HEART DISEASE PREDICTION USING MACHINE LEARNING: A COMPARATIVE ANALYSIS

BY

MD. MURAD HOSSIN
ID: 191-15-2547
AND

MD. RIFAT BHUIYAN
ID: 191-15-2375

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

Supervised By

Al Amin Biswas
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project/internship titled **"Heart Disease Prediction Using Machine Learning: A Comparative Analysis"**, submitted by Md. Murad Hossin, ID No: 191-15-2547 and Md. Rifat Bhuiyan, ID No: 191-15-2375 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 01 February 2023.

BOARD OF EXAMINERS

4/2/23

Chairman

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

Tania Khatun

Internal Examiner

Tania Khatun (TK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

1.2.23

Internal Examiner

Ms. Lamia Rukhsara (LR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Al Amin Biswas, Senior Lecturer, Department of CSE,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Alim 31.a.2023

Al Amin Biswas
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Jahangir 01.02.2023

Mohammad Jahangir Alam
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Murad

Md. Murad Hossin
ID: 191-15-2547
Department of CSE
Daffodil International University

Rifat

Md. Rifat Bhuiyan
ID: 191-15-2375
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Al Amin Biswas, Senior Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning (ML)*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

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

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

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

ABSTRACT

Heart disease is one of the main causes of death worldwide and the most dangerous ailment. Early identification of cardiovascular disease will reduce mortality. The medical establishment has struggled in recent years to accurately anticipate cardiac disease. According to recent data, one person dies every minute from heart disease. Data science is needed to comprehend the vast volumes of new healthcare data. KNN, LR, AdaBoost, XGB, RF, GB, SVM, and DT machine-learning algorithms are used to forecast cardiac disease. Using these algorithms, we could analyze a person's heart disease risk based on dataset attributes. This study used two types of data. The first heart disease dataset had 918 patient records, 11 attributes, and one target. This dataset combines five well-known cardiac datasets. The second dataset on cardiovascular disease included 70000 patient records, 11 characteristics, and a single goal. This research offers a comparison study by investigating the efficacy of numerous machine learning methods. For our first and second datasets, Gradient Boost (GB) was the most accurate, with 91.80% and 74.50%, respectively. Considering the results of the trial, the Gradient Boost (GB) algorithm has the highest level of accuracy, which is 91.80%, compared to other models and studies being done at the time. A realistic web application is also developed.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Introduction

The primary reason for a heart attack is artery blockage. It is also known as cardiovascular disease and arterial high blood pressure [1], among others. About 26 million people worldwide are afflicted by cardiovascular disease [2]. This number is projected to increase considerably in the future years if proper precautions are not implemented [3]. In addition to adopting a healthy lifestyle and controlling one's food, the correct timing of diagnosis and a detailed analysis are also crucial aspects that can eventually save lives [4]. Hence, this study has taken a modest step toward preventing the deaths of heart failure patients as well as provides a strategy for enhancing the diagnostic accuracy of patients based on their healthcare history.

During the period, the majority of patients undergo many tests that can burden them with additional time, physical activity, and costs [5]. As revealed by prior research, the most prevalent causes of cardiovascular disease include improper diet, tobacco use, excessive sugar consumption, obesity, and excess body fat [3], [6], while arm and chest pain are the most prevalent symptoms [7]. Notably, these reasons are separate; effective analysis of this kind of data might enhance the diagnosis process and benefit cardiac surgeons. Previously, a number of techniques, like Extreme Learning Machines [8], cardiovascular disease classification [9], and classification algorithms based on machine learning [1], were applied to improve the heart failure diagnosis process by various researchers. This study aims to improve the effectiveness of classifications by performing tests with numerous machine-learning techniques in order to make more efficient utilization of the medical datasets gathered from diverse sources.

1.2 Motivation

Increased mortality is attributable to heart disease. Typically, heart disease is not identified in its early stages. Until the blockage in the vessels surpasses a threshold, the patient typically does not exhibit severe symptoms. Angiography and other invasive procedures for diagnosing heart disease are expensive and dangerous. Using a decision support system, non-invasive testing can diagnose heart disease. This could limit the possibility of human error in detecting cardiac illness. A judgment support structure can aid in the diagnosis of heart disease before it becomes urgent. If heart disease is found early enough, a person can avoid heart failure and live longer. In disease detection systems, precision is of the utmost significance. The primary purpose of achieving this analysis is to establish a highly accurate technique for detecting heart illness based on medical factors and machine learning techniques that can assist healthcare professionals.

1.3 Rationale of the Study

As just a result of technological advancement, which has led to an improvement in the standard of living and quality of life, there is a growing emphasis on health care today. Moreover, the current financial situation has necessitated establishing a long-term health care method that utilizes the most of the current resources. This research is intended to forecast the occurrence of cardiac disease utilizing several characteristics. The originality of this work is in the concept of developing a highly accurate predictive machine learning algorithm for heart disease utilizing numerous datasets. In this experiment, two distinct datasets were utilized to determine the top essential features and to forecast heart disease.

1.4 Research Questions

These questions are the focus of this investigation:

- Question 1: Does the machine learning model give the same results for people with heart disease when compared to different datasets from different sources?
- Question 2: How can machine learning algorithms be used to get accurate results from datasets that are unknown?

1.5 Expected Output

- Analyze, clean, select, and transform numerical and categorical characteristics.
- To determine the most important characteristics of two distinct datasets for predicting heart disease.
- Compare various classification methods for properly classifying heart disease diagnoses in unseen examples.
- Predict heart disease by employing several machine learning approaches on two datasets.
- Contrast the obtained outcomes with those discovered in the scientific literature.
- A web application capable of making accurate predictions based on user input.

1.6 Project Management and Finance

Effectively working to make sure that all of the components of this project management strategy are managed. This comprised the analysis of project goals, schedule development, teamwork, risk assessment, monitoring, implementation, outcomes, enhancing the project's efficiency and effectiveness, the status of the report, updates and future work were all included in this. The Daffodil International University provided financial assistance for this project. The funding source also supported the study's planning, implementation, management, research, analysis of the data, writing, reviewing, approval of the report and publication of the paper.

1.7 Report Layout

The subsequent sections in this research project are organized as follows: In Chapter 2, outline the foundation for our research and explore related projects, comparable discoveries, and the scope and depth of the concerns, including challenges. In Chapter 3, we discuss the study topic, methodology, and data collection method, statistical analysis, and proposed models. The project's overall technique is explained in detail. In Chapter 4, the efficiency of heart disease forecasting is presented alongside the appropriate results and discussion. In Chapter 5, The classification results and comparison outcomes for each dataset are discussed. Describe the effect on society, sustainability, and the environment. Chapter 6 contains concluding remarks as well as a discussion of future work considerations.

CHAPTER 2
Background

2.1 Preliminaries

To get things started, let's talk about the topic of this research and the goals that it aims to accomplish. The background study consists of a discussion of the investigated region in addition to the most recent information around the topic of the study, prior research conducted on this topic, and facts of historical significance pertaining to the topic. In the ideal case scenario, this component of the study should sufficiently explain the topic's history and background material. As per the requirements of our proposal, we will investigate and identify the impediments being addressed, as well as review the research and identify ways to resolve the problems and difficulties. This research will address the issue faced by individuals attempting to identify the disease, and it will analyze the merits of the research for the overall benefit of humankind.

We will describe our relevant attempt, the project, and any challenges we encountered while performing this investigation. There will be discussion of further connected research articles, and the "Related Works" section will include connection to these articles. In the subsequent sections, In the overview part, we will go over the highlights of the project, and in the challenge section, we will describe the difficulties we encountered within our studies.

2.2 Related Works

Kedia et al., 2021 [10] have established a large number of machine learning techniques, the like decision trees, XGB, random forest, and logistic regression. Kaggle's Cardiovascular Disease dataset, which has 70000 samples with 11 input features, was used to train the analysis of these algorithms. Final accuracy figures for the four models were calculated as follows: XGB accuracy was 72.70%, RF accuracy was 69.18%, DT accuracy was 62.87%, and LR accuracy was 72.39%.

Bashir et al., 2019 [11] applied LR (SVM), Random Forest, Naive Bayes, and Decision Tree, Logistic Regression to forecast cardiac illness from UCI Heart Disease data. The ultimate accuracy for the following models was 82.22% for DT, 82.56% for LR, 84.85% for SVM, 84.24% for NB, and 84.17% for RF, respectively.

Hamdaoui et al., 2020 [12] employed K-Nearest Neighbor, Random Forest, Support Vector Machines, Decision Tree, and Naive Bayes Classifier to predict cardiovascular disease utilizing the dataset of heart disease from UCI. Accuracy, recall, and precision are among the performance metrics. The final accuracy for the following models was, respectively, 84.28% for NB, 81.23% for KNN, 81.42% for SVM, 77.14% for RF, and 82.28% for DT.

Mohan et al., 2021 [13] used Random Forest, Decision Tree, Support Vector Machine, and K-Nearest Neighbor to forecast cardiovascular disease utilizing the cardiovascular disease dataset from Kaggle. The final accuracy for the following models was, respectively, 63.40% for KNN, 71.00% for RF, 68.40% for DT, and 72.50% for SVM.

Using the heart disease data of UCI, Pouriyeh et al., 2017 [14] employed Radial Basis Function (RBF), K-Nearest Neighbor, Naive Bayes, Support Vector Machine, Decision Tree, and Multi-Layer Perceptron to predict heart disease. Final accuracy was 77.55% for the DT model, 83.49% for the NB model, 83.16% for the KNN model, 82.83% for the MLP model, 83.82% for the RBF model, and 84.15% for the SVM model.

Ouf and ElSeddawy, 2021 [15] have developed many machine learning techniques, including Linear Discriminant Analysis, Logistic regression, Decision Tree, Naive Bayes, Neural Network, K-Nearest Neighbor, Random Forest, and Support Vector Machine. Analyses of such techniques are trained by Kaggle and the UCI Dataset on Heart Disease. The final accuracy for all models was determined to be, respectively, 71.25 % for LR, 71.06 % for SVM, 69.61 % for KNN, 65.11 % for RF, 58.44 % for DT, 57.51 % for NB, 62.24 % for Linear Discriminant Analysis, and 71.82 % for NN. The final accuracy for the UCI dataset for all models was determined to be, respectively, 81.97% for LR, 86.21 % for

SVM, 88.52 % for KNN, 89.01 % for RF, 79.31% for DT, 80.33 % for NB, 80.33 % for Linear Discriminant Analysis, and 86.21 % for NN.

Dwivedi, 2016 [16] used a variety of methods for predicting heart disease utilizing the Heart Disease Data of UCI, including Naive Bayes classifiers, Support Vector Machines, Logistic Regression classifiers, Artificial Neural Networks, K-Nearest Neighbors, and classification trees. Overall accuracy was 84.00% for the ANN model, 82.00% for the SVM, 83.00% for the NB, 85.00% for the LR, 80.00% for the KNN, and 77.00% for the classification tree.

Shah, Patel and Bharti, 2020 [17] utilized the UCI Heart Disease Data Set to predict heart disease utilizing K-Nearest Neighbor, Naive Bayes, Random Forest, Decision Tree, and Algorithms. For the following models, the ultimate accuracy was 88.15 % for NB, 80.26 % for DT, 78.94% for KNN, and 84.21 % for RF.

Singh and Kumar, 2020 [18] have worked on multiple machine learning techniques, including Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Logistic Regression. Examination of these computational structures is taught by the UCI Database of Heart Disease, which has 303 examples with 14 input attributes. The ultimate accuracy for the following models was found to be 83.00% for SVM, 80.00% for LR, 79.00% for DT, and 87.00% for KNN, respectively.

Jagtap et al., 2019 [19] have developed a variety of machine learning methods, including Nave Bayes, Support Vector Machines, and Logistic Regression. Analysis of these algorithms is trained by Kaggle and Cleveland Foundation medical research, particularly in the Dataset on Heart Disease. The final accuracy for the three models was determined to be, respectively, 64.40% for SVM, 61.45% for LR, and 60.00% for NB.

2.3 Comparative Analysis and Summary

The assessed accuracy is not up to the mark, which is a significant limitation that has been identified in previous investigations. As evidenced, the widely used machine learning methods have been rarely utilized. The past study that was employed to forecast heart illness for a person using ML approaches was therefore thoroughly described. Utilizing a certain dataset and the ML techniques clarified in the subsequent section, The goal of this project is to enhance previous findings. The results section includes information on how each model performed. However, the dataset and models chosen for this investigation are derived from prior studies. Earlier research, decision trees, random forests logistic regression, and support vector machines have been found as the most widely used machine learning approaches. The Kaggle dataset used in this study had previously been available through the UCI machine learning archive. The accuracy and details of previous experiments for cardiac disease prediction were measured. The comparison research is included in the results section to help readers understand how well the classifiers performed in this study and in earlier research.

2.4 Scope of the Problem

Following study area has been identified based on an analysis of existing literature:

- Researchers only look at certain datasets, even though many different machine learning models can be employed to evaluate many different datasets.
- Using a variety of feature selection techniques, researchers have extracted important features from the dataset. Hybrid techniques for feature selection can be created to pick crucial features in order to attain a higher degree of classification precision.
- Most researchers examined the influence of missing data on a classification system by randomly inserting fictitious missing values into datasets. Additional techniques are available to give missing values. In addition, there was no accessible

mechanism for systematically analyzing the influence of missing data on the dataset.

- Various decision support systems for cardiovascular disease have been suggested, each with varying degrees of precision. The majority of researchers eliminated records with missing values from the dataset and did not utilize them for training. In addition, the problem of missing values and the method for feature selection were not addressed jointly.
- In recent decades, a number of automated systems for identifying cardiovascular disease have been presented, employing diverse methods of machine learning to enhance the performance and accuracy of the diagnostic process. This has introduced a new dimension to the process of medical diagnosis, but there is still potential for improvement.

2.5 Challenges

The collection of data is the most challenging activity we need to complete. In an effort to improve the trustworthiness of our predictions, when seen from the perspective of Bangladesh, the collection of data regarding health-related issues is almost impossible. Because of this, we are unable to manually collect data and instead rely on data obtained from open sources. Open-source data has various difficulties, like missing information, irrelevant characteristics, licenses, etc. After we had collected the dataset, we were required to preprocess the dataset, which was yet another laborious task. Applying a machine learning model is another problem because the data is open source and a variety of researchers have already gained a fair level of accuracy. Therefore, the machine learning model will not be able to incorporate current findings if the preprocessing procedure is not carried out effectively. The most difficult challenge is to demonstrate that our machine learning model performs better as well as other models.

CHAPTER 3
Research Methodology

3.1 Research Subject and Instrumentation

It is clear that the information is the most significant aspect of the test that we are going to take. Finding reliable information as well as an efficient approach or model is a critical part of the investigational job that we do, and it is of the utmost importance that a professional do so. Additionally, we need to look at earlier exam papers that are analogous to the ones we have now. At that time, we will have to select one path forward from among the following alternatives:

- What kinds of data should be gathered?
- How do we know the knowledge we've gathered is accurate?
- Does each piece of information require the same structure?
- How would you recommend labeling every piece of data?

3.2 Dataset Utilized

3.2.1 Dataset 1: Heart Disease

The dataset utilized in current research (the Heart Disease Dataset) was acquired using the Kaggle platform [20]. However, the data was initially provided in the UCI data repository for machine learning, although with a greater number of attributes and fewer occurrences. In order to achieve better model results, we opted for the adjusted data on the Kaggle platform. This dataset was made by putting together different datasets that were already out there but had never been put together before. There are five heart datasets in this data collection. put together based on 11 similar characteristics. This represents the greatest dataset on heart disease research that has been made available so far.

Next, we will describe the attributes and provide a summary of the dataset. The dataset includes 918 observations or patient records that will be used for classification using 12 features, only one of which points to the result that determines whether or not the individual has been diagnosed with heart disease. The dataset includes eleven independent variables and a single dependent variable, which is the target (heart disease). Visit Table 3.1 for a more comprehensive illustration and overview of the facts. This table shows the attributes, data types, and descriptions of the dependent and independent properties.

TABLE 3.1: DESCRIPTION OF DATASET 1 ATTRIBUTES

Attribute Name	**Data Type**	**Description**
Age	Integer	This characteristic includes a patient's age (in years).
Sex	String	This attribute includes the patient's gender as a string, [M = male, F = female].
Chest pain type	String	This characteristic includes, in string format, the type of chest pain reported by the patient [TA = typical angina, ATA = atypical angina, NAP = non-anginal pain, Asy = asymptomatic].
Resting BP	Integer	The patient's blood pressure while at rest in mmHg
Cholesterol	Integer	Serum cholesterol in mm/dL
Fasting BS	Integer	Blood sugar during fasting [1 = if fasting BS > 120 mg/dL, 0 = otherwise]
Resting ECG	String	Electrocardiogram (ECG) result [Normal = Normal, ST = having ST-T wave abnormalities (T wave inversions and/or ST elevation or depression >0.05 mV), LVH = demonstrating probable or definite left ventricular hypertrophy according to Estes' criteria]
Max HR	Integer	This attribute specifies the patient's maximal heart rate [Numeric number between 60 and 202].
Exercise Angina	String	Angina due to exercise [Y = yes, N = no]
Old Peak	Float	In comparison to rest, exercise causes ST depression.
ST-Slope	String	Slop or the peak exercise ST segment [Up = upsloping, Flat = flat, Down = down sloping]
Heart Disease	Integer	Binary Target, [Class 1 = heart disease, Class 0 = normal]

3.2.2 Preprocessing of Dataset 1

Among the most crucial and significant stages in machine learning is dataset preprocessing, and it must be performed prior to model creation for optimal results and to eliminate noisy data.

The experimental dataset has both categorical and numerical characteristics. The categorical features are made up of string data, while the numerical features are made up of numbers. For instance, the attribute of 'sex' in Table 1 contains the values M and F, which represent men and females, respectively. Humans are capable of effortlessly comprehending categorical values, but they are not suitable for training machine learning algorithms. The objective of converting, or data normalization, is to translate this data into integer values suitable for machine learning. There are two methods for transforming category variables to numeric values: encoding of dummy variables and labeling. Encoding of dummy variables employs 0 and 1 regardless of the number of categories to represent the exclusion or inclusion of a category. This research uses label encoding to convert category information into numeric values, hence making the value directly deployable with deep learning and supervised learning methods. The outcomes of applying the encoding technique to the entire dataset using the Python Panda package are shown in Table 3.2.

TABLE 3.2: A SAMPLE DATASET OF ENCODING RESULTS

Age	Sex	ChestPainType	RestingBP	Choles.	FastingBS	Resting ECG	Max HR	Exer.Angina	Oldpeak	ST_Slope	HeartDisease
Before Encoding											
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
After Encoding											
12	1	1	41	147	0	1	98	0	10	2	0
21	0	2	55	40	0	1	82	0	20	1	1
9	1	1	31	141	0	2	25	0	10	2	0

3.2.3 Splitting of Dataset 1

After normalizing the data, it is separated to sets of training and testing using a proportion of 0.8:0.2 for model training and evaluation. Empirically, the ratio of splits is determined based on outcomes assessments utilizing split ratios of 0.75: 0.25, 0.70: 0.30, and 0.85: 0.15. The optimal findings were obtained using a train-test split ratio of 0.80:0.20. Table 3.3 provides the sample size for training and testing. Class 0 represents healthy individuals, and Class 1 represents heart disease patients.

TABLE 3.3: THE NUMBER OF TRAINING AND TESTING SAMPLES FOR DATASET 1

Set	Total Samples	Class 0	Class 1
Training	734	321	413
Testing	184	89	95

3.2.4 Dataset 2: Cardiovascular Disease

For this investigation, a cardiovascular disease dataset from Kaggle [21] was employed. The obtained data set comprises 70, 000 patient records. It includes twelve properties, one of which is a target variable. Ages between 29 and 64 were included for the evaluation. Additionally, their stature and mass are recorded. Gender values of 1 and 0 were assigned to male and female patients, respectively. To determine the influence, blood pressures were analyzed (systolic and diastolic). The results of the patients' cholesterol and glucose tests were classified as normal, above normal, and severely above normal. Cardiac problems are strongly associated with drinking and smoking. Those two factors have binary values assigned to them. The value '1' indicates that he or she is a "smoker/drinker," whilst '0' indicates that he or she is a "nonsmoker/nonalcoholic." The patients who engage in regular physical exercise are denoted with a "1" and the others with a "0." The target attribute is the existence or absence of cardiovascular disease. It is made up of binary values. The number "0" symbolizes normal, whereas the number "1" reflects confirmed cases of cardiac disease. Table 3.4 demonstrates the respective attributes, range of values, and descriptions.

TABLE 3.4: DESCRIPTION OF DATASET 2 ATTRIBUTES

Attribute Name	Description	Range of Values
age	Age	int (years)
gender	Gender	categorical code
height	Height	int (cm)
weight	Weight	float (kg)
ap_hi	Systolic blood pressure	int
ap_lo	Diastolic blood pressure	int
cholesterol	Cholesterol	1: normal, 2: above normal, 3: well above normal
gluc	Glucose	1: normal, 2: above normal, 3: well above normal
smoke	Smoking	binary
alco	Alcoholic	binary
active	Physical activity	binary
cardio	Presence or absence of cardiovascular disease	binary

3.2.5 Preprocessing of Dataset 2

The purpose of data preprocessing is to assure data quality and utility. This phase is crucial since it directly influences our model's capacity for learning.

Scaling the values in the data such that they fall inside the range of 0 to 1 in order to teach the machine learning systems, and scaling all the values before training the models.

3.2.6 Splitting of Dataset 2

In order to facilitate the training procedure, the algorithm for machine learning, the selected column inside the collection is referenced. After that, we separate the collection of data into its training-set and test-set components, each of which is referred to as a sub-dataset. The data collection was partitioned up into learning and examination sections using a ratio of 80% to 20%, respectively. There are 70,000 records in the Cardiovascular Disease dataset, of which about 14,000 items constitute the test dataset, whereas the remainder 56,000 items constitute the training dataset that is given in Table 3.5.

TABLE 3.5: THE NUMBER OF TRAINING AND TESTING SAMPLES FOR DATASET 2

Set	Total Samples
Training	56,000
Testing	14,000

3.3 Statistical Analysis

Descriptive statistics as well as exploratory data analysis are covered in this section. In order to do statistical data analysis, a variety of statistical operations must be performed. It is a type of statistically-based quantitative analysis to try to quantify the data. Survey and observational data are the most common examples of quantitative data.

Exploratory Data Analysis relates to the essential procedure of conducting preparatory work investigations on the data to detect patterns, detect abnormalities, examine theories, and verify statistically supported expectations and visual representations.

3.3.1 Dataset 1

There are only seven numeric attributes out of twelve attributes present in the data that are tabulated below.

TABLE 3.6: DESCRIPTIVE STATISTICS OF HEART DISEASE DATASET

	count	mean	std	min	25%	50%	75%	max
Age	918.0	53.510893	9.432617	28.0	47.00	54.0	60.0	77.0
RestingBP	918.0	132.396514	18.514154	0.0	120.00	130.0	140.0	200.0
Cholesterol	918.0	198.799564	109.384145	0.0	173.25	223.0	267.0	603.0
FastingBS	918.0	0.233115	0.423046	0.0	0.00	0.0	0.0	1.0
MaxHR	918.0	136.809368	25.460334	60.0	120.00	138.0	156.0	202.0
Oldpeak	918.0	0.887364	1.066570	-2.6	0.00	0.6	1.5	6.2
Heart Disease	918.0	0.553377	0.497414	0.0	0.00	1.0	1.0	1.0

In Figure 3.1, we illustrate how the different attributes of the heart disease dataset are distributed.

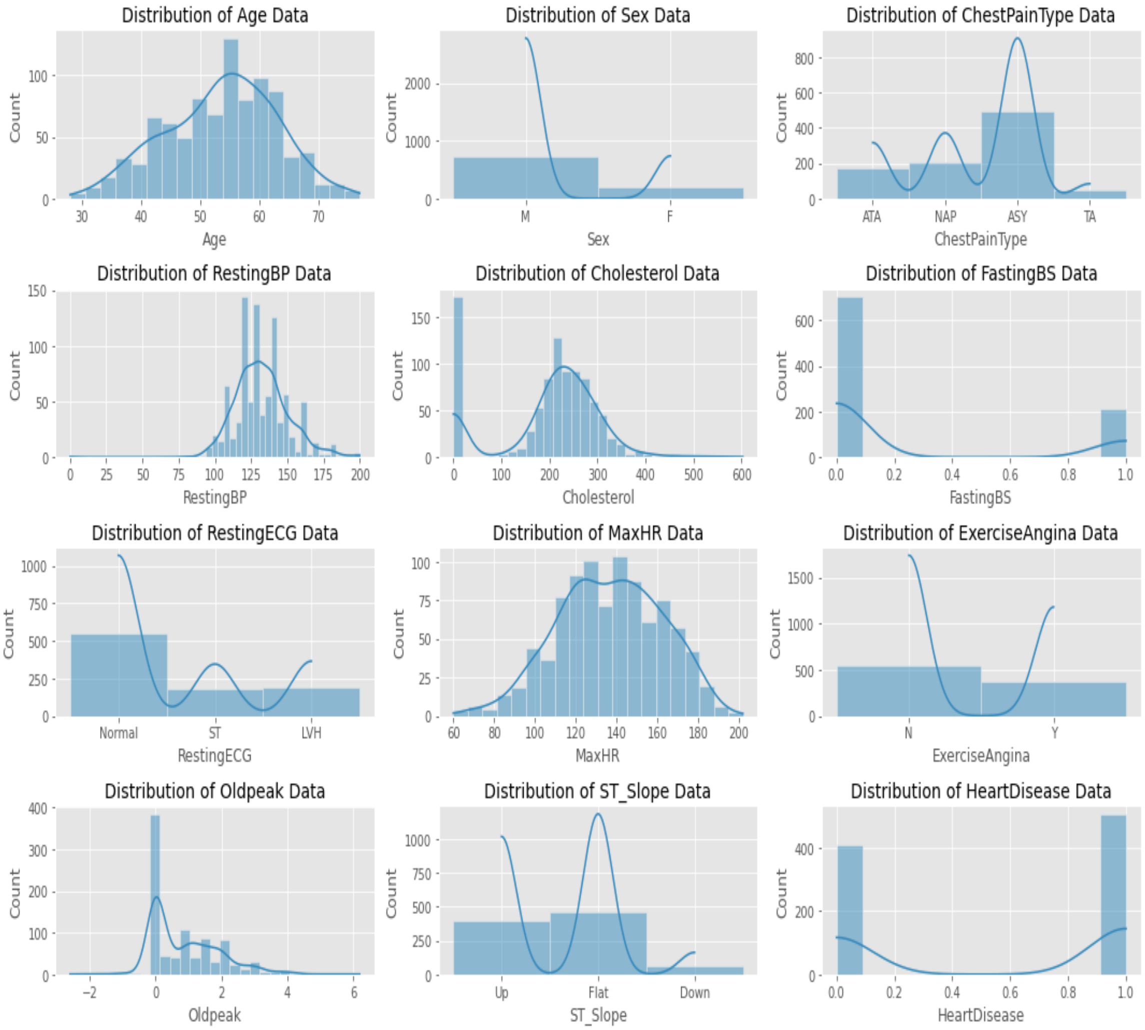
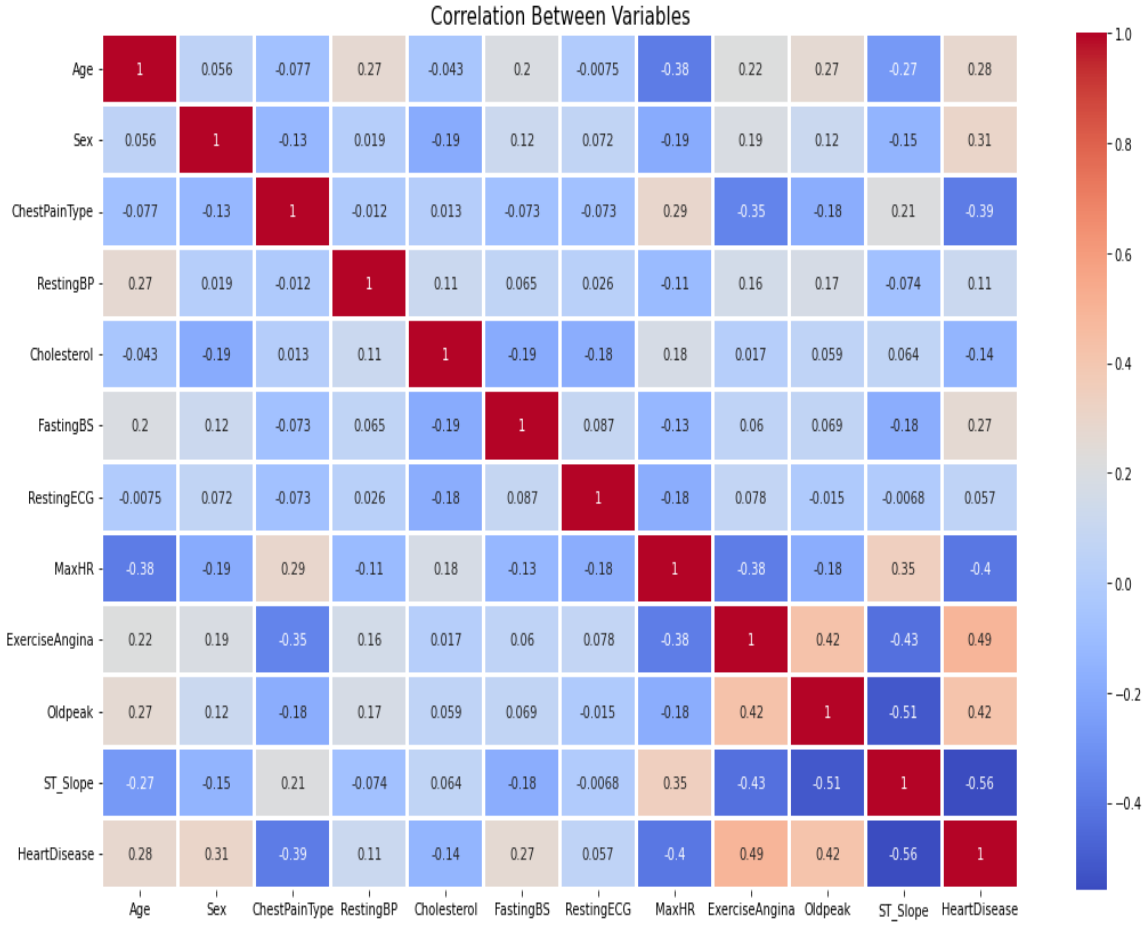


Figure 3.1: Attributes distribution of the heart disease dataset

Figure 3.2 depicts a heatmap of the correlation matrix between variables within the Heart Disease dataset. The heatmap is highly effective for data visualization since it reveals a great deal about the correlations between these 12 aspects of datasets.



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
Age	1	0.056	-0.077	0.27	-0.043	0.2	-0.0075	-0.38	0.22	0.27	-0.27	0.28
Sex	0.056	1	-0.13	0.019	-0.19	0.12	0.072	-0.19	0.19	0.12	-0.15	0.31
ChestPainType	-0.077	-0.13	1	-0.012	0.013	-0.073	-0.073	0.29	-0.35	-0.18	0.21	-0.39
RestingBP	0.27	0.019	-0.012	1	0.11	0.065	0.026	-0.11	0.16	0.17	-0.074	0.11
Cholesterol	-0.043	-0.19	0.013	0.11	1	-0.19	-0.18	0.18	0.017	0.059	0.064	-0.14
FastingBS	0.2	0.12	-0.073	0.065	-0.19	1	0.087	-0.13	0.06	0.069	-0.18	0.27
RestingECG	-0.0075	0.072	-0.073	0.026	-0.18	0.087	1	-0.18	0.078	-0.015	-0.0068	0.057
MaxHR	-0.38	-0.19	0.29	-0.11	0.18	-0.13	-0.18	1	-0.38	-0.18	0.35	-0.4
ExerciseAngina	0.22	0.19	-0.35	0.16	0.017	0.06	0.078	-0.38	1	0.42	-0.43	0.49
Oldpeak	0.27	0.12	-0.18	0.17	0.059	0.069	-0.015	-0.18	0.42	1	-0.51	0.42
ST_Slope	-0.27	-0.15	0.21	-0.074	0.064	-0.18	-0.0068	0.35	-0.43	-0.51	1	-0.56
HeartDisease	0.28	0.31	-0.39	0.11	-0.14	0.27	0.057	-0.4	0.49	0.42	-0.56	1

Figure 3.2: Heatmap of the correlation matrix for dataset 1

3.3.2 Dataset 2

There are twelve numeric attributes out of twelve attributes in the data presented in the table below.

TABLE 3.7: DESCRIPTIVE STATISTICS OF CARDIOVASCULAR DISEASE DATASET

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
age	70000	19468.86	2467.25	10798	17664	19703	21327	23713
gender	70000	1.34	0.47	1.00	1.00	1.00	2.00	2.00
height	70000	164.35	8.21	55.0	159	165	170	250
weight	70000	74.20	14.39	10.00	65.00	72.00	82.00	200
ap_hi	70000	128.81	154.01	-150	120	120	140	16020
ap_lo	70000	96.63	188.47	-70.00	80	80	90	11000
cholesterol	70000	1.36	0.68	1.00	1.00	1.00	2.00	3.00
gluc	70000	1.22	0.57	1.00	1.00	1.00	1.00	3.00
smoke	70000	0.08	0.28	0.00	0.00	0.00	0.00	1.00
alco	70000	0.05	0.22	0.00	0.00	0.00	0.00	1.00
active	70000	0.80	0.39	0.00	1.00	1.00	1.00	1.00
cardio	70000	0.49	0.50	0.00	0.00	0.00	1.00	1.00

In Figure 3.3, we illustrate how the different attributes of the cardiovascular disease dataset are distributed.

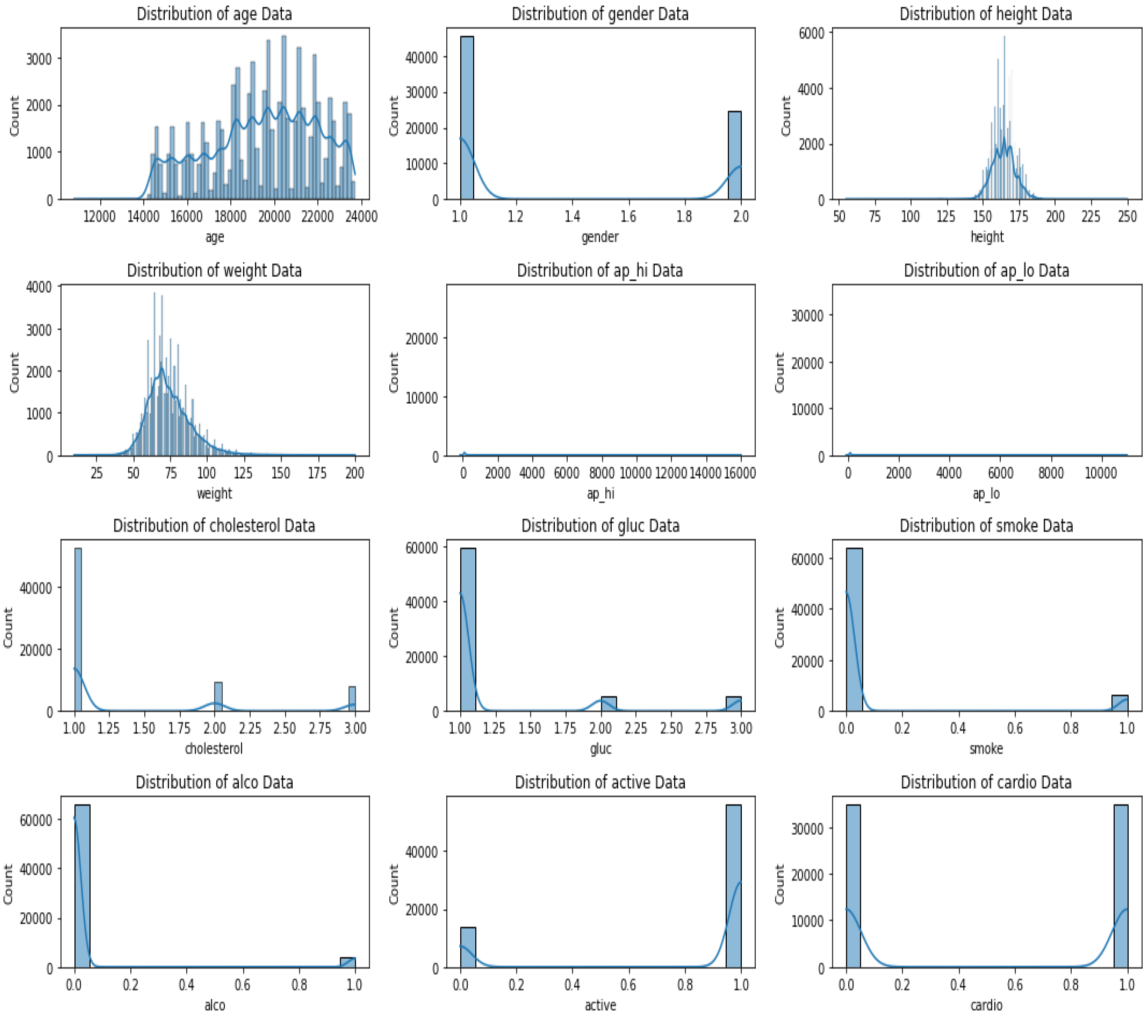
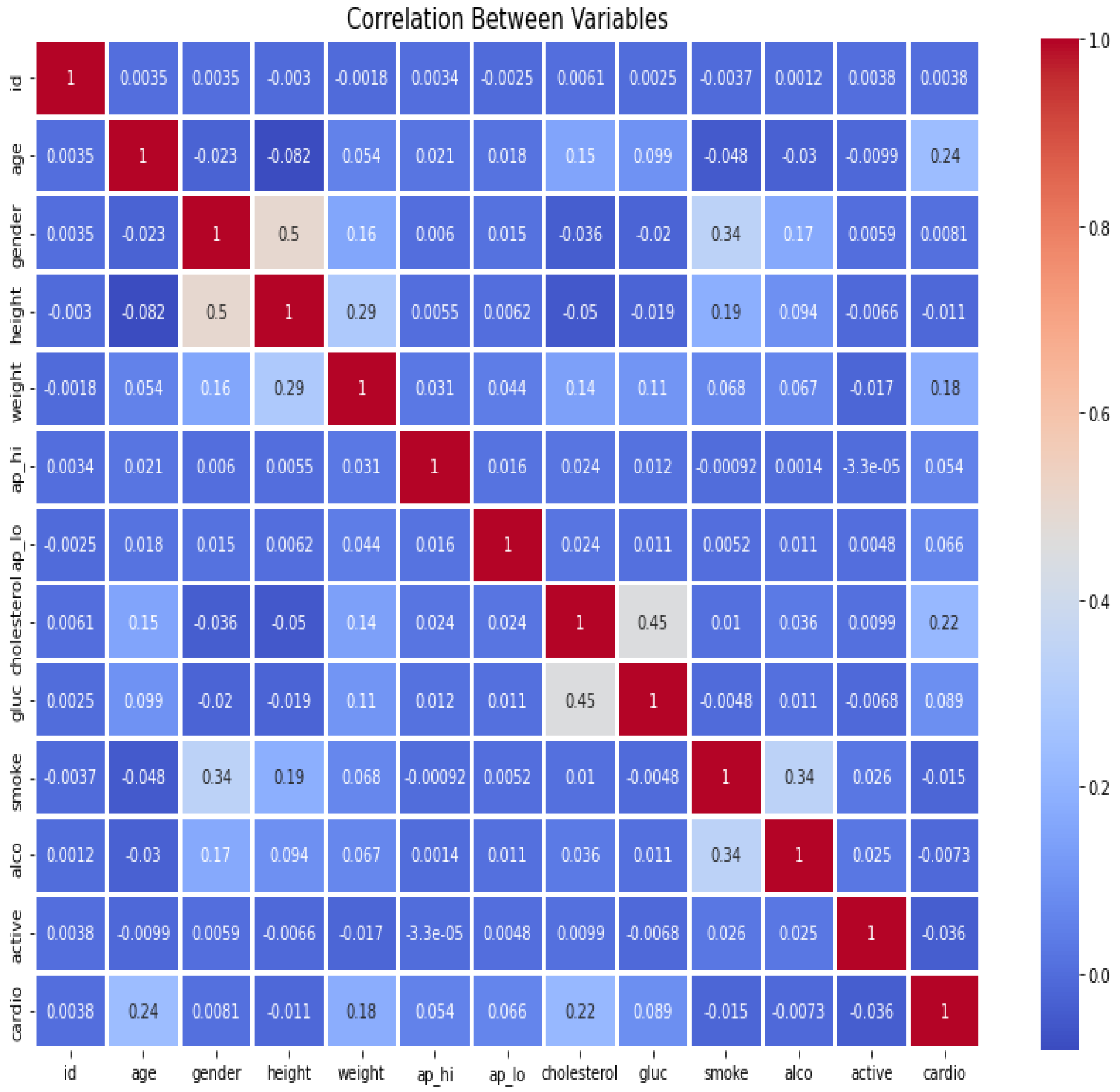


Figure 3.3: Attributes distribution of the cardiovascular disease dataset

Figure 3.4 is a heatmap depicting the correlation matrix between variables in the Cardiovascular Disease dataset. The heatmap is an effective data visualization tool since it shows a great deal about the relationships between these 12 dataset characteristics.



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
id	1	0.0035	0.0035	-0.003	-0.0018	0.0034	-0.0025	0.0061	0.0025	-0.0037	0.0012	0.0038	0.0038
age	0.0035	1	-0.023	-0.082	0.054	0.021	0.018	0.15	0.099	-0.048	-0.03	-0.0099	0.24
gender	0.0035	-0.023	1	0.5	0.16	0.006	0.015	-0.036	-0.02	0.34	0.17	0.0059	0.0081
height	-0.003	-0.082	0.5	1	0.29	0.0055	0.0062	-0.05	-0.019	0.19	0.094	-0.0066	-0.011
weight	-0.0018	0.054	0.16	0.29	1	0.031	0.044	0.14	0.11	0.068	0.067	-0.017	0.18
ap_hi	0.0034	0.021	0.006	0.0055	0.031	1	0.016	0.024	0.012	-0.00092	0.0014	-3.3e-05	0.054
ap_lo	-0.0025	0.018	0.015	0.0062	0.044	0.016	1	0.024	0.011	0.0052	0.011	0.0048	0.066
cholesterol	0.0061	0.15	-0.036	-0.05	0.14	0.024	0.024	1	0.45	0.01	0.036	0.0099	0.22
gluc	0.0025	0.099	-0.02	-0.019	0.11	0.012	0.011	0.45	1	-0.0048	0.011	-0.0068	0.089
smoke	-0.0037	-0.048	0.34	0.19	0.068	-0.00092	0.0052	0.01	-0.0048	1	0.34	0.026	-0.015
alco	0.0012	-0.03	0.17	0.094	0.067	0.0014	0.011	0.036	0.011	0.34	1	0.025	-0.0073
active	0.0038	-0.0099	0.0059	-0.0066	-0.017	-3.3e-05	0.0048	0.0099	-0.0068	0.026	0.025	1	-0.036
cardio	0.0038	0.24	0.0081	-0.011	0.18	0.054	0.066	0.22	0.089	-0.015	-0.0073	-0.036	1

Figure 3.4: Heatmap of the correlation matrix for dataset 2

.

3.4 Proposed Methodology

3.4.1 Machine Learning (ML) Algorithm's Description

This part examined how to apply machine learning methods to the previously provided dataset. In terms of accuracy, F1 score, specificity, sensitivity, and area under the ROC curve, each method's effectiveness was assessed and tabulated. The following is a list of algorithms, each containing a brief explanation:

A. Logistic Regression (LR)

Determine the likelihood that an occurrence belongs to a specific class of events by using logistic regression (LR). [22]. By utilizing the logit function to modify the dependent variable and predicting the logit of the modified dependent variable to the independent variable, a categorical outcome is related to one or more category predictors using logistic regression. Equation (1) depicts the logarithmic function and the probability.

$$\hat{p} = h_{\theta}(x) = \sigma(x^{T}\theta) \tag{1}$$

Where,

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{2}$$

However, Logistic Regression (LR) has the benefit of delivering a probability-based final categorization. In addition, it may encounter the total class separation issue [23].

B. Decision Tree (DT)

A pattern resembling a tree is used in the decision tree method to identify probable consequences, such as event outcomes [24]. The target variables of the tree model can require a discrete set of values. In tree architecture, however, leaves symbolize class labels and branches indicate class label-representing feature joins. The equation for entropy is shown equation (3).

$$E = -\sum_{b=1}^{n} p_{ab}\log_{2}\, p_{ab} \tag{3}$$

The tree structure, with its distinct nodes and edges, is ideally suited for displaying the interaction of the variables. When the transformation of the characteristics is monotonic, the decision tree is effective. Nonetheless, a decision tree does not support linear relationships, and the trees can be unstable at times. If there are a great amount of terminal nodes in a decision tree, it is quite challenging to comprehend the entire tree.

C. Support Vector Machine (SVM)

By evaluating a hyperplane that widens the boundary separation of classes in training data, support vector machines (SVM) categorize data [25]. The term hyperplane may also be written as equation (4).

$$f(x) = a^T x + c \tag{4}$$

Where, a = dimensional coefficient, c = offset.

The ability to choose from a variety of kernels is a benefit of SVM. Complex structured data sets can be handled with the assistance of many kernels. In addition, it has fewer issues with overfitting. Despite the fact that the kernel is the strength of the support vector machine, it is challenging to choose a kernel. In contrast, a huge data set necessitates a substantial amount of computational time [26].

D. Random Forest (RF)

A division of Decision Tree is called Random Forest [27]. The averaging decision trees reduce the variation portion of the model, which consists of high variance and low bias. Unknown samples can be generated by averaging the predictions.

$$I = \frac{1}{N}\sum_{n=1}^{N} f(x) \tag{5}$$

where uncertainty is,

$$\sigma = \sqrt{\frac{\sum_{n=1}^{N} (f(x) - \hat{f})^2}{N-1}} \tag{6}$$

A Random Forest (RF) is a method that applies many decision trees to data, collects predictions from each, and determines the optimal solution. In addition, it is primarily an ensemble learning approach that was founded on the bagging method and is capable of handling missing data values [28].

E. Gradient Boost (GB)

The primary aspects of gradient boosting are the optimization of a loss function, the use of a poor learner to create predictions, and the addition of ineffective scholars to the model for reduce the loss function [29]. The GB approach is among the most advanced machine learning techniques. Problems with machine learning algorithms can be roughly categorized as either bias mistakes or variance errors. Gradient boosting is one of the boosting strategies used to reduce the method's bias error. Unlike the Adaboosting technique, the base estimator of this algorithm cannot be specified. The base estimator of the Gradient Boost algorithm is fixed. The method can be used to forecast both categorical and continuous variables of interest as regressors and classifiers. For classifiers, the cost function is called log loss. MSE is the regressors' cost function. [30].

$$F_m(X) = F_{m-1}(X) + \eta \cdot f_m(X) \qquad (7)$$

Where, F is the ensemble model, f represents the weak learner, η is the learning rate, and X represents the input vector.

F. K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) examines K occurrences within the dataset that are close according to the study. The program will then use its own report to assess the variable y of the examination that should be anticipated [31]. The following equation (8) is utilized to compute the distance between two observations using the Euclidean distance:

$$d(x_i, y_i) = \sqrt{\left(x_{i,1} - y_{i,1}\right)^2 + \cdots \ldots . + \left(x_{i,m} - y_{i,m}\right)^2} \qquad (8)$$

K nearest neighbor uses relatively little processing time due to the fact that it does not need initial training and rather benefits based on the dataset at the time in anticipation. This technique is simple to implement because it only needs two values: the K value and the distance function value. However, it encounters issues when the data set is huge and performs poorly when there are many data dimensions [32].

G. AdaBoost (AdB)

AdaBoost is a procedure that creates a classification iteratively by invoking a base learner at each repetition, which delivers a classifier and assigns the coefficient of weight to it [33]. The final classification decision will be determined by a weighted "vote" of the basic classifiers. If the inaccuracy of the primary algorithms is smaller, its influence in the deciding score will increase. The Adaboost method essentially adjusts the data distribution using the classification improvement of the sample instances from the training set. The revised weights from the amended data are then sent to the lowest classification, and finally, all of the training classifiers are combined. AdaBoost determines its final output using the following function shown in equation (9):

$$C(x) = \operatorname{sign}\left(\sum_{n=0}^{N} \alpha_n W_n(x)\right) \tag{9}$$

Where,

$$\alpha_n = 0.5 \ln\left(\frac{1-\varepsilon_n}{\varepsilon_n}\right), \varepsilon_t = total\ error, W_n(x) = output\ from\ weak\ classifiers$$

Adaboost is less susceptible to overfitting, although it has trouble with unclear data and data containing anomalies.

H. Extreme Gradient Boost (XGB)

In the family of machine learning (ML) methods, XGB (Extreme Gradient Boost), developed by Tianqi Chen in 2014, is a more recent member. Gradient boosting is the principle upon which it is based. It includes both optimizing and machine learning techniques [34,35]. Mathematically, the goal function of the XGB algorithm is shown in equation (10).

$$O(t) = \sum_{i=0}^{n} Q\left(y_i, y'^{t-1} + f_t(x_i)\right) + K \tag{10}$$

Then function of normalization:

$$\operatorname{Nor}(f_t) = \kappa T + 0.5\lambda \sum_{i=0}^{T} W_j^2 \tag{11}$$

Where κ = Influencing factor for the number of leaf nodes

T = Count of leaf nodes

W_j = The j leaf nodes' weight

λ = An excessively controlling factor

K = Constant

XGB performs effectively between both local and large scales datasets, but encounters difficulty when the dataset contains a large number of categorical variables.

3.4.2 Performance Evaluation Metrics

With regard to machine learning, performance metrics relate to measures of the evaluation of an algorithm's performance based on a variety of requirements, such as precision, accuracy, recall, etc. Various effectiveness measures are examined in detail below.

A. *Confusion Matrix*

Confusion matrix is an example of a method for demonstrating how it discriminator becomes confused during prediction.

TABLE 3.8: CONFUSION MATRIX

Actual	**Predicted**	
	+ (1)	- (0)
+ (1)	TP (1,1)	FN (1,0)
- (0)	FP (0,1)	TN (0,0)

In Table 3.8, the true positive value is TP, indicating that the positive coefficient has been accurately classified; the false positive value is FP, indicating that the positive coefficient has been incorrectly classified; the false negative value is FN, indicating that the negative coefficient has been incorrectly classified; and the true negative value is TN, indicating that the negative coefficient has been accurately classified.

Various performance measures can be derived from the confusion matrix. Using Table 6 as an illustration, the definitions of accuracy, F1 score, recall, and precision are shown below.

B. Accuracy

Accuracy is described by the fraction of all forecasts that are made accurately (correctly).

$$\text{Accuracy (Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

C. Sensitivity

The definition of sensitivity is the fraction of actual positive cases that are projected to be positive. It is also known as "recall."

$$\text{Sensitivity (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

D. Specificity

Specificity seems to be the proportion of actual negative situations accurately expected to be negative.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14}$$

E. F1 score

The F1 score merges both accuracy and recall for a separator into a single statistic using their harmonic mean. It's employed in order to contrast the effectiveness of two classifiers. Calculating the F1 score of a model of categorization is as follows:

$$\text{F1 score} = \frac{2(\text{P} * \text{R})}{\text{P} + \text{R}} \tag{15}$$

Where,

P = the precision,

R = the recall of the classification model

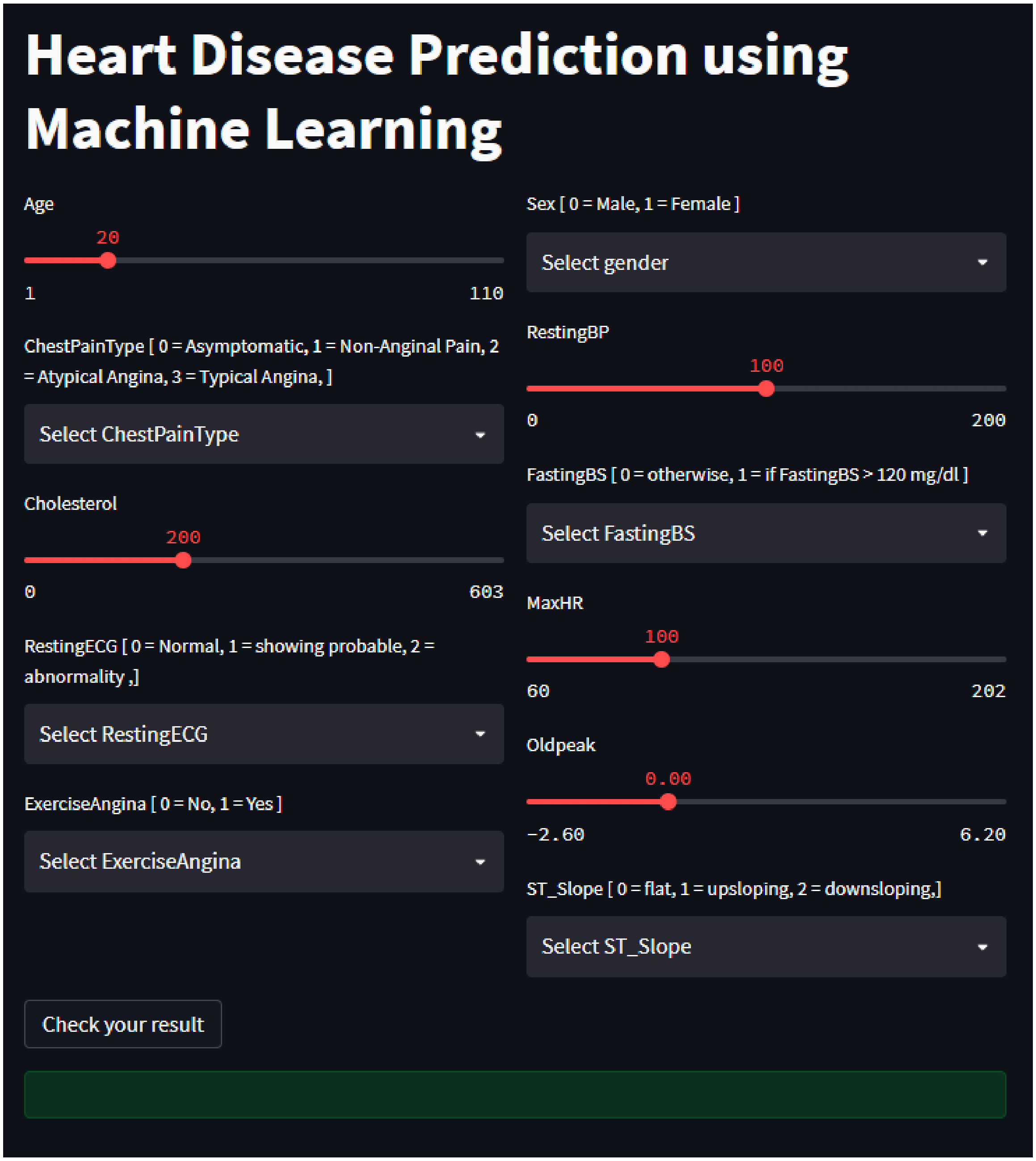
F. AUC-ROC Curve

The relationship between TPR and FPR at various limit settings is represented by the Receiver Operator Characteristic (ROC) curve. An indicator of separability, the Area Under the Curve (AUC), shows how well a model can categorize classes. The greater the AUC, the more classes that can be accurately predicted.

3.4.3 Implementation of Web Application

This section describes the building of a web app for heart disease prediction. The website is created with streamlit. Streamlit is an open-source Python framework for constructing and deploying interactive dashboards and machine learning models for data science [42]. This website can forecast heart disease based on the user's input. Figure 3.5 depicts the input field of a user.

After the highest accuracy model is tested in a Jupyter notebook, a pickle file is created for the development of this web page. The pickle file is then developed with the streamlit framework.



Heart Disease Prediction using Machine Learning

Age

20

1 110

Sex [0 = Male, 1 = Female]

Select gender

ChestPainType [0 = Asymptomatic, 1 = Non-Anginal Pain, 2 = Atypical Angina, 3 = Typical Angina,]

Select ChestPainType

RestingBP

100

0 200

FastingBS [0 = otherwise, 1 = if FastingBS > 120 mg/dl]

Select FastingBS

Cholesterol

200

0 603

MaxHR

100

60 202

RestingECG [0 = Normal, 1 = showing probable, 2 = abnormality ,]

Select RestingECG

Oldpeak

0.00

-2.60 6.20

ExerciseAngina [0 = No, 1 = Yes]

Select ExerciseAngina

ST_Slope [0 = flat, 1 = upsloping, 2 = downsloping,]

Select ST_Slope

Check your result

Figure 3.5: A user's input form

3.5 Implementation Requirements

As mentioned earlier in the methods section, there are a number of implementation-related specifics. Therefore, Python programming is utilized for this endeavor. We are executing Python code in the Jupyter notebook of the Anaconda navigator. In order to employ ML algorithms, the Jupyter notebook is far quicker than any Python IDE tool such as PyCharm or Visual Studio. While developing code, the Jupyter notebook is useful for data visualization and producing graphs, such as histograms and heatmaps of correlated matrices.

Let's go over the steps for implementation:

a) Dataset collection.

b) Importing Libraries: The libraries Matplotlib, Pandas, NumPy, Seaborn, and Scikit-learn were utilized.

c) Exploratory data analysis: to gain deeper knowledge regarding data.

d) Data cleaning and preprocessing: Isnull() and isna() were used to check for null and garbage values.Python functions sum ()In the preprocessing stage, feature engineering was performed on our set of data. We changed category values to numeric ones using the get_dummies() function of the Pandas library.

e) Feature Scaling: At this stage, we normalize the data using standardization by utilizing StandardScalar() and fit_transform() functions from the scikit-learn library.

f) Model selection: Initially, we distinguished X's from Y's. X's are characteristics or input factors of our datasets, whereas Y's are target or dependent variables that are essential for illness prediction. Then, utilizing the train_test_split() function from the sklearn library, we divided our X's and Y's into training and testing splits. We allocated 80% of our records for training purposes and 20% for testing.

g) ML models were implemented, and a confusion matrix was constructed for each model.

h) Utilization of the algorithms with the highest level of accuracy.

i) Develop a web application with the best model.

In this figure 3.6, we have shown that the process that we followed in doing our study is briefly detailed. We are able to understand how to go methodically toward our goal.

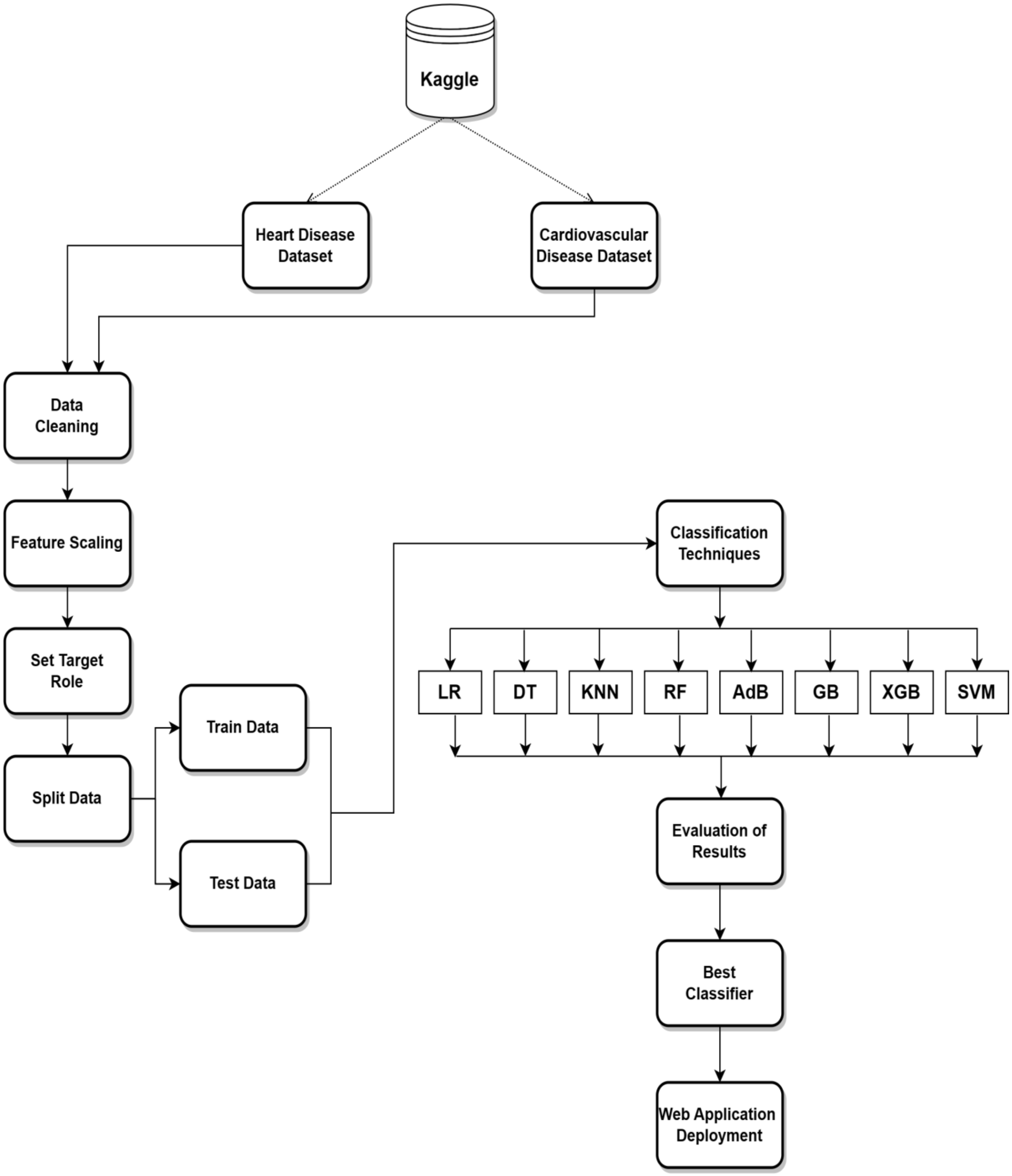


Figure 3.6: Proposed system structure

CHAPTER 4
Experimental Results and Discussion

4.1 Experimental Setup

Classifier performance seems to have been evaluated using supervised classification experiments. Using a collection of features, the classifiers' performance was assessed. Diverse metrics are utilized to evaluate the efficacy of models. In a python environment, various machine learning libraries were used on an Intel Core i5-8400 CPU @ 2.80GHz system to conduct the tests.

This study employs a wide range of tools. They're all free and open source.

1. Google Colaboratory: The Google Colaboratory, commonly referred to as "Colab," is an open solution that combines the Jupyter Notebook's functionality with virtual machines hosted by Google and top-tier hardware. We found that COLAB is also perfectly suited for use in the classroom. Colab was initially created for researchers in AI and data science to exchange reproducible experimentation and descriptions of methodology. The main benefit is that it frees students from having to separately change bundle software and rely on others because they can run instructor-shared notebooks and allows students with sufficient processing capacity to execute advanced AI algorithms simultaneously [36].

2. Python 3.5: A high-level, general-purpose programming language is Python that is commonly used and employed in many fields, such as web development, general programming, software development, machine learning, data analysis, etc. Python is utilized for this investigation due to its adaptability, usability, and community support and extensive documentation [37].

3. NumPy 1.11.3: NumPy is an extremely potent tool that facilitates scientific computing. It has advanced capabilities and can do Fourier transform, algebra, N-dimensional array, etc. NumPy is widely utilized for data analysis and image processing, and numerous additional libraries are constructed on top of it. NumPy serves as the basic stack for these libraries [38].

4. Pandas 0.19.1: Pandas is free, BSD-licensed software that was developed specifically for the Python. It provides a wide range of data analysis capabilities for Python and is the most powerful competitor to the R programming language. In addition to reading CSV files, reading data frames, and indexing Excel files, Pandas also makes it easier to merge, slice, and handle missing data, among other tasks. Pandas' most essential characteristic is their ability to perform time-series studies [39].

5. Seaborn 0.7.1: Seaborn is a package of data visualization tools and was created using the Python programming language. It is a library at a higher level than matplotlib. Seaborn is incredibly user-friendly, appears to be more aesthetically pleasing and instructive [40].

6. SciPy and Scikit-learn 0.18.1: Scikit-learn is a well-known machine learning resource that is a third-party expansion to SciPy, whereas SciPy is a collection of fundamental mathematical operations based on NumPy. The tools and techniques needed for the bulk of machine learning tasks are included in Scikit-learn. Dimensional reduction, regression, clustering, classification, and data preprocessing are made easier by Scikit-learn. Since scikit-learn is Python-based and works with the NumPy library, it is used in this study. This is an incredibly user-friendly [41].

4.2 Experimental Results & Analysis

4.2.1 Results of Dataset 1

A. Prediction Accuracy

Accuracy refers to the correctly predicted values. Figure 4.1 depicts the accuracy of each tested method. The gradient boost method outperformed others with an accuracy of 91.80%. The lowest measure of accuracy obtained by the decision tree is only 75.00%. The Random Forest and Support Vector Machine both achieve an accuracy of 89.60%.

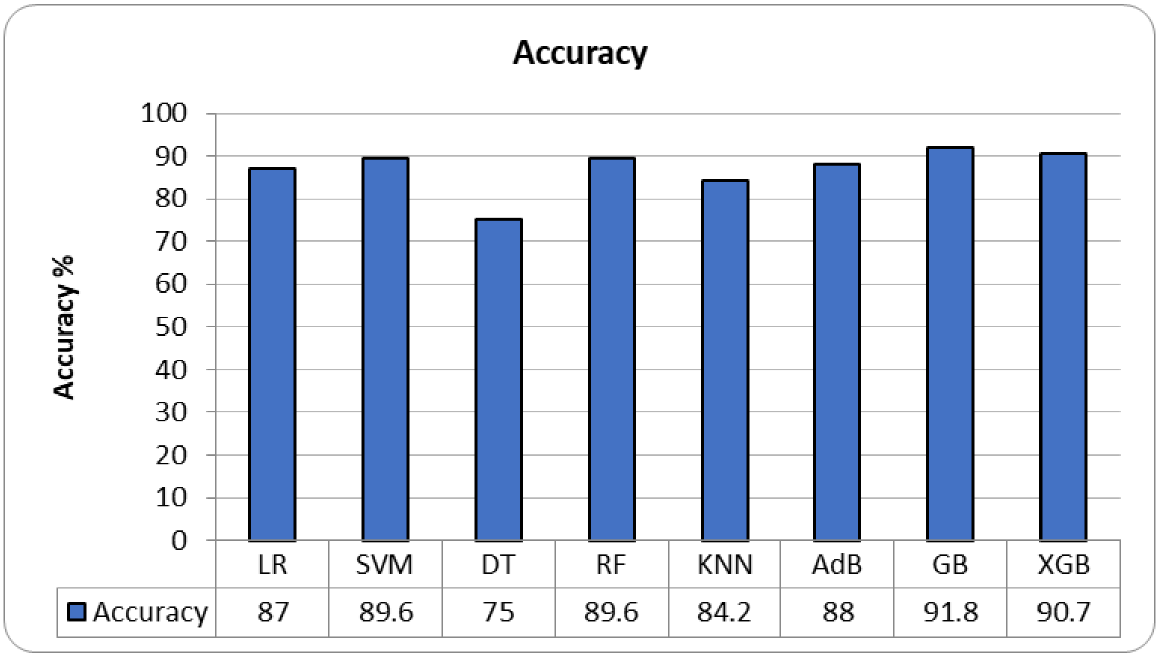


Figure 4.1: Accuracy of different applied model for dataset 1

B. Precision

It represents the actual instances of all positive forecasts that came true. Figure 4.2 depicts the precisions of several algorithms.

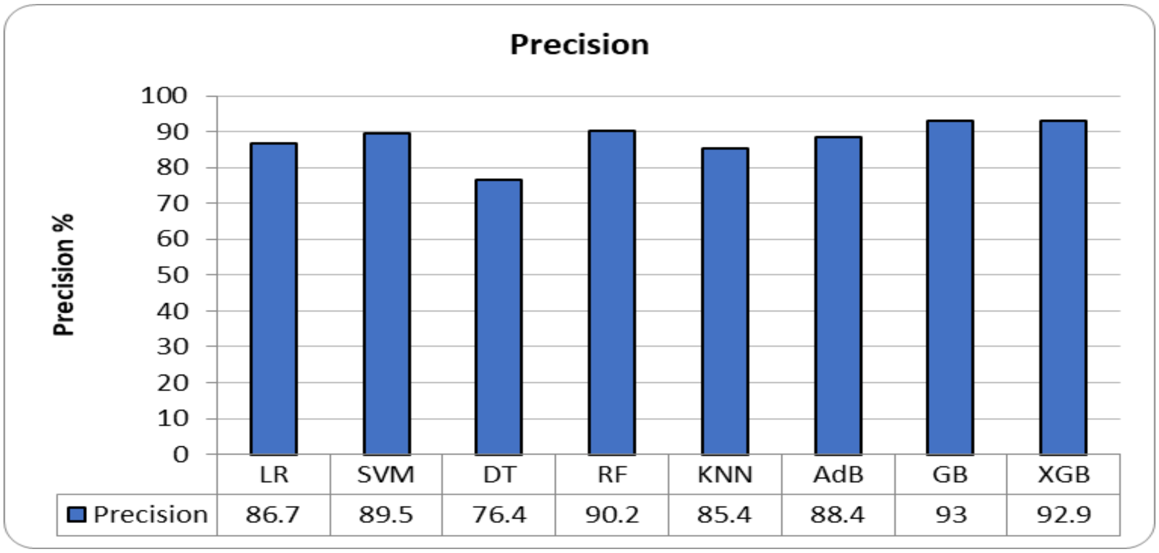


Figure 4.2: Precision of different applied model for dataset 1

C. Recall

It describes the values that were accurately predicted among all positive classifications. Figure 4.3 depicts the recall values among tested algorithms.

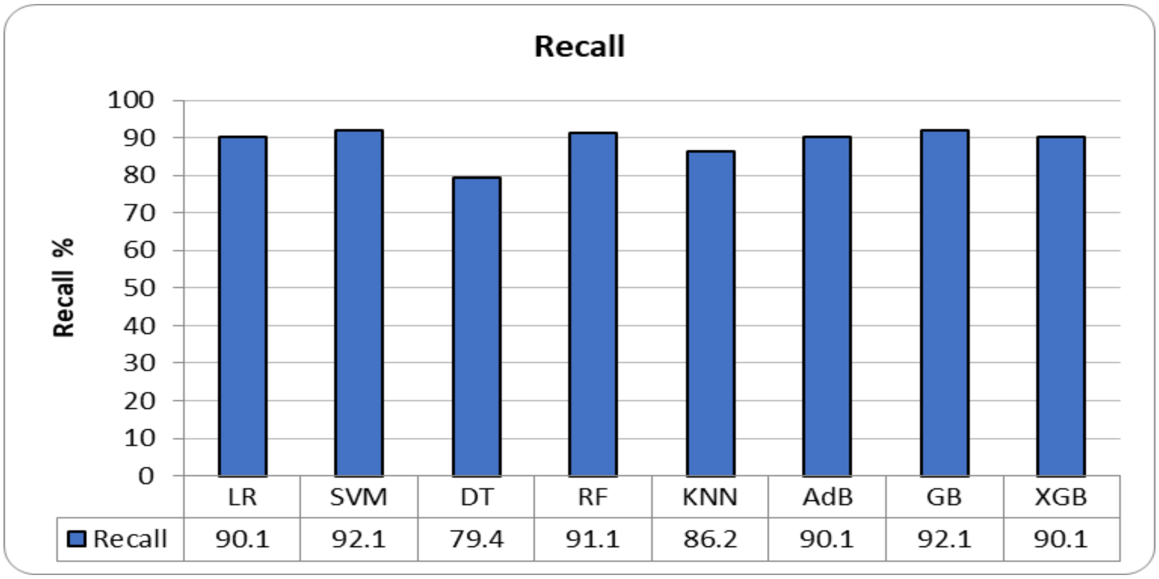
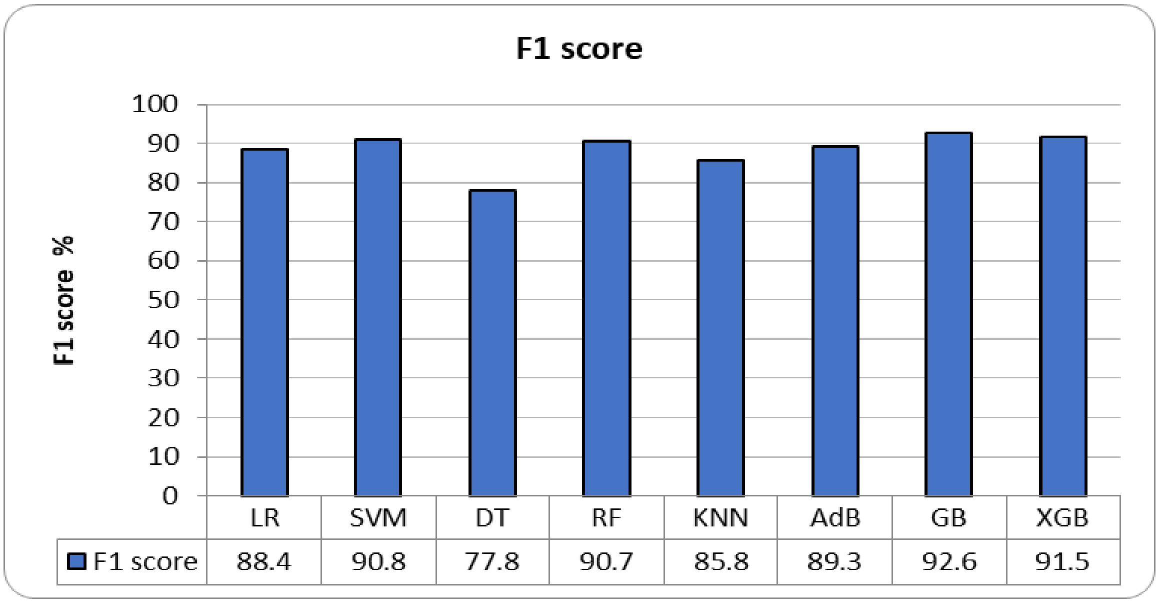


Figure 4.3: Recall of different applied model for dataset 1

D. F1 score

It evaluates Recall and Precision and calculates test accuracy using the Harmonic Mean. Figure 4.4 depicts the F1 score for each algorithm.

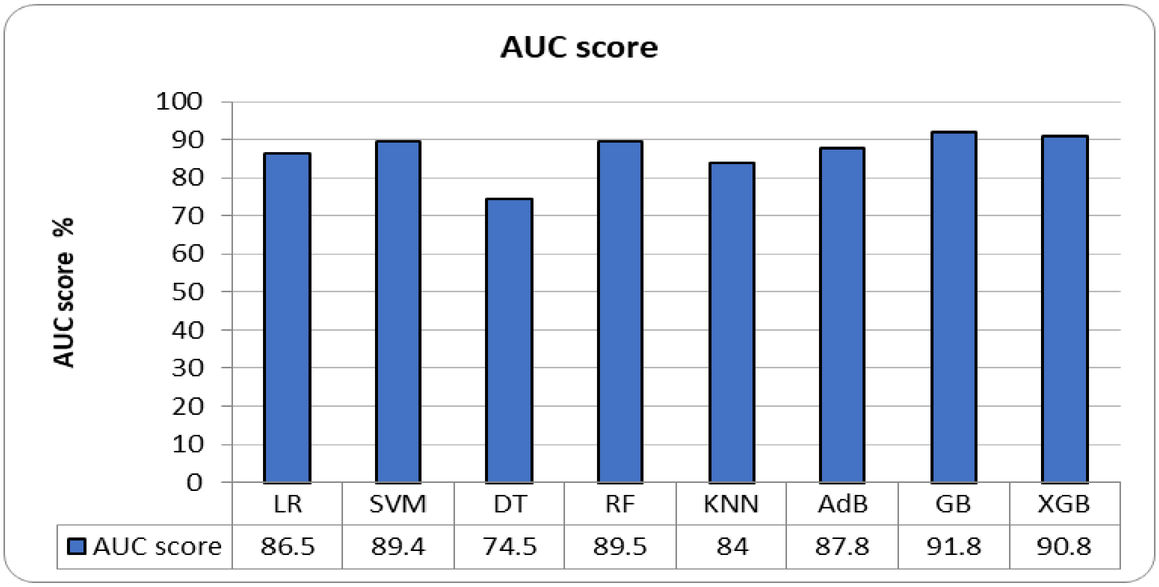


	LR	SVM	DT	RF	KNN	AdB	GB	XGB
■ F1 score	88.4	90.8	77.8	90.7	85.8	89.3	92.6	91.5

Figure 4.4: F1 Score of different applied model for dataset 1

E. AUC score

The ROC curve illustrates the relationship between TPR and FPR at several thresholds. Fig. 4.5 indicates the AUC score across different algorithms.



	LR	SVM	DT	RF	KNN	AdB	GB	XGB
■ AUC score	86.5	89.4	74.5	89.5	84	87.8	91.8	90.8

Figure 4.5: AUC score of different applied model for dataset 1

4.2.2 Results of Dataset 2

A. Prediction Accuracy

Accuracy refers to the values that were successfully predicted. Figure 4.6 shows the accuracy of each approach examined. The gradient boost method surpassed others with a 74.85% accuracy rate. 64.00% is the lowest level of accuracy attained by the decision tree.

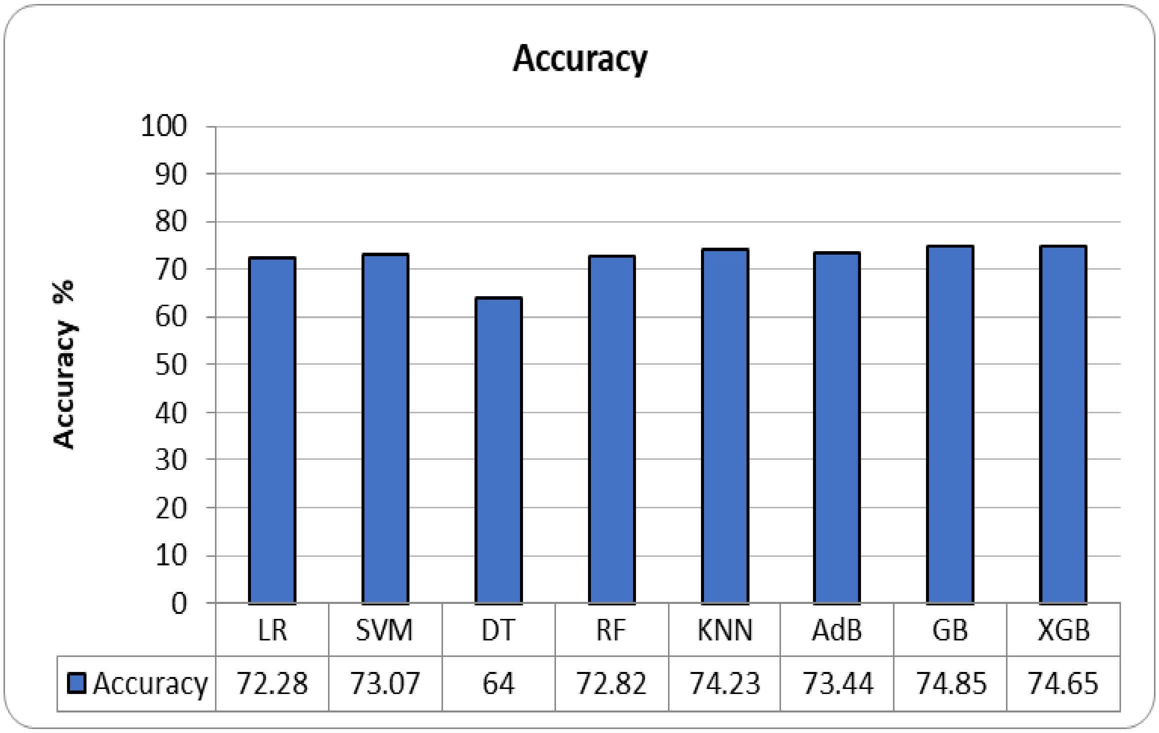


Figure 4.6: Accuracy of different applied model for dataset 2

B. Precision

It depicts the actual cases of all accurate positive projections. Figure 4.7 shown the precision of many algorithms.

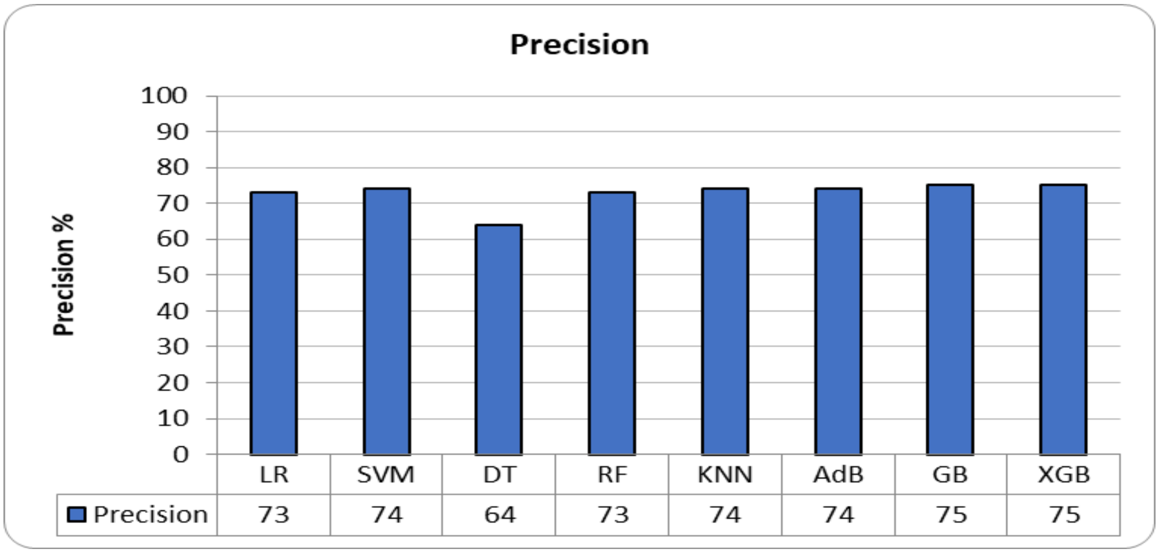


Figure 4.7: Precision of different applied model for dataset 2

C. Recall

It describes the values that all positive categories successfully anticipated. Figure 4.8 illustrates the recall levels for each tested algorithm.

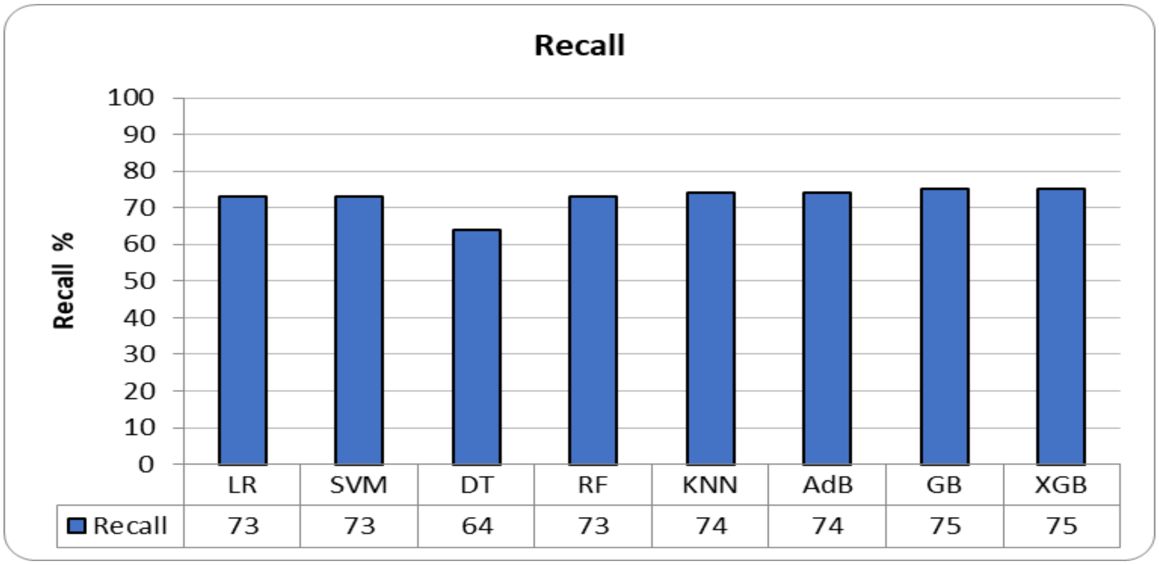


Figure 4.8: Recall of different applied model for dataset 2

D. F1 score

Using Harmonic Mean, it analyzes Precision and Recall and calculates test accuracy. Figure 4.9 illustrates each algorithm's F1 score.

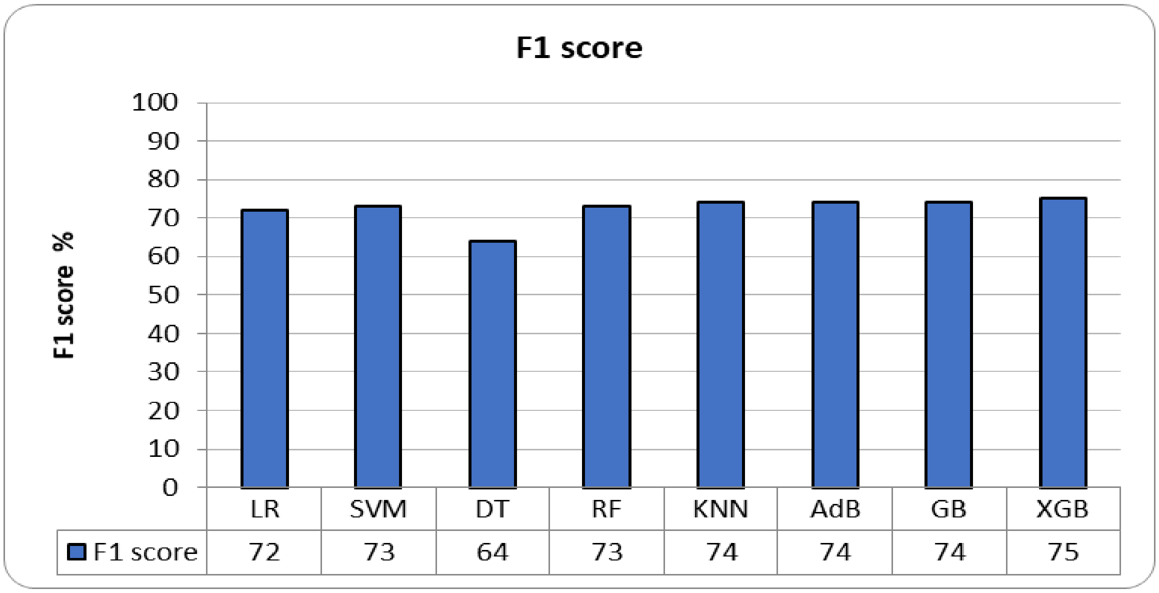


Figure 4.9: F1 Score of different applied model for dataset 2

E. AUC score

The ROC curve depicts the relationship between TPR and FPR at multiple levels. Figure 4.10 displays the AUC score for each algorithm.

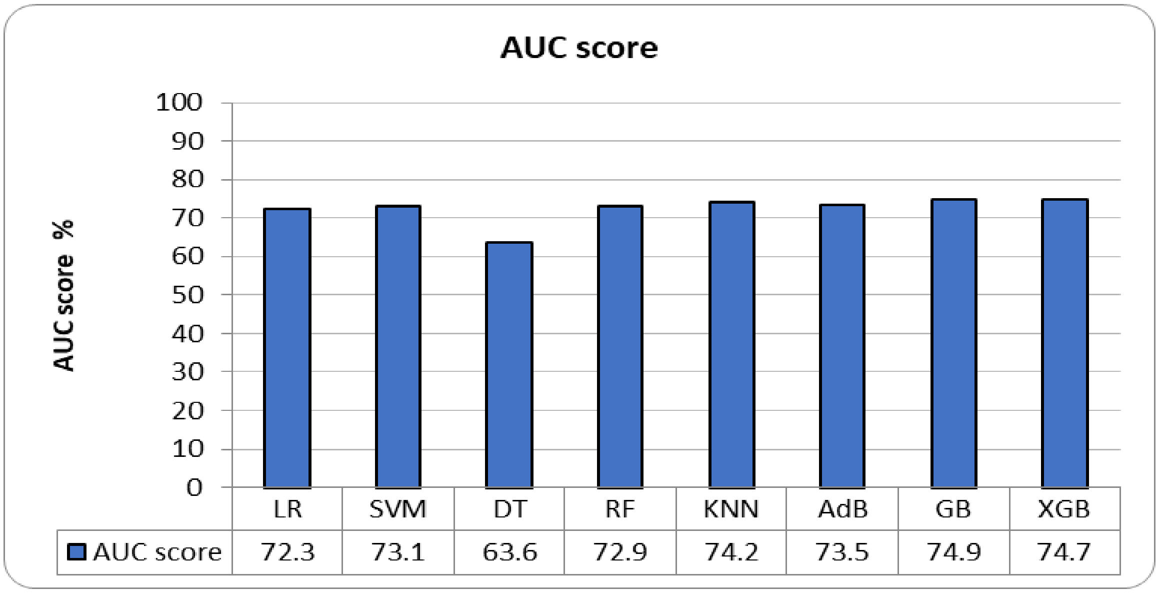
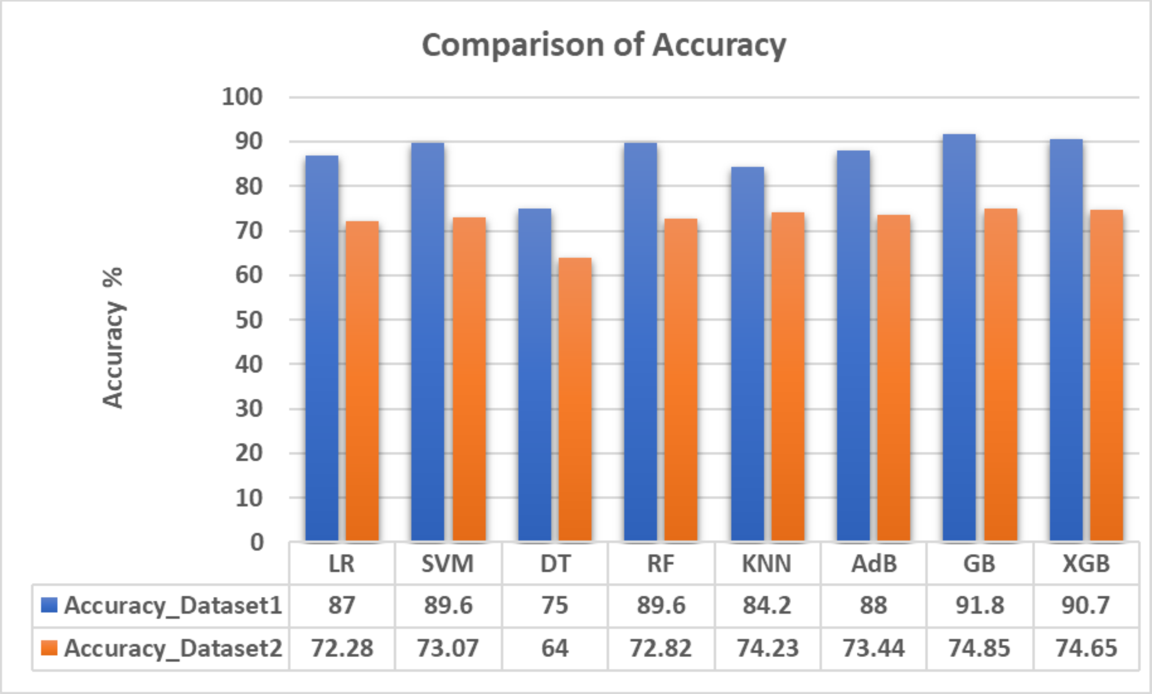


Figure 4.10: AUC score of different applied model for dataset 2

4.2.3 Comparison of Accuracy

Both the dataset of heart disease and the dataset of cardiovascular disease were utilized in the process of carrying out eight different machine learning methods. Comparative accuracy results are depicted in figure 4.11. It is abundantly obvious that the heart disease dataset is beneficial to the performance of all applied machine learning methods. Between the two datasets, Gradient Boost achieved the highest level of accuracy at 91.80%, while the decision tree achieved the lowest level of accuracy at 64.00%. Additionally, it was discovered that Gradient Boost achieved the maximum accuracy for both datasets, but Decision Tree demonstrated the lowest accuracy including all datasets.

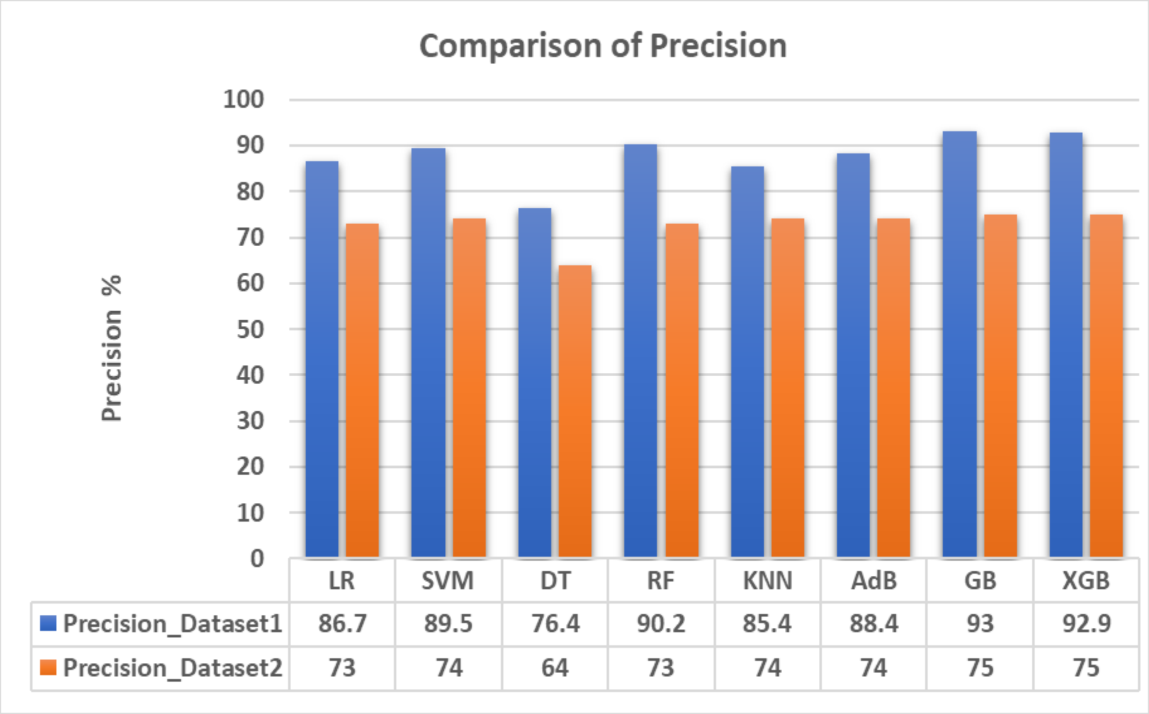


	LR	SVM	DT	RF	KNN	AdB	GB	XGB
🟦 Accuracy_Dataset1	87	89.6	75	89.6	84.2	88	91.8	90.7
🟧 Accuracy_Dataset2	72.28	73.07	64	72.82	74.23	73.44	74.85	74.65

Figure 4.11: Comparison of accuracy for different applied model

4.2.4 Comparison of Precision

Figure 4.12 provides an illustration of a comparison of the precision. It is clear that dataset 1 has a precision that is higher than 76.00%, and dataset 2 has a precision that is highest at 75.00%

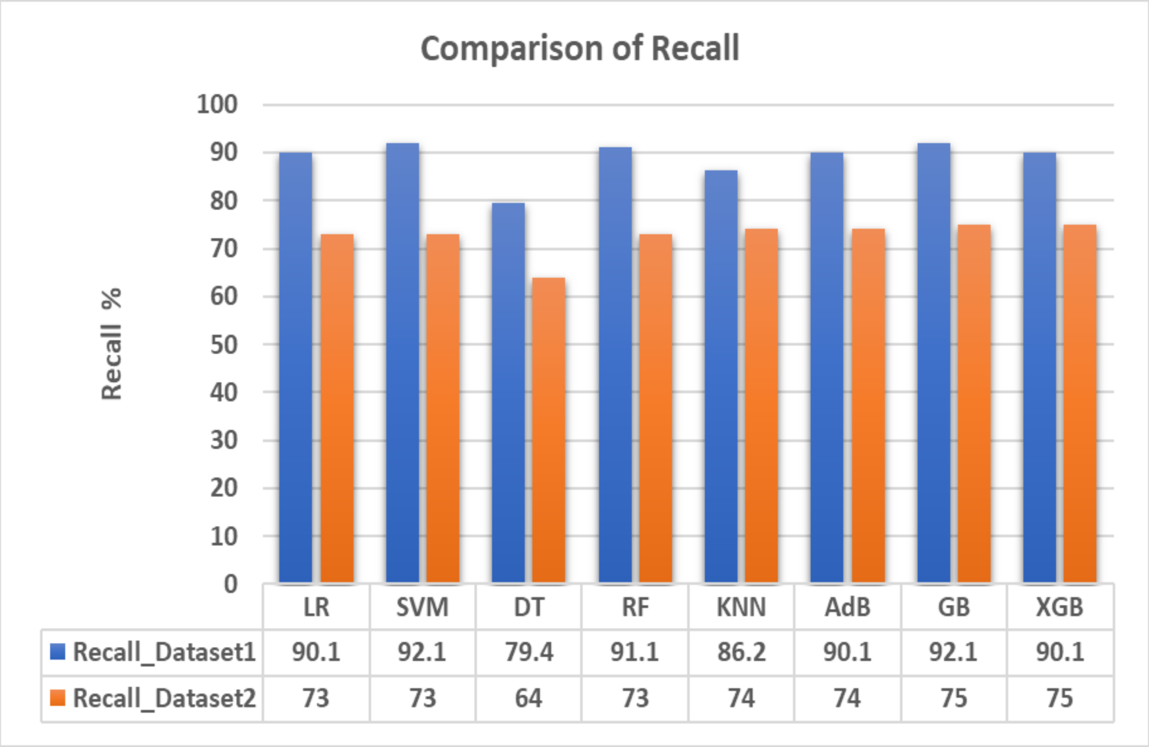


	LR	SVM	DT	RF	KNN	AdB	GB	XGB
🟦 Precision_Dataset1	86.7	89.5	76.4	90.2	85.4	88.4	93	92.9
🟧 Precision_Dataset2	73	74	64	73	74	74	75	75

Figure 4.12: Comparison of precision for different applied model

4.2.5 Comparison of Recall

Figure 4.13 provides an illustration of the recall comparison. The SVM and the GB algorithm both achieved a recall of 92.10% with their respective datasets.

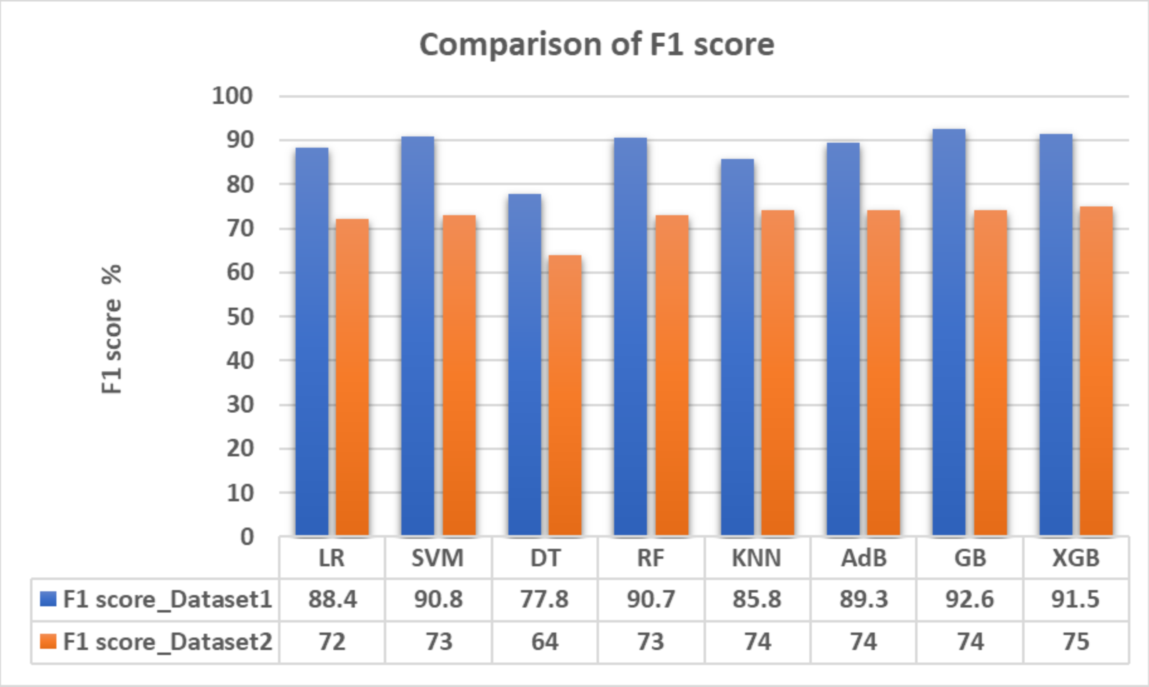


	LR	SVM	DT	RF	KNN	AdB	GB	XGB
🟦 Recall_Dataset1	90.1	92.1	79.4	91.1	86.2	90.1	92.1	90.1
🟧 Recall_Dataset2	73	73	64	73	74	74	75	75

Figure 4.13: Comparison of recall for different applied model

4.2.6 Comparison of F1 score

Figure 4.14 provides an illustration of a comparison of the F1 Score. The F1 scores of KNN, AdB, and GB all achieved 74.00%. GB had the best F1 score among them, coming in at 92.60%, while DT had the worst, coming in at only 64.00%.



	LR	SVM	DT	RF	KNN	AdB	GB	XGB
🟦 F1 score_Dataset1	88.4	90.8	77.8	90.7	85.8	89.3	92.6	91.5
🟧 F1 score_Dataset2	72	73	64	73	74	74	74	75

Figure 4.14: Comparison of F1 score for different applied model

4.2.7 Comparison of AUC score

Figure 4.15 provides an illustration of the comparison of AUC Scores. GB achieved the best possible AUC score of 91.80% in the Heart Disease dataset. In the Cardiovascular Disease Dataset, DT scored the lowest possible AUC percentage, which was 63.60%.

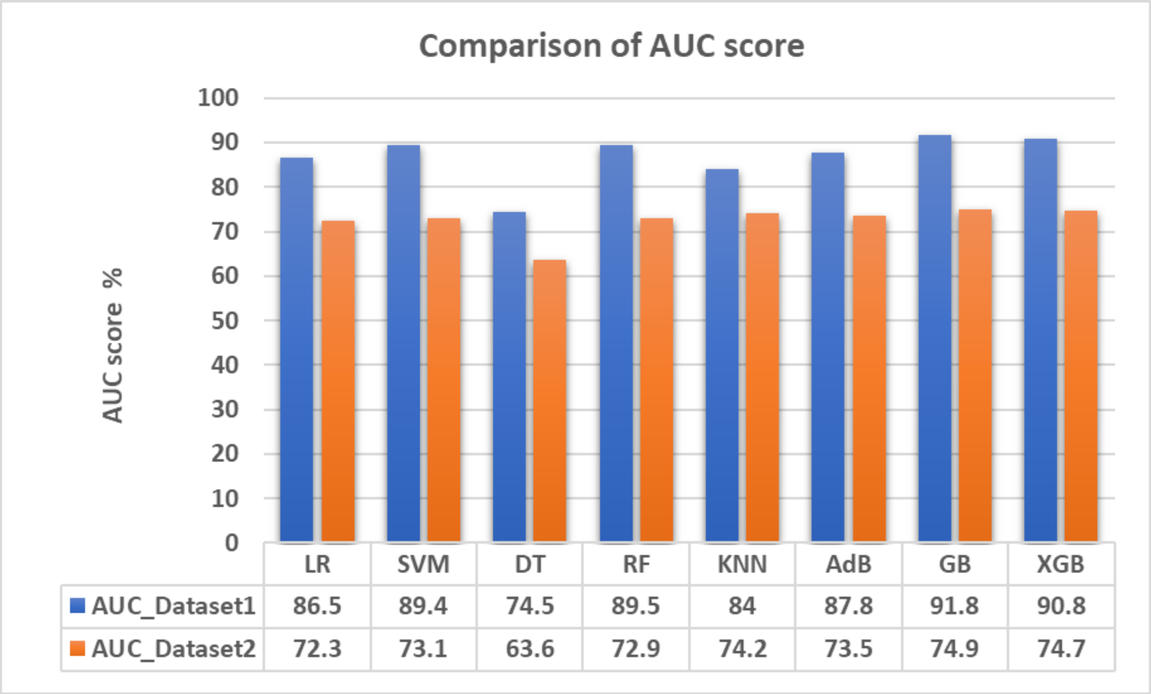


Figure 4.15: Comparison of AUC for different applied model

4.2.8 Performance Analysis of Dataset 1 and Dataset 2

In this analysis, we analyze eight different learning machine algorithms using data from two different datasets, namely Heart Disease and Cardiovascular Disease. Recall, accuracy, F1 score, the area under the curve (AUC) and precision are the metrics utilized to analyze the effectiveness of models. When we tested our suggested models on both the dataset of Heart Disease and the dataset of Cardiovascular Disease, we discovered that the Heart Disease dataset is a better fit for our models than the Cardiovascular Disease dataset.

A. *Best Model of Confusion Matrix for Dataset 1*

Figure 4.16 represents the confusion matrix for Gradient Boost classifiers for 183 instances (20% of the whole Dataset1), where GB model predicts True Negative = 75, True Positive = 94, False Positive = 7 and False Negative =7.

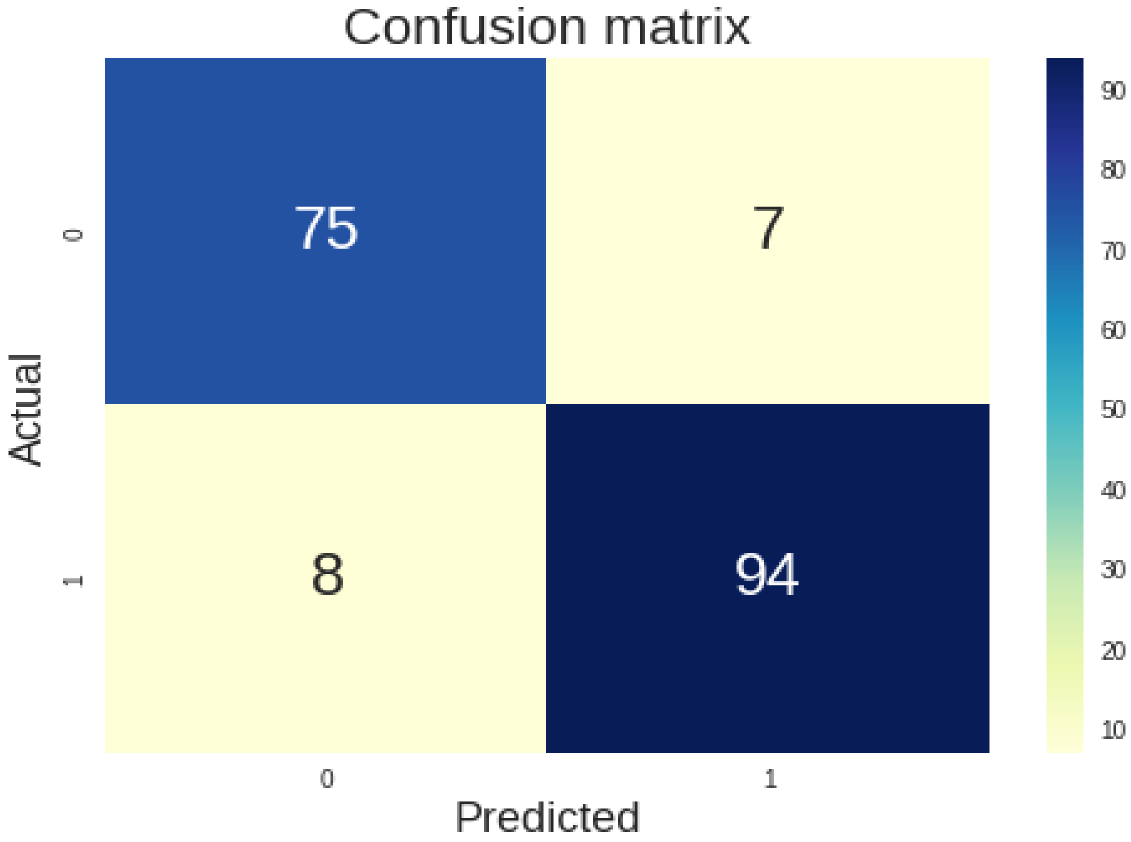


Figure 4.16: Confusion matrix of gradient boost for dataset 1

B. *Best Model of Confusion Matrix for Dataset 2*

The confusion matrix for Gradient Boost classifiers is shown in Figure 4.17 for 14000 occurrences (20% of the whole Dataset2), where the GB model predicts True Negative = 5347, True Positive = 5133, False Positive = 1570 and False Negative = 1950.

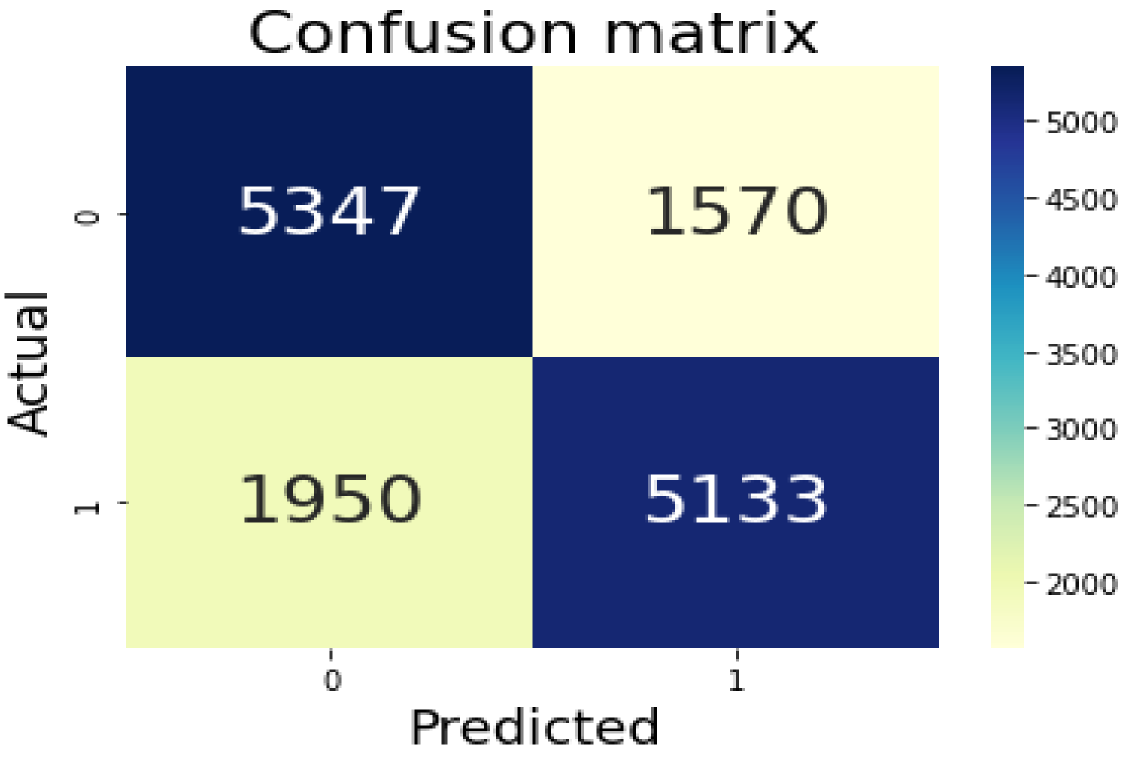


Figure 4.17: Confusion matrix of gradient boost for dataset 2

C. *ROC Curve for Dataset 1 and Dataset 2*

Figures 4.18 and 4.19 respectively, represent the ROC curves for Dataset1 and Dataset2 for a different of machine learning classifiers. When compared to other classifiers, the Gradient Boost classifier has the highest area under the curve with attributes filtration where both figures depict it. Gradient Boost's performance likewise improves, and its Score for Dataset1 and Dataset2 is 0.918 and 0.749, respectively.

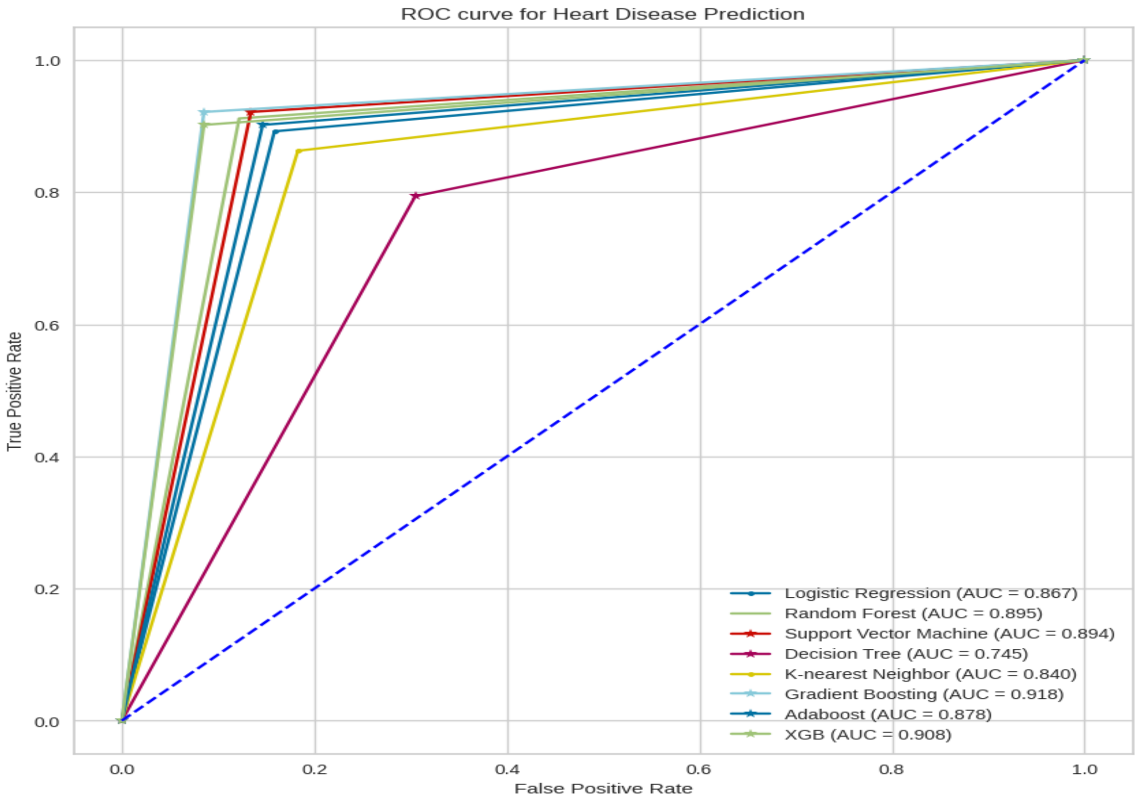


Figure 4.18: ROC curve for dataset 1

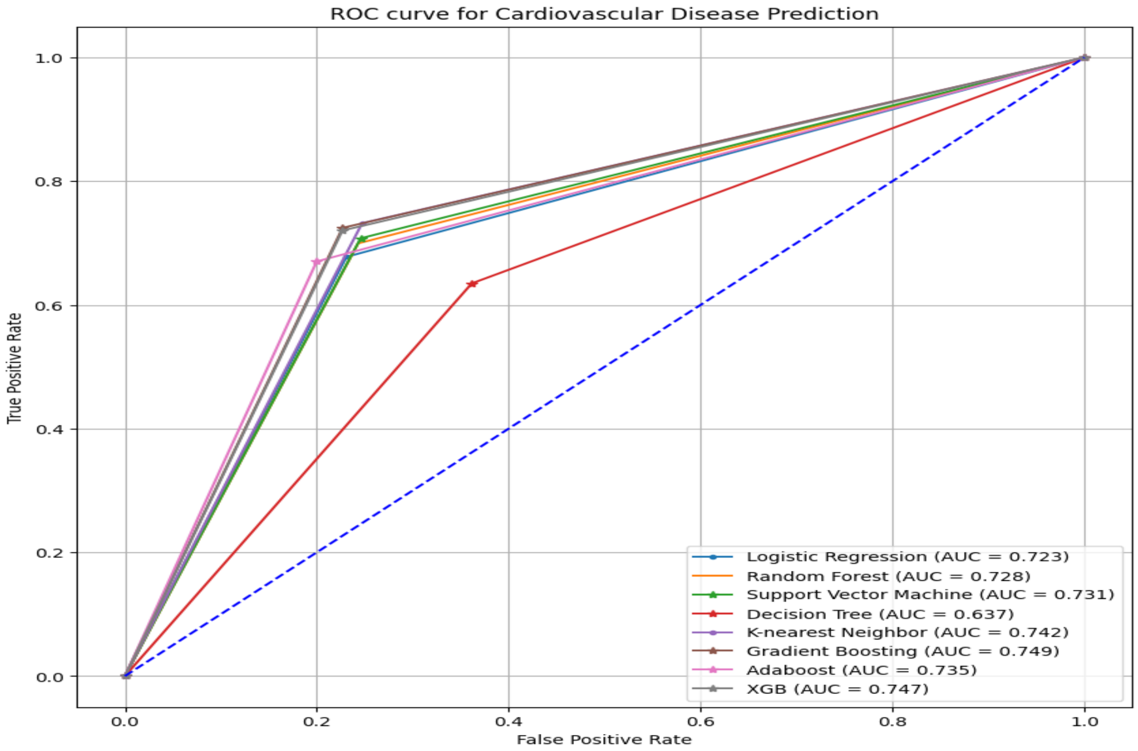
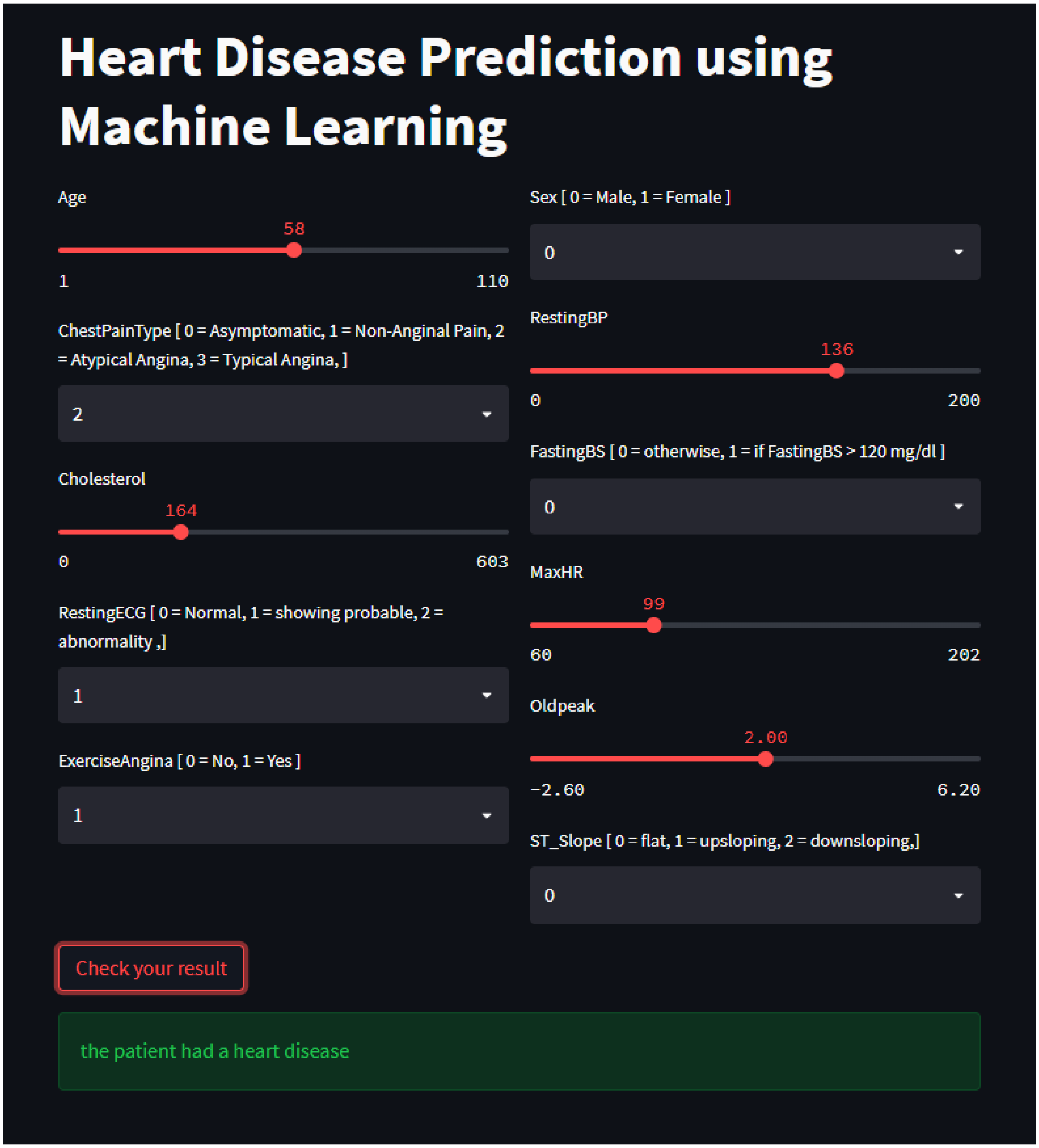


Figure 4.19: ROC curve for dataset 2

4.2.9 Result of Web Application

When a person enters input values and clicks the "Check your result" button, the input parameters are submitted to the machine learning algorithm for the prediction of heart disease. Once the heart disease has been predicted, the result is transferred to the web page via streamlit so that users can view it on the screen. Figure 4.20 represents the output.



Heart Disease Prediction using Machine Learning

Age
58
1 110

ChestPainType [0 = Asymptomatic, 1 = Non-Anginal Pain, 2 = Atypical Angina, 3 = Typical Angina,]
2

Cholesterol
164
0 603

RestingECG [0 = Normal, 1 = showing probable, 2 = abnormality ,]
1

ExerciseAngina [0 = No, 1 = Yes]
1

Sex [0 = Male, 1 = Female]
0

RestingBP
136
0 200

FastingBS [0 = otherwise, 1 = if FastingBS > 120 mg/dl]
0

MaxHR
99
60 202

Oldpeak
2.00
-2.60 6.20

ST_Slope [0 = flat, 1 = upsloping, 2 = downsloping,]
0

the patient had a heart disease

Figure 4.20: Predict outcome using a web application

4.3 Discussion

4.3.1 Comparison Accuracy with Previous Studies

TABLE 4.1: COMPARISON OF RESULTS FROM RECENT STUDIES

Reference	Published Year	Best Model	Accuracy
[10]	2021	Extreme Gradient Boost	72.70%
[11]	2019	Support Vector Machine	84.85%
[12]	2020	Naive Bayes	84.28%
[13]	2021	Support Vector Machine	72.50%
[14]	2017	Support Vector Machine	84.15%
[15]	2021	Random Forest	89.01%
[16]	2016	Logistic regression	85.00%
[17]	2020	Naive Bayes	88.15%
[18]	2020	K Nearest Neighbor	87.00%
[19]	2019	Support Vector Machine	64.40%
Our study	2023	Gradient Boost Extreme Gradient Boost Random Forest Support Vector Machine	**91.80%** **90.70%** **89.60%** **89.60%**

The results of our investigation are presented in a very clear and concise manner in Table 4.1. Previous researchers have achieved a satisfactory level of accuracy in their examinations. However, we have successfully beaten all of them. In addition to this, when compared with earlier research, the accuracy of our four suggested models is the greatest. Acquiring a higher level of accuracy allows us to successfully complete our mission.

CHAPTER 5
Impact on Society, Environment and Sustainability

5.1 Impact on Society

The findings of the research will have a pleasant influence on today's society. Because immediate detection of cardiovascular disease helps to avoid this, and because this will lead to a reduction in the cost of treatment, particularly for those living in low-income countries, were the ones who benefited the most. The number of individuals who pass away as a result of heart disease will go down if early detection of the condition is made possible, and the population as a whole will be in better health.

5.2 Impact on Environment

If people with heart disease had an early diagnosis, the number of people who passed away would be lower. The early detection of this condition enables patients to take preventative measures against it, which in turn enables them to lead healthier lifestyles. Because of this, there will be a really positive influence on the environment because people will no longer be anxious and will be able to go about their daily work. When accurate early warning has finally been achieved, the whole environment will undergo a transformation for the better.

5.3 Ethical Aspects

The collection of data presents various questions, particularly with regard to the ethical implications of the practice. It was said in the previous section (3.1). Treating cardiac disease using machine learning requires a significant quantity of information on the issue that is being resolved. The accumulation of data for the purposes of medical diagnostics requires the storage of a significant amount of personally identifiable information. When this data is made available to the general public, as it was on the Kaggle website that was utilized in this research, it opens the door for malicious actors to use it for their own ends.

For instance, insurance firms might use this information to train models that profile high-risk patients and decide whether or not to extend coverage to such patients, as well as whether or not to charge them a higher premium. Having one's data accessible to the public does, however, increase the risk of developing cardiovascular disease. This is a trade-off.

5.4 Sustainability Plan

The function of diagnostic instruments for cardiac disease should also be examined in the context of medical diagnosis. Diagnostic models for heart disease are not without their limitations and do not provide a perfect answer to the problem of medical diagnosis at this time. Providing a medical expert with a tool that just displays a number on a screen to determine whether a patient is ill or not without any further logic as to why that diagnosis was determined might encourage the medical professional to place an unwarranted amount of faith in the instrument. It is possible that using a model such as DT is better since, despite the fact that it could have a lower classification accuracy than other models, it does convey additional information about how that choice was made. This provides a medical expert with more information to take into consideration while assessing whether or not the diagnosis is accurate.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future

6.1 Summary of the Study

For the purpose of this investigation, in order to forecast cardiac disease, we used two distinct datasets. Both datasets underwent preprocessing in order to produce more accurate results. To acquire a deeper understanding of datasets, it is helpful to visualize their attributes. A number of algorithms for supervised machine learning were developed and put to use in order to determine which machine learning model is appropriate and effective for our dataset. The effectiveness of the approach is analyzed using the performance metric. In order to ensure that our model's accuracy is the highest possible for predicting heart disease, we are comparing each model to check with prior studies.

6.2 Conclusions

As the number of heart disease-related deaths continues to rise, it has become imperative to design a method that correctly and effectively forecasts heart disease. The object of this studies was to identify the optimal machine learning method for detecting heart diseases. For this research, we implemented eight distinct machine learning methods, including Decision Tree, Gradient Boost, Ada Boost, Support Vector Machine, K Nearest Neighbor, Extreme Gradient Boost, Logistic Regression, and Random Forest, to anticipate the occurrence of heart disease. We utilized two publicly accessible datasets via Kaggle. For the dataset of heart disease, GB provided the maximum accuracy of 91.80%, while Decision Tree yielded the lowest test accuracy of 75.00%. When we used ML algorithms to analyze the second cardiovascular disease dataset, we obtained a maximum accuracy of 74.85% with the GB algorithm and the lowest accuracy of 64.00% with the Decision Tree algorithm. Precision, recall, accuracy, F1 score, and AUC score were utilized as performance measurement metrics. In addition to using two distinct datasets, our study effort is far more accurate and efficient than those of prior researchers. In this way, our project differed from that of prior researchers. Following the results of our study, using machine learning models is crucial for the early identification of cardiac disease.

6.3 Recommendations

There are a number of suggestions regarding heart disease classifications, such as:

- Better outcomes can also be achieved by using deep learning algorithms.
- The accuracy of results may improve with larger datasets.

6.4 Implication for Further Study

The work can be improved in the future by establishing a website with the more accurate algorithm and by using a bigger dataset than the one that was utilized within the evaluation. These kinds of improvements will help to provide better results and will assist medical professionals in the accurate and efficient prediction of heart disease.

References

[1] Gjoreski, M., Simjanoska, M., Gradišek, A., Peterlin, A., Gams, M. and Poglajen, G., 2017, August. Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers. In 2017 International Conference on Intelligent Environments (IE) (pp. 14-19). IEEE.

[2] Savarese, G. and Lund, L.H., 2017. Global public health burden of heart failure. Cardiac failure review, 3(1), p.7.

[3] Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R. and Delling, F.N., 2019. Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation, 139(10), pp.e56-e528.

[4] Ramaraj, M. and Thanamani, A.S., 2013. A comparative study of CN2 rule and SVM algorithm and prediction of heart disease datasets using clustering algorithms. Network and Complex Systems, 3(10), pp.1-6.

[5] Gavhane, A., Kokkula, G., Pandya, I. and Devadkar, K., 2018, March. Prediction of heart disease using machine learning. In 2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 1275-1278). IEEE.

[6] Murthy, H.N. and Meenakshi, M., 2014, November. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In International conference on circuits, communication, control and computing (pp. 329-332). IEEE.

[7] Bashir, S., Khan, Z.S., Khan, F.H., Anjum, A. and Bashir, K., 2019, January. Improving heart disease prediction using feature selection approaches. In 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (pp. 619-623). IEEE.

[8] Ismaeel, S., Miri, A. and Chourishi, D., 2015, May. Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis. In 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015) (pp. 1-3). IEEE.

[9] Ekız, S. and Erdoğmuş, P., 2017, April. Comparative study of heart disease classification. In 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.

[10] Kedia, V., Regmi, S.R., Jha, K., Bhatia, A., Dugar, S. and Shah, B.K., 2021, April. Time Efficient IOS Application for CardioVascular Disease Prediction Using Machine Learning. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 869-874). IEEE.

[11] Bashir, S., Khan, Z.S., Khan, F.H., Anjum, A. and Bashir, K., 2019, January. Improving heart disease prediction using feature selection approaches. In 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (pp. 619-623). IEEE.

[12] El Hamdaoui, H., Boujraf, S., Chaoui, N.E.H. and Maaroufi, M., 2020, September. A clinical support system for prediction of heart disease using machine learning techniques. In 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 1-5). IEEE.

[13] Mohan, S.J., Kancharla, S., Illa, M., Arigela, S. and Appasani, M., Effective Detection of Cardiovascular Disease using Machine Learning.

[14] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H. and Gutierrez, J., 2017, July. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE symposium on computers and communications (ISCC) (pp. 204-207). IEEE.

[15] Shimaa Ouf, A.I., 2021. A PROPOSED PARADIGM FOR INTELLIGENT HEART DISEASE PREDICTION SYSTEM USING DATA MINING TECHNIQUES. Journal of Southwest Jiaotong University, 56(4).

[16] Dwivedi, A.K., 2018. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29(10), pp.685-693.

[17] Shah, D. and Patel, S., 2020. Santosh, and K. Bharti,“. Heart Disease Prediction using Machine Learning Techniques, 1, p.345.

[18] Singh, A. and Kumar, R., 2020, February. Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.

[19] Jagtap, A., Malewadkar, P., Baswat, O. and Rambade, H., 2019. Heart disease prediction using machine learning. International Journal of Research in Engineering, Science and Management, 2(2), pp.352-355.

[20] Fedesoriano. Heart Failure Prediction Dataset, 11 Clinical Features for Predicting Heart Disease Events. 2021. Available online: https://www.kaggle.com/fedesoriano/heart-failure-prediction (accessed on 29 July 2022).

[21] Ulianova, S. Cardiovascular Disease Dataset. Available online: https://www.kaggle.com/sulianova/ cardiovascular-disease-dataset (accessed on 29 July 2022).

[22] Khan, M.U., Aziz, S., Bilal, M. and Aamir, M.B., 2019, August. Classification of EMG signals for assessment of neuromuscular disorder using empirical mode decomposition and logistic regression. In 2019 International Conference on Applied and Engineering Mathematics (ICAEM) (pp. 237-243). IEEE.

[23] Christoph, M., 2020. Interpretable machine learning. A Guide for Making Black Box Models Explainable. 2019. URL: https://christophm. github. io/interpretable-ml-book [accessed 2022-03-04].

[24] Ghiasi, M.M., Zendehboudi, S. and Mohsenipour, A.A., 2020. Decision tree-based diagnosis of coronary artery disease: CART model. Computer methods and programs in biomedicine, 192, p.105400.

[25] Zheng, Q., Tian, X., Yang, M. and Su, H., 2019. The email author identification system based on support vector machine (SVM) and analytic hierarchy process (AHP). IAENG International journal of computer Science, 46(2), pp.178-191.

[26] Nurtanio, I., Astuti, E.R., Purnama, I.K.E., Hariadi, M. and Purnomo, M.H., 2013. Classifying cyst and tumor lesion using support vector machine based on dental panoramic images texture features. IAENG International Journal of Computer Science, 40(1), pp.29-32.

[27] Jiang, N., Fu, F., Zuo, H., Zheng, X. and Zheng, Q., 2020. A Municipal PM2. 5 Forecasting Method Based on Random Forest and WRF Model. Engineering Letters, 28(2).

[28] Liu, Y., Wang, Y. and Zhang, J., 2012, September. New machine learning algorithm: Random forest. In International Conference on Information Computing and Applications (pp. 246-252). Springer, Berlin, Heidelberg.

[29] Patel, J., TejalUpadhyay, D. and Patel, S., 2015. Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), pp.129-137.

[30] Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), pp.1937-1967.

[31] Li, Y., Yang, Y., Che, J. and Zhang, L., 2019. Predicting the number of nearest neighbor for kNN classifier. IAENG International Journal of Computer Science, 46(4), pp.662-669.

[32] Arafat, M.Y., Hoque, S., Xu, S. and Farid, D.M., 2019. Machine learning for mining imbalanced data. IAENG International Journal of Computer Science, 46(2), pp.332-348.

[33] Shu, X. and Wang, P., 2015, December. An improved Adaboost algorithm based on uncertain functions. In 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (pp. 136-139). IEEE.

[34] Li, S. and Zhang, X., 2020. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. Neural Computing and Applications, 32(7), pp.1971-1979.

[35] Lai, C.H., Yang, C.T., Kristiani, E., Liu, J.C. and Chan, Y.W., 2019, July. Using xgboost for cyberattack detection and analysis in a network log system with elk stack. In International conference on frontier computing (pp. 302-311). Springer, Singapore.

[36] Nelson, M.J. and Hoover, A.K., 2020, June. Notes on using Google Colaboratory in AI education. In Proceedings of the 2020 ACM conference on innovation and Technology in Computer Science Education (pp. 533-534).

[37] Python Programming Documentation [online]. URL: https://www.python.org/about/ Accessed on 29 July 2022.

[38] NumPy Documentation [online]. URL: http://www.numpy.org/ Accessed on 29 July 2022.

[39] Pandas Documentation [online]. URL :http://pandas.pydata.org/ Accessed on 29 July 2022.

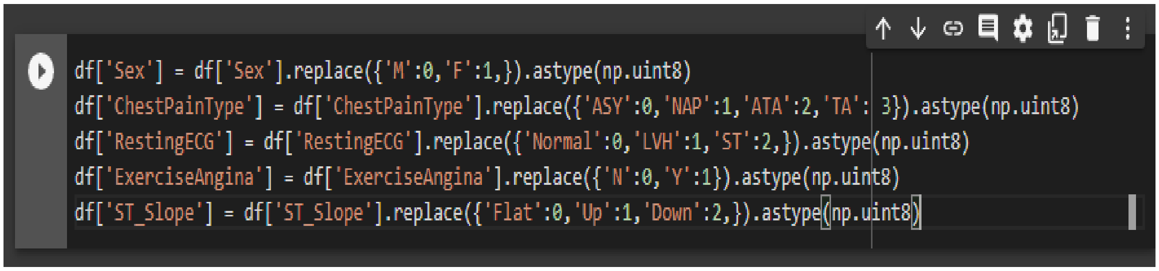
[40] Michael Waksom. An Introduction to Seaborn [online]. URL: http://seaborn.pydata.org/introduction.html Accessed on 29 July 2022.

[41] Fabian Pedregosa. Scikit-learn: Machine Learning in Python [online]. URL: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html Accessed on 29 July 2022.

[42] Getting Started With Streamlit Web Based Application [online]. URL: https://towardsdatascience.com/getting-started-with-streamlit-web-based-applications-626095135cb8 Accessed on 29 July 2022.

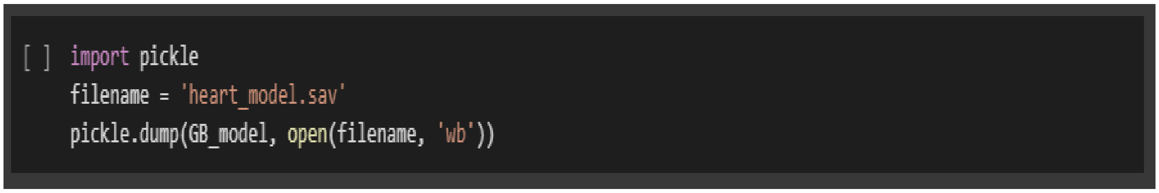
Appendices

Source code of Encoding categorical values



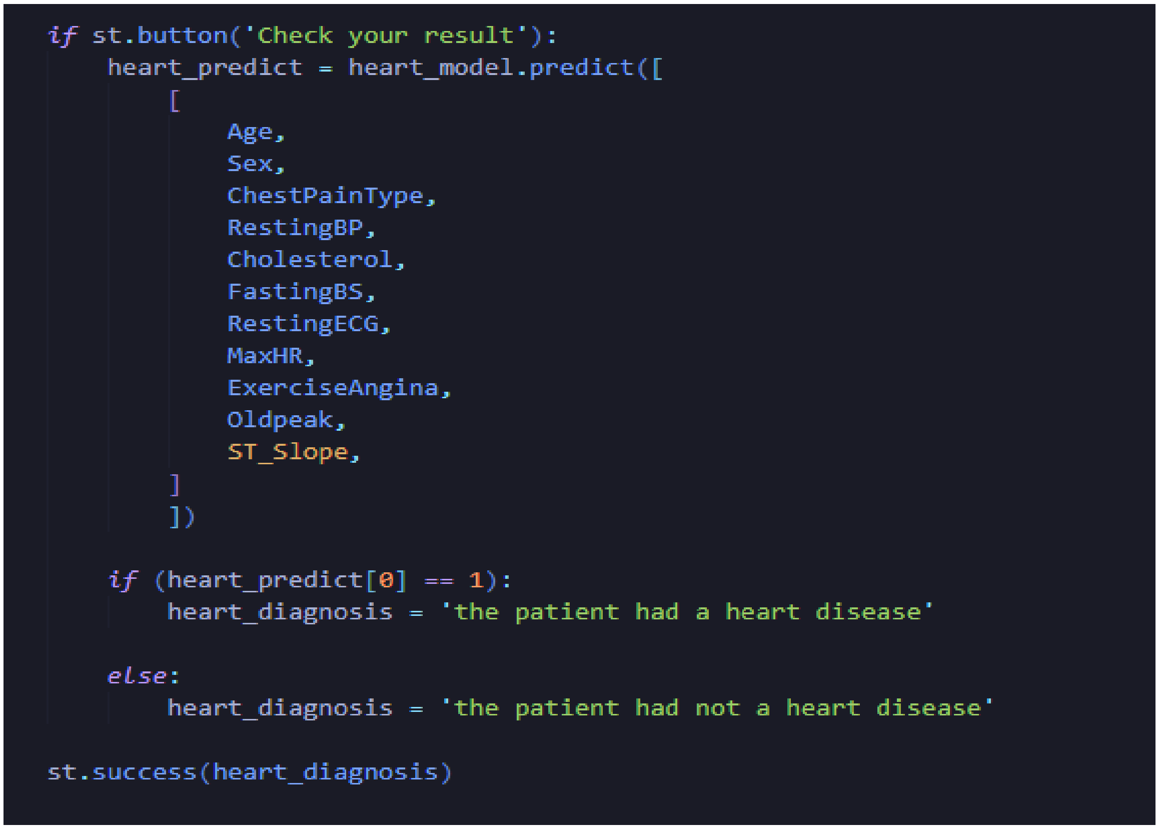
```
df['Sex'] = df['Sex'].replace({'M':0,'F':1,}).astype(np.uint8)
df['ChestPainType'] = df['ChestPainType'].replace({'ASY':0,'NAP':1,'ATA':2,'TA': 3}).astype(np.uint8)
df['RestingECG'] = df['RestingECG'].replace({'Normal':0,'LVH':1,'ST':2,}).astype(np.uint8)
df['ExerciseAngina'] = df['ExerciseAngina'].replace({'N':0,'Y':1}).astype(np.uint8)
df['ST_Slope'] = df['ST_Slope'].replace({'Flat':0,'Up':1,'Down':2,}).astype(np.uint8)
```

Source code of export best model using Pickle



```
import pickle
filename = 'heart_model.sav'
pickle.dump(GB_model, open(filename, 'wb'))
```

Source code of predicting heart disease based on user input



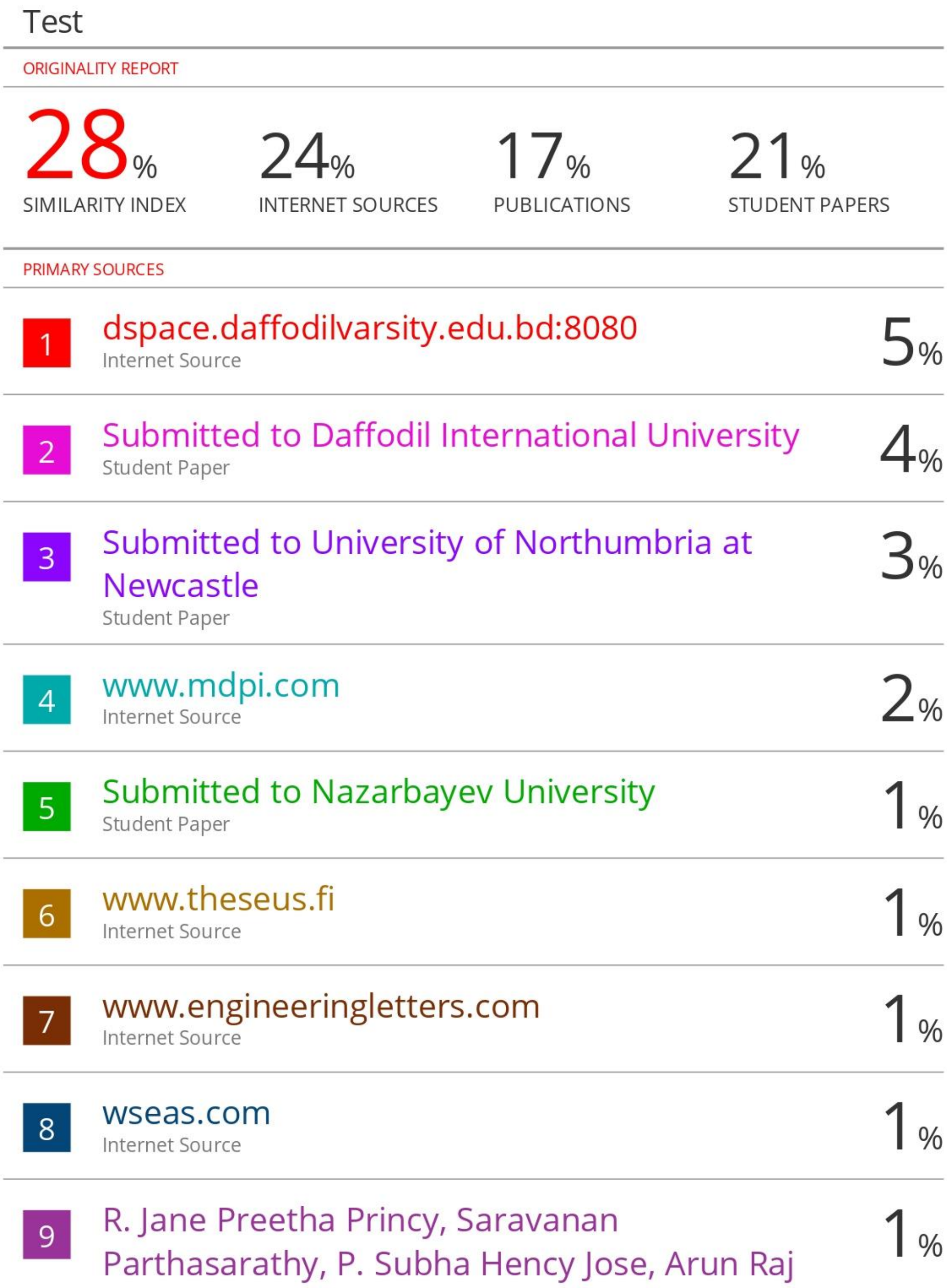
```
if st.button('Check your result'):
    heart_predict = heart_model.predict([
        [
            Age,
            Sex,
            ChestPainType,
            RestingBP,
            Cholesterol,
            FastingBS,
            RestingECG,
            MaxHR,
            ExerciseAngina,
            Oldpeak,
            ST_Slope,
        ]
        ])

    if (heart_predict[0] == 1):
        heart_diagnosis = 'the patient had a heart disease'

    else:
        heart_diagnosis = 'the patient had not a heart disease'

st.success(heart_diagnosis)
```

Plagiarism Report



Test

ORIGINALITY REPORT

28%	24%	17%	21%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	4%
3	Submitted to University of Northumbria at Newcastle Student Paper	3%
4	www.mdpi.com Internet Source	2%
5	Submitted to Nazarbayev University Student Paper	1%
6	www.theseus.fi Internet Source	1%
7	www.engineeringletters.com Internet Source	1%
8	wseas.com Internet Source	1%
9	R. Jane Preetha Princy, Saravanan Parthasarathy, P. Subha Hency Jose, Arun Raj	1%