A Machine Learning-Based Technique for Predicting Heart Disease

BY

Mahmudur Rahman Nahin
ID: 191-15-2589
AND

Sadikuzzaman Shawon
ID: 191-15-2419

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

Supervised By

Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Al Amin Biswas
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/internship titled **"A Machine Learning-Based Technique for Predicting Heart Disease"**, submitted by **Mahmudur Rahman Nahin**(191-15-2589) and **Sadikuzzaman Shawon**(191-15-2419) to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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 *04/02/2023*.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Subhenur Latif
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Sabab Zulfiker
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

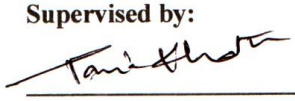
External Examiner

Dr. Md. Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

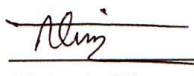
We hereby declare that this project has been done by us under the supervision of **Tania Khatun (Assistant Professor), Lecturer (Senior Scale), Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Al Amin Biswas
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Mahmudur Rahman Nahin
ID: -191-15-2589
Department of CSE
Daffodil International University

Sadikuzzaman Shawon
ID: -191-15-2419
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing in making us possible to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Tania Khatun (Assistant Professor)**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance continual encouragement, constant and energetic supervision, constructive criticism valuable advice r and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head of the Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course mate at 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

Physical diseases including heart disease have been on the rise recently. The subject is well-known in the modern world. The majority of individuals have an issue with heart disease. The discrepancies between the normal and afflicted diagnosis report ratios serve as a gauge of the condition. Heart illness is a condition that has been the subject of several investigations in the past. We have identified a few excellent chances to develop the methodology. We suggest employing efficient algorithm models to forecast dangers and raise early awareness. Our suggested approach is suited for straightforward heart disease predictions and is simple to apply in the actual world. The Kaggle website hosted the dataset. In our model, we have implemented some different classifiers named Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Support Vector Classifier (SVC), Adaboost Classifier (ABC), Naïve Bayes (NB), Decision Tree (DT) algorithms. Random Forest (RF) given an accuracy of 90.22%, Logistic Regression (LR) given accuracy of 89.67%, Gradient Boosting (GB) given accuracy of 89.67%, Support Vector Classifier (SVC) given accuracy of 91.85%, Adaboost Classifier (ABC) given the accuracy 91.30%, Naïve Bayes (NB) given the accuracy 89.67%, Decision Tree (DT) given the accuracy 91.85%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDSGLGA gave the best accuracy of 93.478%. Another voting classifier RDS gave an accuracy of 92.39%. To assign the optimal parameters to each classifier, we employed hyperparameter tuning. The experimental investigation reviewed the results of previous recent studies and found that RDSGLGA performed best, with an accuracy rate of 93.478% in terms of making heart disease predictions.

Keywords: Heart disease, Prediction, Machine Learning, Algorithms, Ensemble Model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cardiac disease refers to a variety of heart problems, such as failure and performance decline, and is the worst aspect of our daily lives. The frequency of this patient is significantly rising. However, finding or recognizing the damaged causes at the time of diagnosis is the key issue. By examining various variables and patient diagnostic records, machine learning may be the most effective portion of a significant aspect in predicting the presence of heart disease from responsive health datasets. In our investigation, we've looked through the patient's diagnosis papers and discovered several key indicators of the sickness. The dataset dealt with the diagnosis of human bodies and determining whether or not they had a cardiac condition. To diagnose the sickness in the body, several other researchers have worked together to develop machine-learning algorithms. However, their method and accuracy were not appropriate nor smooth for predicting heart disease. We suggest our method to increase accuracy in the prediction of sickness in the human body. There are two different kinds of machine learning techniques. One of them is under supervision, while the other is not. Working with labeled data, supervised learning creates outputs from inputs based on examples of input-output pairings. The dataset's training data is used as the working data. Unsupervised learning uses unlabeled data to build models that can make use of previously undetected patterns and information.

1.2 Motivation

Heart disease is becoming more prevalent since it affects the majority of people, and its prevalence rate is rising daily. Diet, exercise, and other lifestyle factors, among others, are the cause of heart disease. According to a 2019 poll by the British Heart Foundation, 80 million women and 110 million men were impacted by coronary heart disease [1]. Alcohol, higher birth weights, and hypertension were also cited as signs of heart disease. A few studies have been conducted to forecast cardiac disease. We do several research on

the prognosis of cardiac illness. The majority of them don't have greater accuracy. As a result, we become more determined, and eventually, we discovered our method's highest accuracy. We have developed a technique to forecast the presence of cardiac disease in questionable or ongoing patients.

In our research, we put up a model to anticipate heart disease in people. Recently, we have observed that this sickness is beginning to harm our society. But we also observed that there is a shortage of knowledge and diagnostic tools. Analyzing a patient's symptoms and diagnosing heart disease are expensive processes in our developing nation. In our work as researchers, we are attempting to use machine learning to address the issue.

1.4 Research Questions

- How are the algorithms in this suggested model functioning?
- What will the success rate of a person who will have heart disease or not?
- How can the early identification of heart disease be predicted?
- What advantages does our suggested model have?
- What potential applications of this work exist in the actual world?
- What is the project's projected future?
- What safety measures are required for this work?
- How can we assess our heart disease prediction model?
- How complicated is this work?
- What qualifications are needed for this job?

1.5 Expected outcome

Heart disease is becoming more prevalent among people. Additionally, nobody is certain if she is impacted or not. We are recommending the best approach for predicting or identifying the condition by looking at the diagnosis report. Our approach can discover heart disease patients, enhance decision-making, and precisely measure the effect. It may quantify life quality while also analyzing connected issues. It can raise people's awareness of heart disease. The suggested model can assess the illness in the smallest amount of time.

1.6 Project Management and Finance

Our suggested model is economical and useful in everyday life. Evaluating cardiac disease may be a significant asset for our country. To apply the prediction process in real life, common tools are required. The greatest results and seamless operation of our model will result from the usage of high-conFigureration tools. However, it is still possible if we utilize simple tools.

1.7 Report layout

The relevant study done by the earlier researchers is covered in Chapter 2. Before beginning the investigation, we need to examine the introduction and motive. As a consequence, we talk about the introduction, which may explain the suggested approach in depth, and the motivation portion, which can explain the forecast. After finishing the Introduction section, we concentrated on relevant research and gathered internal data for our work. In our methodology section, we have chosen machine learning algorithms, applied them to our dataset, and then determined which one is the best. Following the pre-processing phase, we tested the data, and at last, we obtained our desired result, which we may refer to as the comparison one. That was discussed in our final section, which is referred to as the conclusion.

Chapter 2

BACKGROUND

2.1 Preliminaries

For determining the precise layout of cardiac illness, machine learning techniques are applied. In this section, we try to examine the investigations connected to the evaluation examination of the patient's diagnosis report. These models use computations like Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Support Vector Classifier (SVC), Adaboost Classifier (ABC), Naïve Bayes (NB), Decision Tree (DT) algorithms. Machine learning models are put into practice in this section to play out the exploration. Several researchers have used several models in their study; these researchers are mentioned in the segment.

2.2 Related work

We have used a few machine learning classifiers to categorize cardiac illness, and they are appropriate for the task we are proposing. To run decision models, machine learning algorithms that are based on decision tree models are known as "tree structures" [1] [2]. Researcher Reldean Williams, Thokozani Shongwe, Ali N. Hasan, Vikash Rameshar have proposed a machine learning model with KNN 67.21%, NB with 85.25%, DT with 81.97%, SVM with 81.97%, LR with 85.25% accuracy [3]. Researchers Umarani Nagaelli, Debabrata Samanta, Partha Chakraborty have proposed a machine learning model with NB 86%, SVM with 89.4% accuracy [4]. Researcher Prachi Chanchalani, DR. Madan lal Saini have proposed a model of machine learning with SVC 83.70%, NB with 84.50%, LR with 82.35%, ABC with 88.23% accuracy [5]. Noor Basha, Gopal Krishna C, Ashok Kumar P S, Venkatesh P have proposed a model of machine learning with KNN 85%, DT with 82%, SVM with 82%, RF with 81%, NB with 80% accuracy [6].

2.3 Comparative Analysis and Summary

The machine learning model is one that is used a lot these days. To locate our respective job, we had to do a challenging endeavor. All connected works have poor model results and poor accuracy. To identify the dataset's greatest accuracy of prediction, we had to apply a variety of machine learning models. We had to deal with running the models on high-end hardware. To get at the categorization rates, we used a few individual computations. By adding pricey GPUs, complicated models might generate lengthy runtime.

2.4 Scope of the Problem

The issue was simplifying and familiarizing the heart disease diagnosis process. We attempted to provide the best accuracy with our suggested model because there are so many works with machine learning that are linked to it. Although there was little room for improvement in the process, the notion may be put into practice using simple technologies to reduce the number of heart disease diagnoses.

2.5 Challenges

The Kaggle website hosted the dataset. The information was very usable and simple to use. We must manually review the dataset for any missing data when the data gathering is complete. With this dataset, no one has ever as accurate as we are.

Chapter 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrument

To extract the most accuracy from the dataset, we used a variety of algorithms and hybrid models. We required some tools, such as efficient configuration tools with the best GPUs. Python programming language and technologies including Google Collaboratory, Jupiter Notebook, and Anaconda have been utilized. Through the browser, it enables the authoring and execution of any Python code. All tests were performed on a computer running the 64-bit version of Windows 10 Pro on an AMD Ryzen 5 3600 6-core processor clocked at 3.59 GHz with 8 GB of RAM.

3.2 Data Collection Procedure

The dataset was virtually ready for implementation when it was downloaded from Kaggle. There are 12 columns and 918 rows, respectively. The rate of cardiac disease is categorized in the diagnostic column. Every characteristic was crucial for predicting heart disease. Patients are divided into the conditions 0 and 1 in 2 groups. The frequency of these two circumstances has been estimated. The remaining 410 individuals were not suffering from heart disease, leaving 508 people in that stage. Figure 3.1, which is below, displays the ratio. The dataset has been divided into two sections. They undergo testing and training. We have chosen 80% of the candidates for the training portion and 20% for the exam portion.

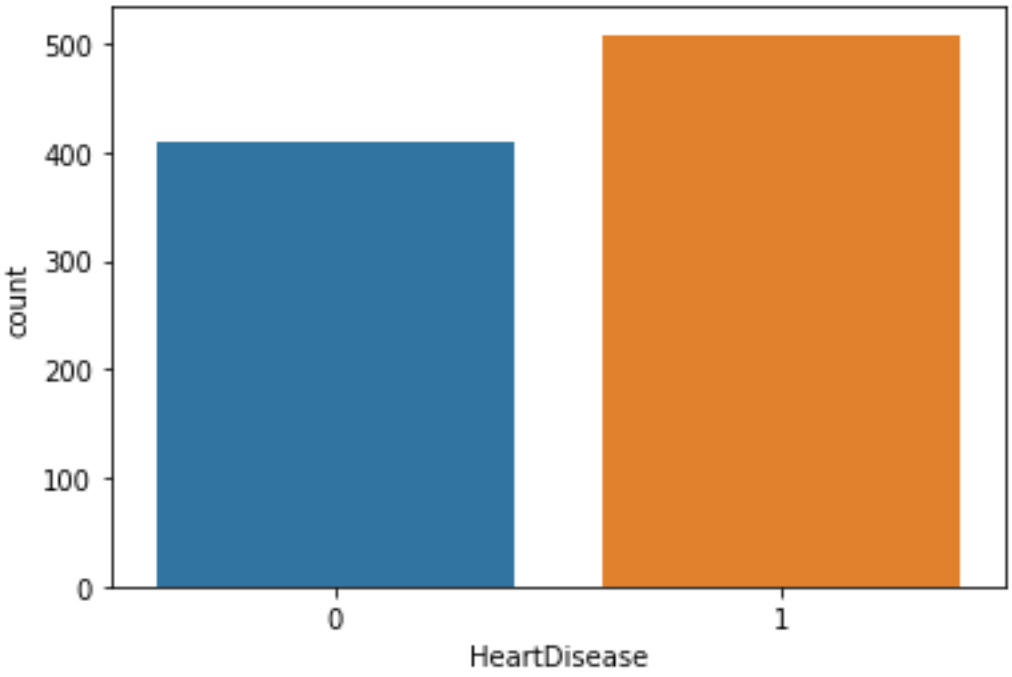


Figure 3.1: Number of target values

There were no missing or inaccurate values in the dataset, which only comprises nominal values. Table 3.1 presents a thorough explanation of the dataset along with its range.

Table 3.1: Details of the dataset

Attributes	Description	Value Range	Types of Values
Age	Age	28 to 77	Integer
Sex	Gender	M and F	Char
ChestPainType	Chest pain type	ASY, NAP, ATA, TA	Char
RestingBP	Resting Blood Pressure	0 to 200	Integer
Cholesterol	Serum Cholesterol	0 to 603	Integer

FastingBS	Fasting Blood Sugar	0 to 1	Integer
RestingECG	Resting electrocardiogram results	Normal, LVH, ST	Char
MaxHR	Maximum heart rate achieved	60 to 202	Integer
ExcerciseAngina	Exercise-induced angina	N, Y	Char
Oldpeak	Oldpeak	-2.6 to 6.2	Float
ST_Slope	The slope of the peak exercise ST segment	Flat, Up, Down	Char
HeartDisease	target	0 to 1	Integer

3.2.1 Categorical Data Encoding

The process of converting categorical data into a numerical value is known as the categorical data encoding method. The categorical encoding strategy was crucial to our investigation since machine learning only accepts and outputs numeric data. To use the categorical data encryption approach, we needed the Sex, ChestPainType, RestingECG, ExerciseECG, ExerciseAngina, and ST Slope columns.

3.2.2 Missing Value Imputation

It involves filling up the blanks or missing data with imputed values that were determined by research with other dataset data. However, it is gratifying that our dataset had no missing values.

3.2.3 Handling Imbalanced Data

It is the process of changing a dataset's class distribution. It manages the data by systematically adding more examples to the dataset. While using the entire dataset as input, the data for minorities is increased.

3.2.4 Feature Scaling

It is a procedure for normalizing the variety of independent data variables. Data with negative values is available. It has undergone scaling modifications.

3.3 Statistical Analysis

Every type of research project needs an analysis section. This section depends on creating and assessing the algorithms I've employed. We must take a few procedures to prepare the dataset for use because we have opted to use a comma-separated value (CSV) file. We have taken a number of measures, including data collecting and pre-processing. We have implemented some different classifiers named Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Support Vector Classifier (SVC), Adaboost Classifier (ABC), Naïve Bayes (NB), Decision Tree (DT) algorithms. Random Forest (RF) given an accuracy of 90.22%, Logistic Regression (LR) given accuracy of 89.67%, Gradient Boosting (GB) given accuracy of 89.67%, Support Vector Classifier (SVC) given accuracy of 91.85%, Adaboost Classifier (ABC) given the accuracy 91.30%, Naïve Bayes (NB) given the accuracy 89.67%, Decision Tree (DT) given the accuracy 91.85%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDSGLGA gave the best accuracy of 93.478%. Another voting classifier RDS gave an accuracy of 92.39%. Hyperparameter tweaking and 10-fold cross-validation have both been employed.

3.4 Proposed Methodology

Flow chart:

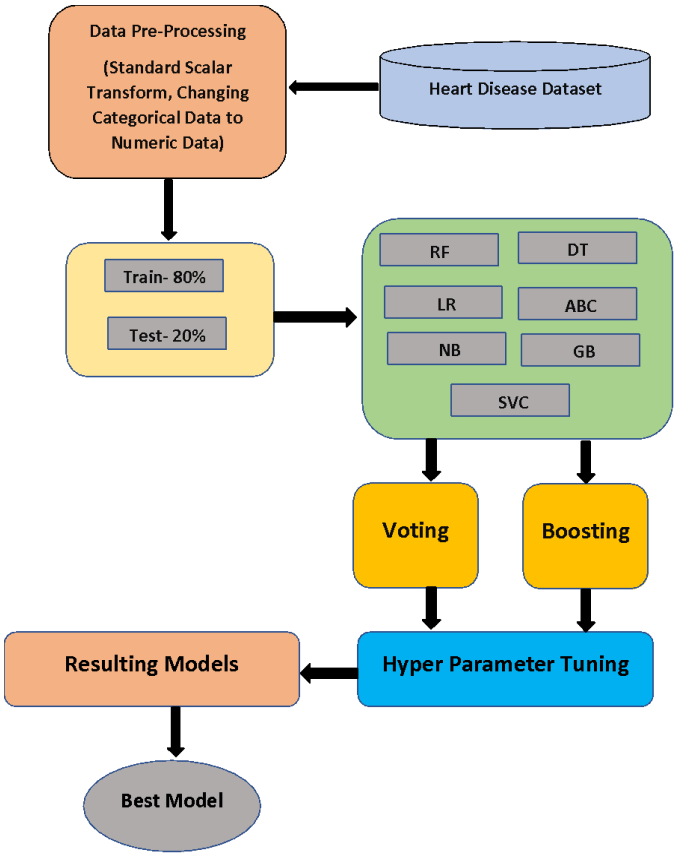


Figure 3.2: Methodology of Heart Disease

In this section, we've predicted heart disease using a process diagram. The dataset for the system's training and testing was initially introduced. Then, data pre-processing techniques like Standard Scaler Transform were used. Categorical data conversion to numeric data. We utilized 80% for the training portion and 20% for the testing portion. After that, we implemented algorithms and assessed the outcomes. Then, in order to get the highest forecast accuracy, we employed ensemble methods. Voting is the name of an algorithmic group. The outcomes of the ensemble algorithms that were used were then assessed. Then we used Hyper Parameter Tuning to verify the outcome. Then, using outcome analysis, we assessed the models that had been put into practice. Figure. 3.2 displays the recommended model technique.

The identification of internal dependencies between two variables, or how one variable changes as a result of the change in another, is referred to as a correlation subplot. The more interdependence between variables suggests that it will be successful to predict one variable from another. It alludes to a deeper comprehension of the dataset and aids in our ability to identify the crucial factors [10]. All of the characteristics that were linked with the anticipated attribute "HeartDisease" were displayed in Figure. 3.3.

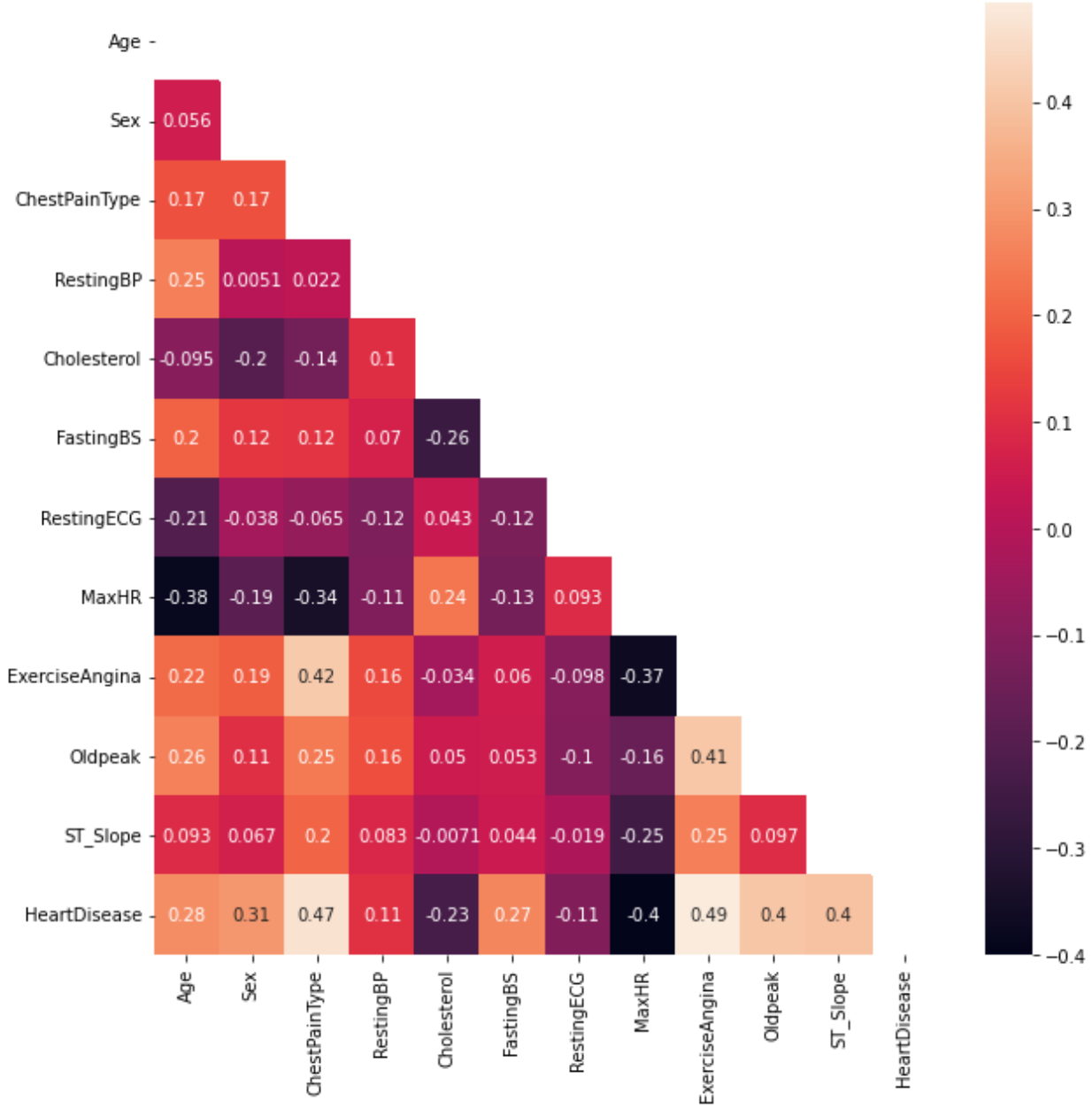


Figure 3.3: Correlated Features of Heart Disease Dataset

3.5 Implementation Requirements

We require data sources in order to examine or train our suggested model. For things to go well, we must clean the dataset. A number of filtering techniques will be used to clean the dataset. Then, data pre-processing techniques like Standard Scaler Transform were used. Categorical data conversion to numeric data. We utilized 80% for the training portion and 20% for the testing portion. After that, we implemented algorithms and assessed the outcomes. Then, in order to get the highest forecast accuracy, we employed ensemble methods. Voting is one of the ensemble algorithms. The outcomes of the ensemble algorithms that were used were then assessed. Then we used Hyper Parameter Tuning to verify the outcome. Then, using outcome analysis, we assessed the models that had been put into practice. The learning process must then be initiated by carrying out the data analysis step. Next, we must put model learning into practice and fit the predictions approach. The models must then be voted on in order to obtain the highest accuracy. The best model may then be chosen for implementation based on accuracy, precision, recall, and F-1 score.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

A supervised learning method, which functions based on training and testing, was employed in this paper. The classification model is built using the training dataset. To obtain the outcome, the generated model is applied to the testing dataset. The machine-learning algorithm will be swiftly illustrated in the following sections.

4.1.1 Classifier Algorithms

We have implemented some different classifiers named Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Support Vector Classifier (SVC), Adaboost Classifier (ABC), Naïve Bayes (NB), Decision Tree (DT) algorithms.

Logistic Regression

A classifier approach based on machine learning called logistic regression (LR) contains two categories for the class label: yes or no, like a binary (0/1) scale. Although it permits the combined value of continuous variables and discrete predictors, logistic regression is appropriate for discrete variables [11]. The idea is depicted in Figure. 4.1 below. Logistic regression adopts the supervised machine learning approach. The fundamental equation is shown below [9] [12].

$$h\Theta(x) = \frac{1}{1 + e - (\beta o + \beta 1X)} \quad \ldots\ldots\ldots\ldots(1)$$

'hΘ(x)' is the output of the logistic function, where 0 ≤ hΘ(x) ≥1

'β1' is the slope

'βo' is the y-intercept

'X' is the independent variable

(βo + β1X) – derived from the equation of a line Y (predicted) = (βo + β1X) + Error.

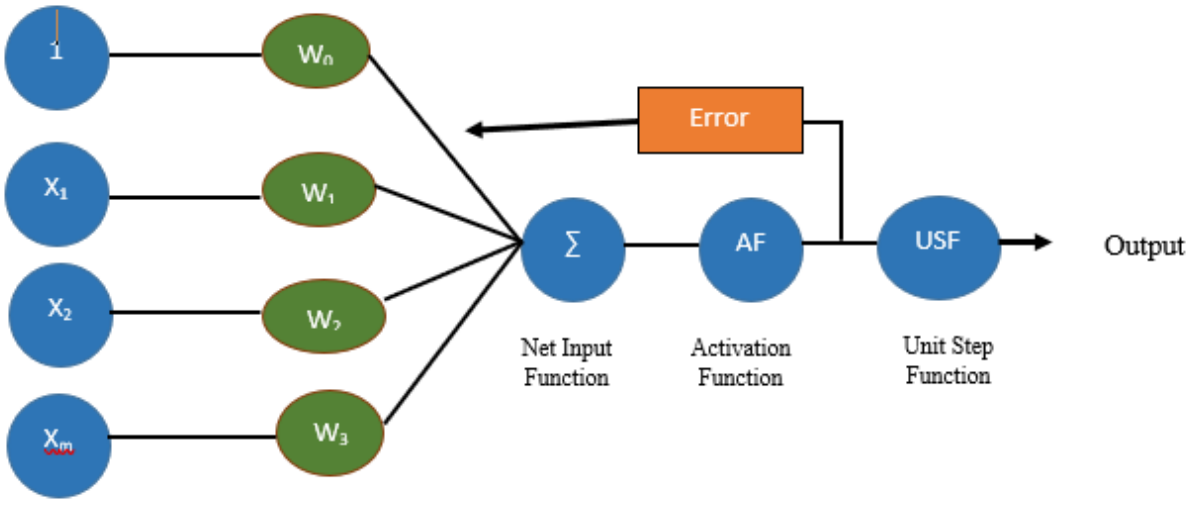


Figure 4.1: Logistic Regression Classifier

Random Forest

Different Decision Tree algorithms make up the Machine Learning (ML) based classifier ensemble approach known as Random Forest (RF) [13] [19]. In order to provide an ideal decision model with more accuracy than the single decision tree model, RF builds several decision trees while the algorithm is being trained. The notion is depicted in Figure. 4.2 below. However, it may be used with big datasets. The mean of all decision tree methods is calculated using the Random Forest algorithm [14] [15] [25] [26]. The Random Forest method estimated the average of two decision tree algorithms.

$$j = \frac{1}{B} + \sum_{b=1}^{B} fb(X') \quad \text{.....(2)}$$

Concerning X = {x1,x2,x3,..... xn} with respect to Y = {y1,y2,y3,..... yn} with the

lower to upper limit is 1 to B.

Sample x′ = mean of the sum of the prediction $\sum^{B}_{b=1} fb(X^{'})$ for every summation.

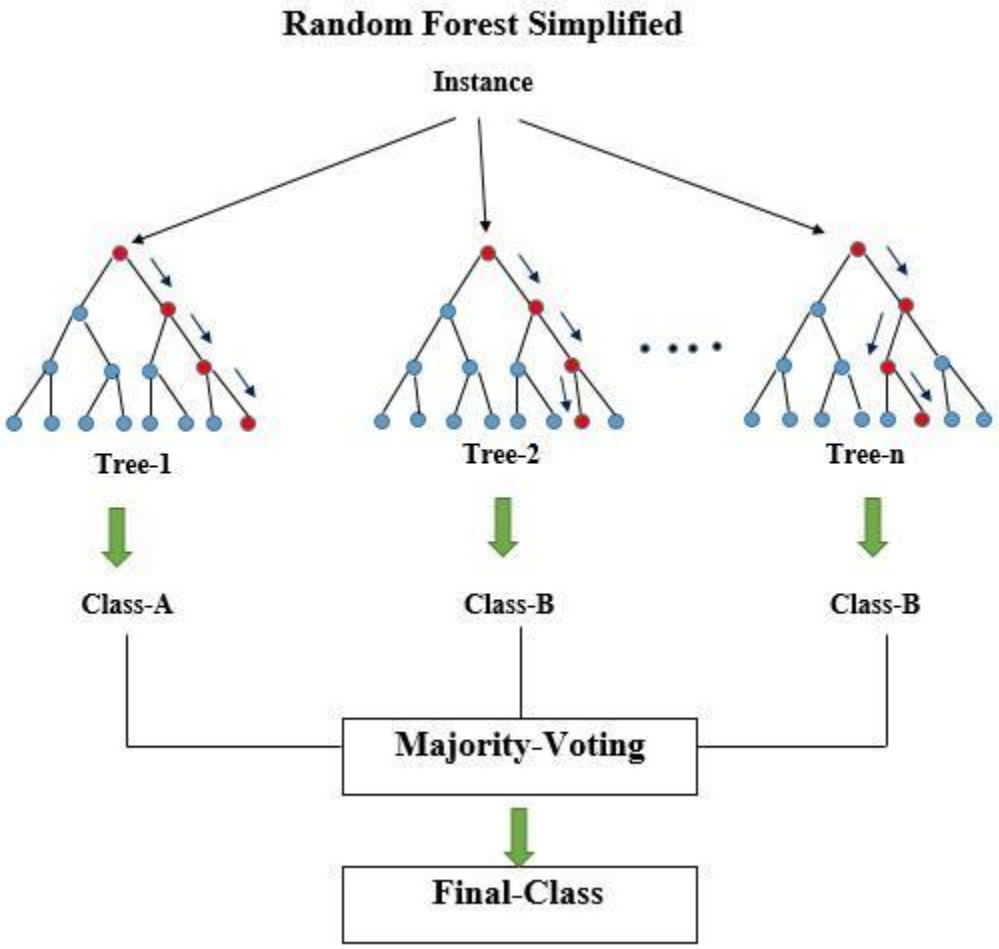


Figure 4.2: Random Forest Classifier

Gradient Boosting

The loss function is the main component of the boosting method known as Gradient Boosting (GB), which is based on Machine Learning (ML). The notion is depicted in Figure. 4.3 below. It works by combining and optimizing weak learners to reduce a model's loss function. To improve an algorithm's performance, overfitting is eliminated [7] [8]. Here fi(x) = loss function with correlated negative gradients (−ρi x gm(X)), m = number of iterations.

Feature increment (i) = 1,2,3, ... , m. Therefore, the optimal function F (X) after m−th iteration is shown below [16].

$$F(X) = \sum_{i=0}^{m} fi(x) \quad \text{.....(3)}$$

Here, gm = the path of loss function's fast decreasing F(X) = Fn − 1(X) the decision tree's target is to solve the mistakes by previous learners [17][18]. The negative gradient for the m^{-th} iteration is shown below.

$$gm = -\left(\frac{\partial L(y, F(X))}{\partial F(X)}\right) F(X) = Fm - 1(X) \quad \text{.....(4)}$$

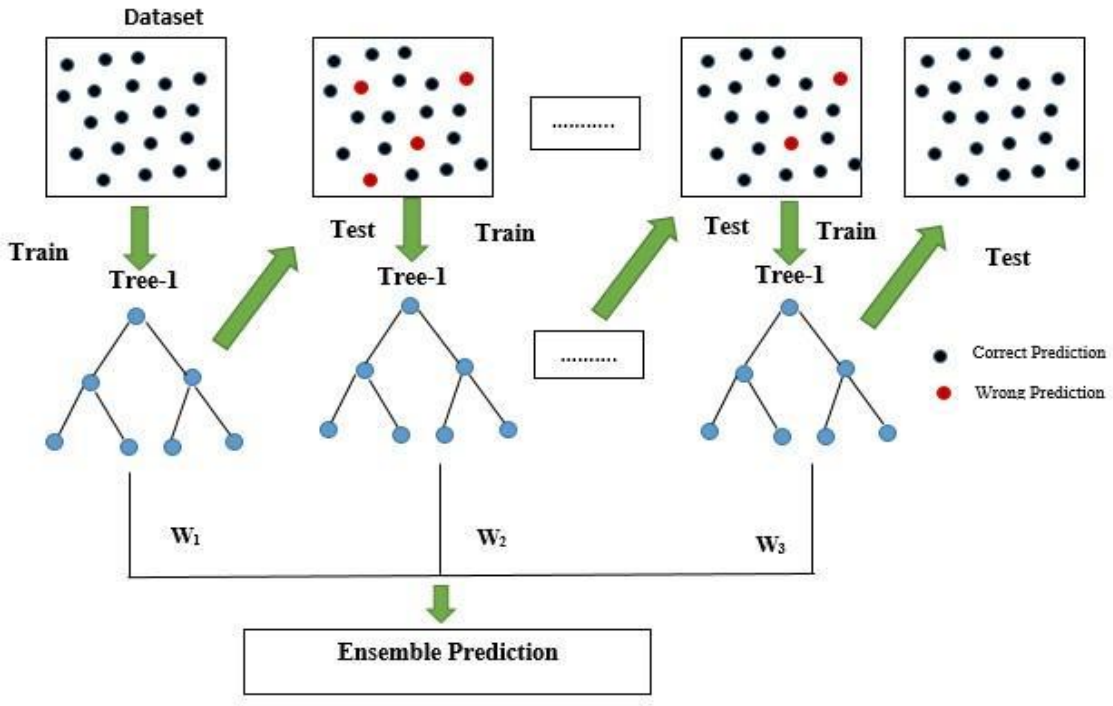


Figure 4.3: Gradient Boosting Classifier

Adaboost Classifier

AdaBoost is a boosting classifier that joins a number of ineffective classifiers to create a powerful classifier. 1000 samples are used by ABC to forecast TA. Weights that differ for classifiers and samples are fixed by ABC. This makes it challenging for classifiers to concentrate on the end outcome. The final formula to achieve TA is,

$$H_k(P) = l - (\sum_{k=1}^{k} a_k h_k(P)) \text{ .....(5)}$$

Here, N=frequency of training data, k = total number of weak classifiers combined to use, hk = output of weak classifier (lowerlimit 1 to upper limit k), ak = weight of classifier. ABC combines sample trainers, fixes the weights of samples and classifiers to geta more accurate and efficient TA. The notion is depicted in Figure. 4.4 below.

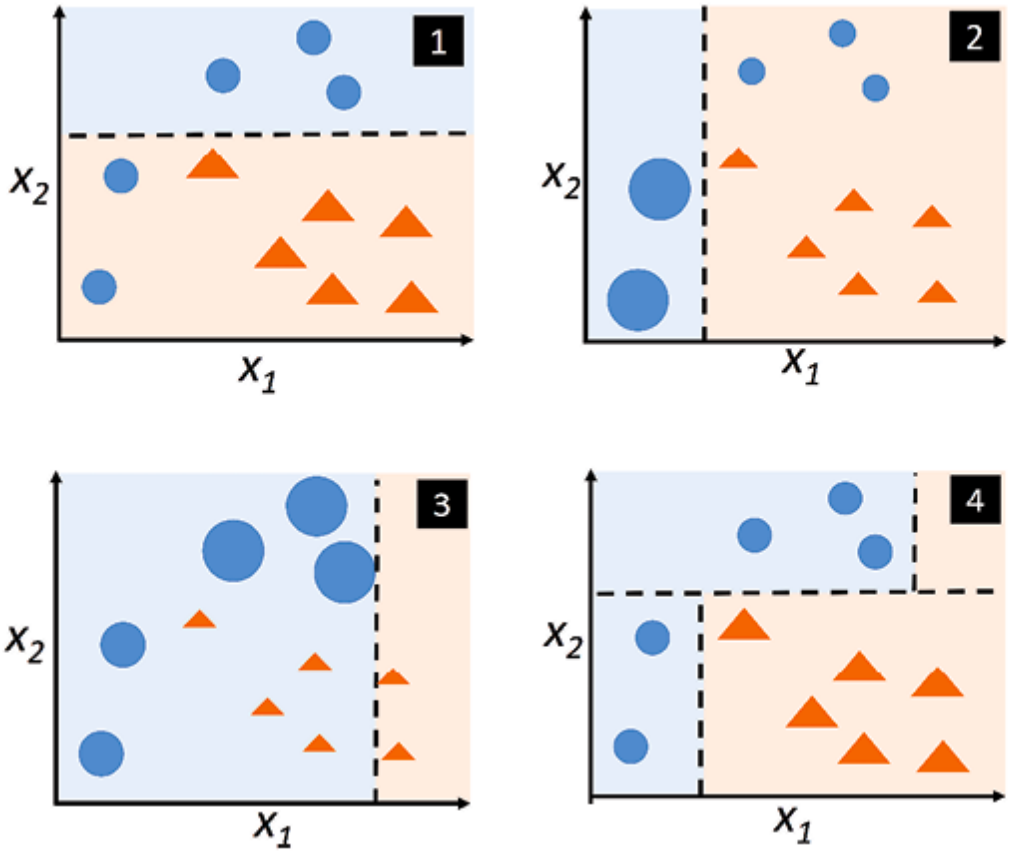


Figure 4.4: Adaboost Classifier

4.1.2 Ensemble Methods of Machine Learning

The term "ensemble approach" refers to the use of several classifiers to turn weak classifiers into strong classifiers by producing the greatest accuracy and effectiveness. It was used in our investigation due to variable handling, bias, and uncertainty since it lowers variances, merges predictions from several models, and narrows the prediction spread [21] [22]. In our investigation, one ensemble approach was employed. Voting and Boosting ensemble modeling was employed.

Boosting

Boosting refers to the technique that uses a weighted average to work with several algorithms and makes the weak learners strong learners boost the accuracy of independent models creating the loss functions [24]. The concept is shown in below Figure 4.5. In our study, the boosting method is applied in the training and testing portion to construct the hybrid model. The equation is shown below [23].

Here, Υ_t = ½-ϵ_t (how much f_t is on the weighted sample).

$$\frac{1}{n}\sum_{i=1}^{n} I(y_j g(x_i) < 0) \leq \prod_{t=1}^{T} \sqrt{1 - 4\Upsilon_t^2} \quad \text{.....(6)}$$

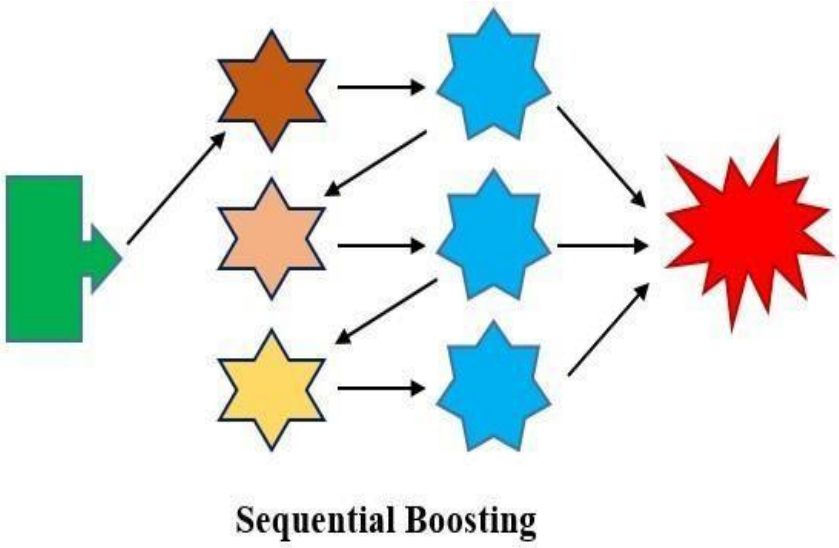


Figure 4.5: Ensemble Boosting

Voting

Voting classifiers are a group of classifiers that are used to forecast the class with the best majority of votes. It implies that the model trains using many models to anticipate outcomes by aggregating the results of voting. The notion is depicted in Figure. 4.6 below. The formula we employed is shown below [24] [20].

Here, w_j = weight that can be assigned to the j^{th} classifier.

$$y' = argmax \sum_{j=1}^{m} w_j p_{ij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

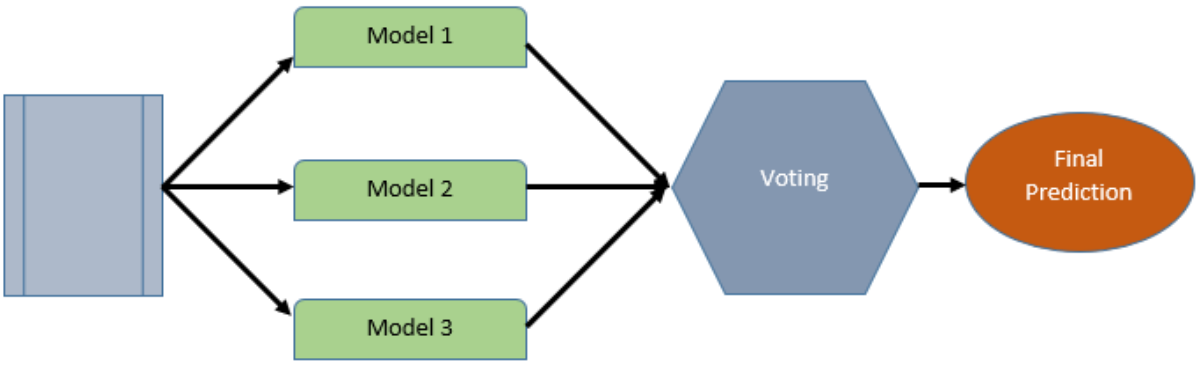
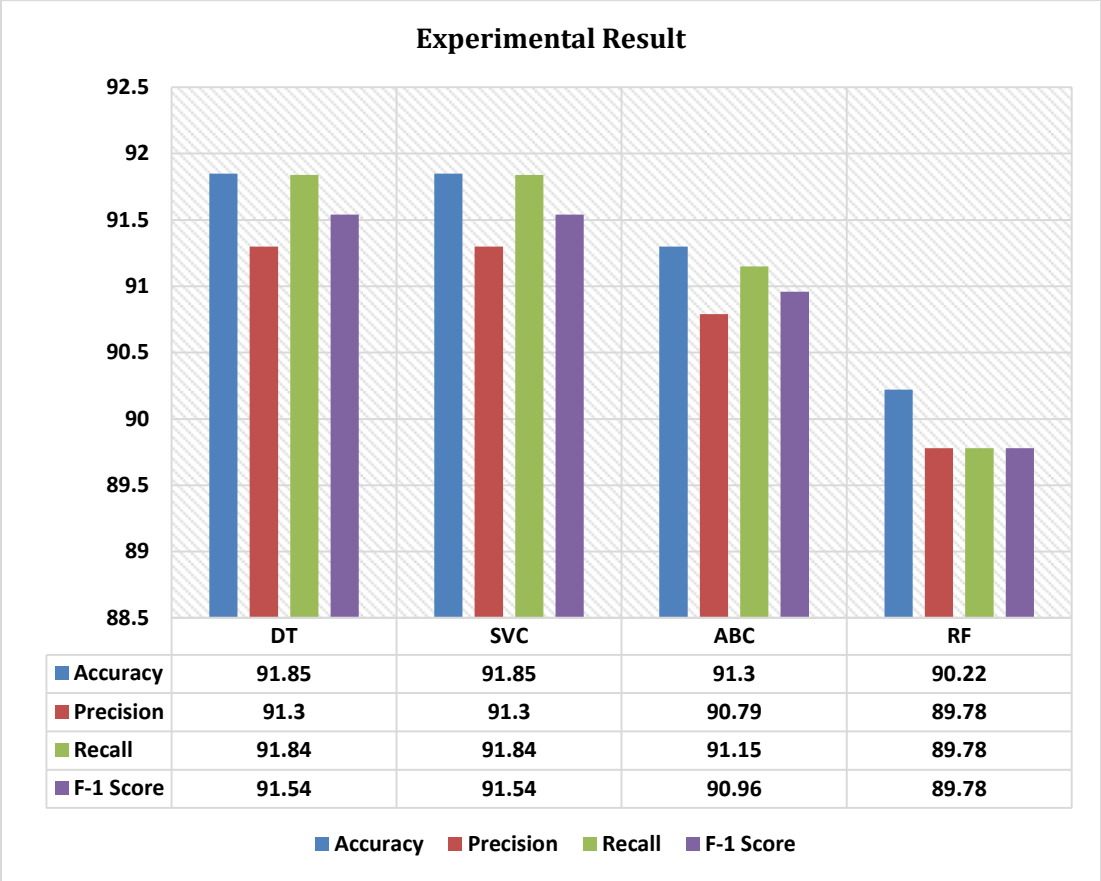


Figure 4.6: Ensemble Voting

4.2 Experimental Result & Analysis

At this point, we had to assess how well the current models performed. To verify the effective performance of our suggested model, we may utilize several performance assessment measurements and approaches. These techniques calculate the total performance based on hypothetical data. In this section, we must present an analysis report based on the results of our machine learning experiments on the targeted dataset for heart disease. We initially put our chosen dataset into practice. Our dataset has been filtered to remove any missing or erroneous values. We put a variety of algorithms into practice and evaluated how well they worked. For our suggested algorithms, we tested

Confusion matrices Accuracy, Precision, Recall, and F-1 Score. These confusion matrices for conventional methods have been measured. We have evaluated for Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Support Vector Classifier (SVC), Adaboost Classifier (ABC), Naïve Bayes (NB), Decision Tree (DT) algorithms. With confusion matrices, we have seen many ensemble approaches in action. We have assessed voting ensemble methods.



	DT	SVC	ABC	RF
Accuracy	91.85	91.85	91.3	90.22
Precision	91.3	91.3	90.79	89.78
Recall	91.84	91.84	91.15	89.78
F-1 Score	91.54	91.54	90.96	89.78

Figure 4.7: Experimental Results of Classifiers

Firstly, we considered the performances of algorithmic classifiers, the best accuracy had obtained at 91.85% Decision Tree (DT). SVC, Adaboost Classifier (ABC), Random Forest (RF) algorithms achieved the accuracy respectively 91.85%, 91.3% and 90.22%. The output is shown in below Figure 4.7. The precision score was Decision Tree (DT), SVC, Adaboost Classifier (ABC), Random Forest (RF) achieved respectively 91.3%, 91.3%, 90.79% and 89.78%. The Recall score was Decision Tree (DT), SVC, Adaboost

Classifier (ABC), Random Forest (RF) achieved respectively 91.84%, 91.84%, 91.15% and 89.78%. The F-1 score was Decision Tree (DT), SVC, Adaboost Classifier (ABC), Random Forest (RF) achieved respectively 91.54%, 91.54%, 90.96% and 89.78%.

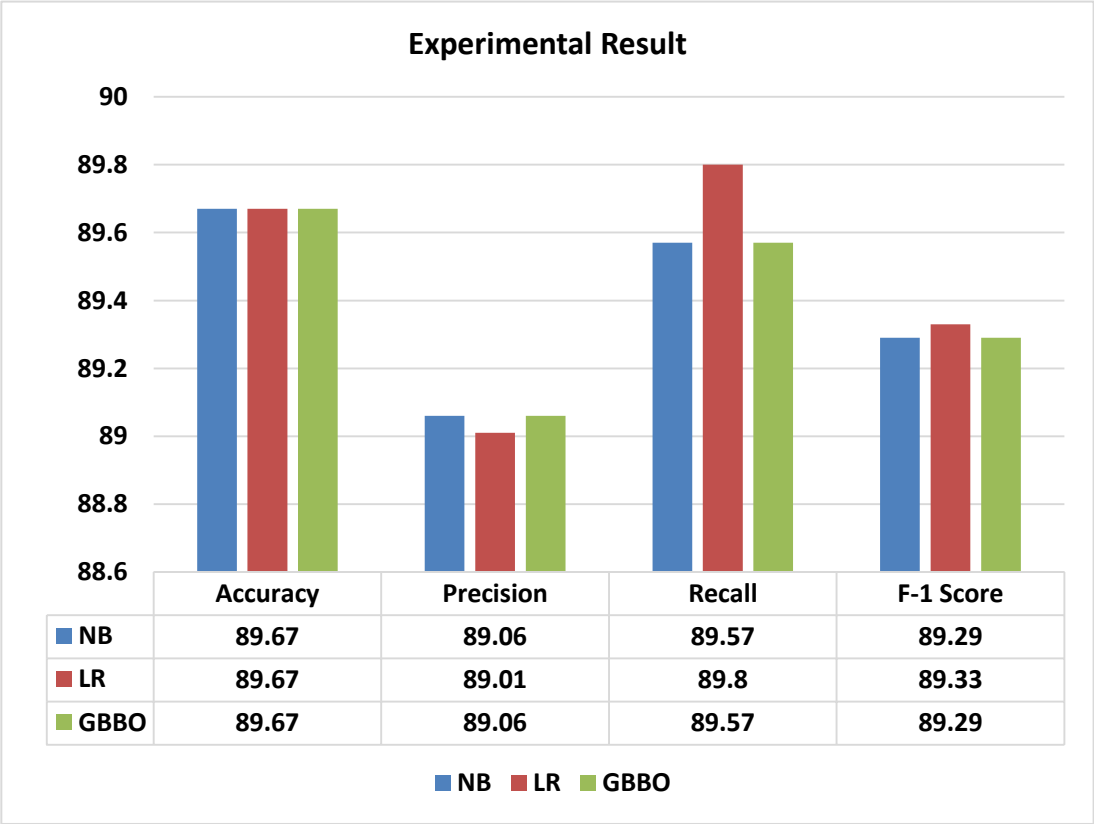


Figure 4.8: Experimental Results of Classifiers

Naïve Bayes (NB), Logistic Regression (LR), Gradient Boosting (GBBO) algorithms achieved the accuracy respectively 89.67%, 89.67% and 89.67%. The output is shown in below Figure 4.8. The precision score was Naïve Bayes (NB), Logistic Regression (LR), Gradient Boosting (GBBO) achieved respectively 89.06%, 89.01% and 89.06%. The Recall score was Naïve Bayes (NB), Logistic Regression (LR), Gradient Boosting (GBBO) achieved respectively 89.57%, 89.8% and 89.57%. The F-1 score was Naïve Bayes (NB), Logistic Regression (LR), Gradient Boosting (GBBO) achieved respectively 89.29%, 89.23% and 89.29%.

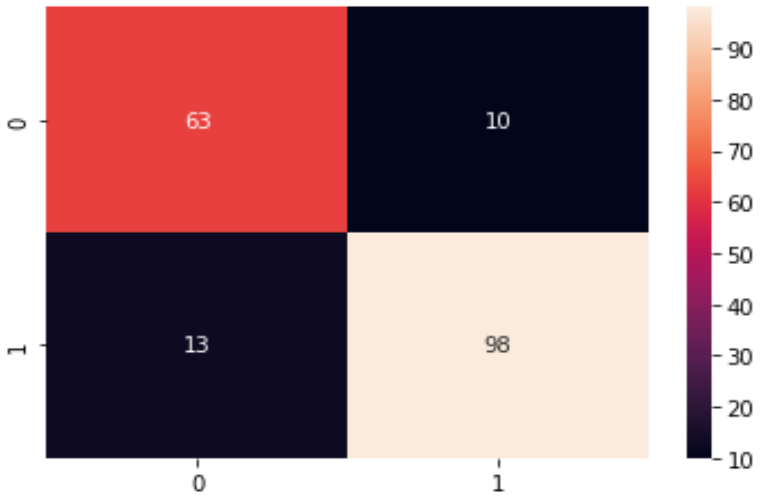


Figure 4.9: Random Forest Confusion Matrix

For Random Forest (RF) the True positive value was 63, False Positive value was 10, False Negative value was 13 and True Negative value was 98. The output is shown in Figure 4.9.

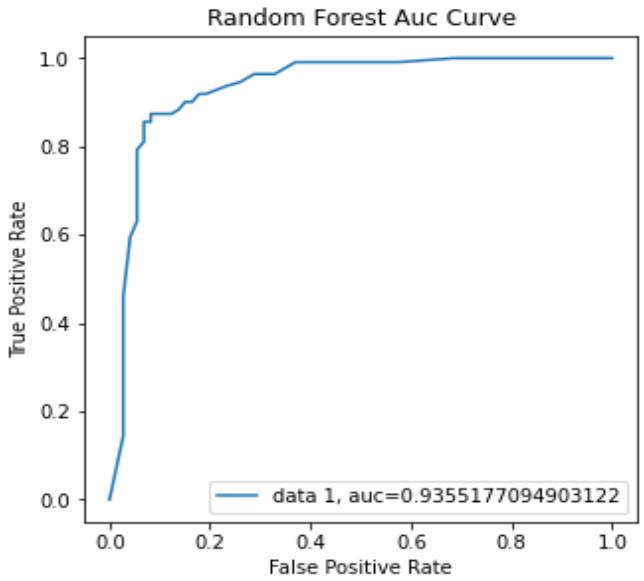


Figure 4.10: Random Forest AUC Curve

For Random Forest (RF) the AUC score was 93.55%. The output is shown in Figure 4.10.

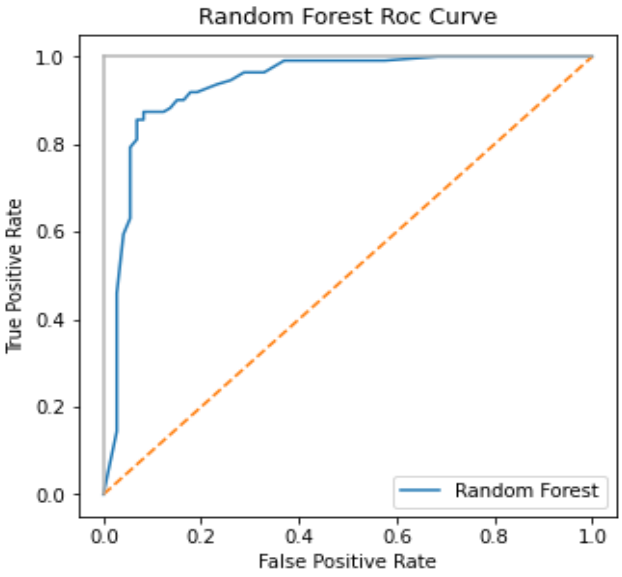


Figure 4.11: Random Forest ROC Curve

For Random Forest (RF) the ROC score is shown in Figure 4.11.

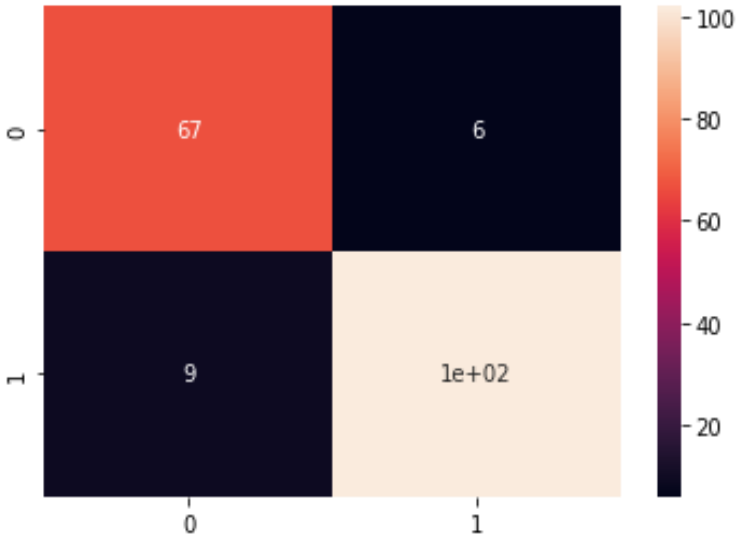


Figure 4.12: Decision Tree Confusion Matrix

For Decision Tree (DT) the True positive value was 67, False Positive value was 6, False Negative value was 9 and True Negative value was 0. The output is shown in Figure 4.12.

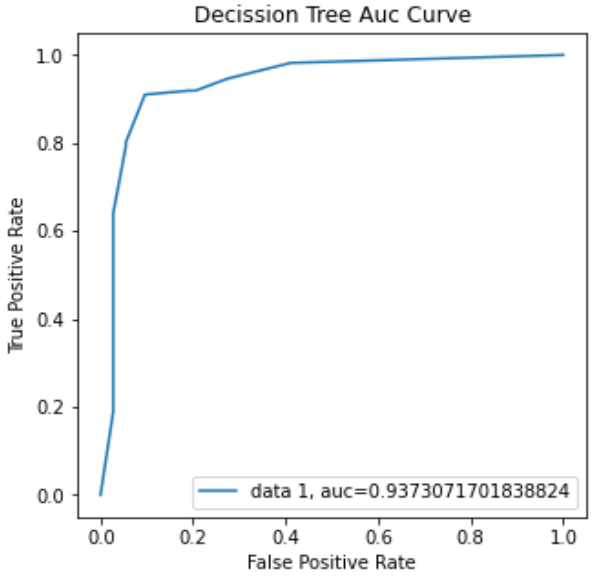


Figure 4.13: Decision Tree AUC Curve

For Decision Tree (DT) the AUC score was 93.73%. The output is shown in Figure 4.13.

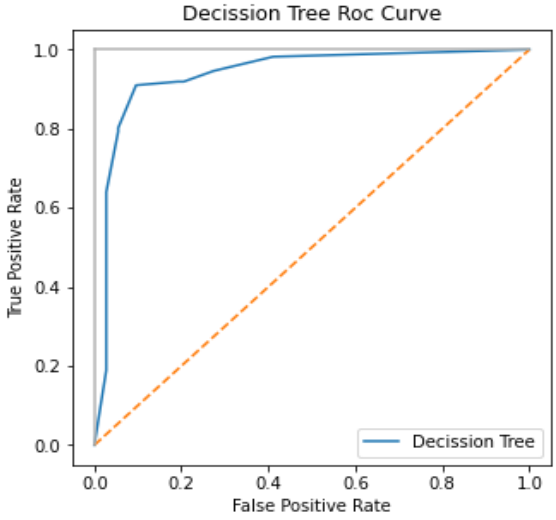


Figure 4.14: Decision Tree ROC Curve

For Decision Tree (DT) the ROC score is shown in Figure 4.14.

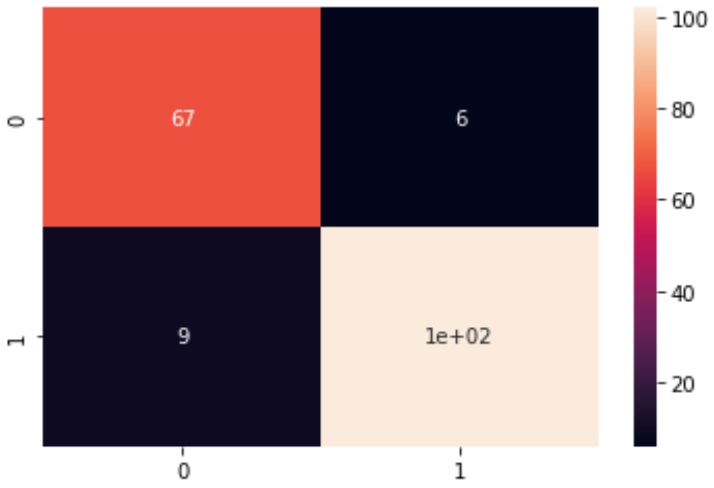


Figure 4.15: SVC Confusion Matrix

For SVC the True positive value was 67, False Positive value was 6, False Negative value was 9 and True Negative value was 0. The output is shown in Figure 4.15.

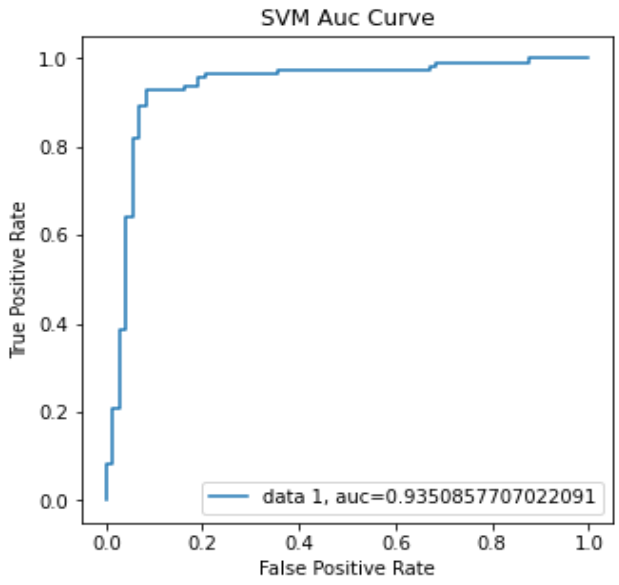


Figure 4.16: SVC AUC Curve

For SVC the AUC score was 93.50%. The output is shown in Figure 4.16.

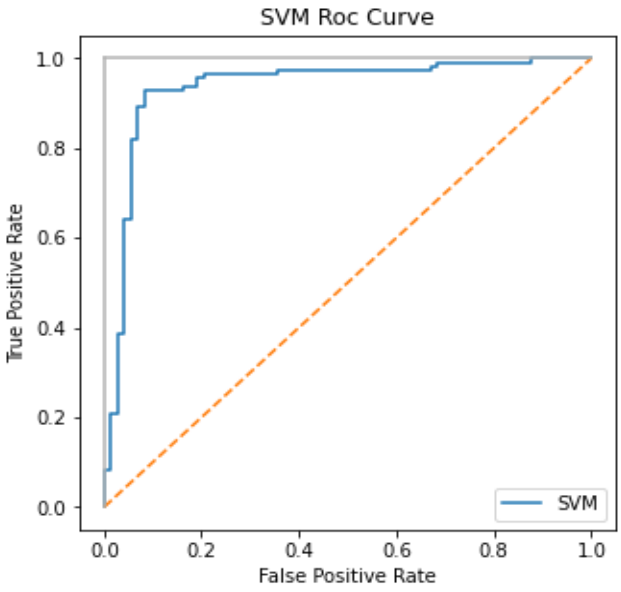


Figure 4.17: SVC ROC Curve

For SVC the ROC score is shown in Figure 4.17.

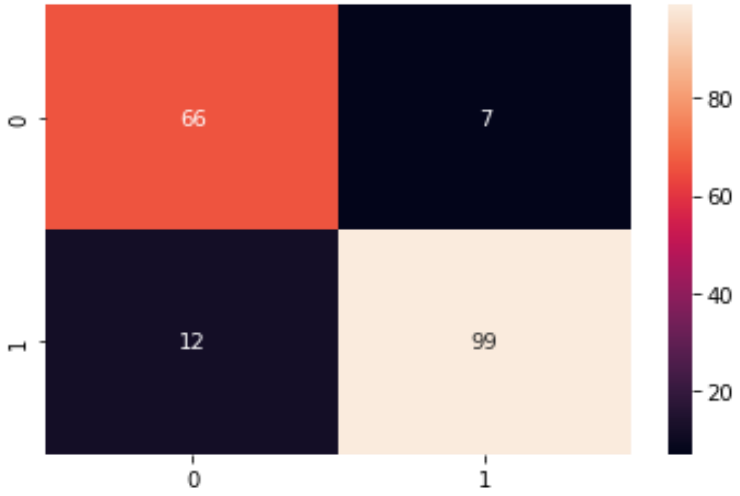


Figure 4.18: Logistic Regression Confusion Matrix

For Logistic Regression (LR) the True positive value was 66, False Positive value was 7, False Negative value was 12 and True Negative value was 99. The output is shown in Figure 4.18.

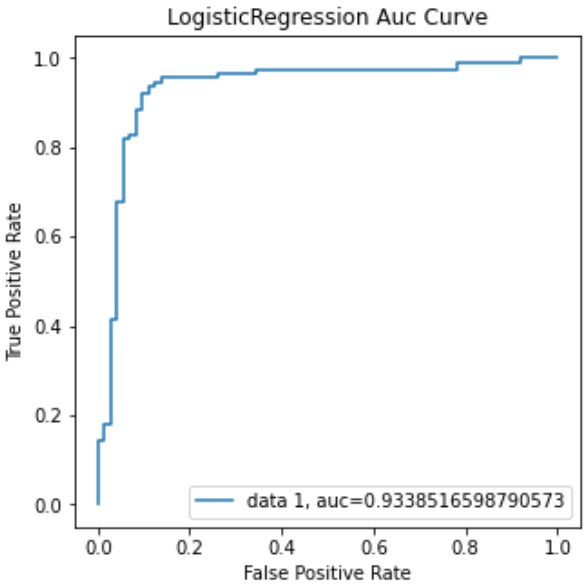


Figure 4.19: Logistic Regression AUC Curve

For Logistic Regression (LR) the AUC score was 93.38%. The output is shown in Figure 4.19.

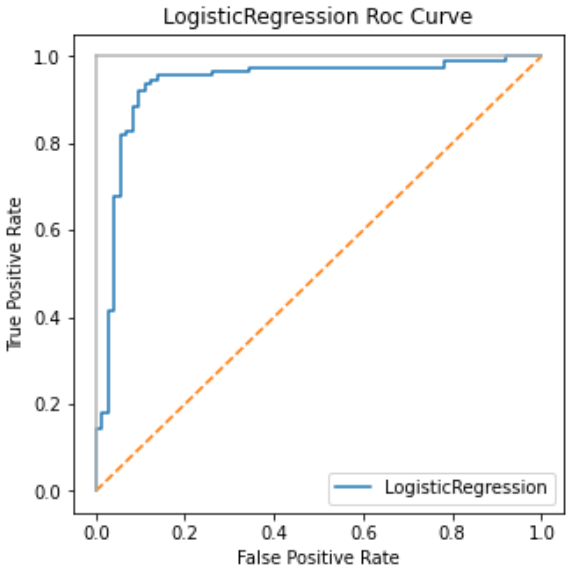


Figure 4.20: Logistic Regression ROC Curve

For Logistic Regression (LR) the ROC score is shown in Figure 4.20.

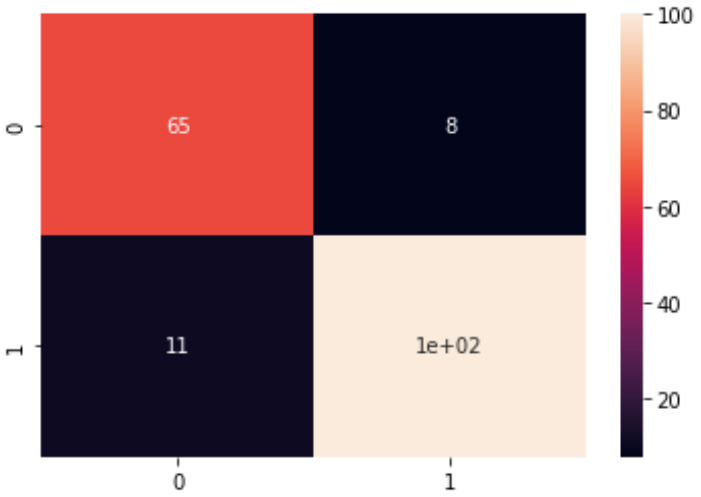


Figure 4.21: Gradient Boosting Confusion Matrix

For Gradient Boosting (GB) the True positive value was 65, False Positive value was 8, False Negative value was 11 and True Negative value was 0. The output is shown in Figure 4.21.

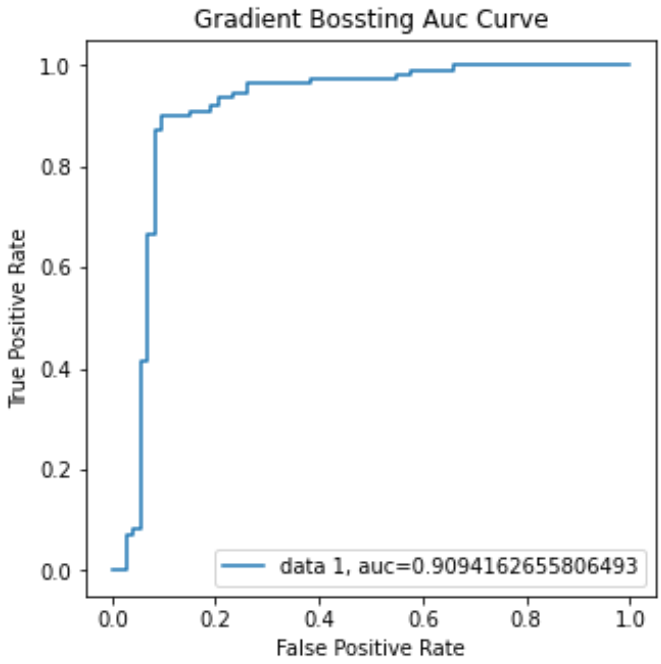


Figure 4.22: Gradient Boosting AUC Curve

For Gradient Boosting (GB) the AUC score was 93.38%. The output is shown in Figure 4.22.

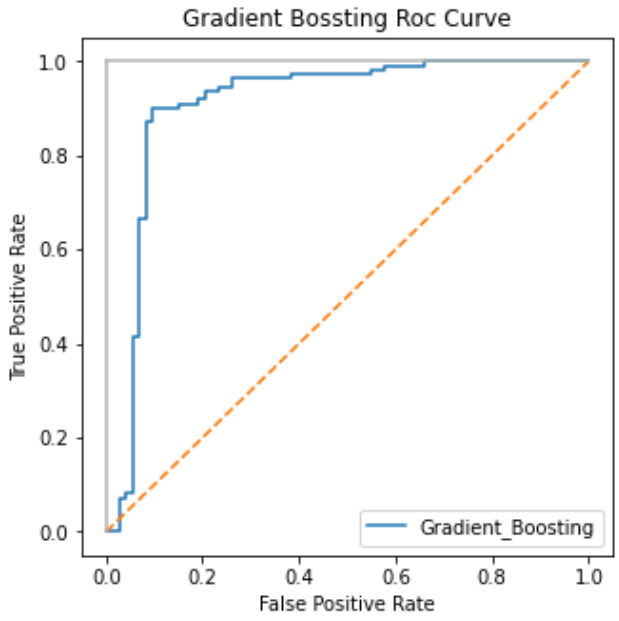


Figure 4.23: Gradient Boosting ROC Curve

For Gradient Boosting (GB) the ROC score is shown in Figure 4.23.

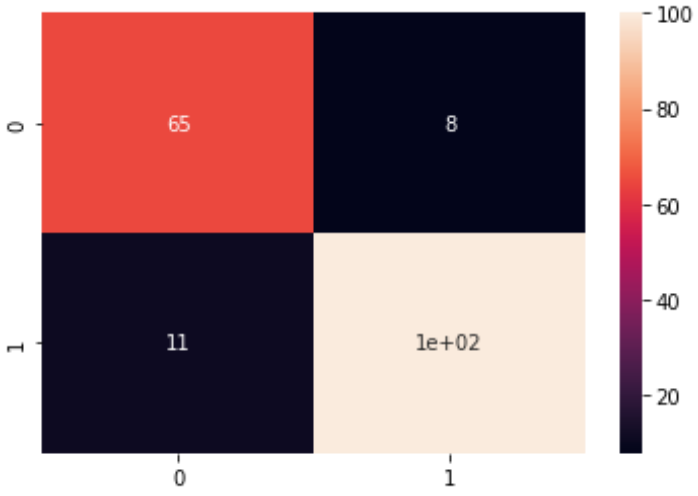


Figure 4.24: Adaboost Classifier Confusion Matrix

For Adaboost Classifier (ABC) the True positive value was 65, False Positive value was 8, False Negative value was 11 and True Negative value was 0. The output is shown in Figure 4.24.

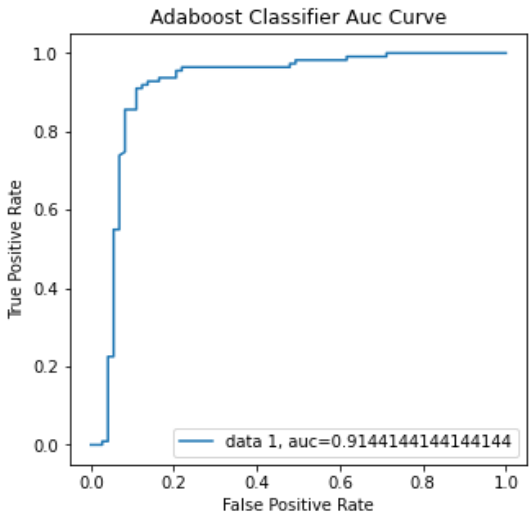


Figure 4.25: Adaboost Classifier AUC Curve

For Adaboost Classifier (ABC) the AUC score was 93.38%. The output is shown in Figure 4.25.

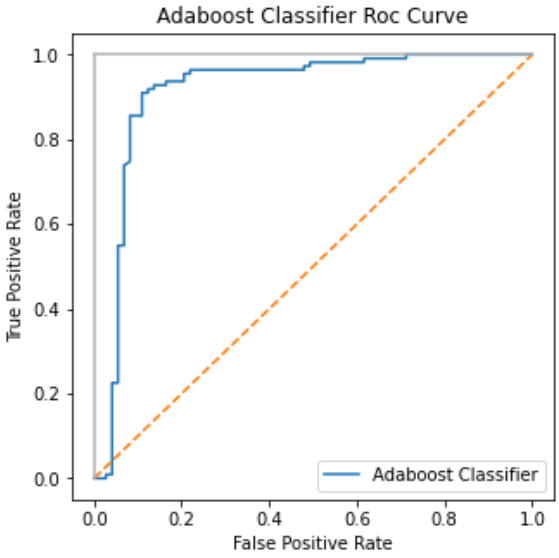


Figure 4.26: Adaboost Classifier ROC Curve

For Adaboost Classifier (ABC) the ROC score is shown in Figure 4.26.

Voting algorithm RDSGLGA performed well with 93.47% accuracy. Another voting algorithm RDS performed 92.39% accuracy.

4.3 Discussion

We shall now define the judicial system of our suggested paradigm. We have taken into account the F-1 score, recall, accuracy, and precision.

4.3.1 Accuracy

It speaks of the proportion of testing data predictions that were correct. Whereas accessibility of the measures with actual measurements is performed by accuracy. It is founded on a solitary variable. Accuracy only addresses deliberate mistakes. It is one of the most straightforward measurement methods for any model.

Accuracy = (TruePositive + TrueNegative)/(TruePositive + FalsePositive + TrueNegative + FalseNegative)

4.3.2 Precision

It speaks about the percentage of positively expected observations that really occurred. The genuine true portion of all the cases where they correctly predicted true are identified by precision. For any type of model, a high recall might also be highly deceptive.

Precision = (TruePositive)/(TruePositive + FalsePositive)

4.3.3 Recall

It speaks about the percentage of positively anticipated observations from a model. High accuracy, however, can occasionally be deceptive. The ratio of projected positives to all positive labels is determined by normally recall.

Recall = (TruePositive)/(TruePositive + FalseNegative)

4.3.4 F-1 Score

It speaks of the precision and recall harmonic means. Both the recall and precision ratios are relevant. We assume the model is quite terrible if the harmonic mean is lower.

$$F - 1 \text{ Score} = 2 * (\text{Recall} + \text{Precision}) / (\text{Recall} + \text{Precision})$$

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

Our suggested approach offers a number of advantages, both economically and socially. To study and ascertain the fundamental aspects or characteristics of a heart disease patient, our model is built on a real-life dataset. The job provides societal benefits, such as the ability to inform people about the prevalence of heart disease and preventative measures. Through accurate diagnosis and frequent checkups, we can recommend early therapy. Because they are more likely to be aware of diseases and be able to predict whether they would be impacted or not. Our approach requires fewer compilations and takes less time. As a result, illness prediction is simple and accurate. With the use of improved diagnosis procedures, we have analyzed the information in our model to determine the root cause of heart disease. We hope that our suggested approach will be adopted and put into practice on a societal level.

5.2 Impact on Environment

The streamlined diagnosis procedures in our suggested paradigm make it particularly useful in remote locations. Through the device model, we can cut down on both time and complexity. Our methodology has no adverse consequences and is simple, so we can guarantee that the environment will gain from it as well. The patients don't need to travel to metropolitan regions to find out if they have heart disease or not. The patient's diagnosis report may be readily supported by the prediction model, which can also forecast potential outcomes. Due to the low cost of diagnosing cardiac illness, patients won't be concerned about it or the expense of local therapies. Because it is less complicated, it can be used by individuals at any level. Our suggested model can make it clear whether or not a patient has cardiac disease. Our suggested model will improve the economic and social climate. If we are given the chance to put our suggested model into

practice, we are confident that it will mark a significant advancement in current medical science technology.

5.3 Ethical Aspects

Before the system is put into operation, we must take some moral safeguards to prevent the disclosure of personal information, diagnostic reports, or humor. Our suggested approach may be used to real-world cardiac disease diagnosis and therapy as well as future research endeavors. We have determined that the issue affects not only a small area or region but also the entire planet. Through the suggested model, any victim or informed person may forecast the rate at which their heart illness will impact them.

5.4 Sustainability Plan

We can guarantee that the technology used in heart disease diagnosis and research throughout the world can accept our suggested model. We are convinced that the victim ladies who can anticipate their likelihood of developing heart disease would find our proposed approach beneficial. We may be inspired and prepared to aid the rural regions if we are provided with the right tools and scope for implementation. We anticipate that our suggested model will be useful and sustainable.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In our fascinating article, we assess the influenced rate of our individuals employing algorithms. With our model, we can successfully forecast the future. The prediction system may benefit from the diagnosing technology. People can gain from understanding if they will have an impact or not. They may erroneously believe that they ought to be aware of cardiac illness. If individuals use our approach, they can quickly identify the different phases of heart disease. Assuming our suggested model can also be beneficial to diagnosis authority. We have utilized a variety of standard algorithms that are quick to develop, need little training, and have excellent accuracy.

6.2 Conclusion

The world we live in today is a contemporary one. The globe is currently a technologically advanced and simple place. The new technology is accessible to anybody in the world. With the aid of technology, what we have suggested is really simple and quick. We have made an effort to simplify the process of predicting heart disease in people. Our innovative models can assist our people. We have to make sure the concept is workable, and we promise to add a lot more features and work on more well-liked topics in the future. We are stating this expectation.

6.3 Implication for Further Study

We have mortality because we are human. In our daily lives, we are impacted by several ailments. While most of us have heart disease, some of us have the necessities for recovery. The therapy and diagnosis technologies are more advanced and precise since we live in a developing country. The process of identifying cardiac illness is now simpler and takes less time thanks to new technology. We have made an effort to provide our folks something fresh. We hope that others will adopt our model. For better performance, we have worked on a few algorithms and want to add more in the future.

REFERENCE

[1] "Heart Disease Statistics", Accessed: December 2022, Available: https://www.bhf.org.uk/-/media/files/research/heart-statistics/bhf-cvd-statistics-global-factsheet.pdf.

[2] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[3] R. Williams, T. Shongwe, A. N. Hasan and V. Rameshar, "Heart Disease Prediction using Machine Learning Techniques," 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 2021, pp. 118-123, doi: 10.1109/ICDABI53623.2021.9655783.

[4] Umarani Nagavelli, Debabrata Samanta, Partha Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models", *Journal of Healthcare Engineering*, vol. 2022, Article ID 7351061, 9 pages, 2022. https://doi.org/10.1155/2022/7351061.

[5] Prachi Chanchalani, DR. Madan lal Saini, 'A Comparative Analysis of Heart Disease Detection Techniques Using Machine Learning', International Journal of Creative Research Thoughts (IJCRT), 2021, Volume 9, Issure 4 April, 2021, ISSN: 2320-2882.

[6] Noor Basha, Gopal Krishna C, Ashok Kumar P S, Venkatesh P 'Early Detection of Heart Syndrome Using Machine Learning Technique' 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019.

[7] V. Lahoura, H. Singh, A. Aggarwal et al., "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, no. 2, p. 241, 2021.

[8] "Breast Cancer Dataset", Accessed: December 29, 2021, Available: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

[9] "What is Correlation in Machine Learning?", Accessed: August 6, 2020, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

[10] L. Mary Gladence, M. Karthi, V. Maria Anu. "A statistical Comparison of Logistic Regression and Different Bayes Classification Methods for Machine Learning" ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, Vol -10, No-14, August 2015.

[11] "What is Correlation in Machine Learning?", Accessed: November 8, 2021, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

[12] "Logistic Regression for Machine Learning", Accessed: August 6, 2021, Available: https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/

[13] Ghosh, Pronab, Asif Karim, Syeda Tanjila Atik, Saima Afrin, and Mohd Saifuzzaman. "Expert cancer model using supervised algorithms with a LASSO selection approach." International Journal of Electrical and Computer Engineering (IJECE) 11, no. 3 (2021): 2631.

[14] Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques." International Journal of DataMining & Knowledge Management Process 8, no. 2 (2018): 01-09

[15] Aljahdali, Sultan, and Syed Naimatullah Hussain. "Comparative prediction performance with support vector machine and random forest classification techniques." International journal of computer applications 69, no. 11 (2013).

[16] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." ArtificialIntelligence Review 54, no. 3 (2021): 1937-1967.

[17] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.

[18] Pasha, Maruf, and Meherwar Fatima. "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection." J. Softw. 12, no.12 (2017): 923-933.

[19] Wang, Yizhen, Somesh Jha, and Kamalika Chaudhuri. "Analyzing the robustness of nearest neighbors to adversarial examples." In International Conference on Machine Learning, pp. 51335142. PMLR, 2018.

[20] Sharma, Ajay, and Anil Suryawanshi. "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure." International Journal of Computer Applications 136, no. 6 (2016): 28-35.

[21] hou, Zhi-Hua. Ensemble methods: foundations and algorithms. CRC Press, 2012.

[22] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.

[23] emmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." Journal of Marketing Research43, no. 2 (2006): 276-286.

[24] Islam, Rakibul, Abhijit Reddy Beeravolu, Md Al Habib Islam, Asif Karim, Sami Azam, and Sanzida Akter Mukti. "A Performance Based Study on Deep Learning Algorithms in the Efficient Prediction of Heart Disease." In 2021 2nd International Informatics and Software Engineering Conference (IISEC), pp. 1-6. IEEE, 2021.

[25] Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Alamin Dhaly, Md Nour Hossain. “A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease”. In April 2022 The 10th International Conference on Computer and Communications Management in Japan.

ggd

ORIGINALITY REPORT

22%	18%	7%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	Submitted to Daffodil International University Student Paper	5%
3	Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Al-Amin Dhaly, Md Nour Hossain. "A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease", The 10th International Conference on Computer and Communications Management, 2022 Publication	2%
4	ir.lib.uwo.ca Internet Source	1%
5	hdl.handle.net Internet Source	1%
6	Submitted to University of Central Lancashire Student Paper	1%
7	dr.ntu.edu.sg Internet Source	<1%