The Heart Disease Prediction using the Technique of Classification in Machine Learning using the concepts of Data Mining

By

Pronay Kumar Saha

191-15-12704

Sudham Chandra Debnath

191-15-12854

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

Supervised by

Md Azharul Islam Tazib
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by
Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH
December 2022

APPROVAL

This Project/internship titled **"The heart disease prediction using the technique of classification in machine learning using the concepts of data mining"**, submitted by **"Pronay Kumar Saha (191-15-12704), Sudham Chandra Debnath (191-15-12854)"** 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 *02/02/2023*.

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

Dr. Sheak Rashed Haider Noori
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Sazzadur Ahamed
Assistant Professor
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 **Md Azharul Islam Tazib, Lecturer, and Department of CSE** at 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:

Md Azharul Islam Tazib
Lecturer
Department of CSE
Daffodil International University

Co - Supervised by:

Foysal

Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Pronay Kumar Saha
ID: 191-15-12704
Department of CSE
Daffodil International University

Sudham

Sudham Chandra Debnath
ID: 191-15-12854
Department of CSE
Daffodil International University

Remove

ACKNOWLEDGEMENT

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

I am really grateful and wish my profound indebtedness to my beloved teacher **Md. Azharul Islam Tazib,** Lecturer, 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 research. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages, appreciation, and respect for another person have made it possible to complete this research.

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

I would like to thank my entire course mate in Daffodil International University, who took part in this discussion and support while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of my parents. They were by my side through the whole varsity life. They meant a lot to me.

ABSTRACT

Machine learning techniques are used in this study to examine data and forecast cardiac illness. The main objective of my work is to examine the data and estimate the proportion of persons who will develop heart disease based on each dataset. We wanted to examine the topic of " The Heart Disease Prediction using the Technique of Classification in Machine Learning using the concepts of Data Mining " and evaluate the machine learning algorithm and approaches for detecting heart disease in this research-based project. The classification methodology found in machine learning and data mining principles is used in this work to propose a technique for predicting cardiac disease. Random Forest Classifier is a method that is commonly employed in machine learning (RFC). This algorithm, along with others like Support Vector Machine, Linear Regression, AdaBoost, Naive Bayes, and K-Nearest Neighbor, aims to deliver better outcomes and forecasts. A method for evaluating the accuracy and error rate is also demonstrated in this study, using performance matrices such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE). Additionally, purely for performance evaluation, all of these algorithms were evaluated on a dataset. Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target are the 14 attributes that we have acquired from our dataset to employ in this prediction technique. In order to predict cardiac illness in this study, I have argued in favor of using the XGBC, GBC, RFC algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter - 1
INTRODUCTION

1.1 Introduction

Among the leading causes of death in the modern world are heart disease, cardiac arrest, and heart attacks. Heart attacks were frequently thought to solely affect the elderly. That notion is now deemed outdated because young and middle-aged people are already having heart attacks as a result of the current situation. Every day, there are more cases of heart illness. About 20 million people die from cardiovascular disease each year around the world, accounting for 32% of all fatalities. In South Asian low- and middle-income countries, heart attacks account for 60% of deaths while coronary heart disease claims the lives of 75% of people. With a 14.31% risk, heart disease is one of the main causes of mortality in underdeveloped countries like Bangladesh. In Bangladesh, the number of heart attack deaths has increased over the past ten years by 48 times for women and 35 times for males. The fact that heart attack incidents are constantly rising among younger generations, however, is what should worry us the most. Predicting any such illnesses is important and concerning. This diagnosis is difficult to make. It needs to be done correctly and efficiently. Which persons are more prone to develop heart disease given specific medical traits is the main subject of the research report. We created a technique to recognize and anticipate cardiac problems based on the patient's medical history. We used nine machine learning techniques, including Linear Regression (LinR) and K-Nearest Neighbors (KNN), to forecast and classify the heart disease patient. In a very beneficial approach, the model's application was regulated to boost the accuracy of heart attack prediction in any individual. By combining Linear Regression (LinR) and K-Nearest Neighbors (KNN), the suggested model was able to accurately predict the symptoms of heart disease in a particular person. When compared to earlier employed classifiers, such as naive bayes, it showed a good degree of accuracy. Applying the offered model to evaluate the likelihood that the classifier will correctly and reliably diagnose cardiac illness has thereby greatly reduced the amount of stress. Given's method for forecasting heart illness improves patient care while being more affordable. Through this investigation, we learned a lot that can be used to predict the heart problems of patients.

1.2 Motivation

We are driven to complete this project based on research. A person now dies from heart disease every minute since it has become such a serious problem in today's world. This ratio takes into account both the male and female categories, and it takes into account individuals between the ages of 25 and 70. This does not mean that persons of different ages won't have cardiac ailments. Predicting the source and progression of the disease is now a very difficult task because this issue may arise at a young age as well. We have covered a number of methods and techniques utilized in this paper to forecast cardiac disorders.

1.3 Objective

Our fundamental objective is to put forth a system for forecasting cardiac disease that makes use of the classification approach in machine learning and the concepts of data mining. The numerous algorithms and methodologies for heart disease prediction will become clearer and better understood as a result of this work. Despite their best efforts during this challenging time, the government and authorities will get unambiguous signals about whether they are doing appropriately. The general public's level of faith in the government's actions will also be examined in this article.

1.4 Expected Outcome

We developed nine machine learning (ML) algorithms and collected information or data from different websites for our study. These data will all be pooled and put via a machine learning system to produce a forecast for heart disease.

1.5 Research Questions

Which machine learning algorithm must be used in Decision Supporting System for heart disease prediction?

1.6 Overview of The Paper

Chapter 1: INTRODUCTION (1.1 Introduction, 1.2 Motivation, 1.3 Objective, 1.4 Expected Outcome, 1.5 Research Questions)

Chapter 2: LITERATURE REVIEW (2.1 Introduction, 2.2 Related Works, 2.3 Scope of the Problem, 2.4 Challenges,)

Chapter 3: RESEARCH METHODOLOGY (3.1 Introduction, 3.2 Research Subject and Instrumentation, 3.3 Research Analysis, 3.4 Data Collection Procedure, 3.5 Implementation Requirements)

Chapter 4: RESULT and DISCUSSION (4.1 Introduction, 4.2 Research Analysis, 4.3 Research Findings)

Chapter 5: CONCLUSION (5.1 Summary of the Study, 5.2 Conclusion 5.3 Further Implication of the Study)

Chapter - 2
LITERATURE REVIEW

2.1 Introduction

The prediction of cardiac disease has been the subject of numerous scientific articles. Modern heart disease, cardiac arrest, and heart attacks are among the leading causes of death. The sole group at risk for heart attacks was frequently thought to be the elderly. This notion has been deemed irrelevant because more and more people in their 20s and 30s are having heart attacks. Every day, more people are developing heart disease. About 20 million people per year die from cardiovascular disease, or 32% of all global fatalities. Heart attacks account for 60% of these deaths, while coronary heart disease is to blame for 75% of deaths in low- and middle-income countries in South Asia. In less developed countries like Bangladesh, the likelihood of dying from cardiac disease is at 14.31%. Over the previous ten years, the death rate from heart attacks has increased 35 times for men and 48 times for women in Bangladesh. People should be especially concerned about the fact that younger generations are progressively having more heart attacks. It is important and concerning to be able to predict such disorders. It's difficult to make this diagnosis. It must be carried out precisely and successfully. Which patient is more likely to have a heart issue, based on a variety of medical factors, is a major emphasis of the research paper. We created a method to estimate the likelihood that a patient will be given a cardiac disease based on their medical history.

2.2 Related Works

We reviewed 20 research papers from many authors and publications. Literature reviews given below,

The research paper's main concern is which persons, given specific medical criteria, are more prone to develop heart disease. The medical histories of 304 different patients, all of diverse ages, are included in the data source that they employ. This dataset on heart illness was discovered in the UCI repository. The (KNN), (LR), and (RFC) algorithms used in this work can aid professionals or health analysts in accurately diagnosing cardiac disease. Our model has an 87.5% accuracy rate. It is also proven that KNN outperforms the other two algorithms we utilized in terms of accuracy (88.52%). [1]

The Heart Disease Prediction Model in this study is able to predict patients' heart disease status based on their clinical data, which is helpful for practitioners and medical staff. Massive volumes of data are produced by the healthcare industry, and it is necessary to mine this data to reveal hidden information so that wise decisions can be made. The source of the thirteen input attributes was the Cleveland Clinic Foundation Heart disease data set. The accuracy of the following methods is given: Decision Tree: 76%; Association Rule: 55%; K-NN: 58%; Artificial Neural Network: 85%; SVM: 86%; Naive Bayes: 69%; Hybrid Approach: 96%. [2]

The aim of this study is to investigate the various data mining techniques that have recently been created for the prediction of heart disease. The results demonstrate that 15-attribute neural networks outperform all other data mining methods. A total of 909 records were obtained from the Cleveland Heart Disease database. (455 records) Testing dataset plus training dataset (454 records). Decision trees, neural networks, and naive bayes all received scores of 90% or higher. The analysis reveals that a neural network with 15 attributes has so far shown a maximum accuracy of 93%. The Decision Tree, on the other hand, has also done well, with 92.62% accuracy, and it uses 15 attributes. [3]

The goal of this research project is to provide an overview of the current knowledge finding techniques in databases using data mining techniques that are applied in modern medical research, particularly in the prediction of heart disease. Naive Bayes, K-NN, and the Decision List technique are three different supervised machine learning methods that have been used to analyze the dataset. a data source 909 records overall with 15 medical features were present in the Cleveland Heart Disease database (factors). (455 records) Testing dataset plus training dataset (454 records). It looks to be the most effective technique because Nave Bayes has the highest percentage of accurate predictions (86.53%). [4]

The study's major focus is on a system for using machine learning and artificial intelligence to detect heart disease. In this study, a python-based application is developed for healthcare research since it is more dependable and helps track and set up many types of health monitoring applications. Using a random forest classifier method, cardiac diseases may be detected with an accuracy of 83%. The random forest approach, which is used to build heart disease detection, is employed in this project to identify heart ailment using Python. [5]

With the help of decision trees, naive bayes, classification modeling techniques, and neural networks, this study tries to predict heart disease. The heart illness database, which is open to the public, can be used to identify different types of heart problems. A total of 909 records were obtained from the Cleveland Heart Disease database. Naive Bayes looks to be the most effective technique, followed by Neural Networks (85.53%) and Decision Trees, with the highest percentage of accurate predictions (86.53%). Decision trees performed the best (89%) at predicting people without heart illness compared to the other two models. [6]

This study's main contribution to this publication is to help non-specialized practitioners make knowledgeable choices about their patients' risk for heart disease. The dataset was collected from the Cleveland Clinic Foundation. In their study report, they suggested an efficient data mining-based heart disease prediction method. They trained and tested the system, with accuracy of 86.3% in testing and 87.3% in training, because this model produces superior results and helps the subject-matter experts. Allow the patient to get early determination results since they can still reason well without retraining. [7]

Using well-known heart disease datasets, such as the Cleveland dataset, this paper presents a complete review of heart disease prediction models that were developed and evaluated. Machine learning prediction models for identification such as ANN, LR, K-NN, SVM, DT, and NB are used. Relief, mRMR, LASSO, and local-learning-based feature selection are the four most advanced feature selection algorithms currently available (LLBFS). A classification system with a 77% accuracy rate was created using machine learning classification algorithms. a diagnosis system that classified data using multi-layer Perceptron and support vector machine (SVM) techniques, with an accuracy rate of 80.41%. [8]

The goal of this research is to provide a framework for heart disease prediction using important risk factors and a range of classifier methods, such as Naive Bayes (NB), Bayesian Optimized Support Vector Machine (BO-SVM), K-Nearest Neighbors (KNN), and Salp Swarm Optimized Neural Network (SSA-NN). In that order, SSA-NN (accuracy: 86.7%, precision: 100%, sensitivity: 60%) and BO-SVM (accuracy: 93.3%, 100% precision, and 80% sensitivity) delivered the best results. This study makes use of the Cleveland dataset. For this project, a data matrix was gathered. The UCI collection contains a large number of heart disease datasets. 76 attributes and 303 records make up the dataset. [9]

Throughout this research, a variety of techniques and technologies for heart disease prognosis have been covered. From the full database, training and testing databases were generated. 20% of the remaining data are used for testing, while the remaining 80% are used for training. The training set of data is trained using four various machine learning techniques. Decision Tree, KNN, AdaBoost, and K-mean clustering. Accuracy rates for decision trees, neural networks, and naive bayes are 93.62%, 94%, and 90.74% respectively. [10]

This study examines several models based on these methodologies and methods, assessing the functionality of each. Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees (DT), Random Forest (RF), and ensemble models are among the models based on supervised learning techniques that researchers are reportedly very interested in. SVMRFE (Recursive Feature Elimination) and gain ratio algorithms choose the top 10 features, and Naive Bayes has an accuracy of 84.1584%. The accuracy of Naive Bayes is 83.49% when all 13 attributes from the Cleveland dataset are employed. 95.9% accuracy was achieved using SVM on the dataset from People's Hospital. [11]

Under this work, machine learning enables implicit, autonomous, learning without programming and improvement through experience. Numerous programs provide a means of automating the development of analytical models for data analysis. The importance of machine learning in the healthcare sector resides in its ability to manage enormous datasets beyond the scope of human skill and then consistently transform the analysis of the data into clinical insights that support physicians in the planning and delivery of care. Data mining is used 59.10%, Weighted Association Rule 60%, Neural Network 65%, Back Propagation technique 65.00%, Naive Bayesian algorithm 73.20%, and. [12]

They describe a method in this study for identifying the existence of heart disease based on the clinical data acquired from patients. The major goal of this study is to use a range of characteristics to develop a heart disease forecasting model. The data source for the experiment was Heart Disease UCI. In the dataset, there are 76 attributes. The use of machine learning prediction models in non-invasive medical decision support systems includes (SVM), (KNN), (ANN), (DT), (LR), (AB), and (NB) (FL). Using a logistic regression classifier, the classification accuracy for heart disease was 77%. Classification accuracy was 87.4%. [13]

Python is used to implement the project's backend, which is a version of an Android app. They chose a data source that has medical data on 70.000 persons. Eighty percent of the dataset is used for training and twenty percent is used for testing. The three algorithms Naive Bayes, Support Vector Machines, and Logistic Regression have all been put to the test. The methods were examined and refined using the Kaggle dataset. Less than 80% of the data are used for training; the remainder are used for testing. A person's likelihood of having heart disease or not is accurately determined by the heart disease prediction app. The system's accuracy for diagnosing heart disease is 73%. [14]

Therefore in study, an intelligent system for predicting heart disease was developed using data mining techniques. naïve bayes, decision trees, and neural networks. We were able to retrieve 909 records with 15 medical parameters (factors) in total from the Cleveland Heart Disease database. (455 records) Testing dataset plus training dataset (454 records). In comparison to naive bayes (47.58%) and decision trees (41.85%), neural networks have a greater rate of accurate predictions (49.34%). Because it offers the largest percentage of accurate predictions (86.12%), followed by Neural Networks (85.68%) and Decision Trees (80.4%), the Naive Bayes model looks to perform better than the other two when the entire population is processed. [15]

Inside this study, they were able to accurately predict heart disease using 6814 patient information and machine learning methods. They employed hybrid random forest and experimented with different feature combinations to achieve an accuracy of 88.7%. They collected a Kaggle dataset in order to forecast cardiac disease. There are records for 4239 patients in the dataset. 95% accuracy and 96.7% recall rate are provided via a support vector machine. The accuracy and recall percentages for the Synthetic Minority Oversampling, Random Forest, and Extra-tree Classifier methods are 91% and 93%, respectively. [16]

UC Irvine's machine learning repository provided a publicly available data collection on heart disease, which was used for testing in this study. There are 303 instances in the Cleveland heart disease dataset with 14 numerical features. This data collection is commonly used for research on heart disease because of its adequate number of traits and cases, absence of noisy data, and limited amount of missing data. They had an 89.2% accuracy rate. The algorithms Naive Bayes, SVM, KNN, NN, J4.8, RF, and GA computed on the complete data set allowed them to reduce the number of features from 14 to 12 without sacrificing accuracy. [17]

Data mining techniques like J48, Naive Bayes, REPTREE, CART, and Bayes Net were employed in this work to predict heart attacks. The study's conclusions show a 94% forecast accuracy rate. The patient data set was assembled using data received from medical professionals in South Africa. Only 11 attributes from the database are used to make the predictions needed to make a diagnosis of heart disease. The investigation makes use of the following algorithms When used to categorize and develop a model to diagnose heart attacks using patient data sets from medical professionals, the J48 algorithm has an accuracy rate of 94.00%, followed by those of the Bayes Net algorithm (95%), the Naive Bayes algorithm (93.2%), the Simple Cart algorithm (95.07% accuracy), and the REPTREE algorithm (95% accuracy). [18]

The authors of this study used machine learning algorithms, which are highly effective at producing results with a high level of accuracy, preventing the formation of cardiac problems in many patients and minimizing their consequences in those who are already experiencing them. Most researchers have used the 303 cases and 76 attributes of the

Cleveland Heart Disease Dataset, which is accessible through the UCI repository. Due to incomplete data, only 14 out of the 76 properties are useable. Logistic Regression, Random Forest, KNN, Ensemble Models with and Without Logistic Regression At 85%, 91.36%, 87.65%, 93.06%, and 92.77% accuracy, respectively. [19]

This study recommends using a range of machine-learning techniques to predict cardiac disease, including logistic regression, naive bayes, support vector machines, k nearest neighbors, random forests, and extreme gradient boost. The second dataset we used included 1190 patient record instances with 11 attributes and one target that were taken from Kaggle. They used a first dataset on heart illness that had 303 record instances and 14 different features. This dataset was taken from the well-known UCI machine learning library. For the study's initial dataset, Support Vector Machine delivered the highest accuracy of 92%. (SVM). The second dataset's accuracy from Random Forest was the highest at 93.12%. [20]

2.3 Scope of the Problem

In this study, linear regression, decision trees, and random forests are used to analyze the data in order to predict heart disease. Even though we obtained data sets in accordance with our research, more data may have been gathered. Accuracy increases with more data. The Linear Regression, Decision Tree, and Linear Regression (LinR) and Kernel Neural Networks (KNN) algorithms were used instead of the SVM and KNN algorithms because they were more accurate for our dataset. The results would be more accurate if we had more time to do our research.

2.4 Challenges

You always have to overcome obstacles in order to finish a task. We don't make any exceptions either because we've encountered so many challenges. Our research discusses the prognosis of cardiac illness. The three members of our group remain outside of Dhaka. Since the start of the Covid-19 scenario, we have been residing in our hometowns, which are many kilometers away, while our university is physically closed as a result of the education ministry's notice. Calling a meeting and reaching a decision regarding the analysis was not always advantageous for us. For group study, having an internet connection has always been a key problem. To consistently avoid similar problems, we divided our tasks among ourselves. However, we were in a Google Meet group meeting as group work is usually preferable while we were evaluating the data in machine learning algorithms. We did our utmost to function as a team. In order to distinguish our task from other tasks, we read a lot of articles that are relevant to our work. We'd also like to add that even though our supervisor has been really helpful to us virtually, we still feel a little unfortunate that we couldn't be personally monitored. These were the problems we ran with while conducting our research.

Chapter - 3

RESEARCH METHODOLOGY

3.1 Introduction

Using the datasets in our research study, we used a variety of supervised machine learning models to forecast heart disease. Once the dataset has been prepared, we use the methods in it. The creation of a dataset and the use of models are the most challenging aspects of the research process. It can be challenging to decide which model best matches the dataset. The entire methodology of our approach will be covered in this part.

Given below the work flow diagram:

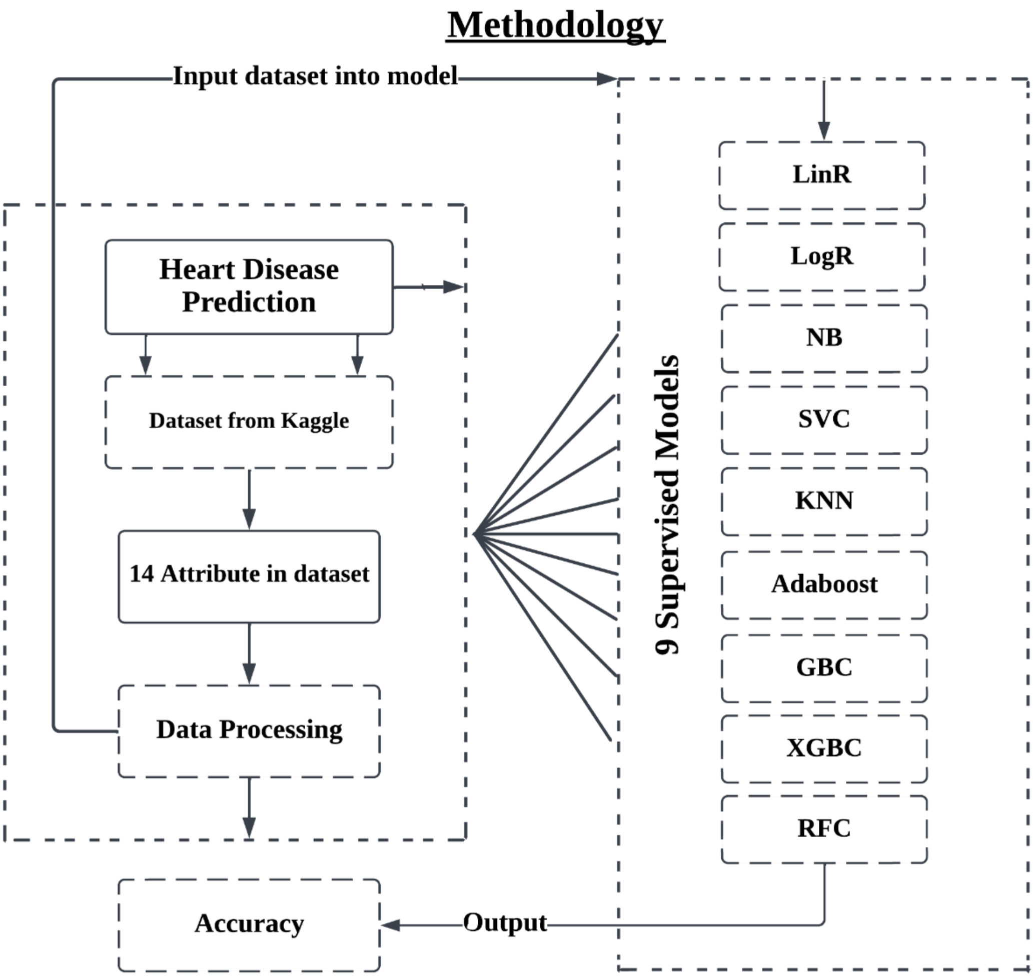


Figure 3.1.1: Flow Chart of Supervised Models

3.2 Research Subject and Instrumentation

We have utilized a range of ML algorithms in our work, named "heart disease prediction," to try and ascertain whether it is possible to predict heart disease with some degree of reasonable accuracy in the future.

3.2.1 Classifying Problem

The problem must be identified at the first and most crucial step. Our input and output variables have been selected. The intended outcome of the detection is displayed by our output variable. We built our model using a variety of ML techniques and compared the outcomes. The collection, preparation, and use of the data were the most challenging aspects of our study article. The objectives and tools that were available to us for formulating our strategy also had a significant influence.

3.2.2 Data Collection

We are using Kaggle to obtain data for our study. More than a thousand data were collected. Our dataset includes 14 more properties in addition to age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target.

3.2.3 Selecting data parameters

Among all tasks, choosing parameters is one of the most difficult. We consider hormone levels, common heart disease signs, physiological symptoms, and medical words when choosing those factors. In order to fit our data into the models, we separated it into 14 attributes.

3.2.4 Select the platform

Our data collection is utilized, and the outcomes of our statistical analysis are determined, using a Google Colaboratory platform. Our home computers can be kept more organized thanks to the Google Colaboratory's increased GPU and TPU resources and ease of use. The language we used to implement our code in the dataset was Python.

3.2.5 Data Preprocessing

We had altered the names of some of the parameters after gathering our data. Initial Key-Errors were encountered when using variables such as "Constrictive_Pericarditis", "Cholesterol", "Resting_Blood_Pressure" and "Resting_Electrocardiographic_Results" after executing the dataset. As a result, we manually transform variables using their short forms, such as "Resting_Electrocardiographic_Results" changing to "restecg", "Constrictive Pericarditis" convert into “cp” etc. We preprocess our variables in this manner. In addition, we manually preprocessed a few of the columns in our dataset. The data collection is then preprocessed to determine the value that is missing. In our data sets, we didn't find any missing values. We then input a model with our dataset.

3.2.6 Train and Testing Dataset

Selecting the test and train methodology is the most important phase in any research effort. Our study's dataset was split into training and testing halves. The models are trained using 800 of the 1026 data, while our model is tested using 226 of the 1026 data. So, we divide our resources into 20% for testing and 80% for training.

The given table is for data validation protocol.

Train and Test	**Dataset**
Train data 80%	800 out of 1026
Test data 20%	226 out of 1026

Table 3.2.6.1: Train and Test Table

3.2.7 Descriptions of used Machine Learning Models in our research

3.2.7.1 Linear Regression Model (LinR)

In order to anticipate values for inputs that are missing from the data set we have, we must find the line that best matches the data points on the plot. If the outputs fall on the line, we may use linear regression to forecast values for the missing inputs.

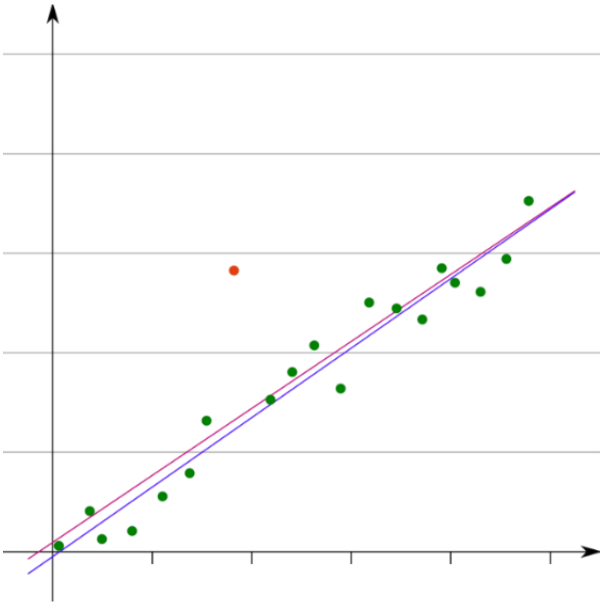
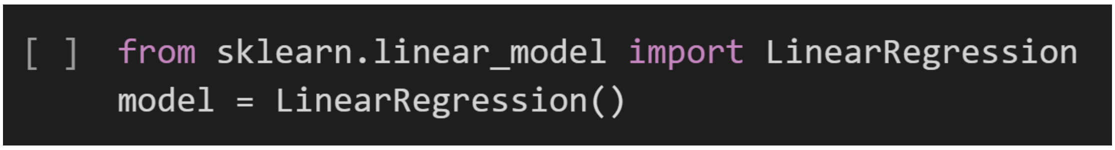


Figure 3.2.7.1.1: LinR Equation

In the Figure 3.2.7.1.1, we can see that how linear regression equation predicts



```
[ ]  from sklearn.linear_model import LinearRegression
     model = LinearRegression()
```

Figure 3.2.7.1.2: LinR Implementation

In the figure 3.2.7.1.2, we can see how the linear regression code is implemented in our work.

3.2.7.2 Logistic Regression (LogR)

To ascertain the relationships between two data components, the logistic regression data analysis method uses math. This relationship is then used to anticipate the value of one of those parameters based on the other. The forecast's outcome often falls into a narrow range, such as yes or no.

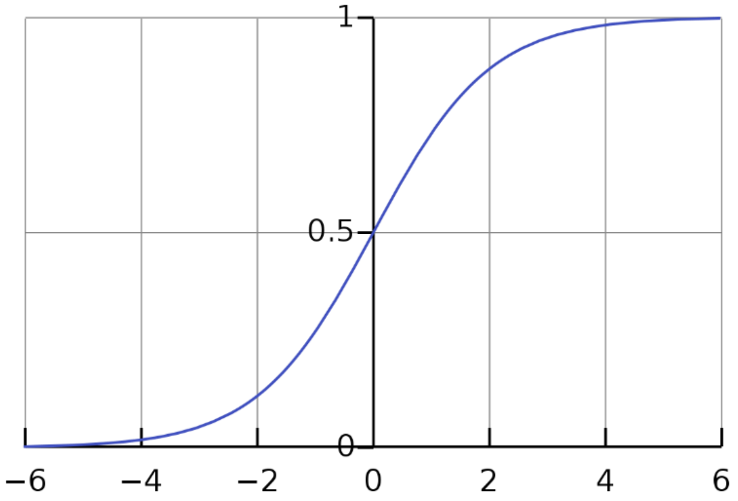
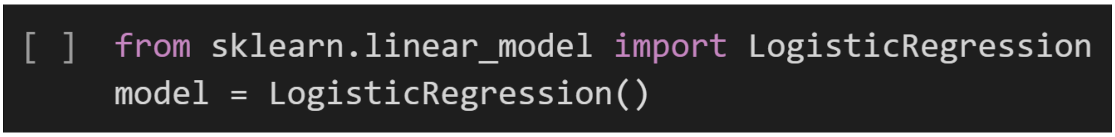


Figure 3.2.7.2.1: LogR Equation

In the Figure 3.2.7.2.1, we can see that how Logistic Regression equation predicts



```
[ ]  from sklearn.linear_model import LogisticRegression
     model = LogisticRegression()
```

Figure 3.2.7.2.2: LogR Implementation

In the figure 3.2.7.2.2, we can see how the Logistic Regression code is implemented in our work.

3.2.7.3 Gaussian Naive Bayes (NB)

Naive Bayes is a kind of classifier that applies the Bayes Theorem. It predicts membership probabilities for each class, such as the likelihood that a particular record or piece of data belongs to a particular class. The category with the highest likelihood is the most likely.

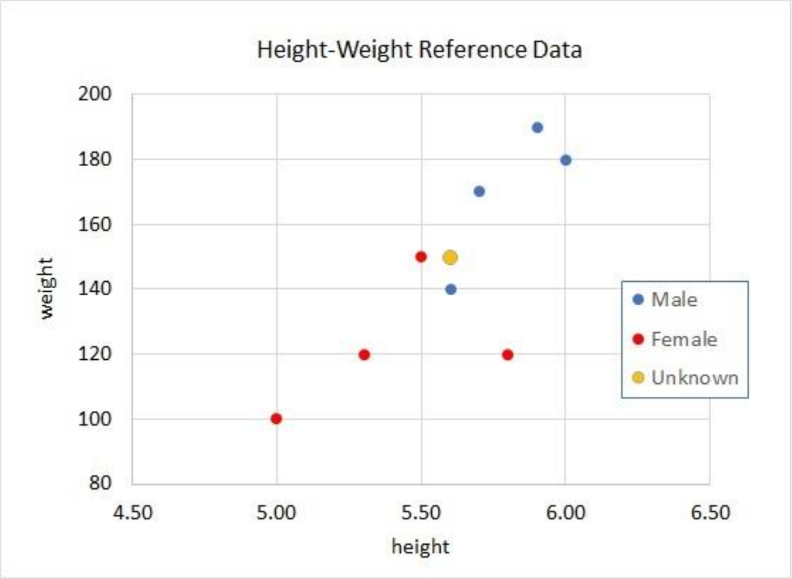
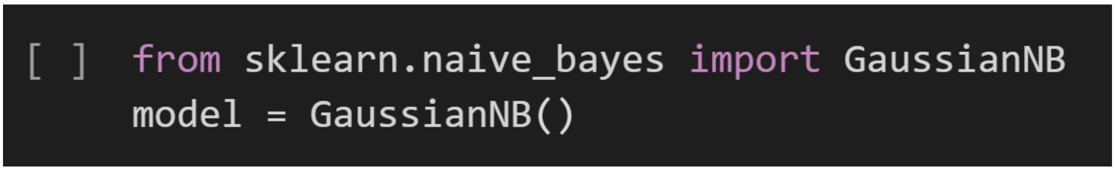


Figure 3.2.7.3.1: NB Equation

In the Figure 3.2.7.3.1, we can see that how Naive Bayes (NB) equation predicts



```
[ ]  from sklearn.naive_bayes import GaussianNB
     model = GaussianNB()
```

Figure 3.2.7.3.2: NB Implementation

In the figure 3.2.7.3.2, we can see that how Naive Bayes (NB) code is implemented in our work.

3.2.7.4 Support Vector Classifier (SVC)

The optimal decision boundary, sometimes referred to as a hyperplane, is found with the help of the Support Vector Classifier (SVC) method. The nearest line from each class is identified by the SVC algorithm. These points are referred to as support vectors.

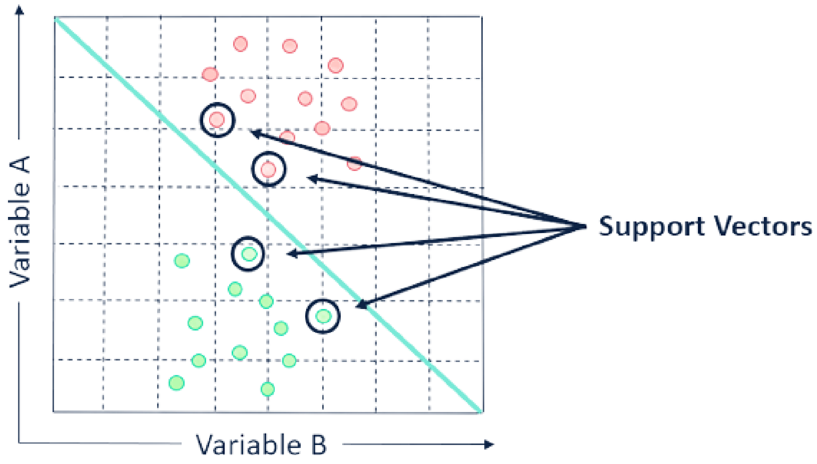
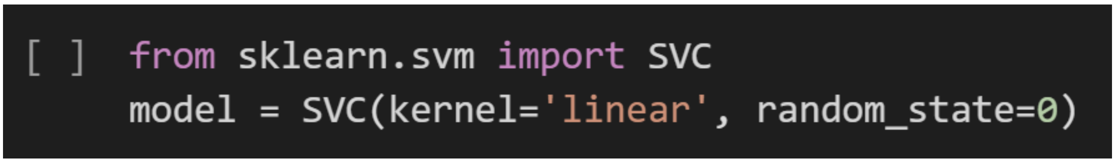


Figure 3.2.7.4.1: SVC Equation

In the Figure 3.2.7.4.1, we can see that how Support Vector Classifier (SVC) equation predicts



```
[ ]  from sklearn.svm import SVC
     model = SVC(kernel='linear', random_state=0)
```

Figure 3.2.7.4.2: SVC Implementation

In the figure 3.2.7.3.2, we can see that how Support Vector Classifier (SVC) code is implemented in our work.

3.2.7.5 K-Nearest Neighbor Classifier (K-NN)

The K-Nearest Neighbors algorithm, often known as KNN or K-NN, is a non-parametric, supervised learning classifier that depends on proximity to provide classifications or predictions regarding the grouping of a single data point.

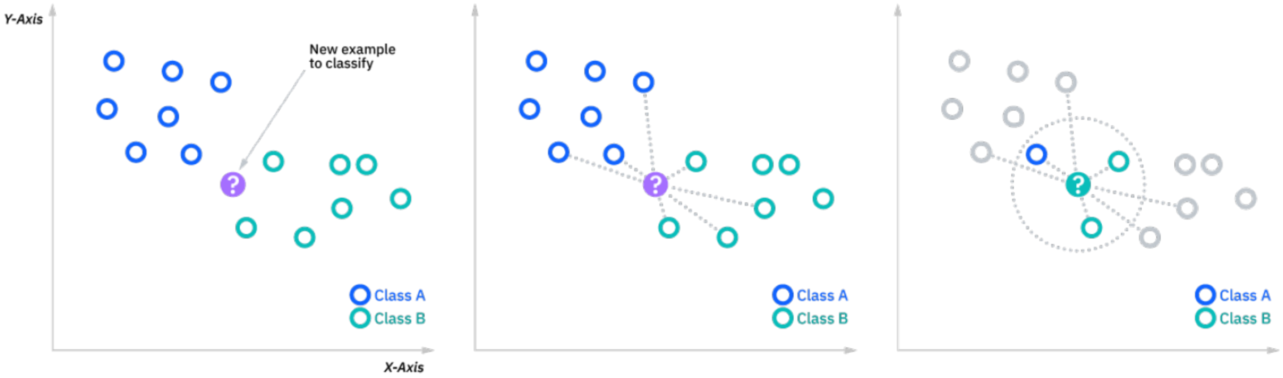
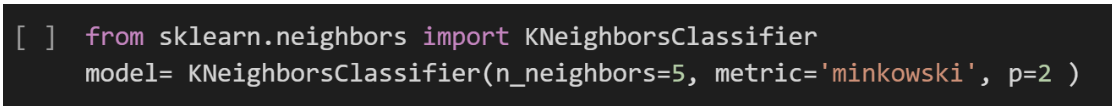


Figure 3.2.7.5.1: K-NN Equation

In the Figure 3.2.7.5.1, we can see that how K-Nearest Neighbors (K-NN) algorithm equation predicts



```
from sklearn.neighbors import KNeighborsClassifier
model= KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2 )
```

Figure 3.2.7.5.2: K-NN Implementation

In the figure 3.2.7.5.2, we can see that how K-Nearest Neighbors (K-NN) code is implemented in our work.

3.2.7.6 AdaBoost Classifier

In order to improve the performance of binary classifiers, the ensemble learning method AdaBoost was first created (sometimes referred to as "meta-learning"). AdaBoost employs an iterative methodology to enhance subpar classifiers by gaining knowledge from their mistakes.

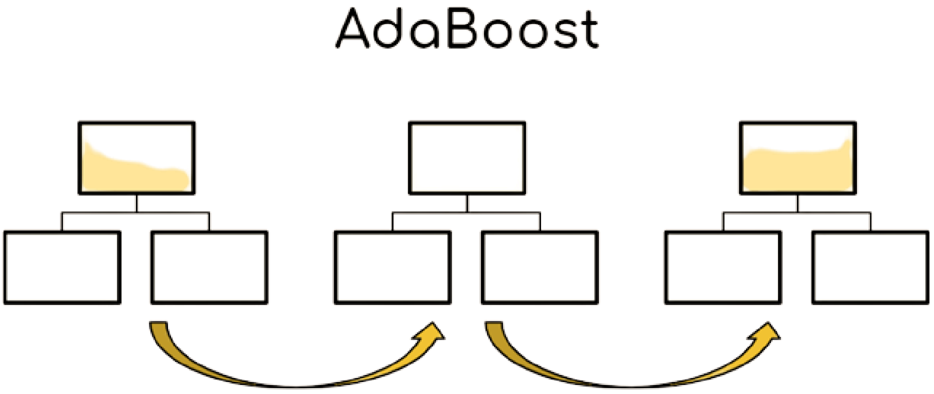
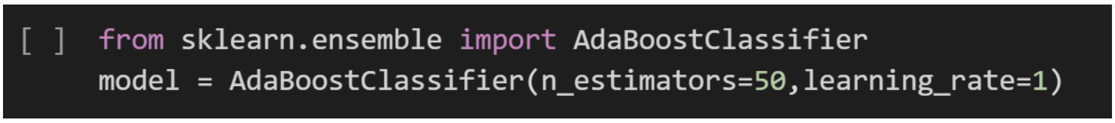


Figure 3.2.7.6.1: AdaBoost Equation

In the Figure 3.2.7.6.1, we can see that how AdaBoost algorithm equation predicts



```
[ ]  from sklearn.ensemble import AdaBoostClassifier
     model = AdaBoostClassifier(n_estimators=50,learning_rate=1)
```

Figure 3.2.7.6.2: AdaBoost Implementation

In the figure 3.2.7.6.2, we can see that how AdaBoost code is implemented in our work.

3.2.7.7 Gradient Boosting Classifier (GBC)

Using the gradient boosting method (as a Regressor), both continuous and categorical target variables can be predicted (as a Classifier). When used as a regressor, Mean Square Error (MSE) is the cost function; when used as a classifier, log loss is the cost function.

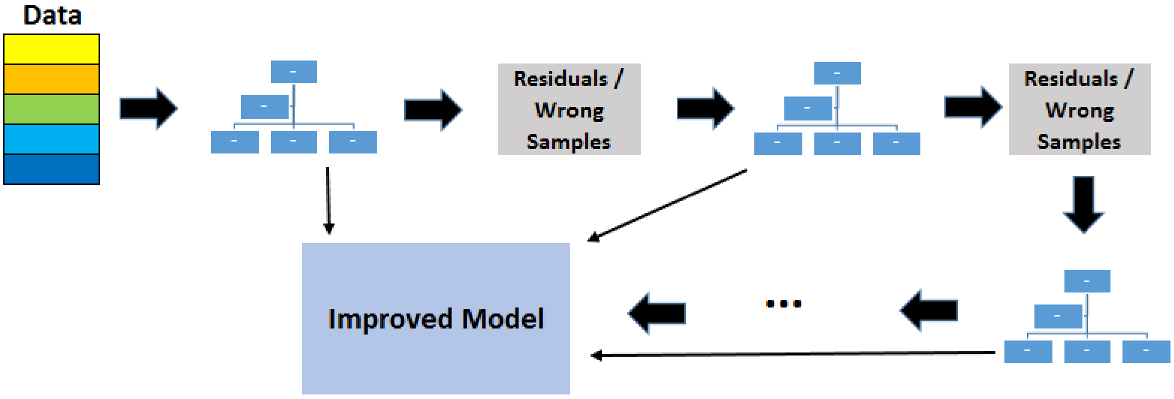
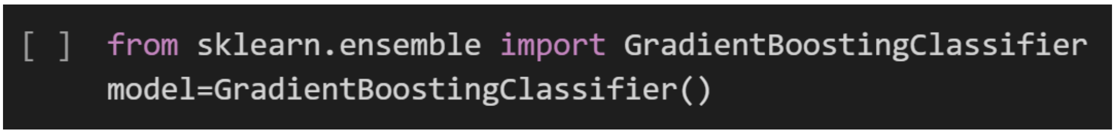


Figure 3.2.7.7.1: GBC Equation

In the Figure 3.2.7.7.1, we can see that how Gradient Boosting Classifier (GBC) algorithm equation predicts



```
[ ]  from sklearn.ensemble import GradientBoostingClassifier
     model=GradientBoostingClassifier()
```

Figure 3.2.7.7.2: GBC Implementation

In the figure 3.2.7.7.2, we can see that how Gradient Boosting Classifier (GBC) code is implemented in our work.

3.2.7.8 XGB Classifier (XGBC)

Open-source software named XGB Classifier uses the gradient boosted trees method extensively and does it effectively. Gradient boosting is a supervised learning technique that combines the predictions of a number of weaker, simpler models in an effort to accurately predict a target variable.

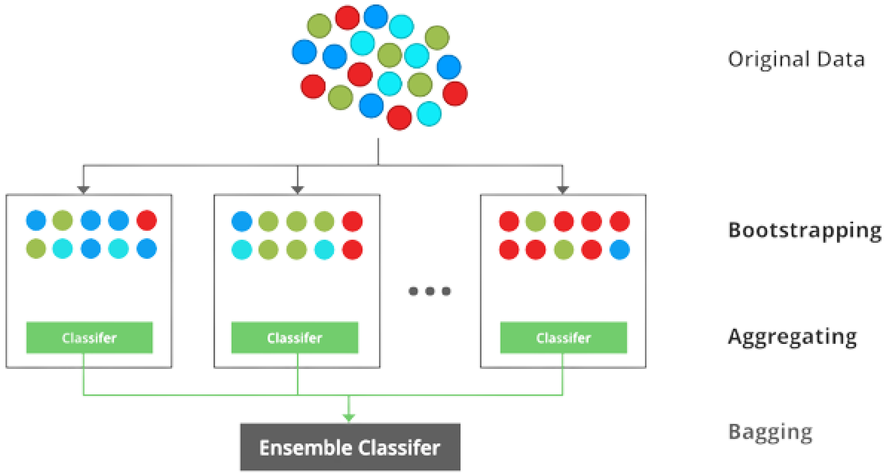
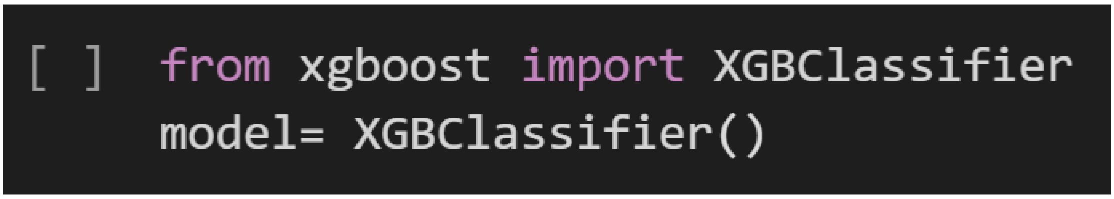


Figure 3.2.7.8.1: XGBC Equation

In the Figure 3.2.7.8.1, we can see that how XGB Classifier (XGBC) algorithm equation predicts



```
[ ]  from xgboost import XGBClassifier
     model= XGBClassifier()
```

Figure 3.2.7.8.2: XGBC Implementation

In the figure 3.2.7.8.2, we can see that how XGB Classifier (XGBC) code is implemented in our work.

3.2.7.9 Random Forest Classifier (RFC)

The Random Forest Classifier algorithm generates a large number of decision trees, which are then combined to produce a more accurate forecast. The Random Forest model is predicated on the notion that a number of uncorrelated models (the various decision trees) perform notably better when combined than when used independently.

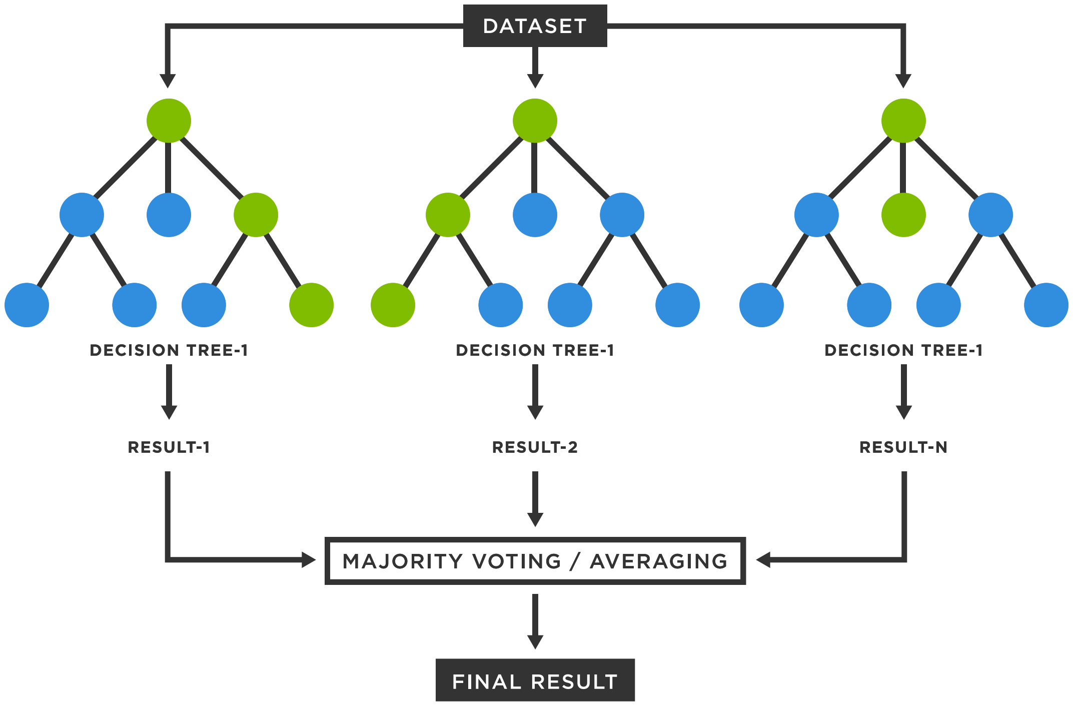
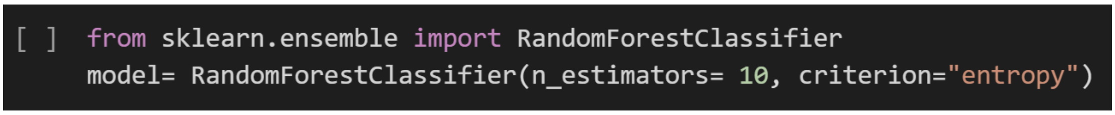


Figure 3.2.7.9.1: RFC Equation

In the Figure 3.2.7.9.1, we can see that how Random Forest Classifier (RFC) algorithm equation predicts



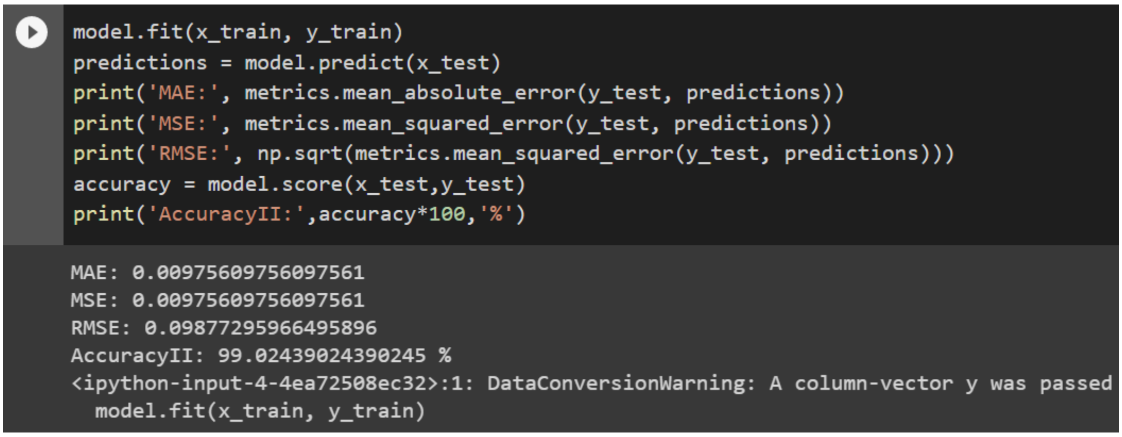
```
[ ]  from sklearn.ensemble import RandomForestClassifier
     model= RandomForestClassifier(n_estimators= 10, criterion="entropy")
```

Figure 3.2.7.9.2: RFC Implementation

In the figure 3.2.7.9.2, we can see that how Random Forest Classifier (RFC) code is implemented in our work.

3.2.8 Implementation of ML models on our dataset

Some of the nine machine learning models we have used in this dataset include Linear Regression (LinR), Logistic Regression (LogR), Gaussian Naive Bayes (NB), Support Vector Classifier (SVC), K-Nearest Neighbor Classifier (KNN), AdaBoost Classifier, Gradient Boosting Classifier (GBC), XGB Classifier (XGBC) and Random Forest Classifier (RFC).

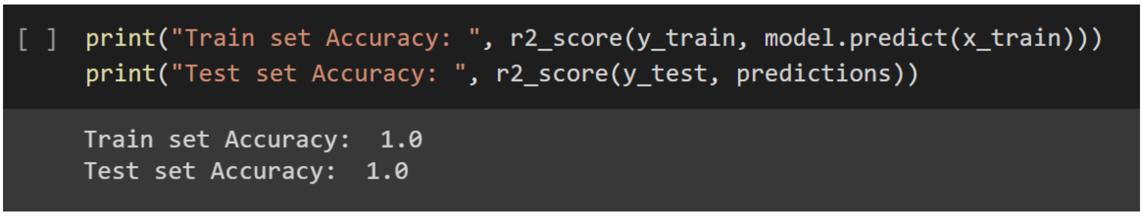


```
model.fit(x_train, y_train)
predictions = model.predict(x_test)
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
accuracy = model.score(x_test,y_test)
print('AccuracyII:',accuracy*100,'%')
```

```
MAE: 0.00975609756097561
MSE: 0.00975609756097561
RMSE: 0.09877295966495896
AccuracyII: 99.02439024390245 %
<ipython-input-4-4ea72508ec32>:1: DataConversionWarning: A column-vector y was passed
  model.fit(x_train, y_train)
```

Figure 3.2.8.1: Codes of Accuracy and Error rates

In the figure 3.2.8.1, we can see how the accuracy and error rate codes find these scores for our ML models.



```
[ ]  print("Train set Accuracy: ", r2_score(y_train, model.predict(x_train)))
     print("Test set Accuracy: ", r2_score(y_test, predictions))
```

```
Train set Accuracy:  1.0
Test set Accuracy:  1.0
```

Figure 3.2.8.2: Codes of Test and Train Accuracy

In the figure 3.2.8.2, we can see how the train and test accuracy codes find these things for DTR in our work.

3.2.9 Verification of Models:

The categorization of algorithms is done based on how precise and accurate each algorithm is. In this paper, we verify our machine learning models while taking accuracy and error rate from results.

3.2.10 Contrasting different machine learning algorithms:

We compare each ML technique we've tried in this section. The algorithms Accuracy, MAE, MSE, and RMSE are present (%) in the graph we utilize to compare all of our applied models.

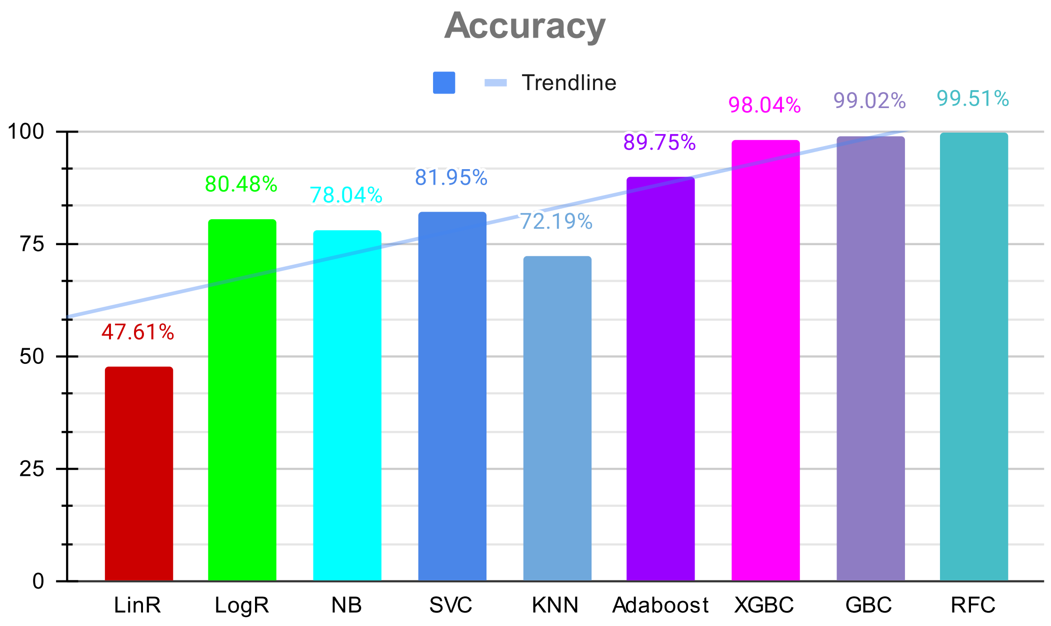


Figure 3.10.1: All Accuracy Graph

The statistical analysis of all the algorithm performances is shown in figure 3.10.1 above. Accuracy of the algorithm is shown in this bar graph. As seen in the graph above, Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC) and XGB Classifier (XGBC) provides the highest accuracy, which is 99.51%, 99.02% and 98.04% on our datasets.

Chapter - 4

RESULT & DISCUSSION

4.1 Introduction:

More than a thousand Kaggle data sets were employed using the machine learning supervised model. In this study, we identified and predicted heart illness using nine machine learning (ML) methods. Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC) and XGB Classifier (XGBC) has the highest accuracy on our dataset. Our accuracy rate for Random Forest Classifier (RFC) is 99.51%. 47.61% is reported as the accuracy of Linear Regression (LinR). Logistic Regression accuracy (LogR) is 80.48%. Naive Bayes (NB) estimate of accuracy is 78.04%. Support Vector Classifier (SVC) offers an accuracy of 81.95%. K-Nearest Neighbor (KNN) has a 72.19% accuracy rate. AdaBoost claims to have an accuracy rate of 89.75%. The Gradient Boosting Classifier (GBC) has a 99.02% accuracy rate. The XGB Classifier's (XGBC) accuracy is 98.04%. Random Forest Classifier (RFC) accuracy is 99.51%. According to our statistics, the Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC) and XGB Classifier (XGBC) performed brilliantly and produced reliable forecasts.

The accuracy graph is given below which represents the best accuracy graph –

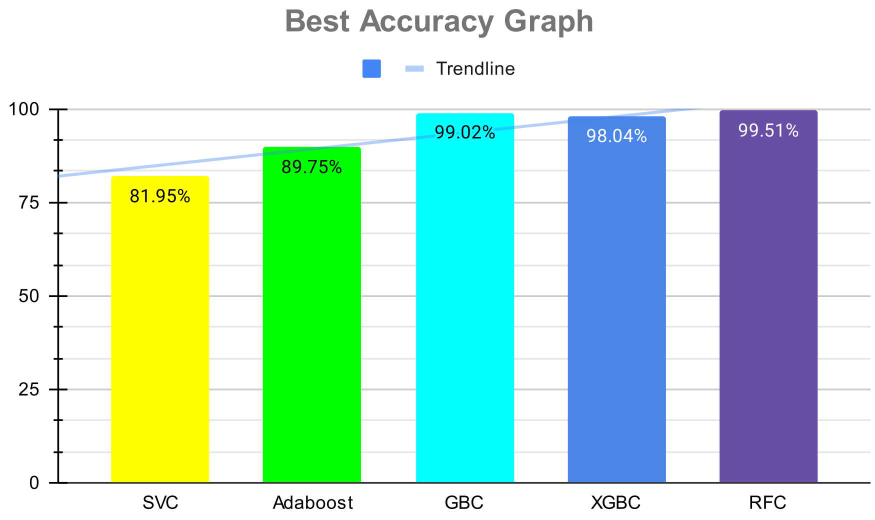


Figure 4.1.1: Best Accuracy Graph

4.2 Research Analysis:

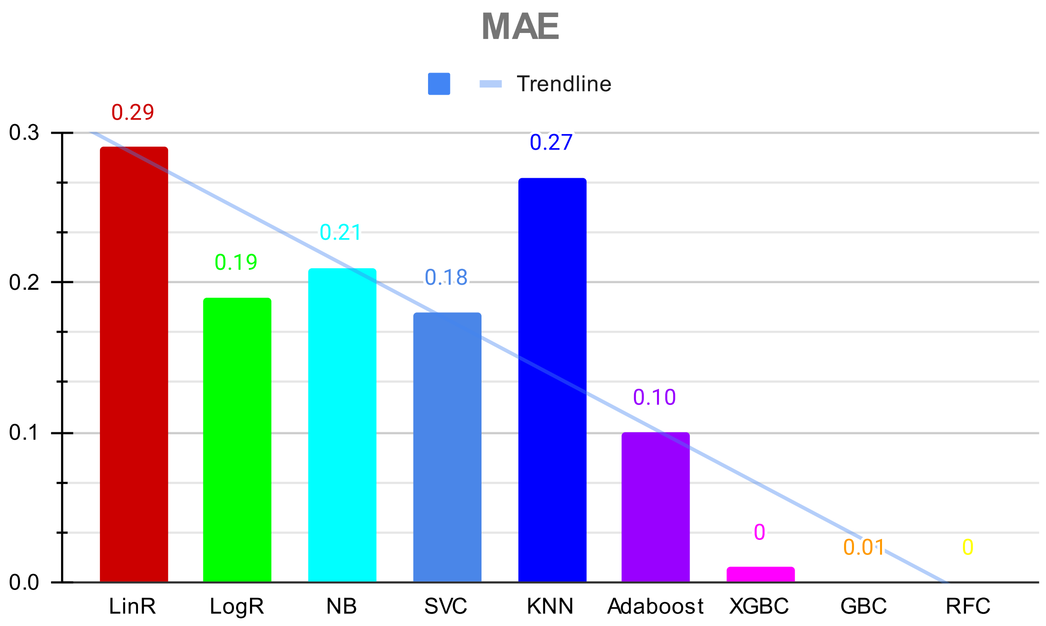


Figure 4.2.1: MAE Graph

Figure 4.2.1 illustrates that whereas LinR has the highest value of MAE, GBC, RFC, and RFC have the lowest values.

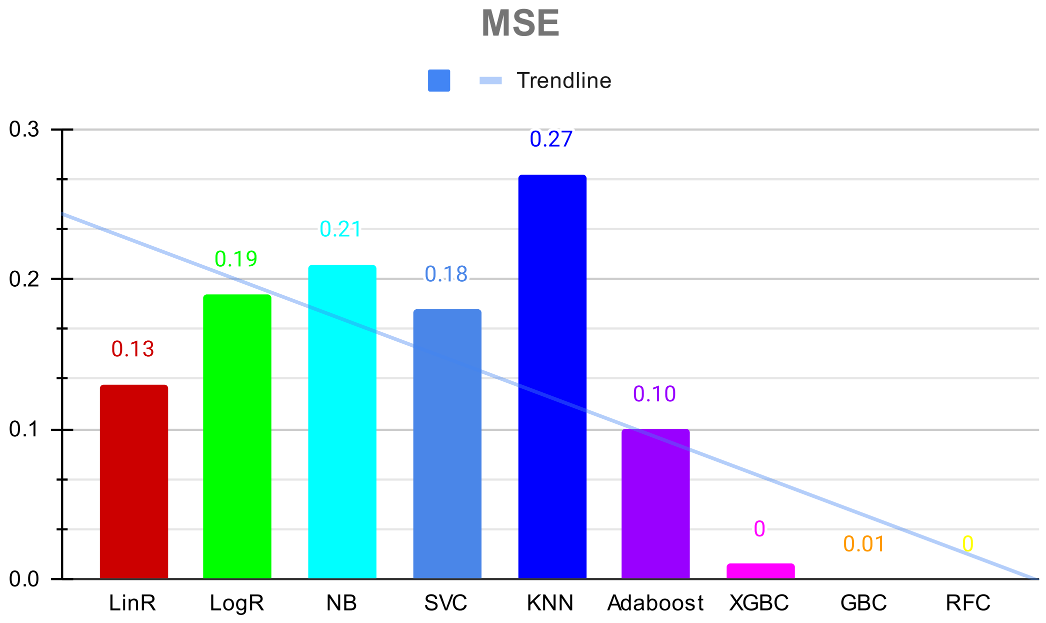


Figure 4.2.2: MSE Graph

Figure 4.2.2 demonstrates that GBC, RFC, and DTC have the lowest MSE, and if we look at greater serial numbers, KNN has that.

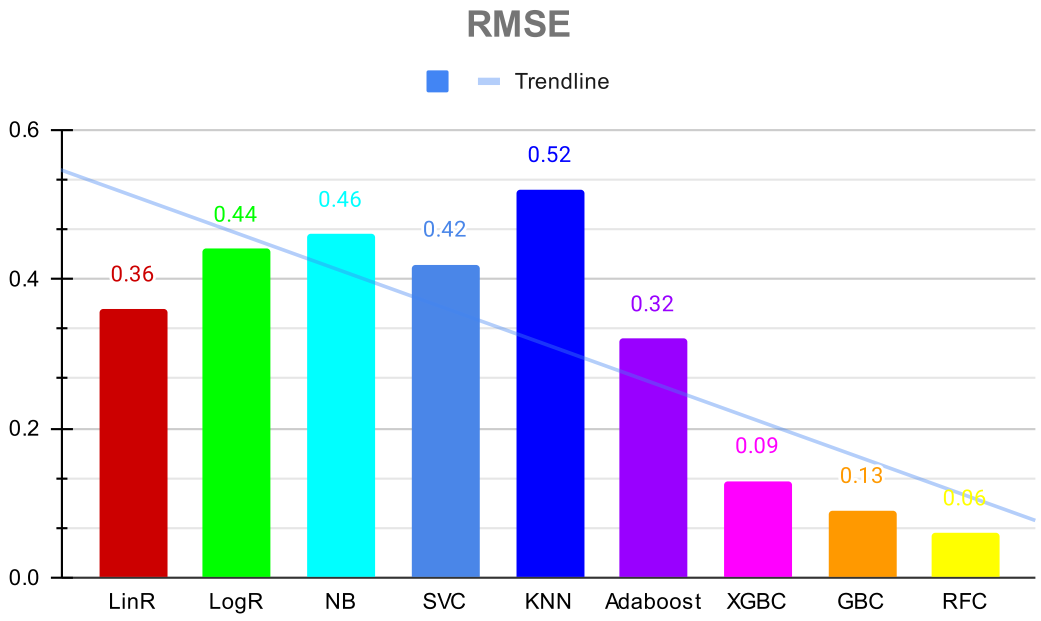


Figure 4.2.3: RMSE Graph

Figure 4.2.3 demonstrates that RFC has the lowest RMSE, and if we look at higher-order serial numbers, KNN has that.

4.3 Research Findings

In the table above, we displayed the results of our nine algorithms, which included values for Accuracy, MAE, MSE, and RMSE. The RFC's Accuracy = 99.51%, MAE = 0.00, MSE = 0.00, and RMSE = 0.00, GBC's Accuracy = 99.02%, MAE = 0.00, MSE = 0.00, and RMSE = 0.09 and XGBC's Accuracy = 98.04%, MAE = 0.01, MSE = 0.01, and RMSE = 0.13 values are the most important for result prediction. Which show that Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC) and XGB Classifier (XGBC) had the best accuracy and did well with the dataset we were given. So, we can say that Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC) and XGB Classifier (XGBC) models are the best for Heart Disease Prediction.

Below is a score table for Accuracy, MAE, MSE, and RMS -

Algorithms	Accuracy	MAE	MSE	RMSE
LinR	47.61%	0.29	0.13	0.36
LogR	80.48%	0.19	0.19	0.44
NB	78.04%	0.21	0.21	0.46
SVC	81.95%	0.18	0.18	0.42
KNN	72.19%	0.27	0.27	0.52
AdaBoost	89.75%	0.10	0.10	0.32
GBC	99.02%	0.00	0.00	0.09
XGBC	98.04%	0.01	0.01	0.13
RFC	99.51%	0.00	0.00	0.06

Table 4.3.1: Score table of Accuracy, MAE, MSE and RMSE

Chapter - 5

CONCLUSION

5.1 Summary of the Study

We wanted to examine the topic of " The Heart Disease Prediction using the Technique of Classification in Machine Learning using the concepts of Data Mining " and evaluate the machine learning algorithm and approaches for detecting heart disease in this research-based project.

5.2 Conclusion

In our research work, we used various ML models or algorithms to predict heart disease. We are gathering Kaggle data for our study. For our research, we gathered over 1000+ data and 9 algorithms. Out of 9 models, Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC) and XGB Classifier (XGBC) techniques performed exceptionally well and provided accuracy of 99.51%, 99.02%, 98.04% on our dataset, whereas Linear Regression (LinR) algorithms provided accuracy of just 47.61%. We determine which ML algorithms are most effective for our dataset through this analysis. The fact that we were unable to gather a large amount of data for this study is a limitation of our publication. We were unable to gather data from Bangladeshi hospitals due to the pandemic. For our research, we are gathering Kaggle datasets for this reason.

5.3 Further Implication of the Study

There is still much work to be done after this investigation. We have a goal to provide more information and depth to this project so that it can be used more broadly. Future studies will allow us to determine how it affects Bangladeshis. We will expand our datasets and gather information from hospitals throughout Bangladesh. Work on improving the dataset of Bangladeshi people's accuracy using ML techniques. We aim to conduct data analysis not just for Bangladesh but also for the entire world.

REFERENCES

[1] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

[2] Purusothaman, G., & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. *Indian Journal of Science and Technology*, *8*(12), 1.

[3] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.

[4] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 43-48.

[5] Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics*, *2*, 100016.

[6] Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early heart disease prediction using data mining techniques. *Computer Science & Information Technology Journal*, *28*, 53-59.

[7] Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia Computer Science*, *85*, 962-969.

[8] G. Chakravarthi1 , S.MD. Jabeer 2 1PG Scholar, Department of CSE, Global College of Engineering and Technology, Kadapa 2Assistant Professor, Department of CSE, Department of CSE, Global College of Engineering and Technology, Kadapa, © November 2021| IJIRT | Volume 8 Issue 6 | ISSN: 2349-6002

[9] Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Informatics in Medicine Unlocked*, *26*, 100696.

[10] Golande, A., & Pavan Kumar, T. (2019). Heart disease prediction using effective machine learning techniques. *International Journal of Recent Technology and Engineering*, *8*(1), 944-950.

[11] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.

[12] Bhajibhakare, M. M., Shaikh, N., & Patil, D. (2019). Heart disease prediction using machine learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *7*(XII), 455-460.

[13] Vardhan, G. H., Reddy, N. S. S., & Umamaheswari, K. M. Heart disease prediction using machine learning.

[14] Prathamesh K, Pratik P, Kaustubh L, Prof. Rovina D, 2022 IRJET, Impact Factor value: 7.529, ISO 9001:2008 Certified Journal

[15] Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications* (pp. 108-115). IEEE.

[16] Lakshmanarao, A., Swathi, Y., & Sundareswar, P. S. S. (2019). Machine learning techniques for heart disease prediction. *Forest*, *95*(99), 97.

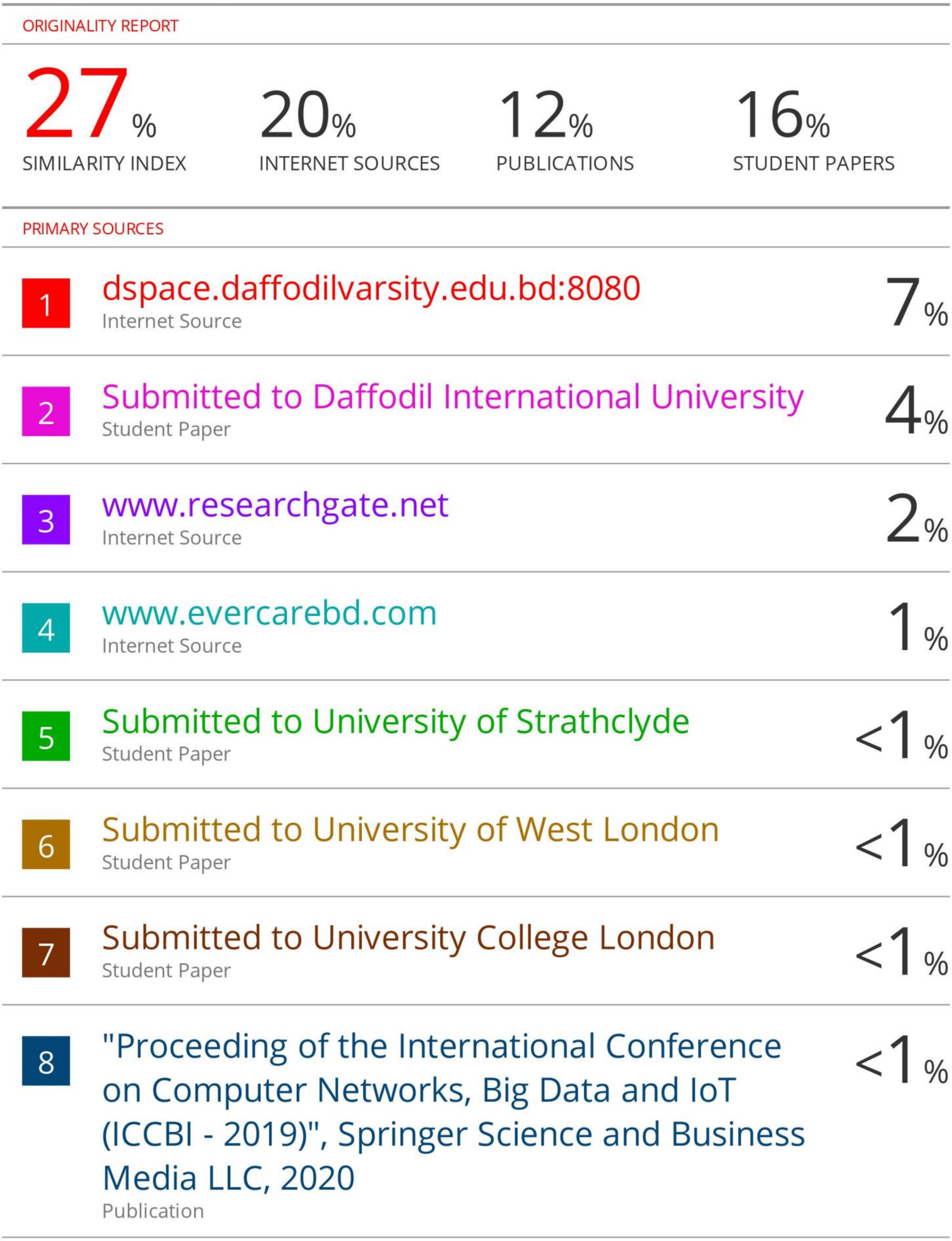
[17] Tarawneh, M., & Embarak, O. (2019, February). Hybrid approach for heart disease prediction using data mining techniques. In *International Conference on Emerging Internetworking, Data & Web Technologies* (pp. 447-454). Springer, Cham.

[18] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and computer Science* (Vol. 2, No. 1, pp. 25-29).

[19] Pandita, A., Yadav, S., Vashisht, S., & Tyagi, A. (2021). Review Paper on Prediction of Heart Disease using Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *9*(6).

[20] Bora, N. I. K. H. I. L., Gutta, S., & Hadaegh, A. (2021). Using Machine Learning To Predict Heart Disease. *The Project Committee In Partial Fulfillment Of The Requirements For The Degree Of Master Of Science In Computer Science, California State University*, 1-18.

Plagiarism Report



ORIGINALITY REPORT

27%	20%	12%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	4%
3	www.researchgate.net Internet Source	2%
4	www.evercarebd.com Internet Source	1%
5	Submitted to University of Strathclyde Student Paper	<1%
6	Submitted to University of West London Student Paper	<1%
7	Submitted to University College London Student Paper	<1%
8	"Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2019)", Springer Science and Business Media LLC, 2020 Publication	<1%