EFFECTIVE HEART DISEASE PREDICTION USING MACHINE LEARNING

BY

Mahmudul Hassan

ID: 152-15-5809

Mahmudul Alam

ID: 152-15-5563

Sabbir Ahmed

ID: 152-15-5564

Reshmi Akter Rima

ID: 152-15-5528

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Prof. Dr. Syed Akhter Hossain

Head

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

May 2019

APPROVAL

This Project titled "Effective Heart Disease Prediction Using Machine Learning Technique", submitted by Mahmudul Hassan, ID No: 152-15-5809 and Md. Mahmudul Alam, ID No: 15215-5563 and Sabbir Ahmed, ID No: 152-15-5564 and Reshmi Akter Rima, ID No: 152-15-5528 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on May 3, 2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Tarek Habib **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Moushumi Zaman Bonny **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Swakkhar Shatabda **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done under the supervision of **Prof. Dr. Syed Akhter Hossain, Head of the Department of CSE**, Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised By:

Prof. Dr. Syed Akhter Hossain

Head of the Department of Computer Science and Engineering

Daffodil International University

Submitted By:

Mahmudul Hassan.

Mahmudul Hassan

ID: 152-15-5809

Department of Computer Science and Engineering

Daffodil International University

md. Mahmudul Alam

Md. Mahmudul Alam

ID: 152-15-5563

Department of Computer Science and Engineering

Daffodil International University

Sabbir Ahmed

Sabbir Ahmed

ID: 152-15-5564

Department of Computer Science and Engineering

Daffodil International University

Reshmi Akter Rima

Reshmi Akter Rima

ID: 152-15-5528

Department of Computer Science and Engineering

Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis successfully.

We really grateful and wish our profound our indebtedness to Prof. Dr. Syed Akhter Hossain, Head of the department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Data Mining and Machine learning” to carry out this thesis. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to Honorable Teachers Nusrat Jahan, Mohammad Masudur Rahman, Montasir Dipto for their kind help to finish our thesis and also to other faculty members of computer science and engineering department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Cardiovascular disease is a leading cause of death in this period. The number of deaths among both males and females increases each day due to heart disease. For example, researchers used machine learning and data mining techniques to support the prognosis of heart disease in healthcare. However, using these techniques there may be a smart system of Heart Disease prediction that is quicker and more proficient than the usual system of diagnosis. The objective of this paper is to use data mining strategies and numerous machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest, and Various Ensemble Method to gain a more accurate measure. We also suggested two of our own models in this paper. In this paper, a UCI repository Cleveland dataset is used. The sole aim of this paper is to use Data Mining and Machine Learning Techniques to find hidden patterns.

TABLE OF CONTENTS

LIST OF FIGURES 6-40

LIST OF TABLES 7-39

CHAPTER 1

INTRODUCTION

1.1 Introduction

In the human body, the heart plays a pivotal role. The heart is a pump circulating blood across the whole body. It beats 60-70 times in one minute on average. If the heart stops functioning normally, our body's other organs will stop working. Heart disease is one of the world's most common diseases. Heart disease is one of the globe's most deadly diseases. As per the latest WHO data released in 2017, deaths from heart disease in Bangladesh entered 14.31 percent of total deaths, and every year cardiovascular disease kills 17.9 million people, 31 percent of global deaths [1]. Heart disease is the major cause of death in the United Kingdom, USA, Canada & Australia, according to CDC (Centrals for Disease Control).

There are several types of heart disease that attack various parts of the body [2].

1. Coronary heart disease
2. Cardiomyopathy
3. Cardiovascular disease
4. Ischaemic heart disease
5. Hypertensive heart disease
6. Heart failure
7. Inflammatory heart disease
8. Valvular heart disease

Common risk factors of heart disease:

1. Age
2. Blood Pressure
3. Total Cholesterol
4. Diabetes
5. Hyper Tension
6. Family History of heart disease
7. Lack of physical exercise
8. Fasting blood sugar

Countries around the world, medical institutions collect vast amounts of cardiovascular disease data. These collected data can be used to explore heart disease problem in varying data mining & machine learning techniques [3]. But these datasets are very noisy. If we use this raw data, a machine learning algorithm cannot give the most accurate result. Before applying this actual data to the machine learning algorithm, it is essential to pre-process this raw data. Data mining is the examination of large data sets to find out hidden & unknown patterns, knowledge, relationships that are difficult to find out without pre-processing these data (Lee, Liao et al. 2000). Machine Learning techniques are used in various disease diagnostic problem & it is very helpful for doctor to predict disease correctly.

1.2 Motivation

The provision of quality services at inexpensive costs is a significant challenge for healthcare institutions (hospitals, medical centers). Decent service requires correct diagnosis of patients and potential treatments. Poor clinical decisions can lead to disastrous unacceptable consequences.

Hospitals also need to mitigate laboratory testing costs. By using optimal computer-based information and decision support systems, they can accomplish these results. Typically, these systems generate enormous amounts of data in the form of numbers, text, charts, and images. Typically, these platforms yield massive amounts of data in the form of numbers, text, charts, and images.

Sadly, this information is infrequently used to help clinical basic leadership. There is an abundance of shrouded information in these data that is to a great extent undiscovered. This brings up a vital issue: "How might we transform information into helpful data that can empower medicinal services specialists to settle on clever clinical choices?" This is the principal inspiration for this examination.

It also helps to reduce the cost of treatment by providing effective treatments. It illustrates the outcomes in both tabular and graphical forms to enhance visualization and ease of interpretation.

1.3 Rationale of the Study

We know, world of technology is moving towards artificial intelligence and machine learning techniques, huge amount of raw data is being produced in real time. This data gives us with the opportunity to analyze this huge amount of data using data mining and machine learning techniques.

So far, the medical data is concerned more and more computerized technology and preservation of patient's data is becoming common in application. Analyzing those data with conventional statistics may give us the answer to HOW questions using machine learning. It can demonstrate the potential solution of those problems. Classification using machine learning and deep learning is being practiced increasingly. These techniques can also reveal hidden patterns or trends.

For using data mining and machine learning techniques, we can classify heart disease patient and non-heart disease patient.

1.4 Research Questions

As of the research project focuses on feature selection and classification of heart disease and non-heart disease patient using machine learning techniques, the followings questions are the things we tried to accomplish in this project.

- Can the Heart Disease Symptoms be classified with convincing level of accuracy?
- Do all the features are influential which are supposed to be dominated?
- Which techniques of ML approaches would be good for classification?

1.5 Expected Output

We are planned to design the research according to the research questions using machine learning algorithms. We expected a benchmark to achieve our goal to reach the following outcomes:

- Analyze the dataset Heart Disease of patient should be classified with good accuracy.
- Find out the correlation of different attributes in the dataset in developing Heart Disease.
- Selection of important features to detect heart disease disorder.
- Compare different machine learning classification methods.

1.6 Report Layout

In the first section of the project report, we have examined the review, motivation of the project, we have discussed our objective and what are the results.

The second section deals intensively with our background study on heart disease and literature survey. We have also enrolled many of the studies in this field.

The third section is about the research methodology we have applied. We also covered shortly the classifier algorithms used in this study.

The fourth chapter consists of a lengthy description of our findings and comparative classifier accuracy studies.

At last, the fifth section examines the rundown, future extent of the examination and talks about further territories for concentrate in the comparable field.

CHAPTER 2

BACKGROUND

2.1 Introduction

In the next few parts we will discuss about heart disease. We will also explore the risks and effects of the heart disease. We discussed about the literature survey in similar field.

2.2 Heart Disease

Coronary illness (Heart Disease) is the main source of death in the U.S. Sooner or later in your life, possibly you or one of your friends and family will be compelled to settle on choices about some part of coronary illness [5]. Coronary illness can strike all of a sudden and expect you to settle on choices rapidly.

Around 610,000 Americans bite the dust from the coronary illness every year—that is 1 in every 4 passing [6]. Coronary illness is the most widely recognized kind of coronary illness, slaughtering in excess of 370,000 individuals every year.

As indicated by the most recent WHO information distributed in 2017 Coronary Heart Disease Deaths in Bangladesh achieved 112,791 or 14.31% of all-out passing. The age balanced Death Rate is 108.99 per 100,000 of populace positions Bangladesh #104 on the planet [7].

Hypertension is the main source of cardiovascular and kidney maladies and over 1.5 billion are relied upon to be influenced by hypertension by 2025; the case being particularly unfavorable in South Asia given that the predominance of hypertension is now at 40 percent [8].

As per the World Health Organization (WHO), one out of three grown-ups more than 35 years has hypertension, however, 50% of them don't know about it, as indicated by a message got from the International Center for Diarrheal Diseases Research, Bangladesh (icddr,b) [9].

Heart disease symptoms depend on types of heart disease [10].

The most well-known Signs of coronary course malady is angina or chest torment. Angina can be portrayed as distress, largeness, weight, throbbing, copying, totality, pressing, or excruciating inclination in your chest. Angina may likewise be felt in the shoulders, arms, neck, throat, jaw, or back.

Different symptoms of coronary artery disease include:

- Shortness of breath
- Palpitations (sporadic heart pulsates, or a "flip-flop" feeling in your chest)
- A quicker heartbeat
- Shortcoming or tipsiness
- Queasiness
- Perspiring

Signs of a Heart Attack

- Uneasiness, weight, greatness, or agony in the chest, arm, or beneath the breastbone
- Uneasiness emanating to the back, jaw, throat, or arm
- Totality, acid reflux, or stifling inclination.
- Perspiring, sickness, spewing, or wooziness.
- Extraordinary shortcoming, tension, or shortness of breath.
- Quick or sporadic pulses

2.2.1 Heart Disease Definition

Any disorder that influences the heart. Now and then the expression "heart disease" is utilized barely and erroneously as an equivalent word for coronary supply route sickness. Coronary illness is synonymous with heart ailment however not with the cardiovascular infection which is any malady of the heart or veins. Among the numerous sorts of heart disease, see, for instance: Angina; Arrhythmia; Congenital coronary illness; Coronary artery disease (CAD); Dilated cardiomyopathy; Heart assault (myocardial localized necrosis); Heart disappointment; Hypertrophic cardiomyopathy; Mitral spewing forth; Mitral valve prolapse; and Pulmonary stenosis [11].

2.2.2 Causes and Risk Factors of Heart Disease

Heart disease causes depend on the types of heart disease.

While disorder will seek advice from completely different heart or vessel issues, the term is commonly accustomed mean harm to our heart or blood vessels by hardening of the arteries (ath-ur-o-skluh-ROE-sis), a buildup of fatty plaques in our arteries. Plaque buildup thickens and stiffens artery walls, which might inhibit blood flow through our arteries to our organs and tissues. Atherosclerosis is additionally the foremost common explanation for disorder. It are often caused by correctable issues, like associate unhealthy diet, lack of exercise, being overweight and smoking.

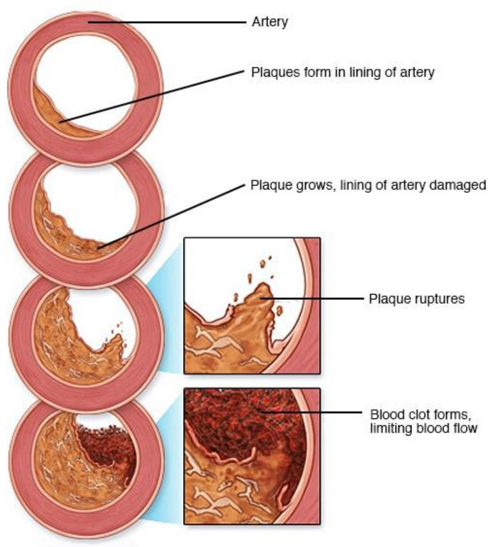

Artery
Plaques form in lining of artery
Plaque grows, lining of artery damaged
Plaque ruptures
Blood clot forms, limiting blood flow

Figure 2.1 Development of atherosclerosis

Causes

- Heart defects you are born with (congenital heart defects)
- Coronary artery illness
- High force per unit area
- Diabetes
- Smoking
- Excessive use of alcohol or caffeine
- Drug abuse
- Stress
- Some over-the-counter medications, prescription medications, dietary supplements and favorer remedies
- Valvular cardiopathy
- Dilated heart condition
- Hypertrophic heart condition
- Restrictive heart condition
- Bacteria, Viruses and Parasites
- Rheumatic fever
- Infections (infectious endocarditis)
- Connective tissue disorders

Risk factors:

Table 2.1 Risk Factors of Heart Disease

Factors	Description
Age	Aging increases our risk of damaged and narrowed arteries and weakened or thickened heart muscle
Sex	Men are generally at greater risk of heart disease. However, women's risk increases after menopause.
Family history	A family history of heart disease increases your risk of coronary artery disease, especially if a parent developed it at an early age (before age 55 for a male relative, such as your brother or father, and 65 for a female relative, such as your mother or sister).
Smoking	Nicotine constricts your blood vessels, and carbon monoxide can damage their inner lining, making them more susceptible to atherosclerosis. Heart attacks are more common in smokers than in nonsmokers.
Certain chemotherapy drugs and radiation therapy for cancer	Some chemotherapy drugs and radiation therapies may increase the risk of cardiovascular disease.
Poor diet	A diet that's high in fat, salt, sugar and cholesterol can contribute to the development of heart disease.
High blood pressure	Uncontrolled high blood pressure can result in hardening and thickening of your arteries, narrowing the vessels through which blood flows.
High blood cholesterol levels	High levels of cholesterol in your blood can increase the risk of formation of plaques and atherosclerosis.
Diabetes	Diabetes increases your risk of heart disease. Both conditions share similar risk factors, such as obesity and high blood pressure.
Obesity	Excess weight typically worsens other risk factors.
Physical inactivity	Lack of exercise also is associated with many forms of heart disease and some of its other risk factors, as well.
Stress	Unrelieved stress may damage your arteries and worsen other risk factors for heart disease.
Poor hygiene	Not regularly washing your hands and not establishing other habits that can help prevent viral or bacterial infections can put you at risk of heart infections, especially if you already have an underlying heart condition. Poor dental health also may contribute to heart disease.

2.3 Related Works

Presently days, heart disease is a standout amongst the riskiest disease of everywhere throughout the world. Data mining & machine learning procedures are utilizing comprehensively for heart disease prediction. The most utilizing machine learning mining methods are Decision Tree, Naïve Bayes, Neural Network, KNN calculation, Artificial Neural Network (ANN), Automatically defined groups bagging algorithm, Support Vector Machine Algorithm have been executed for distributing past previous papers and giving a decent precision level. These Research has been assumed an imperative job in the advanced therapeutic framework. Various works have been done identified with heart disease prediction system by utilizing different data mining techniques and machine learning algorithms by numerous authors. The point of all is to accomplish better precision and to make the system progressively effective with the goal that it can predict the odds of heart assault.

In 2015, Jaymin Patel and Prof. Tejal Upadhyay, Dr. Samir Patel introduced an effective methodology for heart disease prediction utilizing machine learning and Data Mining systems. They utilized Cleveland Dataset from UCI archive. They utilized different Machine Learning Algorithm Like Decision Tree, Logistic Model, Random Forest and got the best outcome in Decision Tree with most minimal test blunder (0.16666667) among these three calculations **[12].**

In 2017, Asst. Educator Sanjay Kumar Sen exhibited a viable Heart Disease prediction paper utilizing machine learning. They have utilized Naïve Bayes, SVM, Decision Tree, K-Nearest Neighbor Machine Learning calculation. Support Vector Machine Provided the Highest exactness (84.1584) **[13].**

Ashok Kumar Dwivedi et al represented to various algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. Among them, Logistic Regression calculation gives better exactness **[14**].

Sujata Joshi and Mydhili They utilize three classification algorithms to be specific decision tree, Naïve Bayes, and K-NN and presents the performance among them. It has discovered the best prediction algorithm regarding its accuracy and error rate on specific dataset **[15].**

Sellappan Palaniappan et al A proficient heart disease prediction system is created by utilizing data mining and machine learning algorithm like Naive Bayes, Neural Network, and Decision Trees was put together by Sellappan Palaniappan et al. Every algorithm has its very own resource for getting the best outcomes. It is web-based, easy to understand and expandable **[16].**

Shadab Adam Pattekari Shadab Adam Pattekari and Asma Parveen introduced a proficient system utilizing data mining modeling method, like Naive Bayes. It is actualized as a web-based application. in this user responds to the predefined questions. It recovers concealed information from put away the database and contrasts the user esteems and prepared informational data **[17].**

2.4 Research Summary

According to previous literature research and study indicates there has been a modest number of studies in this field. The studies has-been fairly successful in their own way. This type of computerized classification problem has been researched on many other diseases. From studying different algorithms to make re-optimization to the existing algorithm to find better results, researchers have gone through many different ways. The observable factor is that although the accuracy has been quite good, yet we have not seen any real implementation of these processes. Probably the idea of consulting a computerized diagnosis system for a disease isn't as convincing as consulting a doctor for the public. But with more accuracy and some experimental periods, a fully automated diagnosis probably would be as normal as consulting a doctor.

2.5 Scope of the Problem

We used different machine learning algorithms and here includes training and testing dataset of the model. We try to search the link between the dataset attributes to find out which attributes are very effective for causing heart disease.

In our country, we see that an automated diagnosis system is a lengthy process in health care. With an improved-symptoms analyzing algorithm, the system can suggest a diagnostic test to the users hence reducing time and cost in big hospitals.

2.6 Challenges

We are very new in the machine learning section. It was difficult for us to cope with it. In this thesis, we used many machine learning algorithms and technique. So, evaluating correctly all the machine learning algorithms was a little difficult. we also faced problems to improve our model using grid search cv, and many techniques. In Ensemble of Ensemble section, we proposed our two models and it was hard for us.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Machine learning, abbreviated as ML, is an area of computer science, especially artificial intelligence that practices statistical methods to provide computers the ability to learn with data, without being explicitly programmed (programmed once only) [18].
Machine Learning (ML) is defined as a technique where large volume data in the database is processed and analyzed to reveal useful patterns/trends and new relations among attributes for achieving various intended goals. Machine learning algorithms are two types [19]:

1. Supervised learning
2. Unsupervised learning

Supervised learning infers outputs from labeled training data consisting of a set of training examples while unsupervised learning surmises unlabeled training data to determine the output. Supervised learning deals with classification problem and unsupervised learning treats with clustering problem. Another type of machine learning approach is reinforcement learning which refers to take decisions based on an environment.

Machine learning algorithms consist of two phases [20]:

1. Training phase
2. Test phase

In the training phase, a machine learning algorithm is applied to train up the system using the given data and in the test phase, new data is given as input in the already built system for output. The system takes the new data and models accordingly. In addition, before the training phase, the data are preprocessed and after the test phase, it may require to post-process the output. Figure 3.1 depicts a comprehensive view of the machine learning procedure.

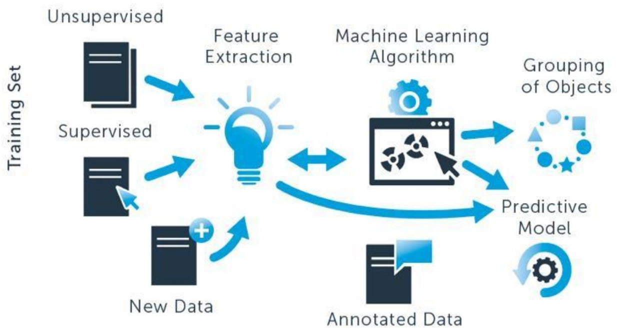


Figure 3.1. Working procedure of Machine Learning (ML) [Image Source: [20]]

Machine Traditional data analysis techniques through a statistical approach have been used. This approach has been very useful and no doubt it will be still be used in the foreseeable future as the storage capacity of modern computers increased, accumulating and preserving various transactional and other types of data became more convenient. Inevitably the size and diversity of the data grew larger and traditional data analysis techniques began to be less effective and inefficient for such a large amount of data warehouses. So, machine learning techniques gained popularity among large companies and researchers.

The latest explosion of Medical data through machine automation and use of computerized technology in the diagnosis and treatment of disease. It has made the machine learning to discover and extracting new patterns and useful knowledge in medical advancement. Although the acceptance of automated classification of disease is still not popular and desirable among the medical community, it is still a research area of the enormous potential for data scientist and researchers around the globe. So, we attempt to explore this concept of machine learning to build a machine-learned system to analyze the heart disease.

3.2 Dataset

For our work we have used data set has been already prepared or collected. The claimer of the dataset benignantly made the dataset available in the UCI Repository from which we obtained access to the dataset [21]. Followings are the information of the creative personal of this dataset.

3.3 Dataset Description & Preprocessing

The dataset has 76 attributes and 303 records. However, only 13 attributes are used for this study & testing.

Table 3.1 Selected Heart Disease Attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 = asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang Continuous Maximum heart rate achieved	Discrete	Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest

Slope	Discrete	The slope of the peak exercise segment; 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Class	Discrete	Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

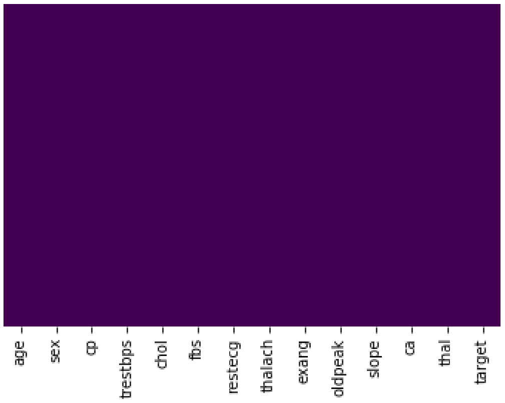


Figure 3.2 Heatmap Of Our Dataset

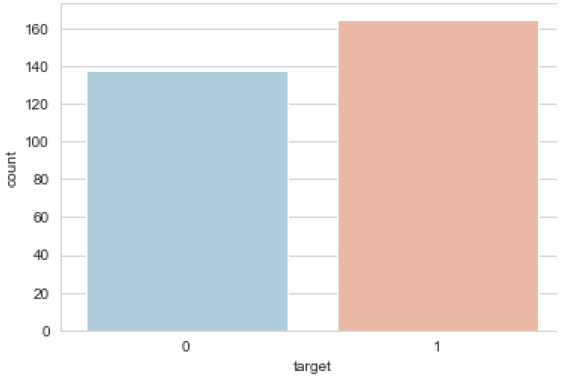


Figure 3.3 Target class

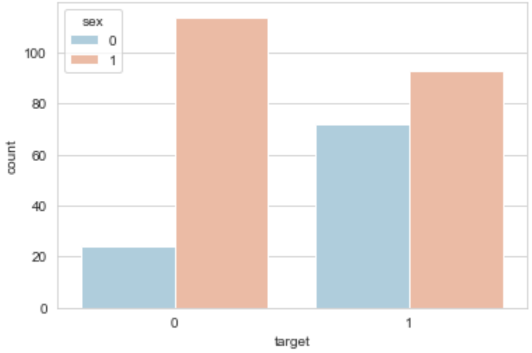
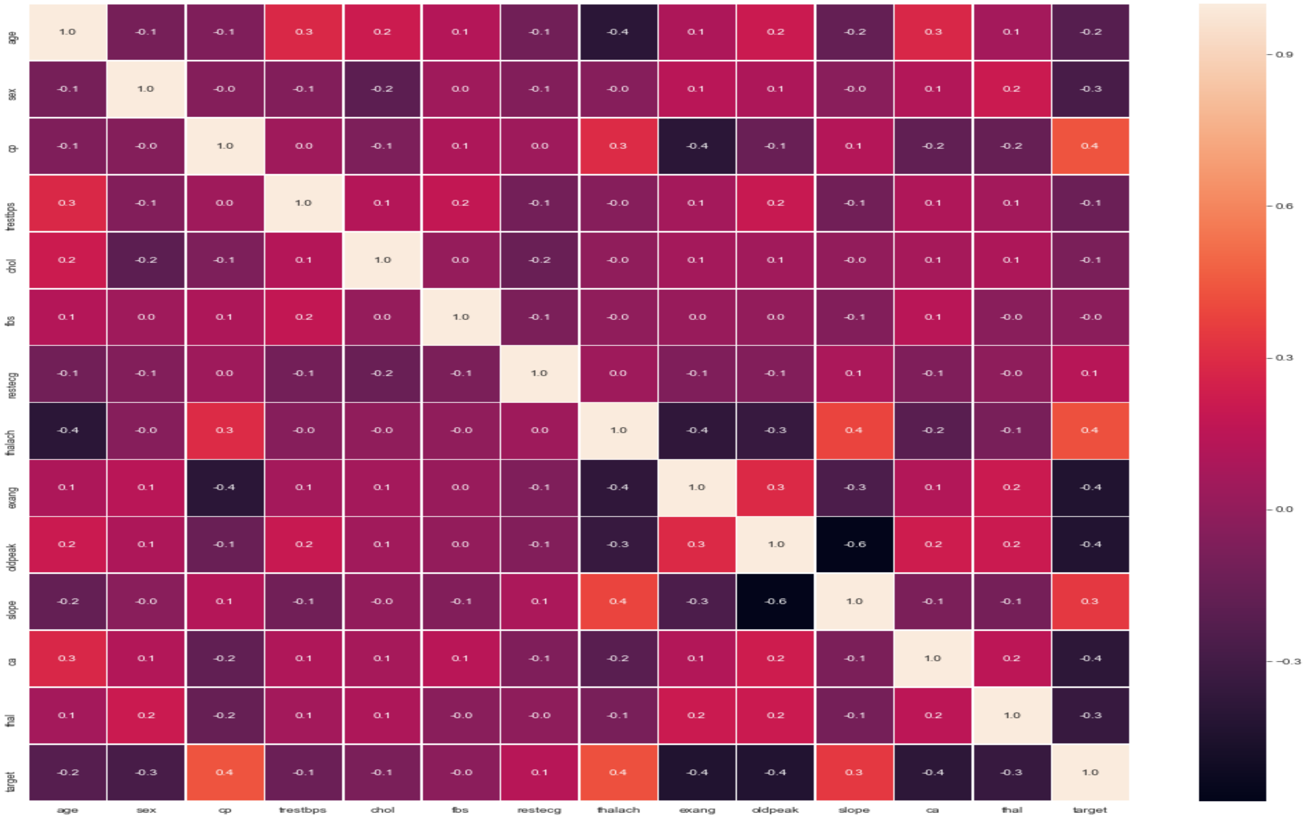


Figure 3.4 Comparison Between Target Class Using Sex

Figure 3.2 shows that there is no missing data in our dataset and figure 3.3 indicates the number of heart disease in our target class and the number of non-heart disease. Figure 3.4 illustrates the comparison between male and female data.

Co-Relations between features:



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.0	-0.1	-0.1	0.3	0.2	0.1	-0.1	-0.4	0.1	0.2	-0.2	0.3	0.1	-0.2
sex	-0.1	1.0	-0.0	-0.1	-0.2	0.0	-0.1	-0.0	0.1	0.1	-0.0	0.1	0.2	-0.3
cp	-0.1	-0.0	1.0	0.0	-0.1	0.1	0.0	0.3	-0.4	-0.1	0.1	-0.2	-0.2	0.4
trestbps	0.3	-0.1	0.0	1.0	0.1	0.2	-0.1	-0.0	0.1	0.2	-0.1	0.1	0.1	-0.1
chol	0.2	-0.2	-0.1	0.1	1.0	0.0	-0.2	-0.0	0.1	0.1	-0.0	0.1	0.1	-0.1
fbs	0.1	0.0	0.1	0.2	0.0	1.0	-0.1	-0.0	0.0	0.0	-0.1	0.1	-0.0	-0.0
restecg	-0.1	-0.1	0.0	-0.1	-0.2	-0.1	1.0	0.0	-0.1	-0.1	0.1	-0.1	-0.0	0.1
thalach	-0.4	-0.0	0.3	-0.0	-0.0	-0.0	0.0	1.0	-0.4	-0.3	0.4	-0.2	-0.1	0.4
exang	0.1	0.1	-0.4	0.1	0.1	0.0	-0.1	-0.4	1.0	0.3	-0.3	0.1	0.2	-0.4
oldpeak	0.2	0.1	-0.1	0.2	0.1	0.0	-0.1	-0.3	0.3	1.0	-0.6	0.2	0.2	-0.4
slope	-0.2	-0.0	0.1	-0.1	-0.0	-0.1	0.1	0.4	-0.3	-0.6	1.0	-0.1	-0.1	0.3
ca	0.3	0.1	-0.2	0.1	0.1	0.1	-0.1	-0.2	0.1	0.2	-0.1	1.0	0.2	-0.4
thal	0.1	0.2	-0.2	0.1	0.1	-0.0	-0.0	-0.1	0.2	0.2	-0.1	0.2	1.0	-0.3
target	-0.2	-0.3	0.4	-0.1	-0.1	-0.0	0.1	0.4	-0.4	-0.4	0.3	-0.4	-0.3	1.0

Figure 3.5 Co-Relations of Dataset

Feature Scaling:

Before applying machine learning algorithms, we normalized our dataset by feature scaling.

3.5 Classification Algorithms

3.5.1 Naïve Bayes

Naïve Bayes is a Bayes Theorem based classification algorithm, which figures a likelihood by tallying the recurrence of qualities and mix of qualities n verifiable data. Bayes hypothesis finds the likelihood of an occasion happening given the likelihood of another occasion that has just happened.

P(B given A) = P(A and B)/P(A)

Naïve Bayes algorithm advantage is, it requires little amount of preparing data for evaluating the parameters fundamental for classification.

3.5.2 Support Vector Machines

Support vector machine is the most prevalent supervised machine learning technique having a pre-characterized target variable which can be utilized for classifier and indicator. For classification, it finds a hyperplane in the component space that separates the classes. The learning data points are described by an SVM model as points in the feature space, mapped in such a way that points belonging to separate classes are segregated as widely as possible. The test data points are then mapped into the same space and classified according to which side of the margin is closest to each other.

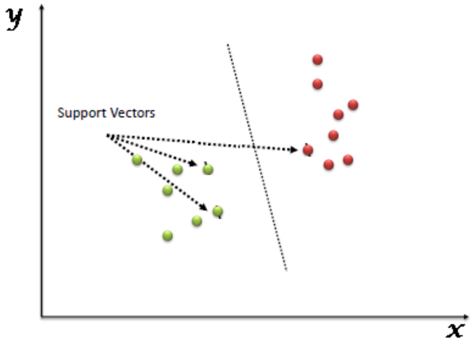


Figure 3.6 Support Vector Machine

3.5.3 Decision Tree

A Decision Tree is a choice help apparatus that utilizes a tree-like diagram or model of choices and their conceivable results including chance occasion results and utility. It is one of the approaches to show a calculation. Choice trees are normally utilized in activities examine, explicitly in a choice investigation to help and recognize a technique that will in all likelihood achieve the objective. It is likewise a well-known instrument in AI. A Decision tree can without much of a stretch be changed to a lot of tenets by mapping from the root hub to the leaf hubs one by one. At long last, by following these tenets, suitable ends can become to.

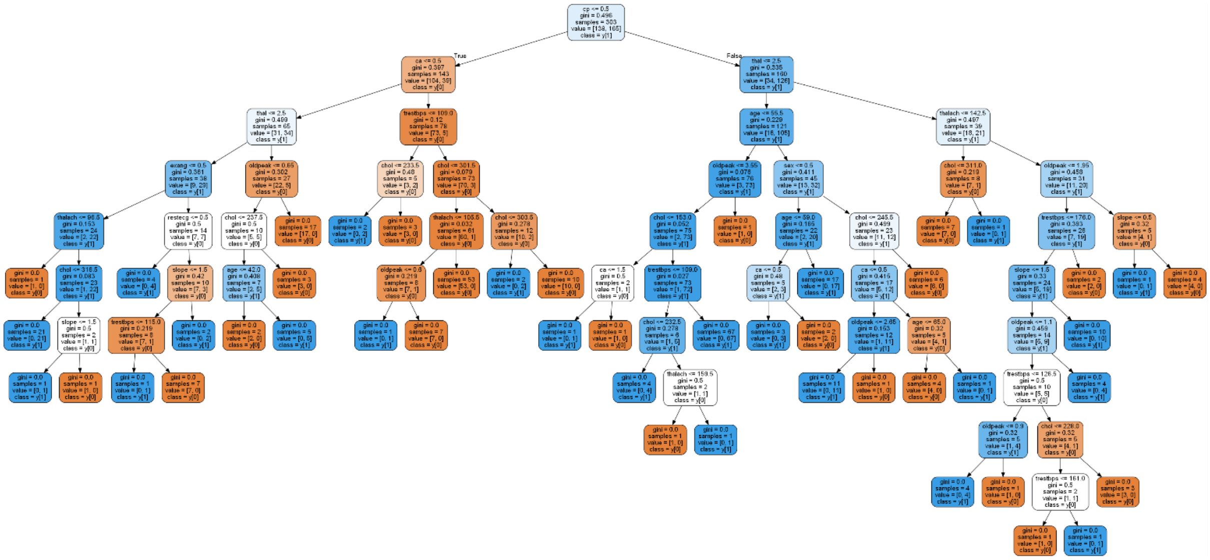


Figure 3.7 Decision Tree

3.5.4 Logistic Regression

Logistic regression is a Boolean classification algorithm. For a given set of independent variables, it is the classification algorithm used to predict binary results. The outcome of the dependent variable is discrete. Logistic regression is named for the function utilized, the Logistic function, at the center of the strategy. Statisticians developed the logistical function, also called the Sigmoid function, to describe the characteristics of population growth in ecology, rising rapidly and maxing out at the environmental carrying capacity. It is an S-shaped curve capable of taking any real-valued number and mapping it to a value between 0 and 1, but never exactly at those limits.

$$1/1+e-x$$

e is the basis of the natural logarithms and x is the transforming value through the logistic function.

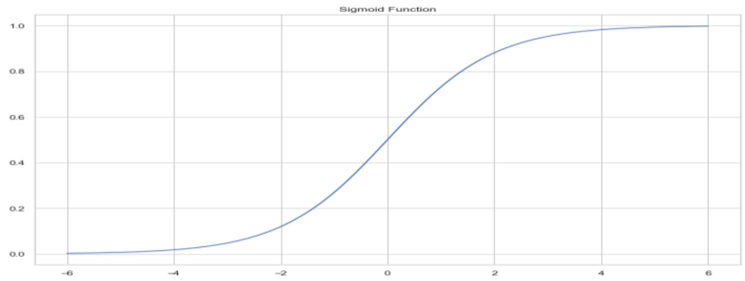


Figure 3.8 Logistic Regression

3.5.5 Random Forest (Ensemble)

Random Forests is an ensemble learning technique for classification and regression techniques. At training time, it develops a number of Decision trees and outputs the class which is the mode of the result classes by individual trees. It also attempts to reduce the problems of high variance and high bias by discovering a natural balance between the two extremes on average.

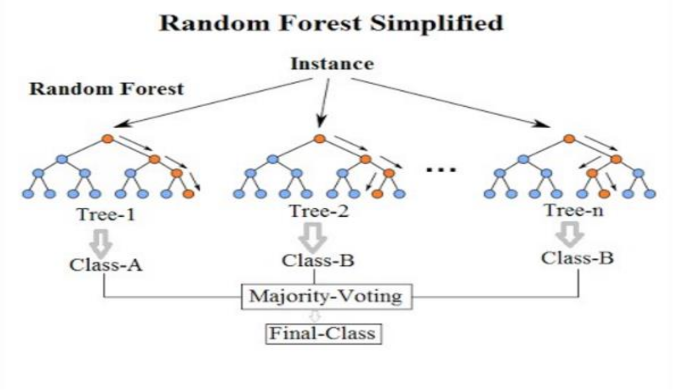


Figure 3.9 Random Forest

3.5.6 Ensemble Model

Ensemble methodologies is a strategy of machine learning which connects numerous basic models to manufacture an appropriate predictive model.

Ada-boost: It concentrates on classification issues and seeks to turn a set of poor classifiers into a powerful one. Comparable to human learning, the algorithm gets to know from past errors by concentrating more on complex issues that previous learning did not get correct.

Bagging: Bootstrap aggregating, otherwise called bagging, is a meta-algorithm for machine learning ensemble designed to improve the soundness and accuracy of the algorithms utilized in statistical classification and regression. It also decreases variance and helps to prevent overfitting.

3.5.7 Gradient Boosting Machine & Extreme Gradient Boosting (GBM & XGB)

It's operates for both classification and regression. Sequentially adding predictors, each of which fits new predictors for residual errors when correcting its predecessor.

Contrast this with AdaBoost: Deflect instance weights at each emphasis.

There are three elements to gradient boosting:

- Loss function to be optimized
- Weak learner to make predictions
- Additive model to add weak learners to minimize the loss function

XG-Boost: One of Kaggle's most loved machine learning algorithms is XG-Boost (Xtreme Gradient Boosting). It can be used for tasks of supervised learning like regression, classification, and ranking. It is built on gradient boosting framework principles and designed to "push the extreme limits of machine computing to provide a scalable, portable and accurate library."

Objective Function: Training Loss + Regularization

3.5.8 Effective Heart Disease Prediction 1

We used Random Forest, Decision Tree, and Logistic Regression in this ensemble model. We combined Decision Tree with Random Forest where Logistic Regression is used as a meta classifier.

3.5.9 Effective Heart Disease Prediction 2

In this ensemble model, first we made Random Forest Model then we fitted the Random Forest model with AdaBoost after that we fitted this combine model into Bagging Classifier.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In the following chapters we will discuss about the results of the guided experiment. We will research and compare the different classifier accuracy and performance. We will show the results in graph and also in tables.

4.2 Experimental Results

Our experiment includes many machine learning algorithms- (Random Forest Classifier. Logistic regression. Naïve Bayes classifier, Decision tree, Support vector machine, Ensemble methods, KNN). The input of our system is the dataset described in Section 3.2 and 3.3. As our work is supervised, the dataset is divided into two set using Hold out method spitting data set:

1. Training Set (80% of data) – 242 data.

2. Test Set (20% of data) – 61 data.

The models are trained using training set and after building model, it is tested on test set.

We obtained the output in terms of confusion matrix as a standard machine learning algorithm.

We measured accuracy, precision, recall with respect to class predictions. The accuracy is measured using the formula in Eq. (9), precision using Eq. (10) and recall using Eq. (11)

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

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

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

The following sections from 4.2.1 to 4.2.4 extensively discusses the results from our study.

4.2.1 Naïve Bayes (Gaussian NB)

After Applying Naïve Bayes (GaussianNB) in the dataset We got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8361

Table 4.1 Classification Report of Naïve Bayes (GaussianNB)

	Precision	**Recall**	**F1 score**	**Support**
0	0.83	0.76	0.79	25
1	0.84	0.89	0.86	36
Micro Avg	0.84	0.84	0.84	61
Macro Avg	0.83	0.82	0.83	61
Weighted Avg	0.84	0.84	0.83	61

Table 4.2 Confusion Matrix of Naïve Bayes (GaussianNB)

	T	**F**
T	19	6
F	4	32

Cross Validation (Average Accuracy) : 0.7976

Accuracy SD: 0.0620

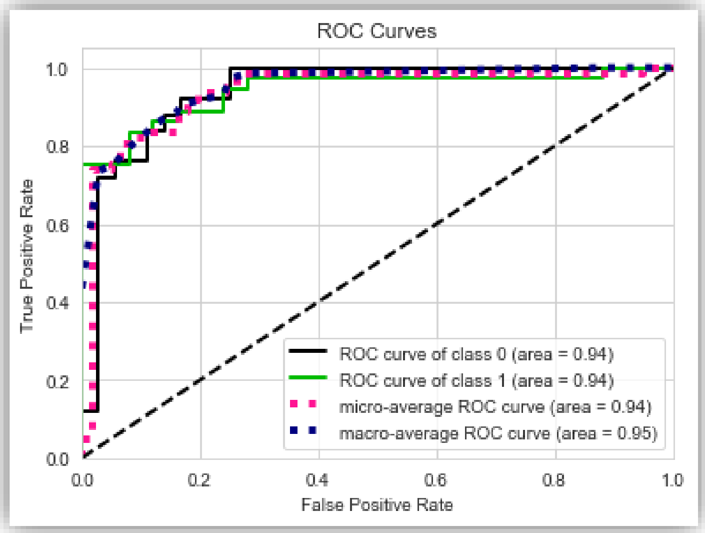


Figure 4.1 Roc Curve of Naïve Bayes (Gaussian NB)

4.2.2 Naïve Bayes (Bernoulli NB)

After Applying Naïve Bayes (Bernoulli NB) in the dataset we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8689

Table 4.3 Classification Report of Naïve Bayes (Bernoulli NB)

	Precision	**Recall**	**F1 score**	**support**
0	0.81	0.88	0.85	25
1	0.91	0.86	0.89	36
Micro Avg	0.87	0.87	0.87	61
Macro Avg	0.86	0.87	0.87	61
Weighted Avg	0.87	0.87	0.87	61

Table 4.4 Confusion Matrix of Naïve Bayes (Bernoulli NB)

	T	**F**
T	22	3
F	5	31

Cross Validation (Average Accuracy): 0.8210

Accuracy SD: 0.0532

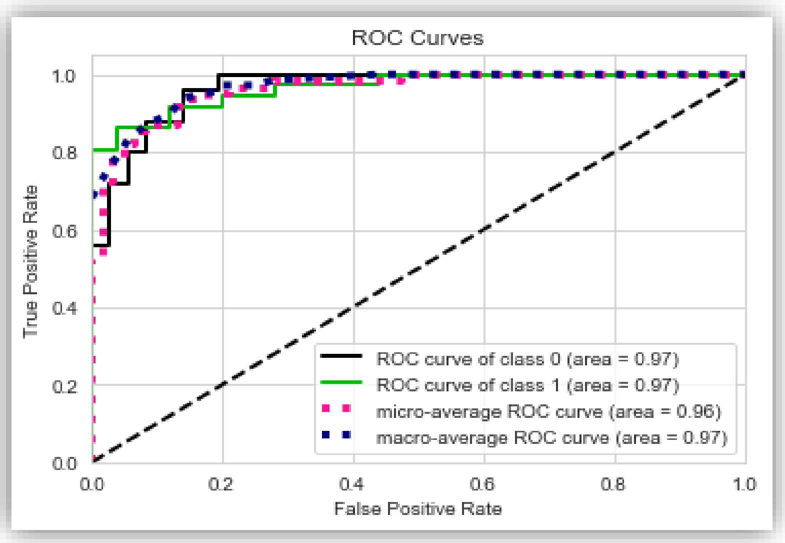


Figure 4.2 Roc Curve of Naïve Bayes (Bernoulli NB)

4.2.3 Support Vector Machine

After Applying Support Vector Machine in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8689

Table 4.5 Classification Report of Support Vector Machine

	Precision	Recall	F1 score	support
0	0.77	0.96	0.86	25
1	0.97	0.81	0.88	36
Micro Avg	0.87	0.87	0.87	61
Macro Avg	0.87	0.88	0.87	61
Weighted Avg	0.89	0.87	0.87	61

Table 4.6 Confusion Matrix of Support Vector Machine

	T	F
T	24	1
F	7	29

Cross Validation (Average Accuracy): 0.5510

Accuracy SD: 0.0165

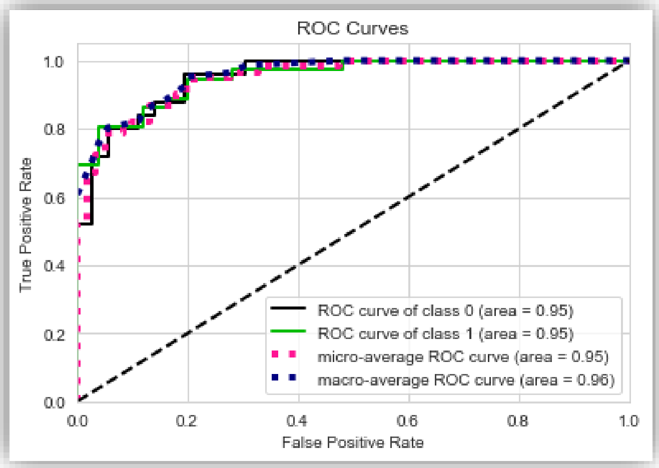


Figure 4.3 ROC Curve of Support Vector Machine

4.2.4 Linear Support Vector Machine (Linear SVM)

After Applying Linear SVM in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.9016

Table 4.7 Classification Report of Linear Support Vector Machine (Linear SVM)

	Precision	Recall	F1 score	support
0	0.85	0.92	0.88	25
1	0.94	0.89	0.91	36
Micro Avg	0.90	0.90	0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61

Table 4.8 Confusion Matrix of Linear Support Vector Machine (Linear SVM)

	T	F
T	23	2
F	4	32

Cross Validation (Average Accuracy): 0.8370

Accuracy SD: 0.00661

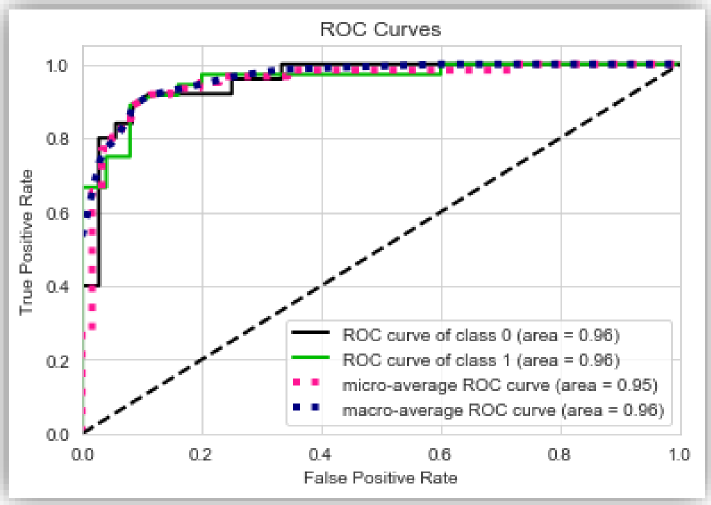


Figure 4.4 ROC Curve Linear Support Vector Machine (Linear SVM)

4.2.5 Polynomial Support Vector Machine (Polynomial SVM)

After Applying Polynomial SVM in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8033

Table 4.9 Classification Report of Polynomial Support Vector Machine (Polynomial SVM)

	Precision	Recall	F1 score	support
0	0.76	0.76	0.76	25
1	0.83	0.83	0.83	36
Micro Avg	0.80	0.80	0.80	61
Macro Avg	0.80	0.80	0.80	61
Weighted Avg	0.80	0.80	0.80	61

Table 4.10 Confusion Matrix of Polynomial Support Vector Machine (Polynomial SVM)

	T	F
T	19	6
F	6	30

Cross Validation (Average Accuracy): 0.7159

Accuracy SD: 0.0660

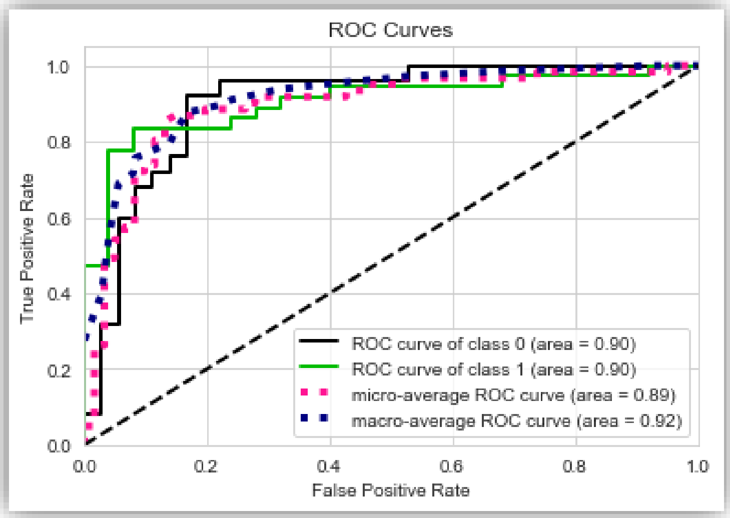


Figure 4.5 ROC Curve Polynomial Support Vector Machine

4.2.6 Radial Basis Function Support Vector Machine (RBF SVM)

After Applying RBF SVM in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8689

Table 4.11 Classification Report of Radial Basis Function Support Vector Machine (RBF SVM)

	Precision	**Recall**	**F1 score**	**support**
0	0.77	0.96	0.86	25
1	0.97	0.81	0.88	36
Micro Avg	0.87	0.87	0.87	61
Macro Avg	0.87	0.88	0.87	61
Weighted Avg	0.89	0.87	0.87	61

Table 4.12 Confusion Matrix of Radial Basis Function Support Vector Machine (RBF SVM)

	T	**F**
T	24	1
F	7	9

Cross Validation (Average Accuracy): 0.5510

Accuracy SD: 0.0165

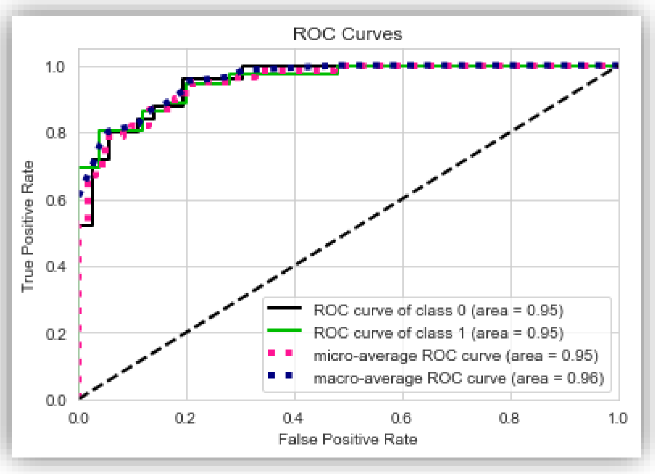


Figure 4.6 ROC Curve Radial Basis Function Support Vector Machine

4.2.7 Grid Search CV On Support Vector Machine

After Applying Grid Search CV on Support Vector Machine in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Best Parameter: {'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}

Accuracy score: 0.9016

Table 4.13 Classification Report of Grid Search CV On Support Vector Machine

	Precision	Recall	F1 score	support
0	0.85	0.92	0.88	25
1	0.94	0.89	0.91	36
Micro Avg	0.90	0.90	0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61

Table 4.14 Confusion Matrix of Grid Search CV On Support Vector Machine

	T	F
T	23	2
F	4	32

Cross Validation (Average Accuracy): 0.7878

Accuracy SD: 0.0577

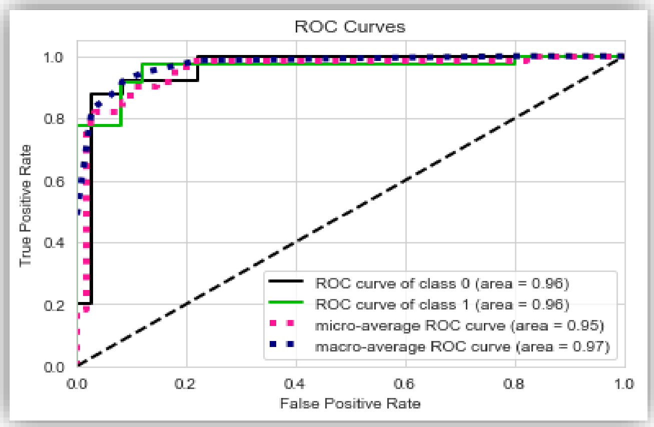


Figure 4.7 ROC Curve Grid Search CV On Support Vector Machine

4.2.8 Decision Tree

After Applying Decision Tree in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8689

Table 4.15 Classification Report of Decision Tree

	Precision	**Recall**	**F1 score**	**support**
0	0.84	0.84	0.84	25
1	0.89	0.89	0.89	36
Micro Avg	0.87	0.87	0.87	61
Macro Avg	0.86	0.86	0.86	61
Weighted Avg	0.87	0.87	0.87	61

Table 4.16 Confusion Matrix of Decision Tree

	T	**F**
T	21	4
F	4	32

Cross Validation (Average Accuracy): 0.7682

Accuracy SD: 0.0678

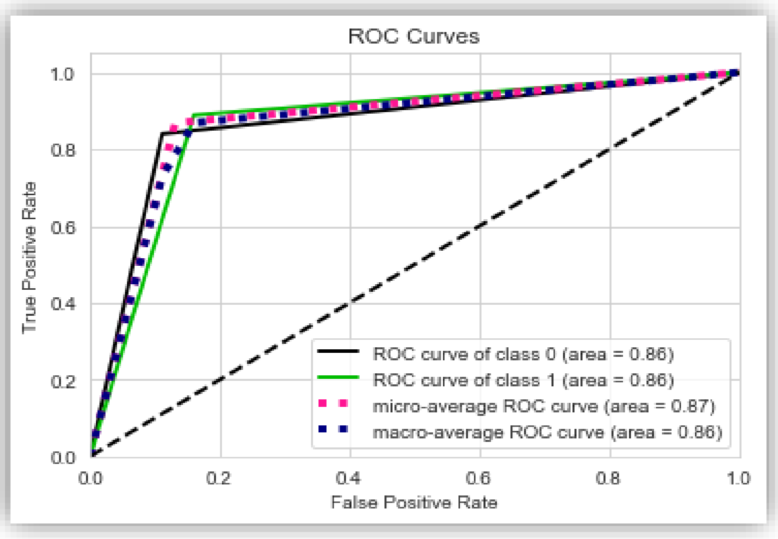


Figure 4.8 ROC Curve Decision Tree

4.2.9 Grid Search CV On Decision Tree

After Applying GridSearch CV on Decision Tree in the dataset we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Best Params: {'min_samples_leaf': 9}

Accuracy score: 0.8197

Table 4.17 Classification Report of Grid Search CV On Decision Tree

	Precision	**Recall**	**F1 score**	**support**
0	0.79	0.76	0.78	25
1	0.84	0.86	0.85	36
Micro Avg	0.82	0.82	0.82	61
Macro Avg	0.81	0.81	0.81	61
Weighted Avg	0.82	0.82	0.82	61

Table 4.18 Confusion Matrix of Grid Search CV On Decision Tree

		T	**F**
T		19	6
F		5	31

Cross Validation (Average Accuracy): 0.7975

Accuracy SD: 0.0603

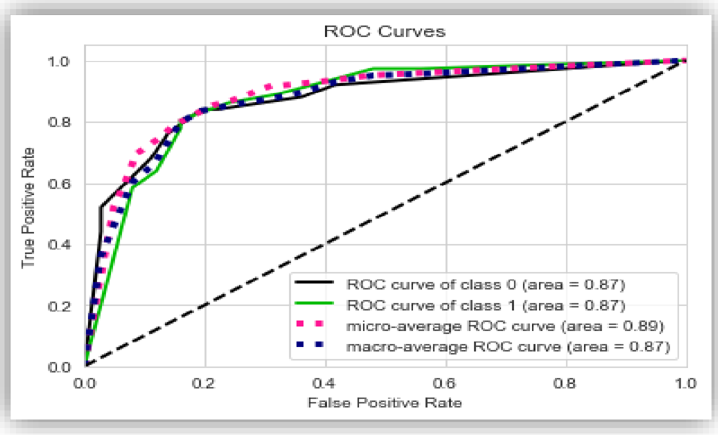


Figure 4.9 ROC Curve Grid Search CV On Decision Tree

4.2.10 Logistic Regression

After Applying Logistic Regression in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8689

Table 4.19 Classification Report of Logistic Regression

	Precision	Recall	F1 score	support
0	0.81	0.88	0.85	25
1	0.91	0.86	0.89	36
Micro Avg	0.87	0.87	0.87	61
Macro Avg	0.86	0.87	0.87	61
Weighted Avg	0.87	0.87	0.87	61

Table 4.20 Confusion Matrix of Logistic Regression

	T	F
T	22	3
F	5	31

Cross Validation (Average Accuracy): 0.8237

Accuracy SD: 0.0676

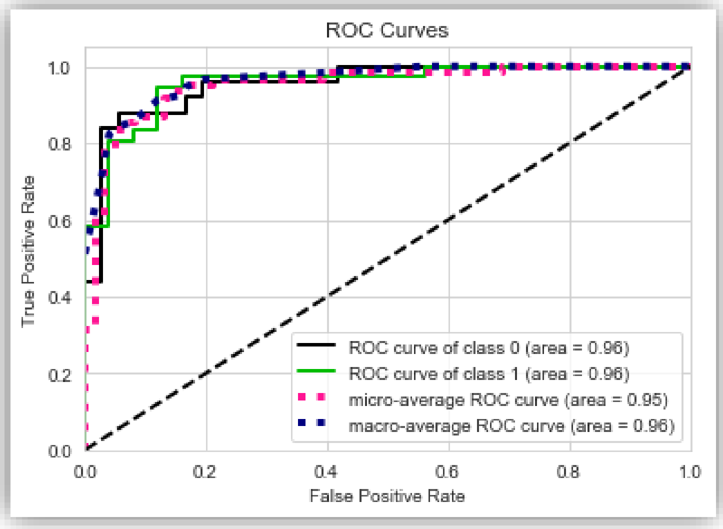


Figure 4.10 ROC Curve Logistic Regression

4.2.11 Grid Search CV On Logistic Regression

After Applying Grid Search CV on Logistic Regression in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Best Params: {'C': 0.10001}
Accuracy score: 0.9016

Table 4.21 Classification Report of Grid Search CV On Logistic Regression

	Precision	**Recall**	**F1 score**	**support**
0	0.91	0.84	0.87	25
1	0.89	0.94	0.92	36
Micro Avg	0.90	0.90	0.90	61
Macro Avg	0.90	0.89	0.90	61
Weighted Avg	0.90	0.90	0.90	61

Table 4.22 Confusion Matrix of Grid Search CV On Logistic Regression

	T	**F**
T	21	4
F	2	34

Cross Validation (Average Accuracy): 0.8303

Accuracy SD: 0.0699

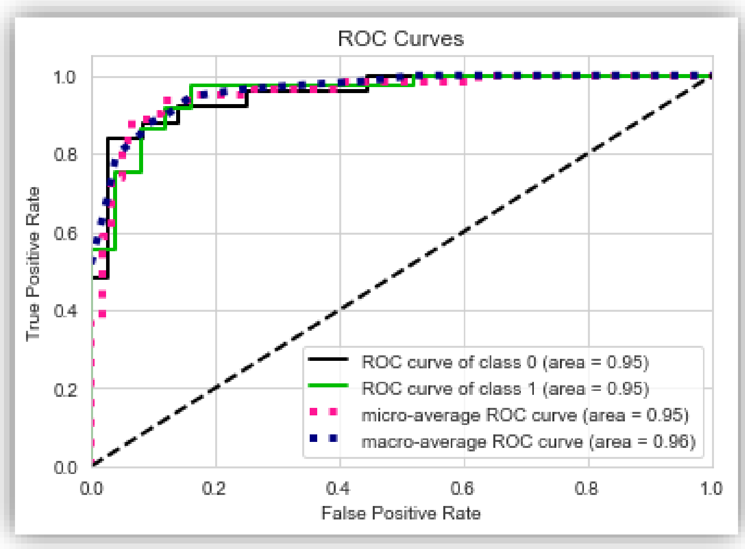


Figure 4.11 ROC Curve Grid Search CV On Logistic Regression

4.2.12 Random Forest

After Applying Random Forest in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8083

Table 4.23 Classification Report of Random Forest

	Precision	**Recall**	**F1 score**	**support**
0	0.72	0.84	0.78	25
1	0.88	0.78	0.82	36
Micro Avg	0.80	0.80	0.80	61
Macro Avg	0.80	0.81	0.80	61
Weighted Avg	0.81	0.80	0.80	61

Table 4.24 Confusion Matrix of Random Forest

	T	**F**
T	21	4
F	8	28

Cross Validation (Average Accuracy): 0.8045

Accuracy SD: 0.0832

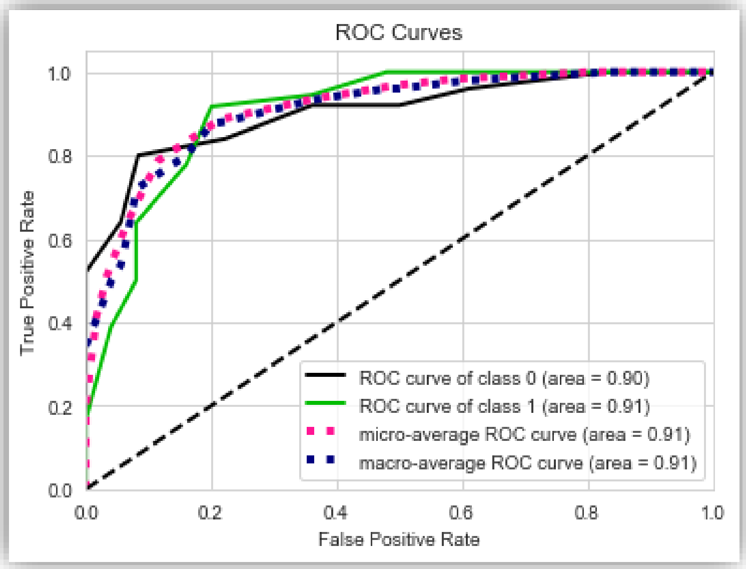


Figure 4.12 ROC Curve Random Forest

4.2.13 Grid Search CV On Random Forest

After Applying Grid Search CV on Random Forest in the dataset we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Best Params: {'bootstrap': True, 'criterion': 'gini', 'max_depth': 3,'min_samples_leaf': 1, 'min_samples_split': 10}

Accuracy score: 0.9180

Table 4.25 Classification Report of Grid Search CV On Random Forest

	Precision	Recall	F1 score	support
0	0.92	0.88	0.90	25
1	0.92	0.94	0.93	36
Micro Avg	0.92	0.92	0.92	61
Macro Avg	0.92	0.92	0.91	61
Weighted Avg	0.92	0.92	0.92	61

Table 4.26 Confusion Matrix of Grid Search CV On Random Forest

	T	F
T	22	3
F	2	34

Cross Validation (Average Accuracy): 0.8282

Accuracy SD: 0.0366

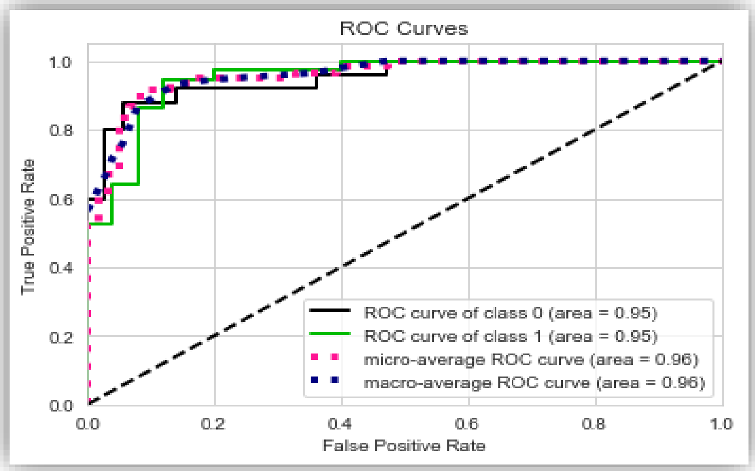


Figure 4.13 ROC Curve Grid Search CV On Random Forest

4.2.14 Extra-Trees (Extremely Randomized Tree)

After Applying Extremely Randomized Tree in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8033

Table 4.27 Classification Report of Extra-Trees (Extremely Randomized Tree)

	Precision	**Recall**	**F1 score**	**support**
0	0.71	0.88	0.79	25
1	0.90	0.75	0.82	36
Micro Avg	0.80	0.80	0.80	61
Macro Avg	0.80	0.81	0.80	61
Weighted Avg	0.82	0.80	0.80	61

Table 4.28 Confusion Matrix of Extra-Trees (Extremely Randomized Tree)

	T	**F**
T	22	3
F	9	27

Cross Validation (Average Accuracy): 0.8013

Accuracy SD: 0.0759

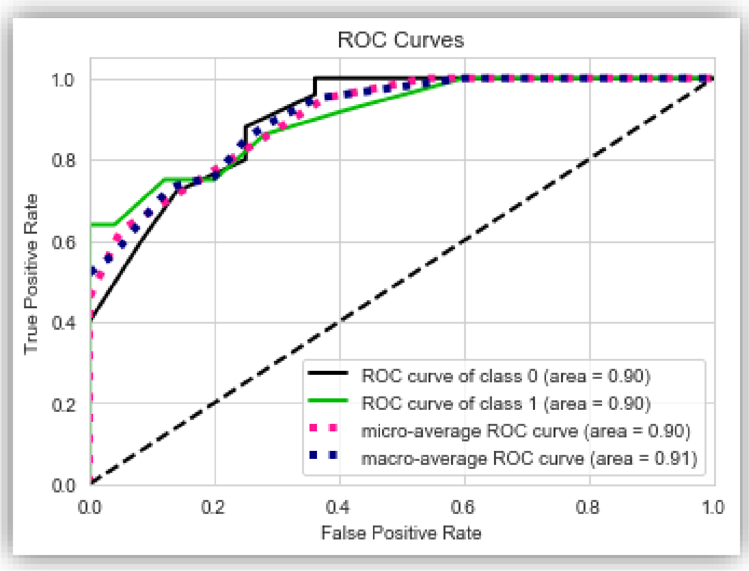


Figure 4.14 ROC Curve Extra-Trees (Extremely Randomized Tree)

4.2.15 AdaBoost (Adaptive Boosting)

After Applying AdaBoost in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8361

Table 4.29 Classification Report of AdaBoost (Adaptive Boosting)

	Precision	Recall	F1 score	support
0	0.83	0.76	0.79	25
1	0.84	0.89	0.86	36
Micro Avg	0.84	0.84	0.84	61
Macro Avg	0.83	0.82	0.83	61
Weighted Avg	0.84	0.84	0.83	61

Table 4.30 Confusion Matrix of AdaBoost (Adaptive Boosting)

	T	F
T	19	6
F	4	32

Cross Validation (Average Accuracy): 0.8100

Accuracy SD: 0.0855

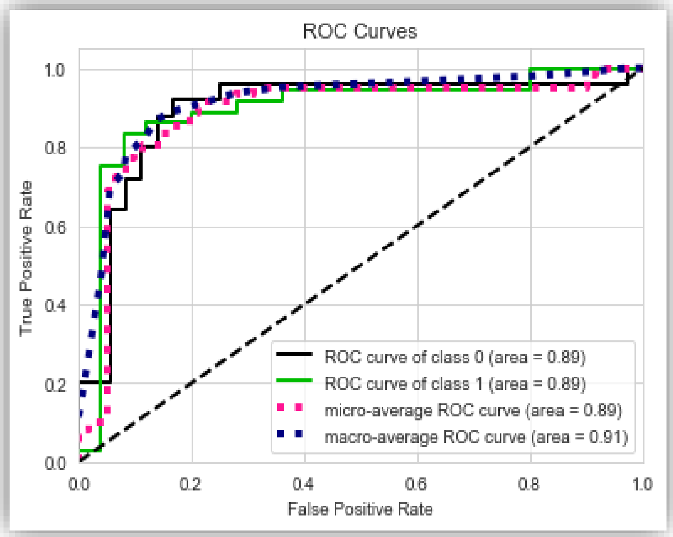


Figure 4.15 ROC Curve AdaBoost

4.2.16 AdaBoost With Random Forest

After Applying AdaBoost with Random Forest in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8689

Table 4.31 Classification Report of AdaBoost With Random Forest

	Precision	**Recall**	**F1 score**	**support**
0	0.81	0.88	0.85	25
1	0.91	0.86	0.89	36
Micro Avg	0.87	0.87	0.87	61
Macro Avg	0.86	0.87	0.87	61
Weighted Avg	0.87	0.87	0.87	61

Table 4.32 Confusion Matrix of AdaBoost with Random Forest

	T	**F**
T	22	3
F	5	31

Cross Validation (Average Accuracy): 0.8076

Accuracy SD: 0.0585

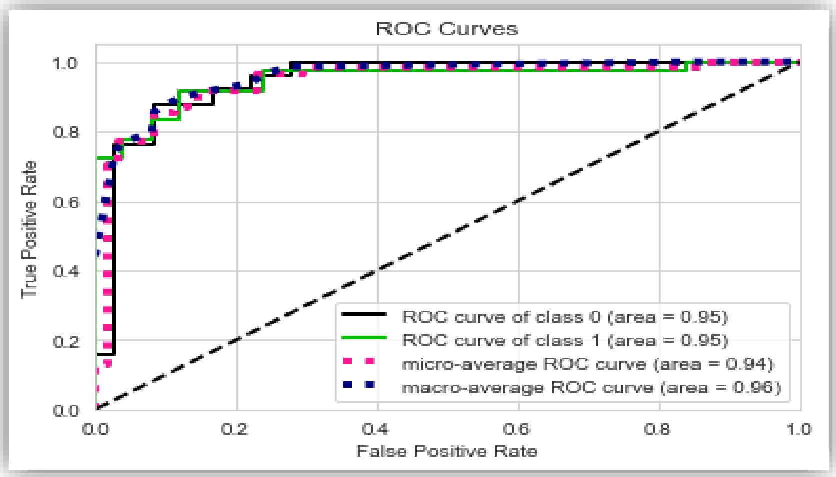


Figure 4.16 ROC Curve AdaBoost With Random Forest

4.2.17 Bagging (Bootstrap Aggregating)

After Applying Bootstrap Aggregating in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8689

Table 4.33 Classification Report of Bagging (Bootstrap Aggregating)

	Precision	Recall	F1 score	support
0	0.84	0.84	0.84	25
1	0.89	0.89	0.89	36
Micro Avg	0.87	0.87	0.87	61
Macro Avg	0.86	0.86	0.86	61
Weighted Avg	0.87	0.87	0.87	61

Table 4.34 Confusion Matrix of Bagging (Bootstrap Aggregating)

	T	F
T	21	4
F	4	32

Cross Validation (Average Accuracy): 0.8037

Accuracy SD: 0.0746

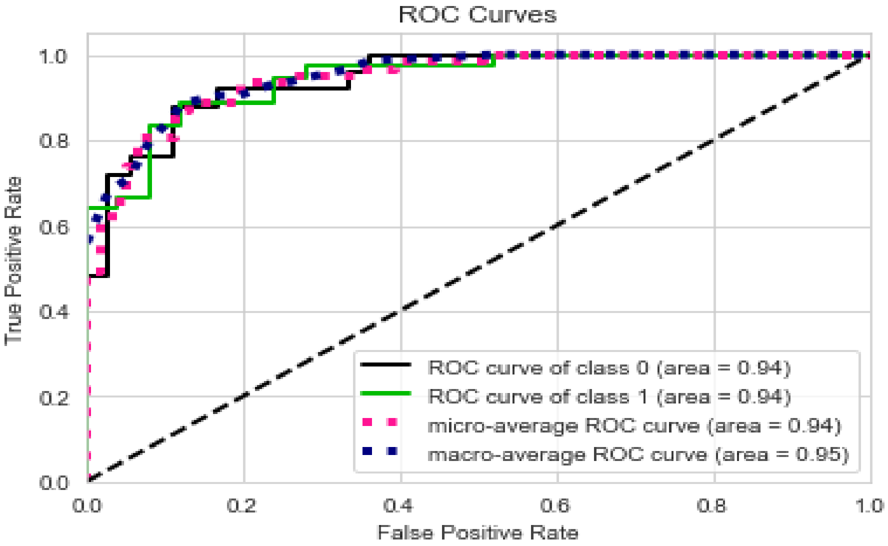


Figure 4.17 ROC Curve Bagging

4.2.18 Gradient Boosting Machine

After Applying Gradient Boosting Machine in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8852

Table 4.35 Classification Report of Gradient Boosting Machine

	Precision	**Recall**	**F1 score**	**support**
0	0.85	0.88	0.86	25
1	0.91	0.89	0.90	36
Micro Avg	0.89	0.89	0.89	61
Macro Avg	0.88	0.88	0.88	61
Weighted Avg	0.89	0.89	0.89	61

Table 4.36 Confusion Matrix of Gradient Boosting Machine

	T	**F**
T	22	3
F	4	32

Cross Validation (Average Accuracy): 0.8137

Accuracy SD: 0.0781

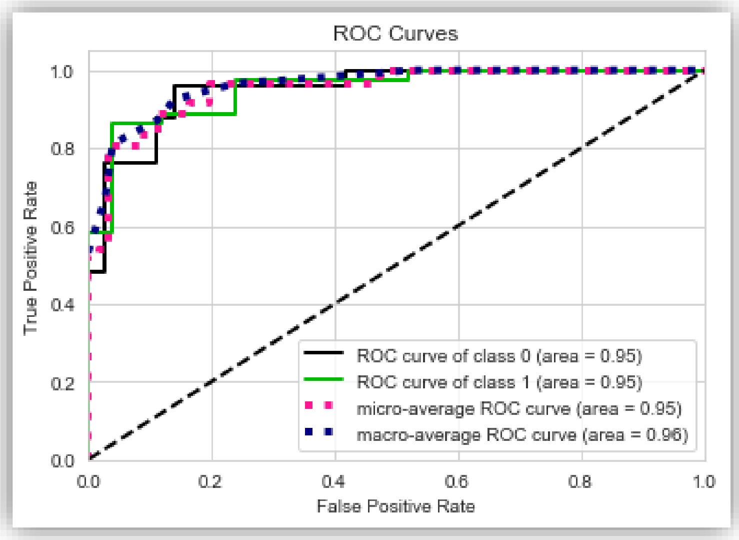


Figure 4.18 ROC Curve Gradient Boosting Machine

4.2.19 Extreme Gradient Boosting

After Applying Extreme Gradient Boosting in the dataset, we got the Following Test Accuracy Score, Classification Report, Confusion Matrix, Cross Validation (Average Accuracy) and Standard Deviation.

Accuracy score: 0.8525

Table 4.37 Classification Report of Extreme Gradient Boosting

	Precision	**Recall**	**F1 score**	**support**
0	0.83	0.80	0.82	25
1	0.86	0.89	0.88	36
Micro Avg	0.85	0.85	0.85	61
Macro Avg	0.85	0.84	0.85	61
Weighted Avg	0.85	0.85	0.85	61

Table 4.38 Confusion Matrix of Extreme Gradient Boosting

	T	**F**
T	20	5
F	4	32

Cross Validation (Average Accuracy): 0.7874

Accuracy SD: 0.0775

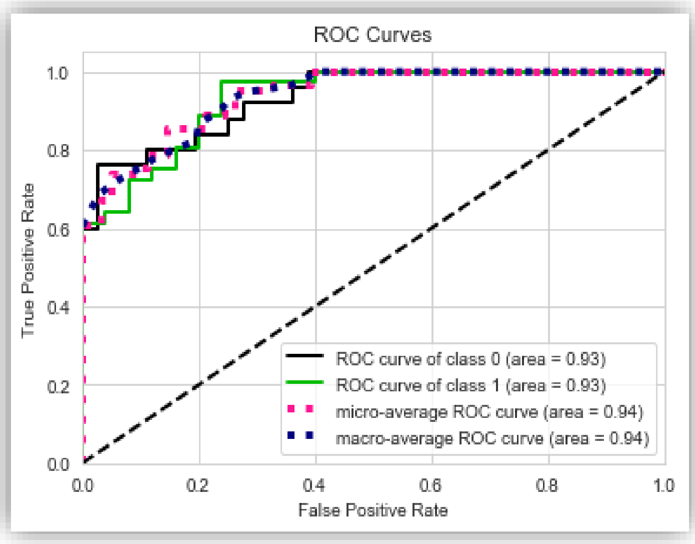
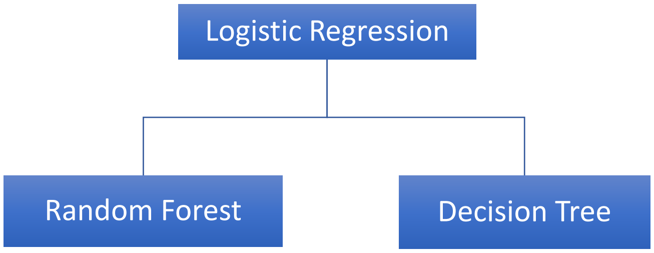


Figure 4.19 ROC Curve of Extreme Gradient Boosting Machine

4.2.20 Ensemble of Ensemble (Effective Heart Disease Prediction 1)

This is our proposed model 1 for Effective heart disease prediction using machine learning. In This Ensemble of Ensemble Model first we create two model using Random Forest and Decision Tree. Then We Combine Both Model and create combined model. Here we use Logistic Regression as a meta classifier and we fitted this meta classifier with the combined model we first created.



Accuracy score: 0.836

Table 4.39 Classification Report of Ensemble of Ensemble (Effective Heart Disease Prediction 1)

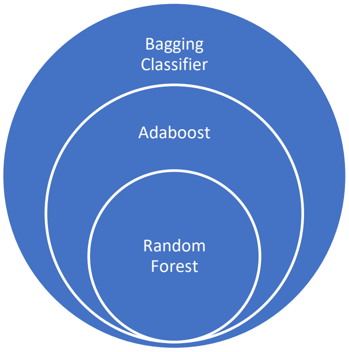
	Precision	**Recall**	**F1 score**	**support**
0	0.78	0.84	0.81	25
1	0.88	0.83	0.86	36
Micro Avg	0.84	0.84	0.84	61
Macro Avg	0.83	0.84	0.83	61
Weighted Avg	0.84	0.84	0.84	61

Table 4.40 Confusion Matrix of Ensemble of Ensemble (Effective Heart Disease Prediction 1)

	T	**F**
T	21	4
F	6	30

4.2.21 Single Classifier (Effective Heart Disease Prediction 2)

This is our proposed model 2 for Effective heart disease prediction using machine learning. In this ensemble model first, we created Random Forest Model then we fitted the Random Forest model with AdaBoost after that we fitted this combined model into Bagging Classifier.



Accuracy score: 0.9344

Table 4.41 Classification Report of Single Classifier (Effective Heart Disease Prediction 2)

	Precision	**Recall**	**F1 score**	**support**
0	0.92	0.92	0.92	25
1	0.94	0.94	0.94	36
Micro Avg	0.93	0.93	0.93	61
Macro Avg	0.93	0.93	0.93	61
Weighted Avg	0.93	0.93	0.93	61

Table 4.42 Confusion Matrix of Single Classifier (Effective Heart Disease Prediction 2)

	T	**F**
T	23	2
F	2	34

Cross Validation (Average Accuracy): 0.8270

Accuracy SD: 0.0664

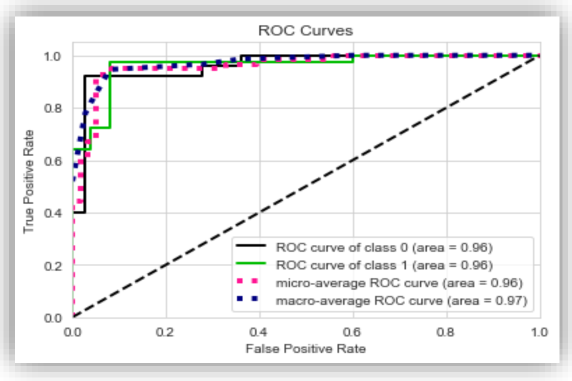


Figure 4.20 ROC Curve Effective Heart Disease Prediction 2

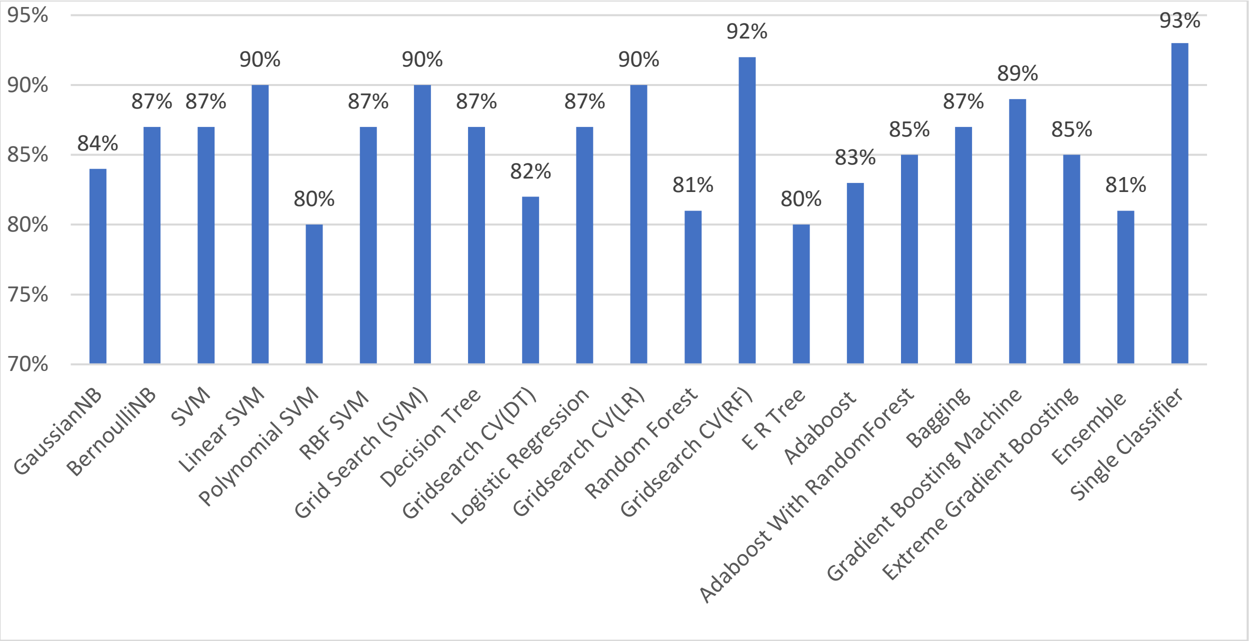


Figure 5.1 Accuracy of all classifiers

4.3 Potential Future Improvement

The study shows that an automated system can be implemented in the clinical diagnosis of Heart Disease Prediction. With more training data and more efficient algorithm this can be a real-life implementation for clinical diagnosis in Bangladesh. Bangladesh medical system can accumulate more clinical data in an organized and structured way where each patient's medical data can be secured which can be later be used in many different studies for other diseases and also incorporate with the existing system for automated classification and detection of those diseases.

4.4 Summary

It tends to be inferred that, AI and information mining system can assume an essential job in anticipating Cardiovascular sickness or Heart related malady. In this exploration work, among every one of the calculations a few calculations performed amazingly well like Naïve Bayes, Linear SVM, Logistic Regression, Random Forest utilizing Grid Search CV and troupes show like Bagging, Boosting and our very own two group demonstrate (Effective Heart Disease Prediction System 1, Effective Heart Disease Prediction System 2). There is still a great deal of extension to improve the general execution.

CHAPTER 5

CONCLUSION AND FUTURE IMPLICATION

5.1 Summary of the Study

This section summarizes the results of our study. Figure 5.1 shows the accuracy obtained from the different models. It can be concluded that our proposed model 2(Effective heart disease prediction 1) gives the most desirable **accuracy of 93.44%.**

5.2 Conclusions

We observed that, we have used 21 different machine learning algorithms among which our proposed model 2 (Effective heart disease prediction 2) had the highest level of accuracy. Although the other algorithms provided very close and accurate result compared to our proposed model 2((Effective heart disease prediction 2)

5.3 Recommendations

Considering the study, it will give us more accuracy if we find more clinical data which are perfectly organized.

5.4 Implication for Further Study

For various other diseases we can use similar techniques and more data on other clinical health problems should be accumulated in order for similar studies. Further studies in the field required for achieving clinical accuracy and reliability.

Web based or android based application can be developed with permission and supervision from appropriate authorities for public usage of the platform

REFERENCES:

[1] World Health Organization Article at https://www.who.int/cardiovascular_diseases/en/

[2] B. L Deekshatulua Priti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm "M.Akhil jabbar* International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[3] Heart Disease prediction using Machine learning and Data Mining Technique By Jaymin Patel, Prof TejalUpadhyay, Dr. Samir Patel.

[4] Brac University Institutional Repository (Effectiveness of data mining in predicting heart diseases at http://dspace.bracu.ac.bd/xmlui/handle/10361/8915

[5] https://health.howstuffworks.com/diseases-conditions/cardiovascular/heart/heart-attack.htm

[6] Heart Disease Facts https://www.cdc.gov/heartdisease/facts.htm

[7] Bangladesh : Coronary Heart Disease https://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease

[8] According to a survey on Dhaka tribune at https://www.dhakatribune.com/health/2018/10/02/3-out-of-4-bd-individuals-run-the-risk-of-developing-cardiac-diseases

[9] According to The Financial Express News at https://thefinancialexpress.com.bd/health/three-out-of-four-bangladeshis-at-risk-of-heart-disease-1538398045

[10] Mayo Clinic https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

[11] Medicine Net Article on heart disease definition at https://www.medicinenet.com/script/main/art.asp?articlekey=31193

[12] Heart Disease prediction using Machine learning and Data Mining Technique by Jaymin Patel, Prof TejalUpadhyay, Dr. Samir Patel.

[13] Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms by Asst. Professor, Computer Science & Eng. Orissa Engineering College, Bhubaneswar, Odisha – India.

[14] Ashok Kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation", Springer, 17 September 2016

[15] Sujata Joshi and Mydhili K.Nair,"Prediction of Heart Disease Using Classification Based Data Mining Techniques", Springer India 2015, volume2.

[16] Sellappan Palaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[17] Shadab Adam Pattekari and Asma Parveen," PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES",International Journal of Advanced Computer and Mathematical SciencesISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[18] Wikipedia, https://en.wikipedia.org/wiki/Machine_learning. [Online]. Last Accessed: 1 April, 2019.

[19] Coursera, https://www.coursera.org/learn/machine-learning. [Online]. Last Accessed: 129 April, 2019.

[20] Introduction to machine learning, http://www.zarantech.com/blog/an-introduction-to-machine-learningwhy-it-matters.[Online] Last Accessed: 1 April 2019.

[21] UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Turnitin Originality Report

Heart Disease Prediction Using Data Mining by Mahmudul Hasan

From heart disease prediction (Spring2019)

Processed on 20-Apr-2019 14:32 +06
ID: 1115909454
Word Count: 6429

Similarity Index	Similarity by Source	
23%	Internet Sources:	11%
	Publications:	11%
	Student Papers:	16%

sources:

1 3% match (publications)
Abdullah Iliyasu, Chastine Fatichah. "A Quantum Hybrid PSO Combined with Fuzzy k-NN Approach to Feature Selection and Cell Classification in Cervical Cancer Detection", Sensors, 2017

2 2% match (student papers from 13-Apr-2019)
Submitted to Daffodil International University on 2019-04-13

3 1% match (Internet from 17-Jan-2019)
http://csjournals.com/IJCSC/PDF7-1/18.%20Tejpal.pdf

4 1% match (Internet from 21-Mar-2018)
https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

5 1% match (publications)
V V. Ramalingam, Ayantan Dandapath, M Karthik Raja. "Heart disease prediction using machine learning techniques : a survey", International Journal of Engineering & Technology, 2018

6 1% match (student papers from 16-Sep-2015)
Submitted to Colorado Technical University Online on 2015-09-16

7 1% match (student papers from 30-Mar-2018)
Submitted to Swinburne University of Technology on 2018-03-30

8 1% match (Internet from 06-Apr-2018)
https://www.safaribooksonline.com/library/view/ensemble-machine-learning/9781788297752/ec772372-bc25-4799-9a0b-578e38b25b5b.xhtml

9 1% match (student papers from 30-Jan-2017)
Submitted to University of East London on 2017-01-30

10 1% match (student papers from 26-Feb-2018)
Submitted to Shanghai World Foreign Language Middle School on 2018-02-26

11 < 1% match (student papers from 23-Sep-2016)
Submitted to University of Wales Institute, Cardiff on 2016-09-23

12 < 1% match (student papers from 25-Mar-2019)
Submitted to Georgia Institute of Technology Main Campus on 2019-03-25

13 < 1% match (Internet from 25-Feb-2019)
https://www.webmd.com/heart-disease/guide/heart-disease-symptoms

14 < 1% match (student papers from 05-Mar-2019)
Submitted to University College London on 2019-03-05

15 < 1% match (Internet from 04-Oct-2018)
https://arxiv.org/pdf/1508.02061.pdf

16 < 1% match (student papers from 21-Feb-2019)
Submitted to University College London on 2019-02-21

17 < 1% match (student papers from 26-Jun-2018)
Submitted to Higher Education Commission Pakistan on 2018-06-26

18 < 1% match (Internet from 25-Mar-2019)
https://acadpubl.eu/jsi/2018-118-7-9/articles/8/22.pdf

19 < 1% match (student papers from 31-Mar-2019)
Submitted to University of Sydney on 2019-03-31

20 < 1% match (Internet from 03-Jul-2018)
https://www.inderscienceonline.com/doi/full/10.1504/IJMPT.2014.062934

21 < 1% match (student papers from 26-Feb-2017)
Submitted to Fiji National University on 2017-02-26

22 < 1% match (student papers from 04-Dec-2012)
Submitted to Colorado Technical University Online on 2012-12-04

23 < 1% match (Internet from 25-Feb-2017)
https://www.rroij.com/open-access/a-student-note-on-cardiovascular-diseases-.php?aid=78103

24 < 1% match (publications)
Dipanjan Sarkar, Raghav Bali, Tushar Sharma. "Practical Machine Learning with Python", Springer Nature, 2018

25 < 1% match (student papers from 02-Nov-2018)
Submitted to University of Sydney on 2018-11-02

26 < 1% match (publications)
Qian Xu, Jianping Hua, Zixiang Xiong, Edward Suh, Edward Dougherty. "Conditional confidence intervals for the true classification error", 2006 IEEE International Workshop on Genomic Signal Processing and Statistics, 2006

27 < 1% match (student papers from 16-Dec-2015)
Submitted to Colorado Technical University Online on 2015-12-16

28 < 1% match (student papers from 14-Sep-2018)
Submitted to University of Lancaster on 2018-09-14

29 < 1% match (Internet from 03-Dec-2016)
https://www.coursehero.com/file/12963934/papers/

30 < 1% match (student papers from 02-Nov-2015)
Submitted to VIT University on 2015-11-02

31 < 1% match (Internet from 05-Mar-2019)
http://assisiaruvithura.blogspot.com/2013/10/life-style-diseases.html

32 < 1% match (publications)
Maria Fernanda Caropreso, Stan Matwin. "Chapter 28 Beyond the Bag of Words: A Text Representation for Sentence Selection", Springer Nature, 2006

33 < 1% match (publications)
Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques", 2008 IEEE/ACS International Conference on Computer Systems and Applications, 03/2008

34 < 1% match (student papers from 19-Aug-2017)
Submitted to Higher Education Commission Pakistan on 2017-08-19

35 < 1% match (student papers from 31-Aug-2018)
Submitted to CTI Education Group on 2018-08-31

36 < 1% match (student papers from 01-Dec-2017)
Submitted to Heriot-Watt University on 2017-12-01

37 < 1% match (Internet from 13-Apr-2019)
https://touchofcarepcp.com/analyzing-the-heart-diseases-2/

38 < 1% match (publications)
Kumar G Dinesh, K Arumugaraj, Kumar D Santhosh, V Mareeswari. "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018

39 < 1% match (student papers from 01-Oct-2018)
Submitted to St. Petersburg College on 2018-10-01

40 < 1% match (student papers from 07-Apr-2019)
Submitted to The British College on 2019-04-07

41 < 1% match (Internet from 28-Oct-2017)
http://crpit.com/confpapers/CRPITV121Shouman.pdf

42 < 1% match (student papers from 07-May-2017)
Submitted to University of South Australia on 2017-05-07

43 < 1% match (student papers from 04-Apr-2019)
Submitted to University of Sydney on 2019-04-04

44 < 1% match (Internet from 20-Oct-2015)
http://facweb.cs.depaul.edu/mobasher/classes/CSC478/Notes/IPython%20Notebook%20-%20Titanic%20-%20Feature-and-Model-Selection.html

45 < 1% match (publications)
Rashmi G Saboji. "A scalable solution for heart disease prediction using classification mining technique", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017

46 < 1% match (student papers from 15-Mar-2010)
Submitted to Colorado Technical University Online on 2010-03-15

47 < 1% match (Internet from 05-Apr-2019)
https://machinelearningmastery.com/logistic-regression-for-machine-learning/

48 < 1% match (Internet from 27-Aug-2014)
http://www.provena.org/stmarys/body.cfm?xyzpdqabc=0&id=147&action=detail&ref=4322

49 < 1% match (Internet from 05-Apr-2019)
https://wwwnc.cdc.gov/eid/article/15/8/08-1237_article

50 < 1% match (Internet from 20-Nov-2018)
https://baadalsg.inflibnet.ac.in/bitstream/10603/91528/14/14_chapter4.pdf

51 < 1% match (publications)
"Scalable Information Systems", Springer Nature America, Inc, 2015

52 < 1% match (publications)
N. ISU. "A Method for Reinforcing Noun Countability Prediction", IEICE Transactions on Information and Systems, 12/01/2007

53 < 1% match (Internet from 23-Nov-2018)
http://ijarcsms.com/docs/paper/volume3/issue8/V3I8-0001.pdf

54 < 1% match (Internet from 02-Feb-2018)
https://lib.dr.iastate.edu/cgi/viewcontent.cgi?amp=&article=7219&context=etd

55 < 1% match (Internet from 14-Apr-2019)

http://www.ijrst.org/public/uploads/author_manuscripts/IJRST_20180410023226_8_04_Published%206456.pdf

56 < 1% match (Internet from 13-Apr-2019)
https://www.grin.com/document/204390

57 < 1% match (publications)
S. R. Bhagya Shree, H. S. Sheshadri. "An initial investigation in the diagnosis of Alzheimer's disease using various classification techniques", 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014

58 < 1% match (publications)
Divya Jain, Vijendra Singh. "Utilization of Data Mining Classification Approach for Disease Prediction: A Survey", International Journal of Education and Management Engineering, 2016

59 < 1% match (Internet from 08-Jun-2017)
http://www.viva-technology.org/New/wp-content/uploads/2016/03/NCRENB-16-Proceedings.pdf

60 < 1% match (Internet from 10-Aug-2018)

http://www.ijaerd.co.in/papers/finished_papers/A%20Comparative%20Study%20of%20Multilayer%20Perceptron,%20Radial%20Basis%20Fu
12997.pdf

61 < 1% match (Internet from 12-Sep-2018)

http://www.ijritcc.org/download/browse/Volume_4_Issues/June_16_Volume_4_Issue_6/1467260373_29-06-2016.pdf

62 < 1% match (Internet from 12-Aug-2018)
https://tituel.org/Papers/FestschriftMitTitel.pdf

63 < 1% match (Internet from 02-Dec-2018)
https://www.datavedas.com/classification-problems-in-python/

64 < 1% match (Internet from 22-Oct-2015)
http://amsdottorato.unibo.it/7058/1/Pasolini_Roberto_tesi.pdf

65 < 1% match (Internet from 21-Jan-2019)
https://edoc.site/buku-data-mining-libre-pdf-free.html

66 < 1% match (Internet from 31-Dec-2016)
https://pdfs.semanticscholar.org/6cdb/dbee4fd5e749288a402f98dd1310889d3e76.pdf

67 < 1% match (Internet from 28-Mar-2016)
http://repository.hkbu.edu.hk/cgi/viewcontent.cgi?amp=&article=1086&context=etd_oa

68 < 1% match (publications)
"Mining Intelligence and Knowledge Exploration", Springer Nature, 2013

69 < 1% match (student papers from 31-Oct-2018)
Submitted to College of Engineering Trivandrum on 2018-10-31

paper text:

56**CHAPTER 1 INTRODUCTION 1.1 Introduction** In **the human body,**

the heart plays a pivotal role. The heart is a pump circulating blood across the whole body. It beats 60-70 times in one minute on average. If the heart stops functioning normally, our body's other organs will stop working.

18**Heart disease is one of the** world's **most common diseases.**