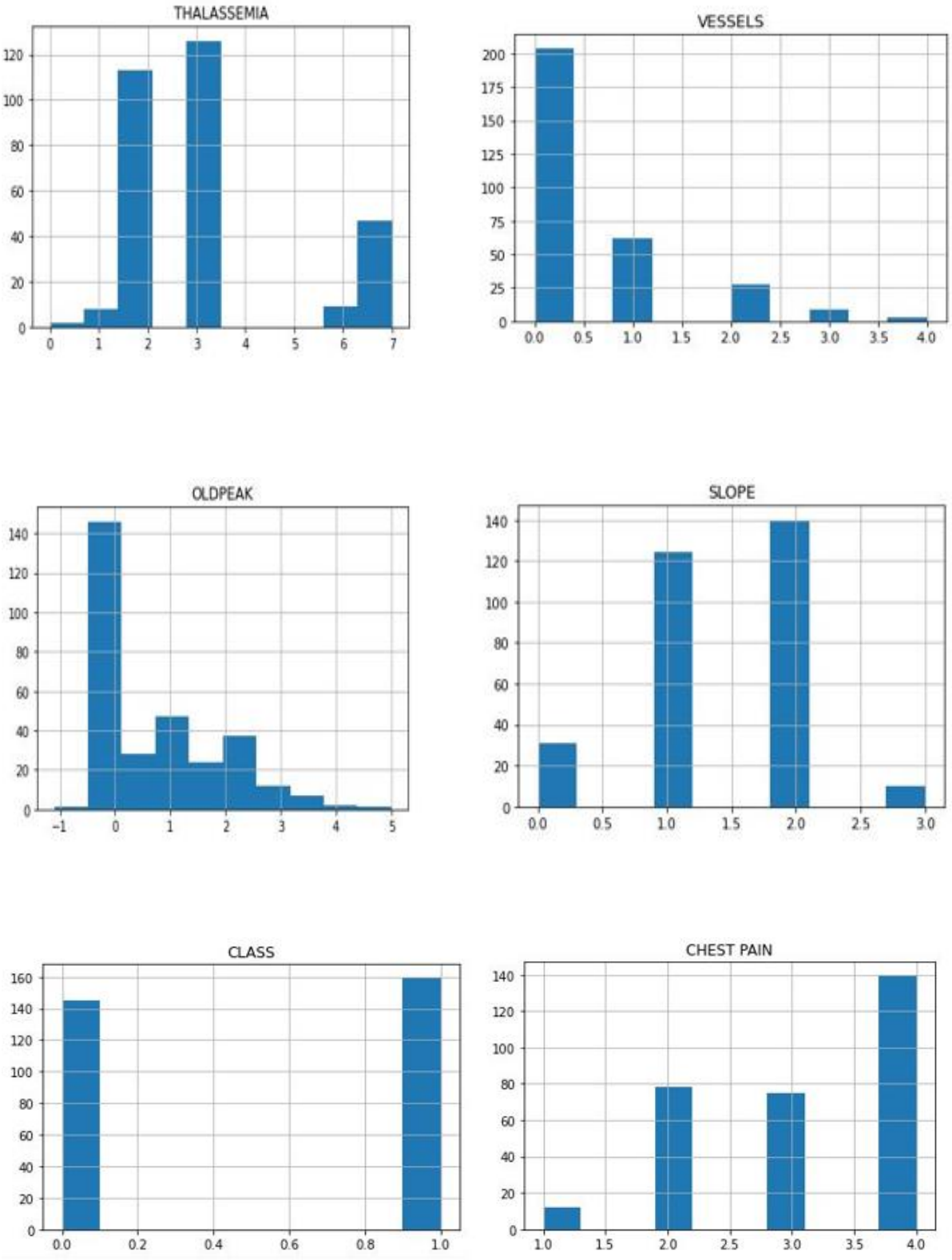


Figure 3.4: Attribute value visualizations with respect to number of instances

Here, we can see that among all 305 instances, 155 instances are classified as heart patient and others 150 instances are classified as absent of heart disease. In these above figures, chest pain

categorized into four categories between 1 to 4 and blood pressure between 94 to 192 and it seen that around 140.The higher number of patients have a positive result of having the disease. In this graph blood sugar and electrocardiograph mainly categorized between two values. Heart rates are categorized between 71 to 202 and among them between 136 to 180 have a higher possibility to be a heart patient.

3.6 CLASSIFICATION ALGORITHMS

Different machine learning classification algorithms and their implementations needed for predicting heart disease. Classification is the process of building a model of classes from a set of records that contain class labels. In this paper we use the following data mining techniques.

3.6.1 Naive Bayes

Naive Bayes Algorithm is used to generate mining models. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification method. The algorithm calculates the probability of every state of each input column given predictable columns possible states. This algorithm is based on statistics. It is used for estimating the probability of a class value during classification and prediction. The common principle of Naive Bayes algorithm is all Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature. Continuous data set is used rather than the categorical data to predict the heart disease easily. As the classifier returns probabilities, it is very easy to apply these results to a large number of tasks than if an arbitrary scale was used [29]. Naïve Bayes theorem provides a way of calculating posterior probability P(c|x) from Class prior probability P(c), predictor prior probability P(x) and Like hood P(x|c). The equation is in below:

$$P(c|x)=P(x|c)P(c)/P(x) \quad \text{..... i}$$

$$P(c|X)=P(x_1|c)\times P(x_2|c) \times \text{.....} \times P(x_n|c) \times P(c) \quad \text{.....ii}$$

3.6.2 Decision Tree

Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting health care professionals in the diagnosis of the world's biggest killer demands higher

accuracy. research seeks to improve diagnosis accuracy to improve health outcomes.[30] In medical field decision trees determine the sequence of attributes. First, it produces a set of solved cases. Then the whole set is divided into training set and testing set. Where a training set is used for the induction of a decision tree. While the testing set is used to find the accuracy of an obtained solution [31].

Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also, on the bases of the training instances the classes for the newly generated instances are being found. Decision Tree is a flowchart like tree structure which contains leaf, nodes and branches [32]. Decision Tree has become popular in knowledge discovery because the construction of decision tree classifier does not require any domain knowledge. Successful decision tree model depends upon the data but in general it has good accuracy. The sample decision tree is shown in Figure 3.5.

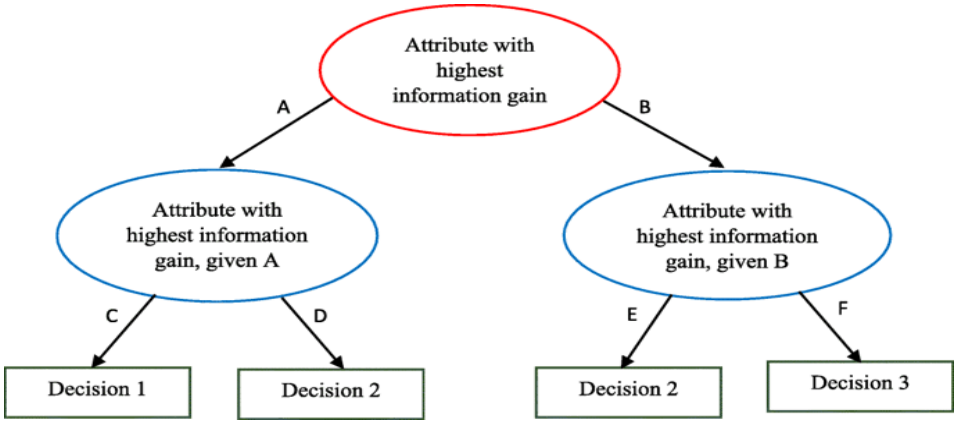


Figure 3.5: A simple decision tree [32]

3.6.2.1 J48

There are different types of decision tree algorithms.J48 algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm, the critical distribution of the data is easily understandable [29]. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. J-48 is a type of decision tree algorithm generate pruned and un-pruned decision tree for classification of data. When a decision tree is built many branches shows anomalies in data due to outliers and tree pruning address this problem in data. This method removes the least reliable branches. It follows the facts in which each attribute of

data can be used by splitting the data into subsets [33]. It uses the concept of information entropy. To make the decision the attribute with highest information gain is used and information gain is basically the difference in entropy.

Information gain is used in C4.5 to collect the information in data sets. And this information is used to take the decisions. It is the mathematical tool that algorithm J48 has used to decide, in each tree node, which variable fits better in terms of target variable prediction. Decision trees can handle both categorical and numerical data. The algorithm works by finding the information gain of the attributes and taking out the attributes for splitting the branches in threes [34]. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

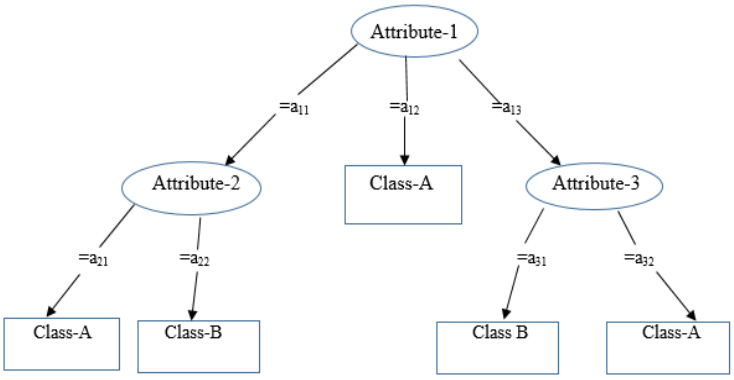


Figure 3.6: A simple structure of J48 Decision Tree [33]

Basic Steps in the Algorithm [33]:

(i) In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.

(ii) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

(iii) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

Counting Gain This process uses the “Entropy” which is a measure of the data disorder. The Entropy of is calculated by

$$\text{Entropy}(\vec{y}) = -\sum_{j=1}^{n} \frac{|y_i|}{|\vec{y}|} \log\left(\frac{|y_i|}{|\vec{y}|}\right)$$

$$\text{Entropy}(j|\vec{y}) = \frac{|y_j|}{|\vec{y}|} \log\left(\frac{|y_j|}{|\vec{y}|}\right)$$

And Gain is

$$\text{Gain}(\vec{y}, j) = \text{Entropy}(\vec{y} - \text{Entropy}(j|\vec{y}))$$

The objective is to maximize the Gain, dividing by overall entropy due to split argument by value j. Because of the outliers this is a significant step to the result. Some instances are present in all data sets which are not well defined and differ from the other instances on its neighborhood. The classification is performed on the instances of the training set and tree is formed. The pruning is performed for decreasing classification errors which are being produced by specialization in the training set. Pruning is performed for the generalization of the tree.

Features of the Algorithm [34]:

(i) Both the discrete and continuous attributes are handled by this algorithm. A threshold value is decided by C4.5 for handling continuous attributes. This value divides the data list into those who have their attribute value below the threshold and those having more than or equal to it.

(ii) This algorithm also handles the missing values in the training data.

(iii) After the tree is fully constructed, this algorithm performs the pruning of the tree. C4.5 after its construction drives back through the tree and challenges to remove branches that are not helping in reaching the leaf nodes.

3.6.2.2 Random Forest

Random forests (RF) are combination of tree predictors using decision tree such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [35]. The generalization error of a forest of tree classifiers

depends on the strength of the individual trees in the forest and the correlation between them. They are more robust with respect to noise. It is a supervised classification algorithm used for the prediction and it is considered as the superior due to its large number of trees in the forest giving improved accuracy than decision trees. Typically, the trees are trained independently

and the predictions of the trees are combined through averaging. Random forest algorithm can use both for classification and the regression based on the problem domain.

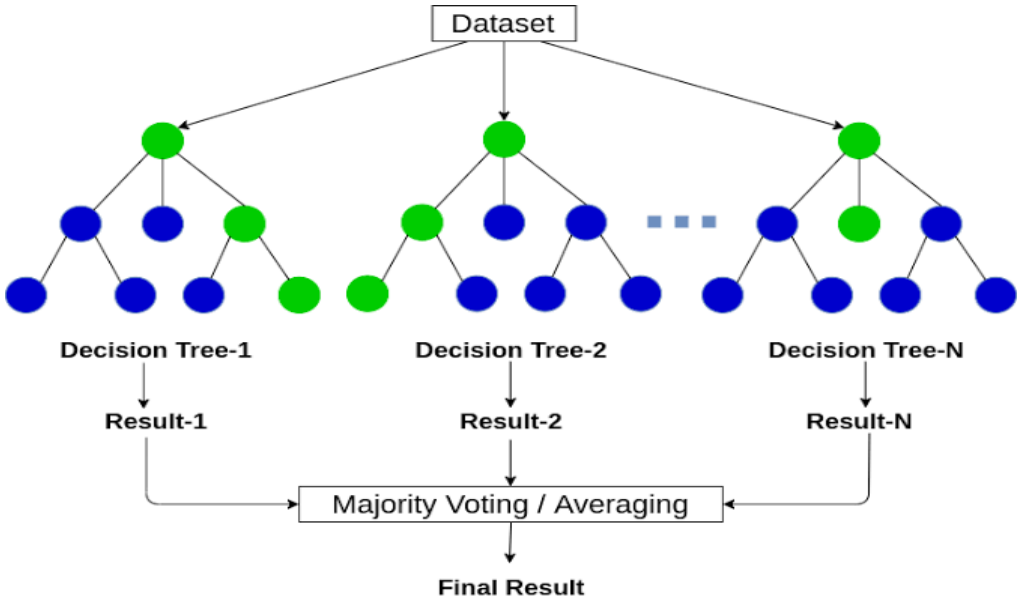


Figure 3.7: A structural view of Random Forest [35]

The random forest algorithm can split into two stages.

- Random forest creation pseudocode.
- Pseudocode to perform prediction from the created random forest classifier.

Random Forest pseudocode:

i. Randomly select “k” features from total “m” features. Where k << m
ii. Among the “k” features, calculate the node “d” using the best split point.
iii. Split the node into daughter nodes using the best split.
iv. Repeat 1 to 3 steps until “l” number of nodes has been reached.

v. Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, observe that it is randomly taking features and observations. In the next stage, using the randomly selected “k” features to find the root node by using the best split approach. The next stage, it will be calculating the child nodes using the same best split approach. Will the first 3 stages until we form the tree with a root node and having the target as the leaf node. Finally, repeat 1 to 4 stages to create “n” randomly created trees. This randomly created trees forms the random forest.

To perform prediction using the trained random forest algorithm uses the below pseudocode.

i. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
ii. Calculate the votes for each predicted target.
iii. Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm, we need to pass the test features through the rules of each randomly created trees. Suppose let’s say we formed 100 random decision trees to from the random forest. Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target votes will be calculated. Suppose the 100 random decision trees are prediction some 3 unique targets x, y, z then the votes of x are nothing but out of 100 random decision tree how many trees prediction is x. Likewise, for other 2 targets (y, z). If x is getting high votes. Let’s say out of 100 random decision tree 60 trees are predicting the target will be x. Then the final random forest returns the x as the predicted target.

3.6.3 Neural Network (Multilayer Perceptron)

Artificial Neural Networks are the human neurons type network structure which consists of number of nodes that are connected through directional links where each node represents a processing unit and the links between them specify the casual relation between them. This classification technique is becoming powerful tool in data mining and may be used for different purposes in descriptive and predictive data mining. Multilayer Perceptron classifier is based on

back propagation algorithm to classify instances of data. It is a feed forward neural network with one input and output layer with several possible hidden layers that are totally interconnected. The back propagation neural network is referred as the network of simple processing elements working together to produce an output. The multilayer feed-forward neural network should be learned by performing the back-propagation algorithm. It should be learned by a set of weights for predicting the class label of tuples. This neural network consists of three layers namely input layer, one or more hidden layers, and an output layer.

The nodes of input layer are connected to the nodes of hidden layers and the nodes of hidden layer may be connects to each other or to an output layer. To make input layer, the inputs are fed simultaneously into the units. These inputs are passed through the input layer and are then weighted and fed simultaneously to a second layer of neuron like units, which is known as a hidden layer. The sample artificial neural network is shown in Figure 3.6. A Neural Network start with an input layer where each node is corresponds to a predictor variable [36].

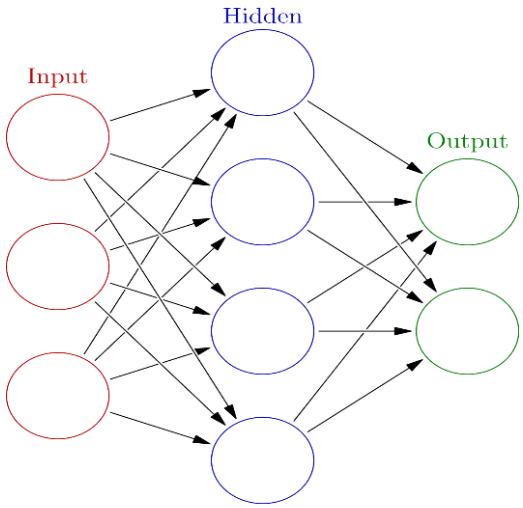


Figure 3.8: Sample Artificial Neural Network

The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used [37]. At the core, back propagation is simply an efficient and exact method for calculating all the derivatives of a single target quantity (such as pattern classification error) with respect to a large set of input quantities (such as the parameters or weights in a classification rule). To improve the classification accuracy, we should reduce the training time of neural network and reduce the number of input units of the network [38]. Throughout training the output from the

MLP, the desired output may not equal for a provided input vector. Difference between the desired and real output is defined as an error. To reduce the entire error of MLP, training uses the magnitude of error signal to adjust the degree of weight in the network.

3.7 DATA MINING TOOL

ANACONDA

Anaconda is a combination of the Python and R programming languages popularly used for scientific computing that aims to simplify package management and deployment. Package versions in Anaconda are maintained by the package management system conda. This package manager was spun out as a separate open-source package as it ended up being useful on its own and for other things than Python. Anaconda is a distribution of Python [39].

Anaconda Navigator uses desktop graphical user interface (GUI) included in Anaconda distribution. Anaconda Navigator allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, then install them to environment, run the anaconda packages and update them. It is available for Windows, macOS and Linux.

The following applications are available by default in Navigator [39]:

- Jupyter Lab
- Jupyter Notebook
- Spyder
- Glue
- Orange
- R Studio
- Visual Studio Code

We used Jupyter Notebook for our data analysis visualization and prediction purpose.

JUPYTER NOTEBOOK

Jupyter notebook is an open-source, interactive web tool used as a computational notebook, which researchers can use for the combination of software code, computational output, explanatory text and multimedia resources in a single document. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. The Jupyter has two components. Users input programming code or text in rectangular cells in a front-end web page. The browser then passes that code to a back-end 'kernel', which runs the code and returns the results. In our research Python programming language is used [40]. Jupyter ships with the Python kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use. The icon of jupyter notebook shown in figure 3.9.

Figure 3.9: User Interface of project Jupyter [39].

Jupyter actually supports four types of extensions [40]:

- Kernel
- Python kernel
- Notebook
- Notebook server

The Jupyter Notebook is quite useful not only for learning and teaching a programming language such as Python but also for sharing your data.

Jupyetr Notebook is commonly used for [39]:

(i) Data cleaning and transformation,
(ii) Numerical simulation,
(iii)Statistical modeling,
(iv)Data visualization,
(v) Machine learning and much more.

3.8 FLOW DIAGRAM OF THE MODEL

The heart disease prediction can be performed by following the procedure which is similar to Figure 3.8 which specifies the research methodology for building a classification model required for the prediction of the heart diseases in patients. The model forms a fundamental procedure for carrying out the heart disease prediction using any machine learning techniques. First, heart diseases patient records are collected that preprocess and extract some features for analyzing further manipulation of it. Then, four different classification algorithms used to measure the performance of heart disease. In figure we represent several steps how to implement our proposed model. Those steps are described briefly as follows:

At first Heart patient records are collected from different hospitals. From there, we have collected 305 heart patient records for analysis. The selected data was checked for noise, inconsistency and missing values. Noises and inconsistencies identified in the dataset were corrected manually, while missing values were replaced with the most similar value. Jupyter notebook used for this purpose.

This research performed df.describe() python statement for viewing some basic statistical details like percentile, mean, standard deviation etc. of each attributes. The correlation among the attributes are also checked using df.corr() statements. In correlation metrics, diagonal values are 1 some are less then 1 and some are negatives.

In splitting portion, we split the original sample into a training set to train the model, and test set to evaluate it. we used 80 percent of data for training the models and 20 percent used for testing purpose. This paper applied Decision tree, Bayesian algorithms (Naive Bayes), Neural Network (Multilayer Perceptron) classifiers on this splitting dataset. After performing classifiers, Performance Comparison was done on classifier's results based on their accuracy. The prediction results varied with respect to classifier algorithms.

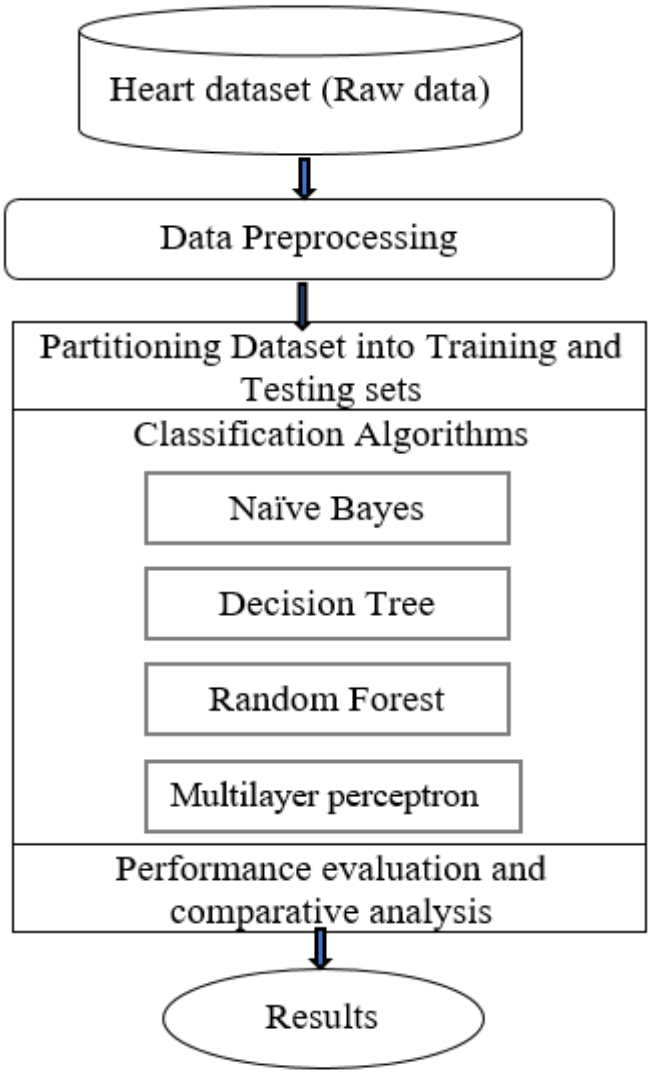


Figure 3.8: Block diagram of proposed system

3.9 PERFORMANCE ANALYSIS

Some common performance measures used to evaluate the accuracy of classification algorithms. These measures were selected because they are widely used to assess performance of classification models. For finding these measurements we need to calculate a confusion matrix. The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. A confusion matrix we can create as follows:

Table 3.2: Confusion Matrix for performance analysis

	Actual		
Predicted		Present	Absent
	Present	True Positives (TP)	False Positives (FP)
	Absent	False Negatives (FN)	True Negatives (TN)

It is used for Classification problem where the output can be of two or more types of classes (Present and Absent). The confusion matrix, is a table with two dimensions ("Actual" and "Predicted"), and sets of "classes" in both dimensions. Actual classifications are columns and Predicted ones are Rows. The Confusion matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on Confusion Matrix and the numbers inside it. Terms associated with Confusion matrix are as follows: -

True Positives (TP): True positives are the cases when the actual class of the data point was True and the predicted is also True.

True Negatives (TN): True negatives are the cases when the actual class of the data point was False and the predicted is also false.

False Positives (FP): False positives are the cases when the actual class of the data point was False and the predicted is True.

False Negatives (FN): False negatives are the cases when the actual class of the data point was True and the predicted is False.

The performance evaluation for each classifier models depend on the following terms: -

Sensitivity (Recall or True Positive Rate)

The first measure is Recall also called sensitivity as shown in Equation, which measures how well a binary classification test correctly identifies a condition probability of correctly labeling members of the target class. Recall is how many relevant items are selected. It is calculated as

the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR).

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}}$$

Accuracy (ACC)

Another important measurement is Accuracy (ACC) which measures the performance of classification. Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

Precision (Positive Predictive Value)

Precision (PREC) is calculated as the number of correct positive predictions divided by the total number of positive predictions. Precision measures probability that a positive prediction is correct. Precision is how many selected items are relevant. It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0.

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Specificity (True Negative Rate)

The fourth measure is Specificity (SP) as shown in Equation, which measures how well a binary classification test correctly identifies the true negative values. Specificity is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR).

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = \frac{\mathrm{TN}}{\mathrm{N}}$$

The last measure is F-measure as shown in Equation bellow, which measures probability that a positive prediction is correct.

$$\mathrm{F-measure} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

ROC CURVE

In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, it can count on an AUC - ROC Curve. When it needs to check or visualize the performance of the multi - class classification problem, it uses AUC (Area under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model's performance. It is also written as AUROC (Area under the Receiver Operating Characteristics).

CHAPTER 4

RESULT AND DISCUSSION

This chapter discusses in detail, the different experiments that were conducted on the Heart disease dataset. The comparative analysis of predicted results of all classifiers based on accuracy and runtime also discussed in this chapter.

4.1 EXPERIMENT WITH DIFFERENT CLASSIFIER ALGORITHMS

In this section different classification algorithm were used for the calculation of accuracy. This paper used Jupyter machine learning software with python programming language for performance measurements of different classifiers. There has different classification algorithm but here worked with Naïve Bayes, Multilayer Perceptron, Decision Tree and Random Forest classifier algorithm.

4.1.1 Experiment with Naïve Bayes Classifier

Naïve Bayes classifier applied on the dataset by using Jupyter Notebook. After performing this algorithm on the dataset, the resultant accuracy was obtained was 72%. Table 4.1 shows the performance of Naïve Bayes classifier. The table showing the Precision, Recall, for both 'Present' and 'Absent'. In Naïve Bayes precision value for present class is 0.78 whereas for absent instance the value is 0.66.

Table 4.1: Performance evaluation of Naïve Bayes classifier

Class	Precision	Recall	F1 Score
Absent	0.66	0.73	0.69
Present	0.78	0.71	0.75

The value of recall also varies for present instances and absent instances, and the value of recall for present instances is 0.71 and for absent instances is 0.73. F1 scores for absent class is 0.69 and for present class is 0.75. Here the True positive value is 25 and True negative values is 19. The Area Under Curve for Naïve Bayes founded 0.74. The confusion matrix generated by using Naïve Bayes classifier is shown in Table 4.2. Following curve shown in figure is for ROC curve for measuring performance.

Table 4.2: Confusion matrix of Naïve Bayes classifiers

	Absent	**Present**
Absent	19	7
Present	10	25

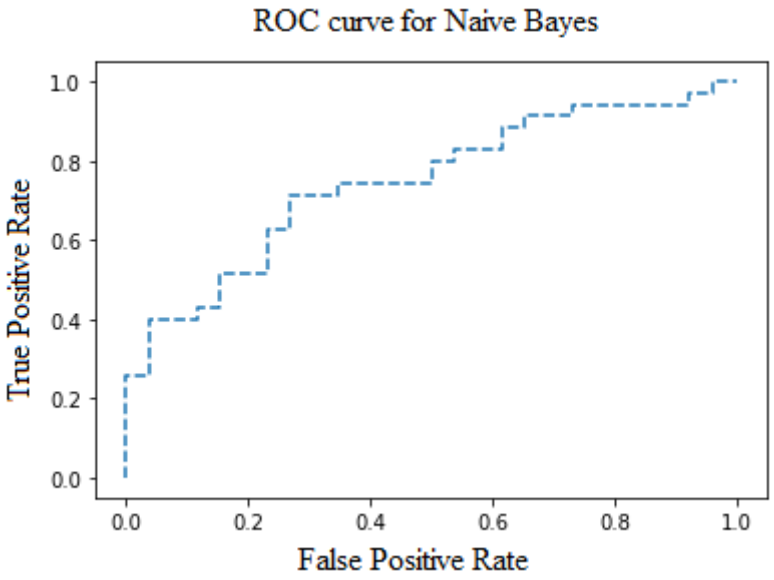


Figure 4.1: ROC curve for measuring performance of Naïve Bayes Classifiers

Figure 4.1 shows Receiver Operating Characteristic (ROC) Curve of Naïve Bayes Classifiers. The Y-axis is stand for true positive value and X-axis is for false positive value. Area under ROC (AUC) for both Class Absent and Class Present is 0.74.

4.1.2 Experiment with Decision Tree Classifier

Decision Tree classifier applied on the dataset by using Jupyter notebook. After performing this algorithm on the dataset, the resultant accuracy was obtained was 64% which is smaller than Naïve bayes classifier. Table 4.3 shows the performance of Decision Tree classifier. The table showing the Precision, Recall, for both 'Present' and 'Absent'. In Decision Tree classifier precision value for present class is 0.72 whereas for absent instance the value is 0.56.

Table 4.3: Performance evaluation of Decision Tree

Class	Precision	Recall	F1 Score
Absent	0.56	0.69	0.62
Present	0.72	0.60	0.66

The value of recall also varies for present instances and absent instances, and the value of recall for present instances is 0.60 and for absent instances is 0.69. F1 scores for absent class is 0.62 and for present class is 0.66. In confusion matrix, True positive value is 21 and True negative values is 18. The Area Under Curve for Decision Tree founded 0.69. The confusion matrix generated by using Decision Tree classifier is shown in Table 4.4.

Table 4.4: Confusion matrix of Decision Tree classifier

	Absent	Present
Absent	18	8
Present	14	21

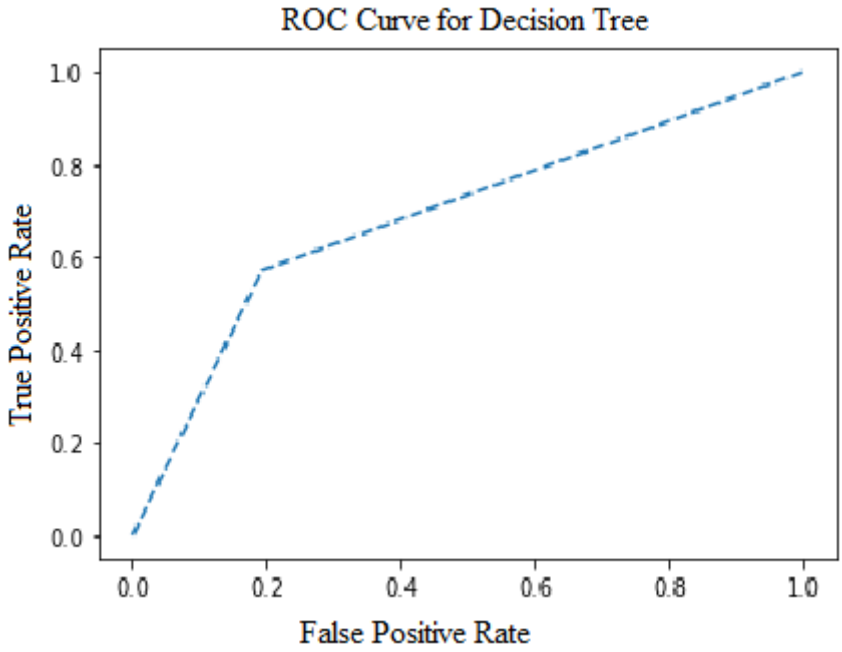


Figure 4.2: ROC curve for measuring performance in Decision Tree

Figure 4.2 shows Receiver Operating Characteristic (ROC) Curve of Decision Tree Classifiers. The Y-axis is stand for true positive value and X-axis is for false positive value. Area under ROC (AUC) for both Class Absent and Class Present is 0.69.

4.1.3 Experiment with Random Forest Classifier

Random Forest machine learning algorithm is applied on the dataset by using Jupyter notebook tool. After performing this algorithm on the dataset, the resultant accuracy was obtained was 80% which is better than Naïve bayes and Decision tree classifiers for predicting the heart disease. Table 4.5 shows the performance of Random forest classifier. The table showing the Precision, Recall, f1-score for both 'Present' and 'Absent' class. In Random forest classifier precision value for present class is 0.81 whereas for absent instance the value is 0.79.

Table 4.5: Performance evaluation of Random Forest

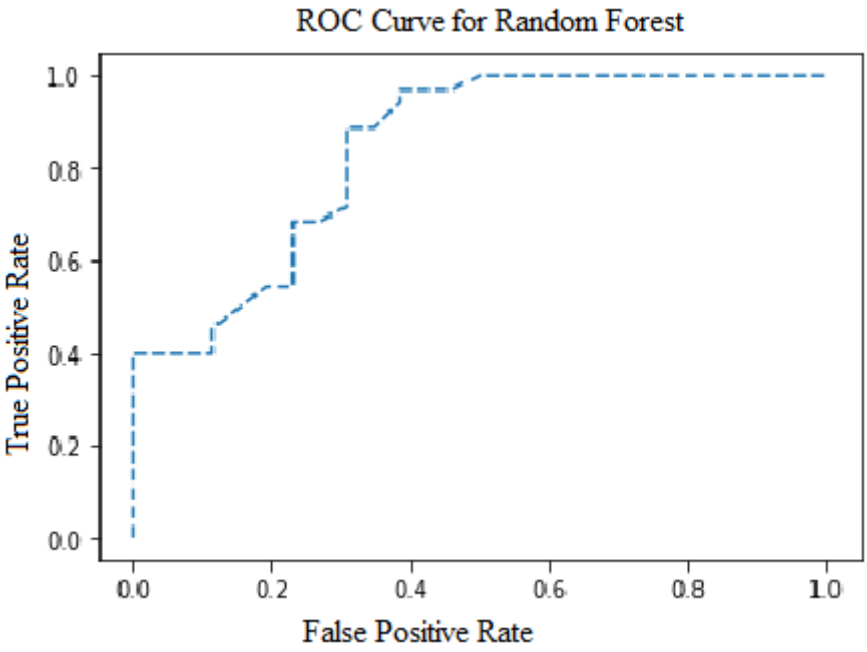
Class	Precision	Recall	F1 Score
Absent	0.79	0.73	0.76
Present	0.81	0.86	0.83

The value of recall also varies for present instances and absent instances, and the value of recall for present instances is 0.86 and for absent instances is 0.73. F1 scores for absent class is 0.76

and for present class is 0.83. In confusion matrix, True positive value is 30 and True negative values is 19. The Area Under Curve we found for random forest is 0.84. The confusion matrix generated by using Random forest classifier is shown in Table 4.6.

Table 4.6: Confusion matrix of Random Forest classifier

	Absent	**Present**
Absent	19	7
Present	5	30



4.3: ROC curve for measuring performance of Random Forest

Figure 4.3 shows Receiver Operating Characteristic (ROC) Curve of Random Forest Classifiers. The Y-axis is stand for true positive value and X-axis is for false positive value. Area under ROC (AUC) for both Class Absent and Class Present is 0.84. Which is better than Naïve Bayes and Decision Tree classifiers for the desired output for predicting the heart disease.

4.1.4 Experiment with Neural Network (Multilayer Perceptron) Classifier

Multilayer Perceptron machine learning algorithm is applied on the dataset by using Jupyter notebook tool. After performing this algorithm on the dataset, the resultant accuracy was obtained was 60% which is smaller than previous classifiers. Table 4.7 shows the performance of Multilayer Perceptron classifier. The table showing the Precision, Recall, f1-score for both 'Present' and 'Absent' class. In Multilayer Perceptron classifier precision value for present class is 0.62 whereas for absent instance the value is 0.57.

Table 4.7: Performance evaluation of Multilayer Perceptron

Class	Precision	Recall	F1 Score
Absent	0.57	0.31	0.40
Present	0.62	0.83	0.71

The value of recall also varies for present instances and absent instances, and the value of recall for present instances is 0.83 and for absent instances is 0.31. F1 scores for absent class is 0.71 and for present class is 0.40. In confusion matrix, True positive value is 29 and True negative values is 8. The confusion matrix generated by using Multilayer Perceptron classifier is shown in Table 4.8.

Table 4.8: Confusion matrix of Multilayer Perceptron classifier

	Absent	Present
Absent	8	18
Present	6	29

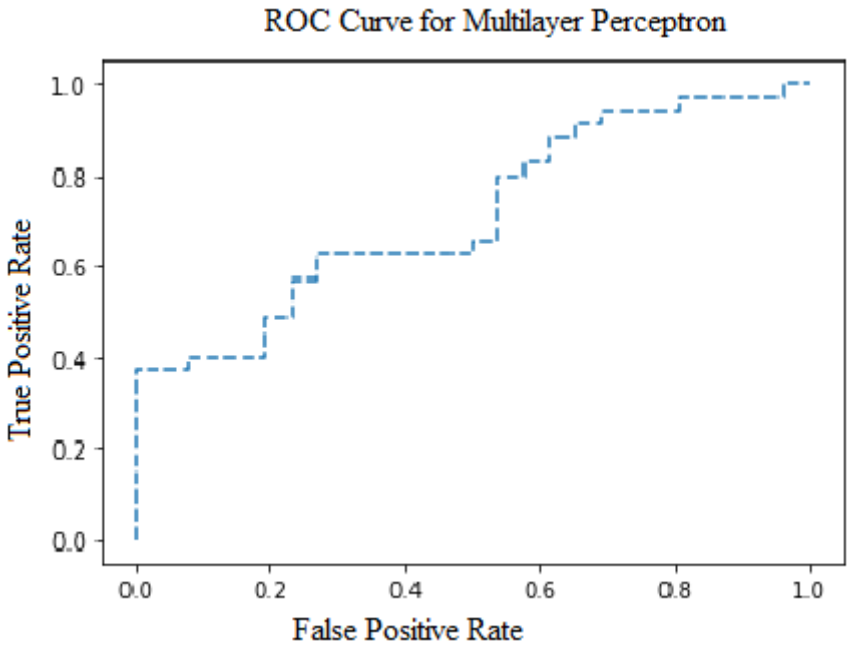


Figure 4.4: ROC curve for measuring performance of MLP

Figure 4.4 shows Receiver Operating Characteristic (ROC) Curve of Multilayer perceptron Classifiers. The Y-axis is stand for true positive value and X-axis is for false positive value. Area under ROC (AUC) for both Class Absent and Class Present is 0.71.

4.2 COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS

In order to evaluate the performance and usefulness of four different classification algorithms for predicting Heart disease Jupyter open-source machine learning tool is used. After performing different classification algorithm in on jupyter notebook, value of different term for each classification algorithm are listed in a table to measure and investigate the performance on the classification methods. Here Table 4.9 shows different performance parameters like Precision, Recall, F-measure, accuracy for different classifiers. In heart disease prediction it is needed to keep the number of True Positives high and number of False Positives low. Early diagnosis is the key factor for a successful treatment of a disease; therefore, classification algorithms are expected to perform well and emphasis is given to select the algorithms having high TP rate. Accuracy is necessary for accurately identifying heart patients as much as possible. Kappa statistic is also used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity.

Table 4.9: Performance measurement of all classifiers

Algorithms	Precision	Recall	F1-Score	Accuracy (%)
Naïve Bayes	0.73	0.72	0.72	72
Decision Tree	0.66	0.64	0.64	64
Random Forest	0.80	0.80	0.80	80
Multilayer Perceptron	0.60	0.61	0.58	60

The table above presents the great accuracy of classifier algorithms on the given heart disease dataset, the lowest accuracy provided by Multilayer Perceptron is 60% and the highest accuracy provided by Random forest is 80%. Accuracy provided by other two classifiers Decision Tree and Naïve Bayes are 64% and 72% respectively. The figure 4.9 showing the graphical representation of accuracy comparison of classifiers.

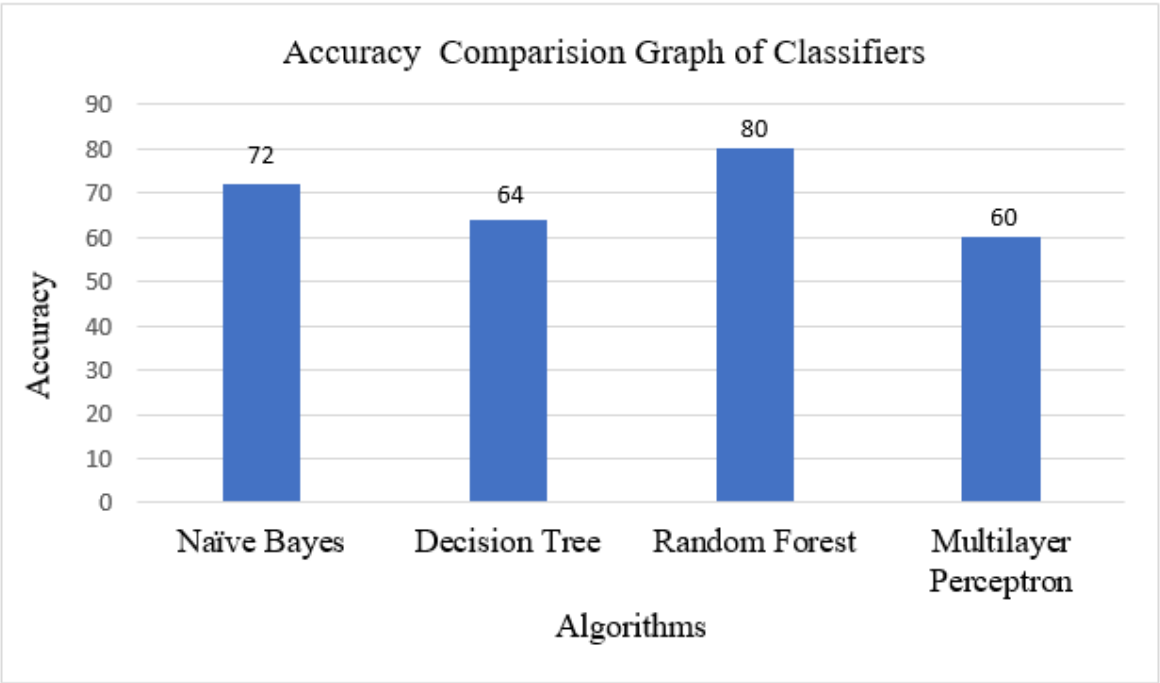


Figure 4.5: Graphical representation of classifiers with their Accuracy

On the other hand, Other performances measures Precision, Recall and F1-score are also used to compare the results also achieve remarkable performance. Precision, Recall and F-measure of Naïve Bayes algorithm are 0.73 ,0.72 and 0.72 respectively. In Decision Tree algorithm, we found Precision value is 0.66, Recall is 0.64 and for F1-score is 0.64. On the other hand, Random forest algorithm provides better rate, this algorithm provides Precision, Recall and for F1-score is 0.80. Multilayer perceptron provided the lower values.

Based on above results and comparisons we found that the Random forest performs the highest Accuracy, Precision, F1-measure value. Naïve Bayes also score the fastest execution time as compare to other classifiers. Multilayer Perceptron provided lowest accuracy for heart disease prediction. So, we can say that, Random forest provided the better accuracy than other classifiers for heart disease prediction.

CHAPTER 5

CONCLUSION

5.1 CONCLUSION

Health care related data are voluminous in nature and they arrive from different sources all of them are not entirely appropriate in quality for using in disease diagnosis. In medical diagnosis various data mining techniques are used for organizing numerous data. In this research, different data mining classifiers algorithms were applied to improve the early detection of heart diseases. This paper deals with real dataset for training and testing purpose. Generally, four classifiers were utilized using Jupyter Notebook web-based machine learning tool to predict better accurate results and performance analysis. Comparisons of these algorithms are based on the performance factors of classification accuracy. Random Forest classifiers algorithm highest accuracy among all that is 80% is considered as the best classifier algorithm based on performance analysis in the diagnosis of heart disease patients. The experimental results show that we can produce short but accurate prediction list for the heart patients by applying the predictive models to the records of incoming new patients.

5.2 FUTURE WORK

The outcomes of this thesis may be used as assistant tool to help in making more consistent diagnosis of heart diseases. This study will also work to identify those patients which needed special attention. As a future work, we will use the research described here as a foundation for the development of effective prediction system to enhance medical care and reduce costs. We will significantly extend the functionality of the current research. It will also possible to plug in the prediction system into other systems such as access control security system to support the security fundamentals of healthcare systems and suggest useful treatment according to disease level.

REFERENCES

[1] Smita1, P. Sharma. 2014. Use of Data Mining in Various Field: A Survey Paper. *IOSR Journal of Computer Engineering (IOSR-JCE)*.**16**(3): 18-21.

[2] P. Ahmad, S. 2015. Qamar Techniques of Data Mining In Healthcare A Review. *International Journal of Computer Applications*. **120**(15): 45-50.

[3] C. McGregor, C. Christina, J. Andrew. 2012. A process mining driven framework for clinical guideline improvement in critical care Learning from Medical Data Streams. *13th Conference on Artificial Intelligence in Medicine (LEMEDS),* **7**(2): 13-14.

[4] P. S. Qamar, S. Rizvi, Ahmad. 2015. Techniques of data mining in healthcare. *International Journal of Computer applications*. **120**(3): 120-127.

[5] S. Jayien, M. K. Nair. 2015. Prediction of Heart Disease Using Classification Based Data Mining Techniques. *Scientific Research Journal*. **2**(2): 206-210.

[6] Dhanya P, Varghese, Tintu P B. 2015. A Survey On Health Data Using Data Mining Techniques. *International Research Journal of Engineering and Technology (IRJET).* **2**(7): 20-25.

[7] J.Yanqing, H.Y., J.T.,P.D. and A. Mansour. 2011. Mining Infrequent Causal Associations in Electronic Health Databases. *11th IEEE International Conference on Data Mining Workshops*. **3**(2): 219-222.

[8] S. Patel, H. Patel. 2016. Survey of Datamining Techniques Used in Healthcare Domain. *International Journal of Information Sciences and Techniques (IJIST)*. **6**(2): 44-45.

[9] Anbarasi, E. A., N. M. 2010. Enhanced Prediction of Heart Disease with Feature Subset Selection. *International Journal of Engineering Science and Technology*. **2**(2): 106-110.

[10] AH Chen. 2011. HDPS: Heart Disease Prediction System. *IEEE*. **3**(2): 6-10.

[11] Chaitrali, S. D., S. S. Apte. 2012. A data mining approach for prediction of heart disease using neural networks. *International Journal of Computer Engineering & Technology*. **3**(3): 15-17.

[12] Joyti soni,U.A. 2011. Intelligent and Effective heart diseases prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*. **3**(6): 35-37.

[13] S. Pal, V. Chandra. 2013. Data Mining Approach to Detect Heart Dieses. *International Journal of Advanced Computer Science and Information Technology*. **2**(4): 55-58.

[14] S. Ahmed., P. Akiz. 2012. Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences.* **2**(3): 66-71.

[15] F. Rashedur, M. Rahman. 2017. Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. *Journal of Software Engineering and Applications*. **10**(2): 223-230.

[16] My Chau Tu, D. S. 2009. A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. *Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing,* **2**(3): 12-14.

[17] Aqueel Ahmed, S. A. Hossain. 2012. Data Mining Techniques to Find Out Heart Diseases. *International Journal of Innovative Technology and Exploring Engineering(IJITEE).* **1**(1): 20-25.

[18] L. Moreno.,V. Fateh. 2015. Association Rules: Problems, solutions and new applications. *Actas del III Taller Nacional de Minería de Datosy Aprendizaje.* **2**(1): 313-315.

[19] S. Dangare, A. Chaitrali, S. Dangare. 2014. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications. **47**(10): 888 – 975.

[20] N. Kumaravel, K. Sabullah., N. Nair. 2016. Automatic diagnoses of heart diseases using neural network. *In Proceedings of the Fifteenth Biomedical Engineering Conference,* **3**(2): 319-322.

[21] B.Venkatalakshmi, M. S. 2016. Heart Disease Diagnosis Using Predictive Data mining. *International Journal of Innovative Research in Science.* **3**(3): 345-445.

[22] D. Methaila. 2016. Early Heart Disease Prediction Using Data Mining Techniques. *CCSEIT, DMDB, ICBB, MoWiN, AIAP*. **3**(2): 53-59.

[23] Mamta Sharma, F.Khan. 2017. Comparing Data Mining Techniques Used For Heart Disease Prediction. *International Research Journal of Engineering and Technology (IRJET).* **4**(6): 30-35.

[24] Isra'a Ahmed Zriqat. 2016. A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. *International Journal of Computer Science and Information Security (IJCSIS).***14**(12).

[25] S. Saravanakumar and S.Rinesh. 2014. Effective Heart Disease Prediction using Frequent Feature Selection Method. *International Journal of Innovative Research in Computer and Communication Engineering*. **2**(3).

[26] Marjia Sultana. 2015. Analysis of Data Mining Techniques for Heart Disease Prediction. Department of Computer Science and Engineering, Jahangirnagar University. **3**(2): 29-32.

[27] H. Benjamin, F. David. 2018. Heart disease prediction using data mining techniques. *International Journal of Semantic Computer*.**10**(3): 200-205.

[28] P. S. UshaRani. 2016. A Novel Approach for Imputation of Missing Attribute Values for Efficient Mining of Medical Datasets-Class Based Cluster Approach. *arXiv preprint arXiv.* **3**(2): 312-315.

[29] Korting, T. S. 2014. C4.5 algorithm and Multivariate Decision Trees. *Image Processing Division, National Institute for Space Research—INPE.* **6**(3): 44-47.

[30] Kundra. 2015. A novel technique to predict heart diseases in data mining. *International Journal of Advanced Trends in Computer Applications*. **2**(7): 10-14.

[31] Vaghela, N. Bhatt and D. Mistry. 2015. A Survey on Various Classification Techniques for Clinical Decision Support System. *International Journal of Computer Applications*. **116**(23): 37.

[32] D. Larose. 2015. Discovering Knowledge in Data: An Introduction to Data Mining. *New Jersey: John Wiley & Sons.*

[33] Q. Rizvi. J. 1995. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

[34] Armstrong. 2012. Illusion in Regression Analysis. *International Journal of Forecasting*. **28**(3): 689

[35] A. Liaw and Matthew Wiener. 2002. Classification and Regression by Random Forest. *R News*. **2**(3): 18-22.

[36] T. C. Corporation. 2005. Introduction to Data Mining and Knowledge Discovery. *Two Crows Corporation.*

[37] Rohit Arora, Suman. 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. *International Journal of Computer Applications.* **54**(13): 0975-8887.

[38] Margaret H Dunham. 2003. Data mining: Introductory and Advanced Topics. *Published by Pearson Education.*

[39].Anaconda .*wikipedia*. Retrieved 22 April 2020.

[40] An Introduction to Jupyter notebook. *Real python*. 2018. Retrieved 2018-11-11.

[41] P. Reuteman, B. Pfahringer, E. Frank. 2007. Proper: A Textbook for Learning from Relational Data with Propositional and Multi-instance Learners. *17th Australian Joint Conference on Artificial Intelligence. Springer-Verlag,* **4**(3): 207-210.