

**CARDIOVASCULAR DISEASE DETECTION USING MACHINE LEARNING
ALGORITHMS**

BY

**MARIA BINTE BELAL
ID: 191-15-12698**

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

Supervised By

Fahad Faisal

Assistant Professor

Department of CSE
Daffodil International University

Co-Supervised By

Dr. Md Zahid Hasan

Associate Professor & Program Director MIS

Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project titled “**CARDIOVASCULAR DISEASE DETECTION USING MACHINE LEARNING ALGORITHMS**”, submitted by Maria Binte Belal 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 26 Jan,2023.

BOARD OF EXAMINERS



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

Chairman



Subhenur Latif
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mohammad Monirul Islam **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Dewan Md Farid
Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

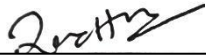
We hereby declare that, this project has been done by us under the supervision **Fahad Faisal, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:




Fahad Faisal
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Dr. Md Zahid Hasan
Associate Professor & Program Director MIS
Department of CSE
Daffodil International University

Submitted by:



Maria Binte Belal
ID: 191-15-12698
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 **Fahad Faisal, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining & Machine Learning*" 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 and other cardiovascular disorders have surpassed all others as the leading cause of mortality worldwide during the last several decades. Given the many potential causes of heart disease, it is essential to develop effective, efficient methods for making an early diagnosis and taking prompt action to treat the illness. In the healthcare industry, data mining has become more popular as a method for evaluating massive datasets. Researchers use a variety of machine learning and data mining approaches to analyze large, complicated medical datasets to help healthcare practitioners in making heart illness predictions. This research proposes a model that makes use of various supervised learning methods, including the Decision Tree, the Random Forest, the K-Nearest Neighbor, the XG Booster, the Support Vector Machine, the Gaussian Naive Bayes, the Bernays Naive Bayes, and the Logistic Regression, as well as two hyper-parameter optimization strategies, the Grid Search CV and the Randomized Search CV, and three feature selection strategies, the Univariate selection, the Model It makes use of the preexisting UCI collection of people with heart illness. Keeping score in Cleveland. The dataset has 1025 samples with 14 different characteristics. All of these are essential for the proper operation of different algorithms. The goal of this research is to determine how likely it is that participants will develop heart disease. The findings suggest that the Univariate selection method provides the maximum reliable outcomes.

TABLE OF CONTENTS

CONTENTS

	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v

CHAPTER

CHAPTER 1: INTRODUCTION **1-7**

1.1 Cardiovascular Disease Types	1-3
1.2 Prevalence of Cardiovascular Diseases	3
1.3 Machine Learning	3-5
1.3.1 Supervised Learning	4
1.3.2 Unsupervised Learning	4
1.3.3 Reward-Based Learning	4
1.4 Motivation	5-6
1.5 Objective	6
1.6 Expected Outcome	6-7
1.7 Report Layout	7

CHAPTER 2: LITERATURE REVIEW **8-11**

CHAPTER 3: METHODOLOGY **12-32**

3.1 Description of medical dataset used	12-15
3.2 Data Pre-processing	16-18
3.3 Algorithms Used	18
3.3.1. Decision Tree	18-19
3.3.2 Random Forest Algorithm	19-20

3.3.3	K Nearest Neighbor Classification(KNN)	20-21
3.3.4	Logistic regression	21-22
3.3.5	XG Booster	22-23
3.3.6	Support Vector machine	24-23
3.3.7	Gaussian Naive Bayes	25-26
3.3.8	Bernoulli Naive Bayes	26-27
3.4	Hyper-Parameter Optimization	27-28
3.4.1	Grid Search CV	28
3.4.2	Randomized Search CV	28-29
3.5	Feature Selection	29-30
3.5.1	Univariate Selection	30-31
3.5.2	Model-Based Feature Selection	31
3.5.3	Recursive Feature Eleimination	31-32
CHAPTER 4: RESULTS AND DISCUSSION		33-48
CHAPTER 5: LIMITATION		49
CHAPTER 6: CONCLUSION AND FUTURE PROSPECTS		50
REFERENCES		46-49

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1. Distribution of the instances.	14
Figure 3.1.2. Gender distribution of the instances.	15
Figure 3.1.3. Age Distribution for people with heart disease	15
Figure 3.2.1. Heat map of cross-correlation values.	17
Figure 3.2.2. Attribute bar plots.	18
Figure 3.5.1. Feature importance	30
Figure 4.1. Confusion matrices for all the classifiers.	34-40
Figure 4.2. ROC curve for all models.	32-48

LIST OF TABLES

TABLES	PAGE NO
Table 3.1.1 Attributes and details of dataset of heart disease	12-13
Table 4.1. Performance metrics of different ml algorithms using default hyperparameter	41

CHAPTER 1

INTRODUCTION

One of the most important organs in a human body is the heart. The blood arteries are the conduits that allow blood to be pushed through the circulatory system. It is essential for the body's multiple organs to have a healthy circulatory system since it carries blood, oxygen, and other substances to all of the body's organs. The importance of the heart as a part of the cardiovascular system cannot be overstated. Inadequate function of the heart may lead to diseases that are deadly as well as other significant health issues.

1.1 Cardiovascular Disease Types

Diseases that affect both the heart and the blood arteries are included in the group of conditions that are commonly referred to as cardiovascular diseases, or CVD for short. Conditions such as coronary artery disease (CAD), which may lead to symptoms such as chest pain and even a heart attack, are included in the category of cardiovascular disease (commonly known as a heart attack). Plaque, a waxy material, may build up within the coronary arteries and lead to coronary heart disease (CHD), which is another kind of heart ailment. CHD is defined by the buildup of this waxy substance. These are the arteries that feed the muscle of the heart with oxygenated blood, and they are located in the heart. Atherosclerosis is the medical name for the condition that takes place in these arteries when plaque starts to build up in them. It might take years for the first signs of plaque to appear. This plaque has the ability to either shatter or become more rigid over time (break open). Plaque that has hardened over time progressively narrows the coronary arteries, which in turn reduces the quantity of oxygen-rich blood that can pass through those arteries and reach the heart. If it breaks, the surface of this plaque has the potential to create a clot in the bloodstream. The majority of the time, a substantial blood clot is able to completely cut off the blood flow to a coronary artery. The repercussions of the ruptured plaque over time include the hardening or constriction of the coronary arteries. If the blood flow is not quickly restored, the damaged heart muscle will begin

the degenerative process of wasting away. If treatment for a heart attack is delayed for even a short period of time, the patient may develop major health complications or possibly pass away. Around the globe, heart attacks are a leading cause of mortality in both men and women. The following is a list of common symptoms that occur during a heart attack. [1]

1. Chest pain

It is the most typical warning sign that a heart attack is about to occur. It is possible for a person to experience pain, tightness, or pressure in their chest while they are experiencing a heart attack or have a blocked artery in their body.

2. Nausea, Indigestion, Heartburn and Stomach Pain

Some of these symptoms are among the warning signs of a heart attack that are often neglected by people. In most cases, women are the ones who are affected by these symptoms more often than males do.

3. Pain in the Arms

Most of the time, the discomfort begins in the chest and gradually spreads to the arms, particularly the left arm.

4. Feeling Dizzy and Light Headed

factors resulting in a loss of equilibrium

5. Fatigue

You shouldn't ignore simple tasks that make you feel fatigued.

6. Sweating

A stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, cardiac arrhythmia, congenital heart disease, valvular heart disease, aortic aneurysms, peripheral artery disease, and venous thrombosis are a few examples of additional cardiovascular disorders that are extremely prevalent. Certain irregularities in the way the circulatory system operates can give rise to heart conditions, or existing heart conditions can be made worse by certain choices in lifestyle, such as smoking cigarettes, adhering to certain dietary patterns, or leading a sedentary lifestyle, amongst other things. The effective treatment and management of heart diseases are made possible by the early diagnosis of heart disorders. In this particular circumstance, early discovery is of the utmost importance. When it comes

to the prevention of heart disease, having a comprehensive awareness of both its origins and its repercussions is generally helpful.

1.2 Prevalence of Cardiovascular Diseases

According to some estimates, cardiovascular illnesses are responsible for the deaths of 17.5 million people all over the globe. More than seventy-five percent of fatalities that occur in countries with middle-income and low-income levels are attributable to cardiovascular illnesses. In addition, heart attacks account for 80% of all deaths caused by cardiovascular disease. [2] A growing number of people in India are diagnosed with cardiovascular diseases on an annual basis. At the moment, there are over 30 million people in India who are afflicted with heart disease. More than 2 million open-heart surgeries are performed every year in India. The fact that the number of patients who need coronary treatments has been growing at a rate of 20% to 30% in recent years is a very alarming trend. [3] The remaining aspects of the thesis are discussed in the following paragraphs. In the next section, you will find an explanation of some well-known data mining techniques for predicting heart disease. In this section, we take a look at some of the most common data mining strategies that are used for the sake of doing data analysis. In Section 4, a summary of the methodologies and conclusions of past research on the diagnosis and prognosis of heart illness is provided. In Section 5, we will cover both the positive and negative aspects of doing a literature study. The conclusion may be found in Section 6, which also includes a discussion on the essay's path going forward.

1.3 Machine Learning

The primary goal of machine learning, which is a subfield of artificial intelligence that is gaining increasing prominence, is to create systems and then give those systems the ability to learn. After that, make use of these algorithms to make predictions about the future based on data from the past. Machine learning algorithms need to be trained on a training dataset before they can be used to generate a model. The system makes predictions about heart disease by utilizing the most recent input data. The input dataset is analyzed via the lens of machine learning in order to unearth hidden patterns, after which models are developed. It generates precise forecasts based on the raw data

provided. [4] Using the newly provided data, the algorithm makes predictions about heart disease, and then evaluates how accurate those predictions are. The following categories include the many methods of profiting from machines:

1.3.1 Supervised Learning

The model is trained with the help of a dataset that contains labels. It takes in information and then outputs outcomes. Datasets go through a process of categorization before being separated into training and test sets. Our model is trained using the training dataset, and the test dataset gives additional information that is used to evaluate how well the model predicted the data. The results produced by the models are included in the dataset. Categorization and regression are two examples of its applications.

1.3.2 Unsupervised Learning

The dataset that is being utilized for training does not include any labels or categories. The goal is to unearth hidden data patterns as quickly and efficiently as possible. Training is given to the model so that it may learn to recognize patterns. It is able to easily predict hidden patterns in each new dataset that is supplied, but as it examines the data, it utilizes the datasets to draw inferences about the hidden patterns. The dataset that we are using for this approach does not display any replies. Examples of unsupervised learning techniques include the clustering approach and other similar methods.

1.3.3 Reward-Based Learning

The software is able to learn from experience rather than from a tagged dataset or outcomes that are tied to data since it does not utilize either of these. The structure of the aforementioned method, which improves the way it appears based on how it interacts with the world around it, finds, via analysis of and experimentation with a variety of possibilities, how to address the method's deficiencies and create the desired result. Estimating the chance of developing heart disease is a common use of classification strategies, which are often employed in supervised learning approaches.

Research on a wide range of diseases, including liver disease [5], Parkinson's disease [6], heart disease [7], breast cancer [8], lung disease [9], and others, has benefitted from the application of a number of different machine learning strategies over the course of the past few decades in the healthcare industry [5, 6]. The many approaches that were used in order to make accurate diagnoses of the illnesses each generated favorable results overall. Patients have a number of expectations, the most important of which is that their illnesses, particularly cancers, would be appropriately and quickly detected. The fact that the technique for detecting demands specific training and skill shouldn't come as much of a surprise to anybody. We believe that the use of machine learning and data mining may lead to an improvement in the accuracy of the approach, a decrease in the number of incorrect diagnostics, and, eventually, the provision of high-quality treatment to patients.

The primary goal of this investigation is to use a wide range of data mining and machine learning strategies in order to come up with an original and accurate model for CAD identification. The traditional machine learning techniques of Decision Tree, Random Forest, K-Nearest Neighbor, XG Booster, Support Vector machine, Gaussian Naive Bayes, Bernoulli Naive Bayes, and Logistic Regression, as well as the three feature selection techniques known as Univariate selection, Model based feature selection, and Recursive Feature Elimination, are all extremely well known. were evaluated, as well as two hyper-parameter optimizations, such as Grid Search CV and Randomized Search CV, which were carried out. We feel that the approach that was presented may be useful in triage, reduce the need for professional counsel, and ultimately result in time and cost savings when diagnosing coronary artery disease (CAD)

1.4 Motivation

Today, heart disease is a leading cause of death in both men and women, making it a significant public health issue that warrants careful consideration. According to the World Health Organization (WHO), heart disease is responsible for the deaths of 17.9 million people each year, which accounts for 31% of all deaths. Although there are methodologies and techniques for machine learning accessible for the prediction of cardiac issues, there are now no suitable models available that are able to do so in a more accurate and quicker manner. At the moment, there is no reliable automated technology

that may assist in the diagnosis of heart disease or minimize the severity of its symptoms. Therefore, it will be a significant achievement to use algorithms for machine learning in order to lessen the impact of the condition on a daily basis. It has the potential to greatly delay the development of cardiac conditions, as well as enhance the quality of life for individuals who already have them. The fundamental objective of this research project is to develop a model that can anticipate the occurrence of cardiac issues. In addition, the purpose of this research is to determine the classification method that has the maximum degree of accuracy in predicting the illness that was described before.

1.5 Objective

The proposed research investigates the aforementioned four classification algorithms in order to conduct performance analysis and make predictions about heart disease. The following is a list of the goals that the work being done on the present project aims to accomplish:

1. The primary purpose of this research is to develop a diagnostic tool that may help doctors detect cardiac problems at an earlier stage.
2. The purpose of this investigation is to make an accurate diagnosis of whether or not a patient has cardiac disease.
3. The purpose of this research is to assess whether or not it is probable that the patient will be diagnosed with any cardiovascular heart diseases. The medical parameters that will be considered include the patient's gender, age, chest discomfort, fasting sugar level, and other similar factors.
4. In addition, the purpose of this study is to identify the most effective categorization approach for determining whether or not a patient has cardiac illness.

These are the most important objectives that we have set for ourselves.

1.6 Expected Outcome

After the project is finished, the following results are anticipated to have occurred: 1. The medical professionals will have access to a tool that is quick and simple to use for the diagnosis of cardiovascular disease.

2. It is possible to diagnose heart disease at an extremely early stage.

3. There will be less of a need for elaborate and pricey diagnostic equipment.

1.7 Report Layout

The structure of the report was the point at which everything started; how was it constructed?

In essence, work was done on four different chapters, and they are as follows:

The first chapter of the book is “Introduction”

Discussion of the project's rationale, aims, and projected outcomes may be found in the first chapter, which provides a description of the project. Following that is a discussion on the format of the report.

“Literature review” was discussed in Chapter two.

Methodology was Mentioned in Chapter three. All the necessary design requirements of the project are discussed in this chapter.

Chapter four shows Overall results that were found from this project and all results were also thoroughly discussed

Limitations of the project was discussed in Chapter five.

The project is concluded with future scope at Chapter six.

CHAPTER 2

LITERATURE REVIEW

The use of various machine learning algorithms on medical datasets of various diseases, such as the detection of different types of cancer, has resulted in amazing progress being achieved. The basic clinical methods for evaluating heart sickness include the electrocardiogram (ECG), echocardiogram, cardiac computed tomography (CT) scan, blood tests, cardiac catheterization, Holter monitoring, and cardiac magnetic resonance imaging. In this section, we will review the study on the detection of cardiac diseases, with a particular focus on CAD. We will do this by applying data mining, machine learning, and other data mining approaches. [10]

It was recommended by Shu et al. to use quantitative computerized TCM in conjunction with representation-based approaches in order to forecast cardiac disease (HD).

Appropriate classifiers (a total of 11 algorithms) were fine-tuned via the use of a probabilistic collaborative agency-based technique. The block and the ProCRC classifier FHB + LCB + NBB were combined to create the highest level of accuracy, which was 88.01%. [11] Pawiak presented an innovative method for researching cardiac diseases, which included an analysis of the ECG data and a model that was based on evolving neural networks. Following preprocessing that included normalization and feature extraction, the proposed model was applied to a dataset including information on heart illness. The four major classifiers that were used were SVM, KNN, PNN, and RBFNN. According to the collected data, the evolving neural system that made use of SVM fared the best when applied to a 17-class ECG dataset. It achieved an accuracy of 90%. The recommended approach also required an extremely short amount of time for responses. [12] From 2012 till 2017, Alizadehsani et al. explored a variety of machine learning approaches in order to identify coronary artery disease (CAD). [13] Other researchers focused their attention on the detection of heart illness by using rule mining techniques. The Particle Swarm Optimization (PSO) method was used in order to build rules for a

dataset including information on cardiac disease. The accuracy achieved as a result was 87%. [14]

In addition, the powerful algorithms known as decision trees (DT) may be used for the correct diagnosis of heart illness. Abdar made use of four different decision tree algorithms in order to provide rules for the dataset on heart illness that were understandable and crystal clear (C5.0, CHAID, CART, and QUEST). The research results indicated that decision trees may have a good performance and develop simple rules for the dataset. It was determined that C5.0 had the greatest performance since it had the highest level of accuracy (85.33%). [15] In addition to using NN, K-Nearest Neighbor (K-NN), C5.0, and SVM, Abdar et al. analyzed a dataset that had 270 records and was connected to heart disease. In order to choose characteristics that had a statistically significant impact ($p = 0.05$ or above), logistic regression was utilized. They discovered that the C5.0 algorithm produced the greatest results when linked with the qualities of choice, with an accuracy of 93.02%. Clustering was the method that Verma and colleagues used in order to arrive at a diagnosis for heart illness. In order to create their model, they made use of K-means clustering, PSO search, and the selection of the correlation-based feature subset (CFS). [16] According to the findings of Verma and colleagues, the suggested model performed very well when multinomial logistic regression (MLR) was used, achieving an accuracy of 88.40%. [17] An unsupervised model-based clustering method was employed by Hinchcliff et al. to investigate the involvement of the heart in systemic sclerosis. This was done so that the debate may go further. The model that was used resulted in the dataset being segmented into several categories, which in turn showed certain correlations that were not previously known to exist between the samples. [18] A powerful method of machine learning known as fuzzy systems has the potential to be used in the creation of an intelligent diagnostic model. The unique model that Zou and Deng introduced was one that was based on fuzzy concept lattice. [19] In order to diagnose coronary heart disease, Lahsasna et al. created a rule-based fuzzy system (FRBS). The performance of the model was improved by using a multi-objective evolutionary algorithm in conjunction with an ensemble classifiers approach (ECS). [20] The A-FRBS classifier achieved the highest possible accuracy on the test data, which was 84.44%. In addition, Hassan et al. proposed a fuzzy expert soft

system for predicting CAD. This system was effective in their eyes since it was a model that they could comprehend, hence they deemed it to be successful. [21] Paul and colleagues recommended using a rule-based fuzzy system that was built on a weighted adaptive technique for the purpose of calculating the likelihood of acquiring heart disease. [22]

Using efficient methods of ensemble learning may make it possible to boost the performance of traditionally-implemented algorithms. Radial Basis Function (RBF), K-NN, Naive Bayes (NB), Discrete Trees (DT), Multilayer Perceptron (MLP), Support Vector Machines (SVM), and Single Conjunctive Rule Learner were the seven machine learning methods that Pouriyeh et al. utilized on a dataset for heart disease (SCRL). On top of optimization, methods for stacking, boosting, and bagging ensemble learning were applied. Out of all of these methods, the boosting-based technique of SVF fared the best. [23] A gradient boosting classifier was used in order to make a prediction about CAD cases that had physiological importance. In order to test this model, a dataset of 252 patient records was used. Scores for accuracy, sensitivity, positive predictive, specificity, and negative predictive were, respectively, 52.7%, 84.6%, 78.2%, 63.0%, and 68.30% for the recommended approach. [24]

For the purpose of diagnosing CAD using ECG data, an innovative model known as stacked CNN-LSTM was presented. The effectiveness of machine learning algorithms was significantly improved by the use of a simple stacking. This hybrid model correctly predicted not just data that were not particular to a single subject, with an accuracy of 99.85%, but also data that were specific to a single subject, with an accuracy of 95.76%. [25]. The accuracy of the suggested approach was 89.70% when the size of the total ensemble was 10, which was an improvement over the performance of earlier techniques in the dataset on heart disease. The algorithm had been put to use in order to conduct an analysis of four distinct datasets that were related to the illnesses of the breast, diabetes, heart, and hepatitis. [26]. In addition, the effect that machine learning strategies have on the categorization of supraventricular and ventricular ectopic beats was looked into. Using a recently created hierarchical cardiac classification method that makes use of random projection and the SVM ensemble technique, each recording from Data Set 2

(DS2 as a training data set) was given the Association for the Advancement of Medical Instrumentation (AAMI) standard. This system had an accuracy rate of 99.90%, high performance, and was somewhat fast. [27]

CHAPTER 3

METHODOLOGY

In the interest of assisting both patients and medical professionals working within the field of medicine, this research makes an effort to make a prognostication about the likelihood of acquiring heart disease as a possible cause of computerized heart disease prediction. This research study investigates both the analysis of datasets and the use of a variety of machine learning strategies to the data gathering in order to fulfill the requirements of the aim. This research demonstrates, among other things, that some characteristics are more significant than others when it comes to establishing an expectation of better accuracy. Because not all of a patient's attributes will necessarily have a substantial influence on the result, this might save a patient money on numerous research that they have to do. [28]

3.1 Description of medical dataset used

The Cleveland database, which is located in the UCI repository and includes patient data on persons who have cardiac disease, was used for the purpose of this research. It includes a real dataset with 1025 data instances and 14 distinct variables (13 predictors and 1 class), some of which include age, sex, resting blood pressure, and others. The class is "age" (Table 3.1.1).

TABLE 3.1.1. ATTRIBUTES AND DETAILS OF DATASET OF HEART DISEASE

Age	Continuous	Age in years
Sex	Discrete	0=female 1=male
Chest pain type	Categorical	Typical angina=1, Atypical angina=2, Non-anginal pain=3, Asymptomatic=4
Resting blood pressure	Continuous	Resting blood pressure (mmHg)

Cholesterol	Continuous	Cholesterol (mg/ml)
Fasting blood sugar	Discrete	Lower than 120 mg/ml=0, Greater than 120 mg/ml=1
Rest ecg	Categorical	Normal=0, ST-T wave abnormality=1, Left ventricular hypertrophy=2
Max heart rate	Continuous	Max heart rate (bpm)
Exercise induced angina	Discrete	Yes=1, No=0
Slope	Categorical	Up sloping=1, Flat=2, Down sloping=3
Vessels colored by fluoroscopy	Categorical	Zero=0, One=1, Two=2, Three=3, Four=4
thalassemia		No=0, Normal=3, Fixed Defect=6, Reversible Defect=7
Target	Discrete	Yes = 1 No = 0

A bar plot, similar to the one shown in Figure 3.1.1, is also used to represent the distribution of the instances included in the data set. The dataset included the information of around one hundred people who had heart illness.

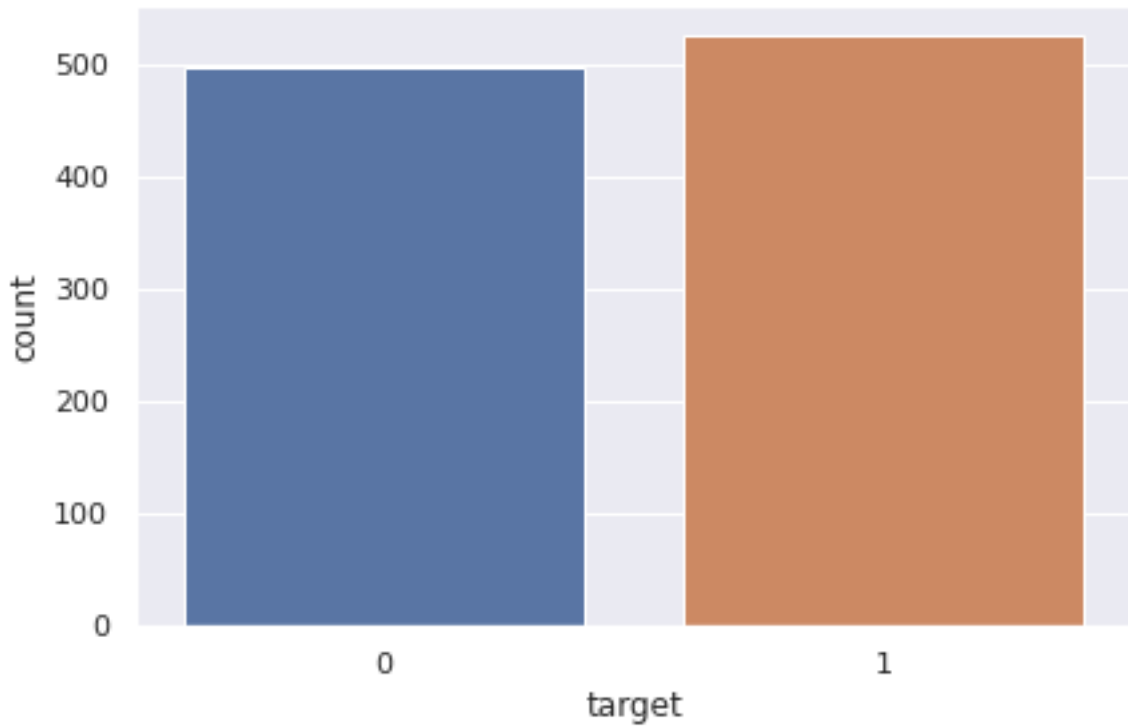


Figure 3.1.1. Distribution of the instances.

In yet another bar plot, which can be seen in Figure 3.1.2, the gender distribution of the instances that make up the data set is shown. It is unmistakable that there were about sixty percent more men (almost 600) in the sample than there were females (400).

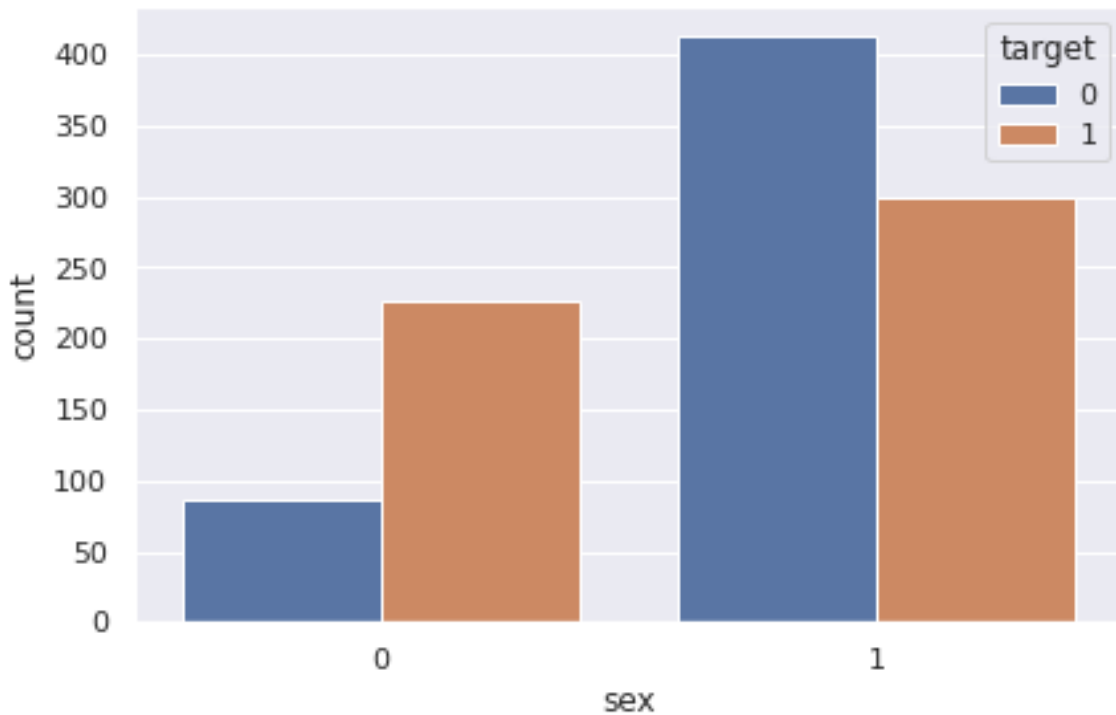


Figure 3.1.2. Gender distribution of the instances.

The difference in age between individuals who have cardiac disease and those who do not is seen in Figure 3.1.3. It was shown that people between the ages of 50 and 65 had a disproportionately high prevalence of heart disease patients.

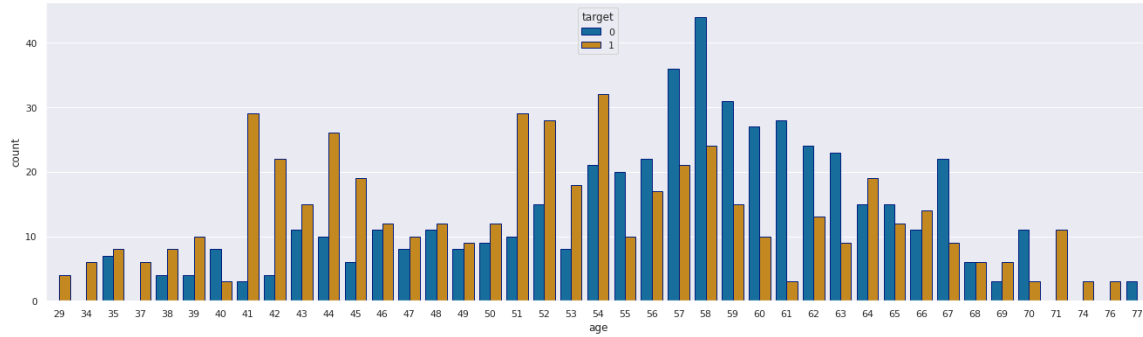


Figure 3.1.3. Age Distribution for people with heart disease.

3.2 Data Pre-processing

The data that was collected from the actual world included a considerable quantity of information that was either missing or noisy. These data have been preprocessed in order to prevent the difficulties that were mentioned and to create reliable predictions. The sequential chart for our proposed model is shown in Figure 3.2.1 below.

In general, the data obtained comprises noise as well as values that are missing. These data need to have any noise removed from them, and any missing values need to have replacements found for them in order to provide an accurate and useful output.

In order to comprehend the data that is shown in Figure 3.2.1, a measure that was developed to determine the degree to which each metric and the goal diagnostic are connected was developed. It should be pointed out that time was the sole factor that had the strongest links to the desired characteristic. This makes it much easier to build an overall perspective on the data that is being processed.

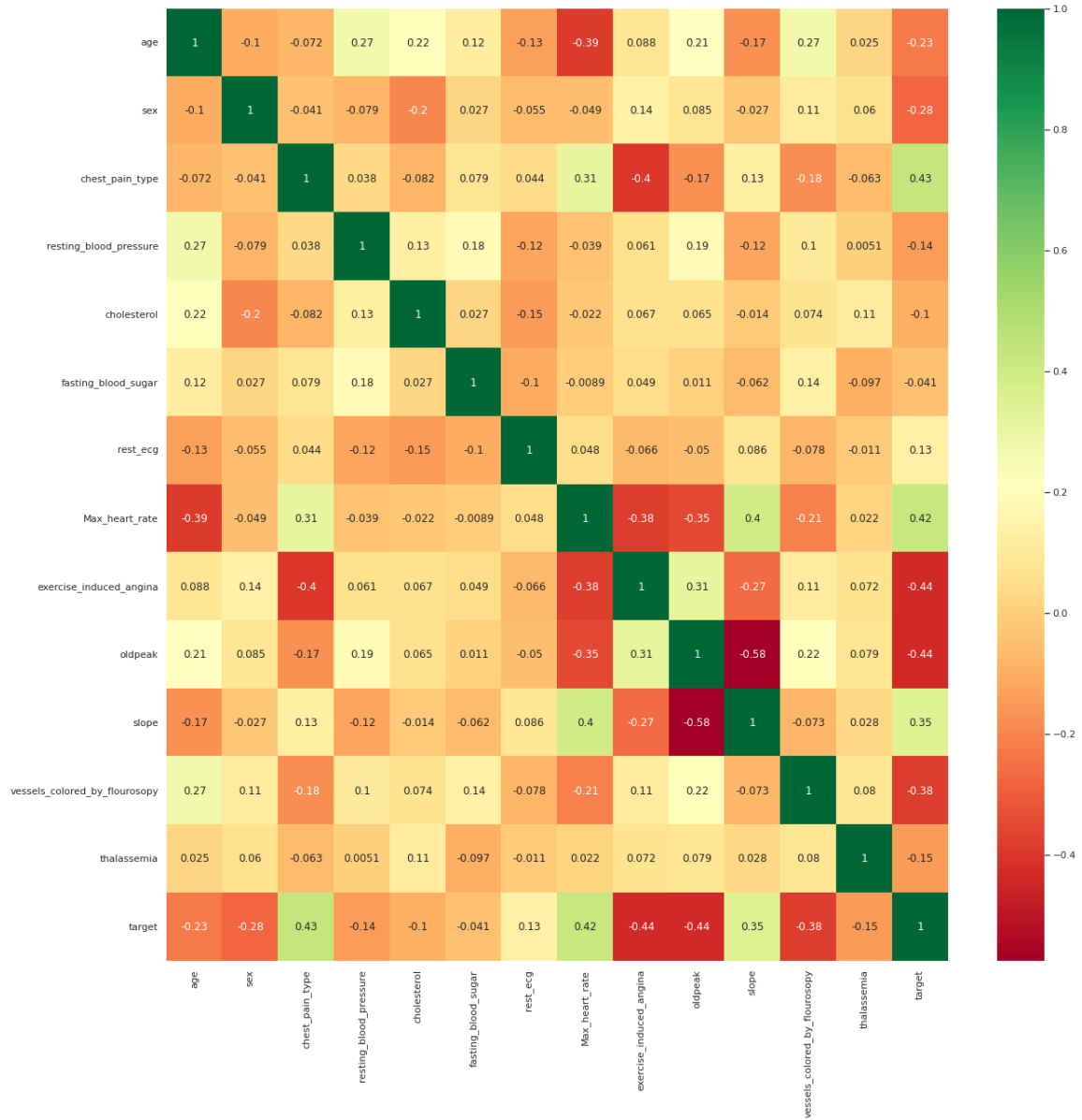


Figure 3.2.1 Heat map of correlation values.

In addition, bars were made for the characteristics data visualization in order to have a sneak peek at the distribution of the data, as can be seen in Figures 3.2.2. It should be underlined that the normal distribution applies to all continuous characteristics, since this is an important point.

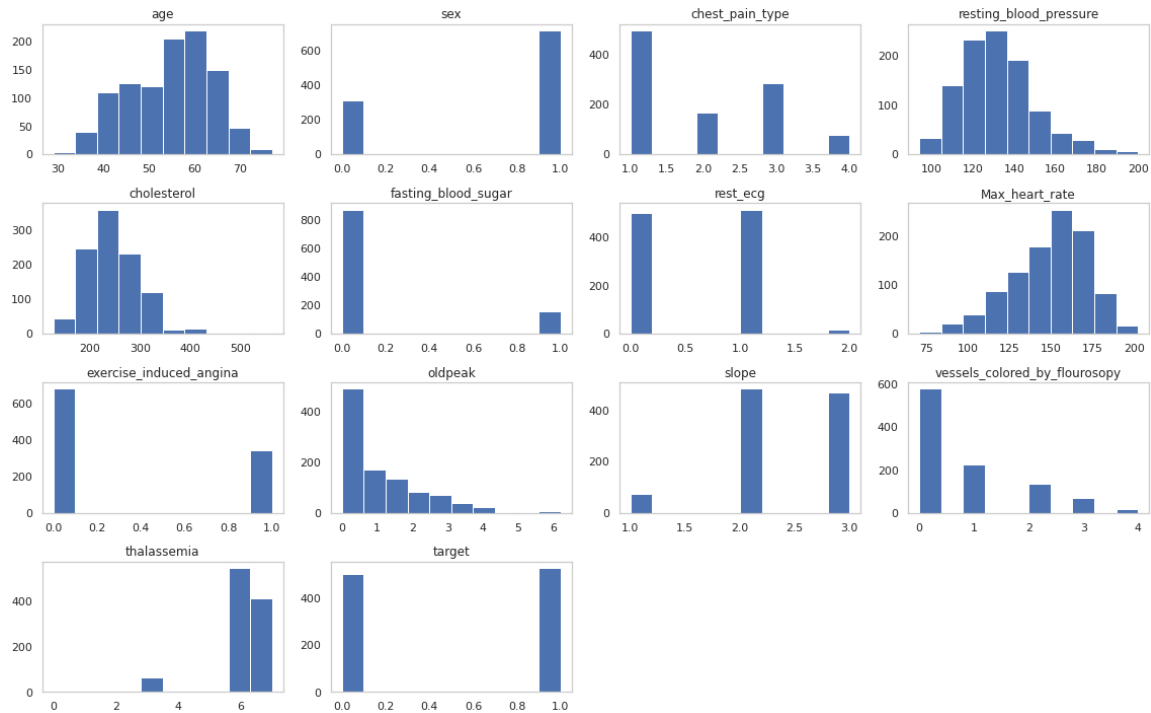


Figure 3.2.2. Attribute bar plots.

3.3. Algorithms Used

3.3.1 Decision Tree

Decision trees, which are a kind of classification procedure, may make use of either categorical or numerical data. The use of decision trees results in the production of structures that resemble trees. Decision trees are a basic and common method that are used for the management of medical information. It is not difficult to do data analysis when using a graph that has a tree structure. The decision tree model conducts data analysis by using three nodes as its foundation.

This method divides the data into two or more groups that are identical based on the indications that are considered to be the most significant. In order to divide the data in accordance with the entropy of each characteristic, the predictors with the highest information gain or the lowest entropy are utilized:

$$\text{Entropy}(S) = \sum_{i=1}^c -P_i \log_2 P_i,$$

$$\text{Gain Index}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

The findings are simpler and more straightforward to read and understand. This method is superior to others in terms of precision since it conducts its analysis on the dataset using a network that resembles a tree. However, given that just a single characteristic is considered in the decision-making process at any one moment, the data may be overclassified. [29]

3.3.2 Random Forest Algorithm

When doing supervised classification, the algorithmic method known as the random forest is often used. In the context of this method, a forest consists of a great number of trees. The prediction for a model is based on the category that received the most votes in a random forest, which is determined by the class expectation emitted by each tree. The employment of additional trees in the random forest classifier leads to a higher level of accuracy being generated. The following is a rundown of the three common approaches:

- Forest RC (random blend)
- Forest RI (random input choice)
- Combination of forest Resource Inventory and Forest Resource Census

It is utilized for classification issues as well as regression issues, although it is particularly useful for the former and can deal with variables that are not present. In addition, the results are either unexpected or it takes a long time to create predictions since they need a large amount of data sets and numerous trees. The accuracy of the

random forest approach was able to be improved to 91.6% thanks to the Cleveland dataset. Using the People's dataset, we were able to achieve an accuracy of 97%. [30]

The equation for a Random Forest can be represented as:

$$y_{\text{pred}} = \text{majority_vote}(T1(x), T2(x), \dots, Tn(x))$$

Where:

- y_{pred} is the final predicted label
- $T1(x), T2(x), \dots, Tn(x)$ are the predictions of the individual decision trees

The predictions made by each individual decision tree are combined into a single output by the function known as majority vote. While dealing with classification issues, it is possible to define it as the mode of the predictions, whereas when dealing with regression issues, it is possible to define it as the mean of the predictions.

The selection of the decision tree in Random Forest may be done by bootstrapping the dataset, and for each tree, a random collection of characteristics is chosen to divide the nodes in the tree. The procedure in question is referred to as the random subspace method.

The Random Forest approach may increase the accuracy of the decision tree by lowering the variation produced by the associated characteristics in the data. By averaging the predictions of many different decision trees, it also helps to decrease the overfitting that might be induced by the decision tree.

3.3.3 K Nearest Neighbor Classification(KNN)

The Knearest neighbors algorithm is a kind of technique that may be used for supervised classification. The classification is done by using the closest neighbor approach. Learning of this kind is known as "instance-based learning." The Euclidean distance is what is used to determine how far apart two attributes are in relation to one another. In order to designate another point, it makes use of a number of points that have been given names. After the data have been sorted into groups according to their similarities, K-NN may be used to complete the dataset by supplying missing values. Following the completion of

the task of completing the missing values, the data set is next put through a variety of prediction procedures. The accuracy of the results may be improved in a number of different ways by combining these methods. The K-NN method is simple to implement since it does not need the creation of a model or the formulation of any additional assumptions. Problems involving categorization, regression, and search are all amenable to being solved using this technique. Despite being the most straightforward approach, K-Nearest Neighbors suffers from accuracy issues caused by noisy and irrelevant feature combinations. In the research carried out by Pouriyeh and colleagues, using the parameter $K=9$ led to an accuracy of 83.16 percent. [23]

The equation for KNN can be represented as:

$$y_{\text{pred}} = \text{mode}(y_1, y_2, \dots, y_k)$$

Where:

- y_{pred} is the predicted label of a new data point
- y_1, y_2, \dots, y_k are the labels of the K closest training data points

mode is the function that returns the label that occurs the most often among the K data points that are located the closest together.

The KNN algorithm is a straightforward and effective method that performs well for classification and regression issues. It is especially useful for problems involving limited datasets and situations in which the decision boundaries are not clearly defined. It is simple to build, does not call for any assumptions to be made about the way the data is distributed, and is straightforward to understand.

3.3.4 Logistic regression

In spite of the fact that it has its roots in the eighteenth century, logistic regression (LR) analysis has developed into a statistical technique that is being more used in medical research, notably over the course of the previous two decades. In circumstances when the probability of a binary (dichotomous) outcome must be predicted from one or more independent (predicting) factors, it is generally accepted as the ideal statistic.

LR is used to establish if or not an event took place, as opposed to establishing when it did so (time course information is not used). Research in the field of health sciences often makes use of this method because it lends itself particularly well to models that include states of illness (whether well or ill) and decision-making (yes or no). If the variable to be predicted falls into more than two categories, more complicated forms of logistic regression are required to handle the issue. These more complex forms of logistic regression are known as polychotomous or multinomial logistic regression. [31]

The equation for logistic regression can be represented as:

$$P(y = 1|x) = 1 / (1 + e^{(-w^T \times (x - b))})$$

Where:

- $P(y=1|x)$ is the predicted probability of the output y being in class 1 given the input features x
- w^T is the weight vector
- x is the feature vector
- b is the bias term
- e is the Euler's number

The logistic function, which is often referred to as the sigmoid function, is a function that translates the anticipated probability between 0 and 1.

3.3.5 XG Booster

Extreme Gradient Boosting, or XGBoost for short, is a software package that is open-source and designed for gradient boosting on decision trees. It is often used in machine learning contests as well as in industry since it was developed for the efficient and consistent management of huge datasets. It is well-known for its strong performance as well as its ability to deal with huge numbers of features and manage missing data.

The XGBoost algorithm utilizes a technique known as ensemble learning called gradient boosting as its foundation. It is able to produce a strong learner that is capable of making

correct predictions by combining a number of weak learners, also known as decision trees.

The objective function is the primary equation that is used while working with XGBoost. This function is defined as the sum of the loss function and a regularization term. The loss function calculates the percentage of variance that exists between the value that was anticipated and the value that actually occurred. By including a penalty term for models that have substantial weights, the regularization term contributes to the prevention of overfitting.

The objective function in XGBoost can be represented as:

$$\text{Objective} = \text{Loss}(y, y_{\text{pred}}) + \Omega(f)$$

Where:

- y is the true label
- y_{pred} is the predicted label
- f is the decision tree
- $\Omega(f)$ is the regularization term

The regularization term $\Omega(f)$ can be represented as:

$$\Omega(f) = \gamma T + 1/2 \times \lambda \times \sum(w^2)$$

Where:

- γ is the parameter for controlling the complexity of the tree
- T is the number of leaves in the tree
- λ is the L2 regularization term
- w is the weight of each leaf node

Adjusting the parameters of the decision trees is one of the goals of the XGBoost algorithm, which is meant to maximize the value of the objective function.

3.3.6 Support Vector machine

Support Vector Machine, or SVM for short, is a technique for supervised learning that is useful for solving issues involving classification and regression. The fundamental concept underlying support vector machines (SVM) is to locate a hyperplane, which may be thought of as a line or plane in high-dimensional space, that effectively divides the data points into distinct categories. The objective is to locate the hyperplane that minimizes the margin, which may be thought of as the distance that separates the hyperplane from the support vectors that are the data points that are closest to it from each of the classes.

SVMs are especially helpful in situations in which there are more samples than there are features, as well as in circumstances in which the classes are not entirely linearly separable from one another. In this scenario, Support Vector Machine (SVM) makes use of a method known as the kernel trick, which transfers the input data into a higher-dimensional space in which the classes may be linearly separated from one another. Kernels such as linear, polynomial, and radial basis function (RBF) are often used in support vector machines (SVM).

A mathematical representation of SVM may be found in the form of a quadratic optimization problem with linear constraints. The purpose is to determine the values of the parameters that, given to the restrictions, provide the lowest possible value of the objective function.

SVM is useful for a broad variety of applications in a variety of domains, including bioinformatics, image and voice recognition, and text categorization, to name a few.

The basic equation for a linear Support Vector Machine (SVM) classification problem is:

$$w^T \times x + b = 0$$

Where:

- w^T is the weight vector

- x is the feature vector of a data point
- b is the bias term

The equation acts as a representation of the hyperplane that categorizes the data points into their respective groups. Finding the values of w and b that maximize the margin, which is the distance between the hyperplane and the data points from each class that are closest to it and are referred to as support vectors, is the purpose of this exercise.

3.3.7 Gaussian Naive Bayes

Classifying data points is the job of the probabilistic algorithm known as Gaussian Naive Bayes, which applies Bayes' theorem. It is a straightforward technique that may be used for the solution of issues involving binary as well as multi-class classifications. Given the class label, the fundamental assumption of Naive Bayes is that each of the characteristics may be considered independent of the others. Although the statistics from the actual world often contradict this assumption, which is why it is referred to be "naive," the assumption may nonetheless be useful in a number of circumstances.

In order to determine the probability of a feature given a class label, the technique makes use of the probability density function that is associated with the Gaussian distribution. A data point's projected label is determined by selecting the class label that has the best possibility of being correct.

The Gaussian Naive Bayes algorithm can be represented mathematically as follows:

$$P(y|x) = P(x|y) \times P(y) / P(x)$$

Where:

- $P(y|x)$ is the probability of the class label y given the feature vector x
- $P(x|y)$ is the likelihood of the feature vector x given the class label y
- $P(y)$ is the prior probability of the class label y
- $P(x)$ is the prior probability of the feature vector x

When the characteristics being analyzed are continuous and the data follows a normal distribution, the Gaussian Naive Bayes algorithm is at its most effective. It is also a smart option when there are few data points available but a huge number of attributes to choose from.

3.3.8 Bernoulli Naive Bayes

Classifying data points is the job of the probabilistic algorithm known as Bernoulli Naive Bayes, which applies Bayes' theorem. It is a modification of the Naive Bayes method that was developed with binary characteristics in mind from the very beginning.

Bernoulli Naive Bayes represents binary data with a Bernoulli distribution, in contrast to the Gaussian Naive Bayes model, which models continuous features with a Gaussian distribution. This signifies that it is assumed that each feature is binary (that is, true or false, or 0/1), and the likelihood of the feature being true is modeled individually for each class label.

The Bernoulli Naive Bayes algorithm is identical to the Gaussian Naive Bayes algorithm in terms of its fundamental equation, which is as follows:

$$P(y|x) = P(x|y) \times P(y) / P(x)$$

Where:

- $P(y|x)$ is the probability of the class label y given the feature vector x
- $P(x|y)$ is the likelihood of the feature vector x given the class label y
- $P(y)$ is the prior probability of the class label y
- $P(x)$ is the prior probability of the feature vector x

However, the likelihood of the feature vector x given the class label y is now modeled with a Bernoulli distribution:

$$P(x|y) = (p_y)^x \times (1 - p_y)^{(1-x)}$$

Where:

- p_y is the probability of the feature being true for the class label y

- x is the feature value (0 or 1)

Following the estimation of the likelihood and prior probability of each class label, the likelihood and prior probability of the class label with the greatest probability is selected as the label that will be predicted.

Bernoulli Naive Bayes is especially helpful in situations in which the characteristics being considered are binary and the data being considered is sparse. It is especially helpful in situations in which there is a high number of characteristics and the data is binary.

3.4 Hyper-Parameter Optimization

When training an algorithm for machine learning, the parameters known as hyper-parameters are ones that cannot be changed. They can be used to determine the structure of the model, which includes the number of hidden layers and the activation function, or they can be used to evaluate the efficiency and precision of model training, which includes the learning rate (LR) of stochastic gradient descent (SGD), batch size, and optimizer parameters. Both of these applications are possible with the help of these variables (hyp). One way to think about the HPO is as the execution phase of the model.

By adopting the HPO technique, the hyper-parameters of a machine learning model are automatically improved. This removes the need for humans to participate in the feedback loop of the machine learning system. HPO calls for a significant amount of computer resources as a substitute for human work, especially in situations in which numerous hyper-parameters are optimized at the same time. Because of the difficulties associated with making the most efficient use of computing resources and planning out search areas, a great deal of research has been done on HPO, specifically on algorithmic frameworks and toolkits.

HPOs have the following goals: to improve the accuracy and efficacy of neural network training; to lower the barrier for research and development; to reduce the cost of menial tasks performed by artificial intelligence (AI) specialists; and to improve the plausibility

of the selection of the hyper-parameter set and training outcomes. In this study, we use the following two types of hyper-parameter optimization.[32]

3.4.1 Grid Search CV

The HPO approach that is most essential is the grid search. An exhaustive search is carried out on the set of hyperparameters that has been supplied by the user. Users are responsible for the generation of all candidates; thus, they should have a fundamental understanding of these hyper-parameters. Grid search is useful when there are several hyper-parameters but only a limited amount of space to search.

The most straightforward search strategy, known as grid search CV, produces the most accurate predictions. If the user is provided with sufficient resources, they will always be able to choose the optimal combination. It is simple to do grid search in parallel due to the fact that every trial may be conducted individually and without respect to the chronological sequence of events. There is no connection between the results of one research and those of any other trials that have been conducted. It is feasible to devise allocation strategies for computing resources that are very flexible. Because an increase in the number of hyper-parameters that need to be modified results in an exponential rise in the amount of computing resources used, grid search is doomed to fail due to the curse of dimensionality. [33]

3.4.2 Randomized Search CV

Randomized search CV is a significant improvement that stands in stark contrast to grid search CV. A random study of hyper-parameters selected from certain distributions of probable parameter values is what this term refers to. The search process will continue until it either achieves the level of accuracy that was considered to be acceptable or until the allotted money has been depleted. Grid search is analogous to random search; however, random search has been shown to provide superior results due to the following two benefits:

- In contrast to a grid search, in which the budget for each set of hyper-parameters is a fixed amount equal to B/N , where B is the overall budget and N is the number of hyper-

parameters, independent budget assignments are able to be formed based on the distribution of the search space. As a result of the unequal distribution of some hyper-parameters, random search may prove to be more effective.

- Even if using a random search to find the ideal is not likely to be successful, it is a given that spending more time will enhance the odds of finding the optimal collection of hyper-parameters. This is because more time equals more chances. This kind of thinking is referred to as Monte Carlo approaches, and it is used in circumstances involving multidimensional deep learning that involve dealing with huge quantities of information.

Even while grid search often has a great deal more success than random search, it nevertheless calls for a significant amount of processing. It is recommended to use random search in the beginning stages of HPO so that the search space may be swiftly narrowed down. After then, it is recommended to use a guided algorithm in order to get a more accurate result. [33]

3.5 Feature Selection

The process of picking a subset of relevant characteristics to be used in the creation of a model is referred to as feature selection. It is a method that may enhance the performance of machine learning models by lowering the dimensionality of the data, getting rid of features that are useless, and simplifying the structure of the model. It is possible to accomplish this goal via the use of a variety of ways, including filter methods, wrapper methods, and embedded methods. The objective is to choose a subset of features that effectively captures the underlying issue, while at the same time minimizing overfitting and enhancing the model's capacity to be interpreted by humans.

An indicator of how much a particular feature contributes to the overall performance of a machine learning model is referred to as its "feature significance." It enables us to identify which features are most relevant for a certain activity, and it may be used for feature selection, feature engineering, and the interpretability of models. The relevance of the features in the dataset is seen in Figure 3.5.1.

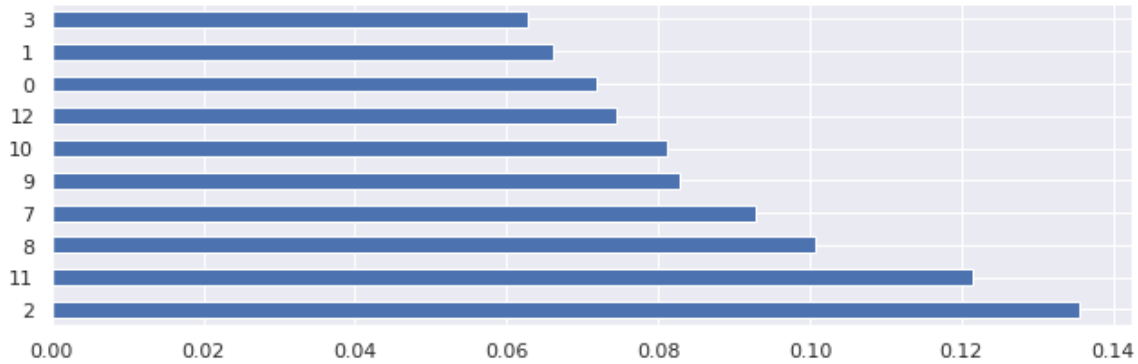


Figure 3.5.1. Feature importance

There are several different types of feature selection methods, among them, most commonly used three techniques were used in this study.

3.5.1 Univariate Selection

A sort of filter mechanism known as univariate selection may be used for feature selection. It does this by examining the connection that exists between each characteristic and the target variable in order to arrive at an overall relevance score for each feature. The objective here is to identify the characteristics that have the most significant correlation with the variable of interest.

The p-value is a typical statistical measure that is used for univariate selection. Its purpose is to evaluate the likelihood that the association between a characteristic and the target variable is the result of random chance. Features that have low p-values (usually less than 0.05) are regarded as statistically significant and are kept, whilst those that have high p-values are taken out of consideration.

The SelectKBest method in scikit-learn is an example of univariate selection. This method enables users to choose a predetermined number of top-performing features based on the results of a variety of statistical tests, such as the chi-squared test or the f classif test (which determines the ANOVA F-value between a label and a feature for classification problems).

It is crucial to note that while univariate selection is simple and fast to perform, it does not take into account the connection between characteristics, which, depending on the activity, might be quite significant. As a consequence, it's important to experiment with a variety of approaches and evaluate the outcomes of each one.

3.5.2 Model-Based Feature Selection

A specific kind of wrapper approach for feature selection is referred to as model-based feature selection. An method for machine learning is used to assess the feature subset. This approach works by first training a model with many distinct feature subsets, and then choosing the feature subset that yields the highest level of performance. This method takes into account the connections between the features, as well as the connections between the features themselves, as well as the connections between the features and the target variable.

Recursive Feature Elimination (RFE), which begins with all features and iteratively eliminates the feature that contributes the least to the model's performance, is an example of model-based feature selection. This process continues until a certain number of features are left. Another example of this would be the `SelectFromModel` class found in `scikit-learn`. This method requires an estimate as an argument and then picks the features from the input data whose coefficients or significance are non-zero or greater than a certain threshold.

Because it involves training a model for each subset of data, model-based feature selection demands much more computing resources than univariate feature selection does. This is an important point to keep in mind. On the other hand, it is seen as being more strong since it takes into consideration the connection between traits, which may be essential for certain activities. As a consequence, it's important to experiment with a variety of approaches and evaluate the outcomes of each one.

3.5.3 Recursive Feature Elimination

One form of wrapper approach for feature selection is known as recursive feature elimination, or RFE for short. It is an iterative procedure that begins with all of the

features and eliminates the feature that contributes the least to the model's performance until a certain number of features are left. This process continues until a certain number of features are left. The procedure starts with training a model with all of the characteristics, after which the features are ranked according to the relevance of each one. After that, the feature that contributes the least will be eliminated, and the procedure will be repeated with the other features until the required number of features has been obtained. When ranking the features, one of the most frequent methods to do so is based on the estimator's determination of the coefficients or relevance of each feature.

RFE is often used in combination with other algorithms, such as support vector machines (SVMs) or logistic regression. However, it is compatible with any supervised learning technique that has a feature significance characteristic. RFE is especially helpful when there are a high number of features since it helps to reduce characteristics that are redundant or unnecessary, which in turn may enhance the model's performance and make it easier to comprehend. However, it may be quite computationally costly, particularly when the number of features is very high. Additionally, it is possible that it is not always the best choice; thus, it is always a good idea to test out a variety of approaches and compare the outcomes.

CHAPTER 4

RESULTS AND DISCUSSION

The findings of lengthy simulations, which were done after an analysis of many different machine learning approaches, were then utilized to make forecasts about whether or not people will suffer from cardiovascular conditions. Other performance indicators, including as accuracy, specificity, F1-score, precision, and sensitivity, as well as the area under the receiver operating characteristic curve (ROC-AUC), are calculated and analyzed. ROC-AUC is an abbreviation for receiver operating characteristic area under the curve. The associated mathematical formulae that are presented below make use of the abbreviations TP, TN, FP, and FN, which stand, respectively, for True Positive, True Negative, First Probability, and First Negative, respectively.

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}}$$

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

$$\text{Specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}}$$

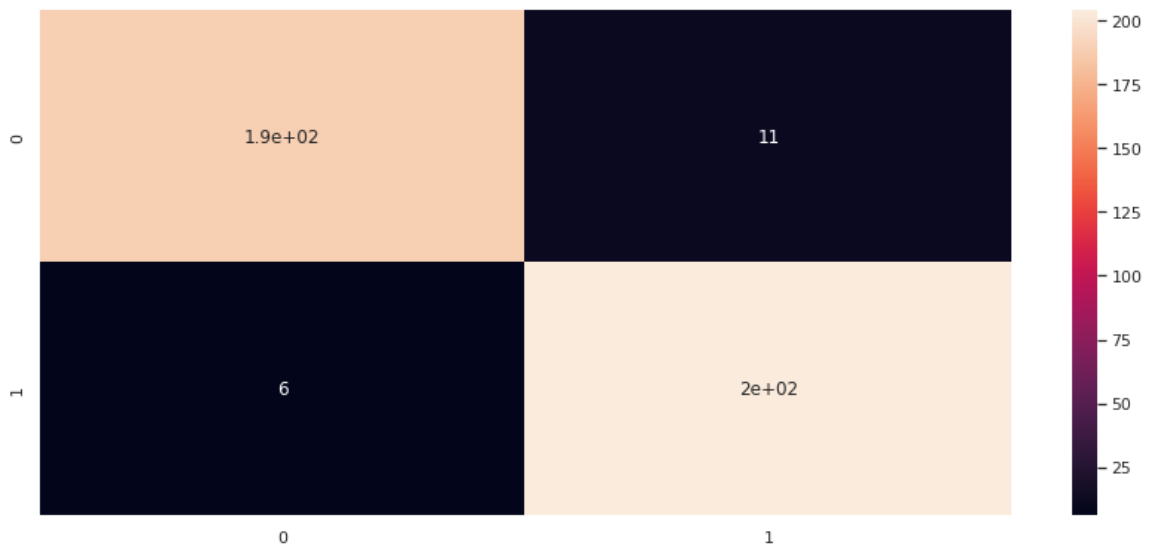
In order to make an accurate forecast, the machine learning model has to go through the process of hyperparameter optimization, also known as HPO. During the course of this inquiry, the Randomized search CV and Grid Search CV techniques for hyperparameter optimization were used to fine-tune our data. This investigation was carried out using a computer that had a CPU from Intel's 9th generation Core i5 family and a memory capacity of 16 gigabytes.

Following the setting of the hyperparameters by the use of the approach that we proposed, the machine learning models were trained to keep the bias as low as feasible in order to prevent overfitting.

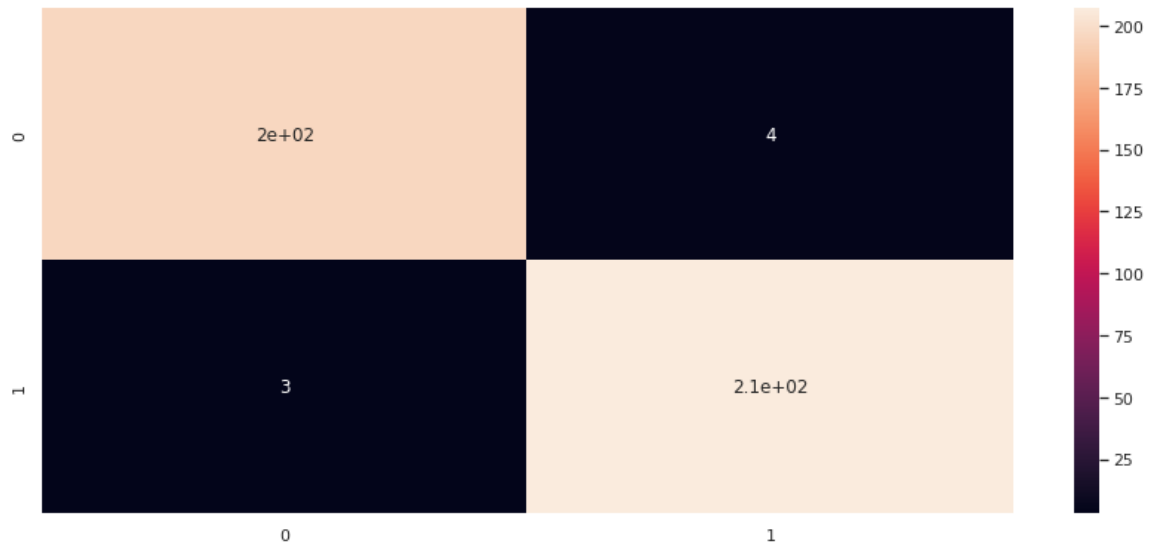
After then, cross-validation was used to conduct the evaluations, with the goals of removing any potential of data loss and ensuring that there is as little fluctuation as is practically possible.

Confusion matrices for each classifier are presented for examination in Figure 4.1. On the other hand, the findings that are shown in Table 4.1 for each of the ML algorithms' performance metrics for default hyperparameter (DHP) change. In addition, the best performing algorithms for each of these three hyperparameter tweaking strategies are shown in Table 4.1.

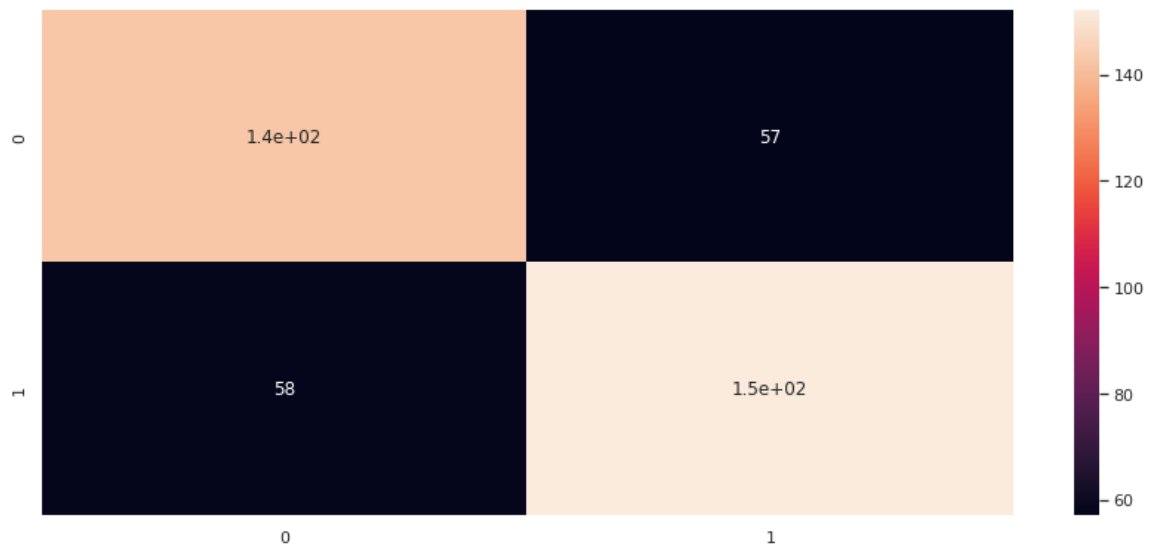
a)



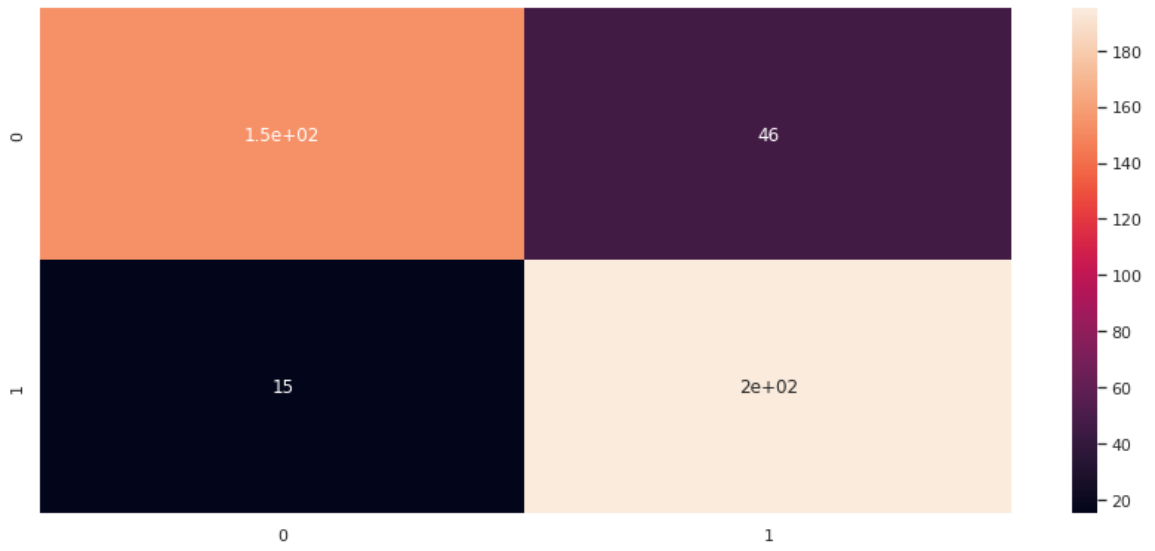
b)



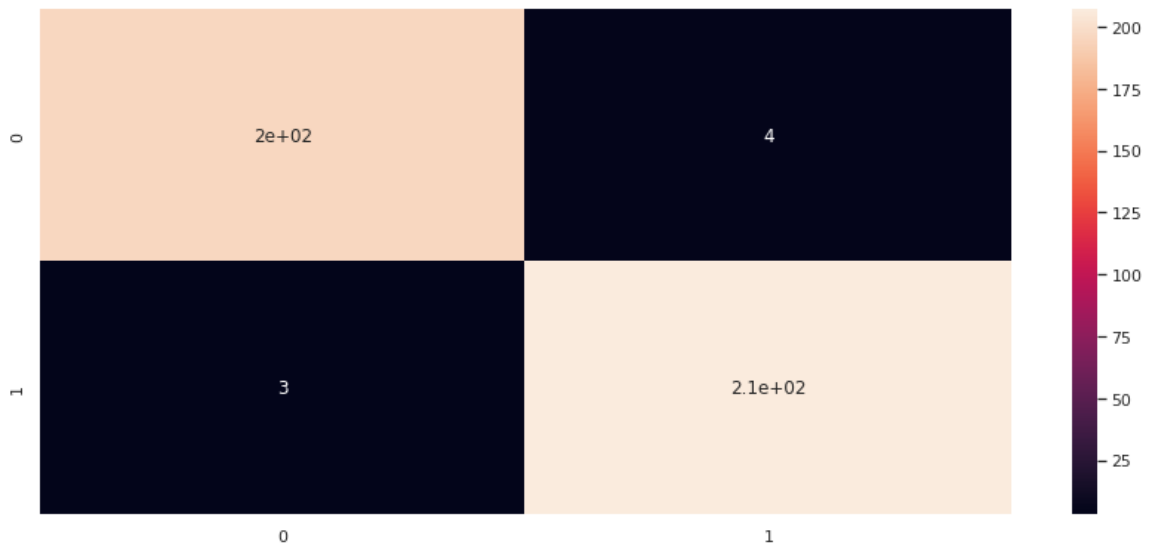
c)



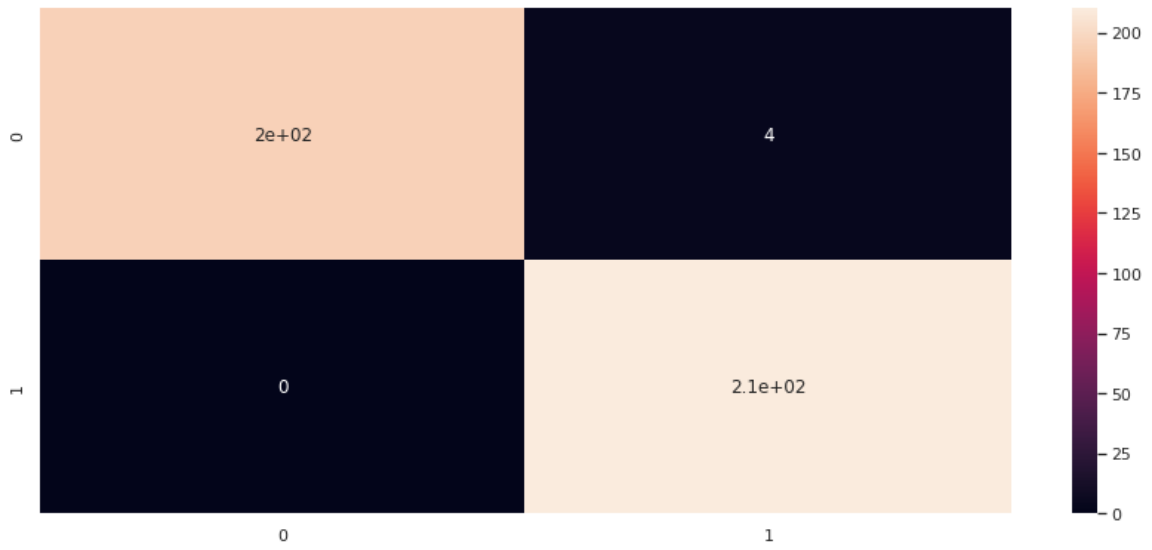
d)



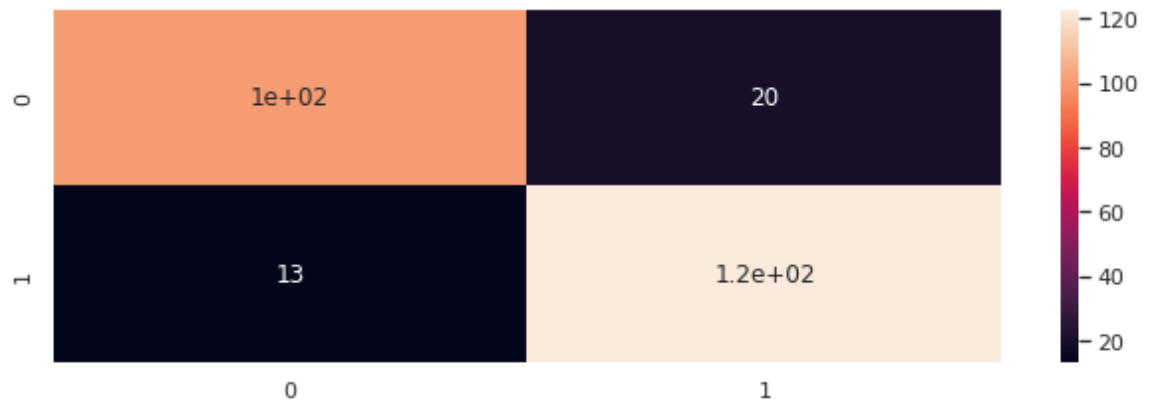
e)



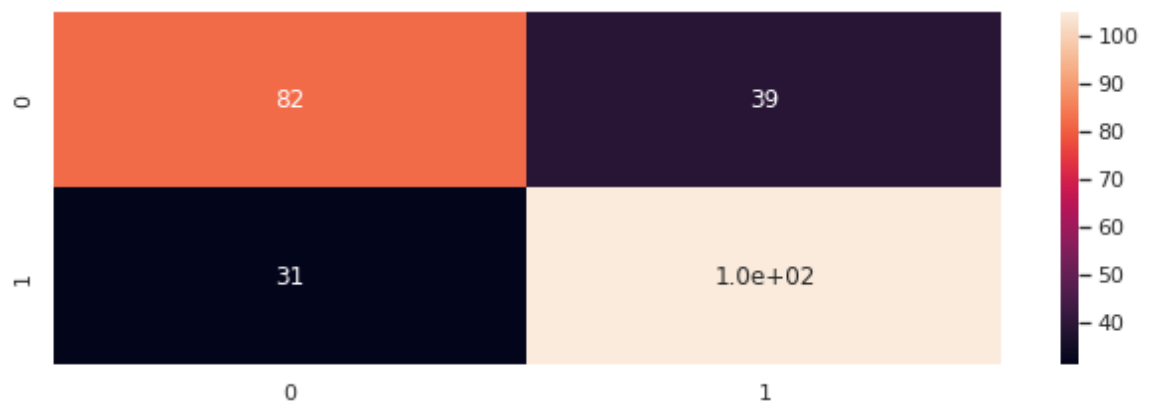
f)



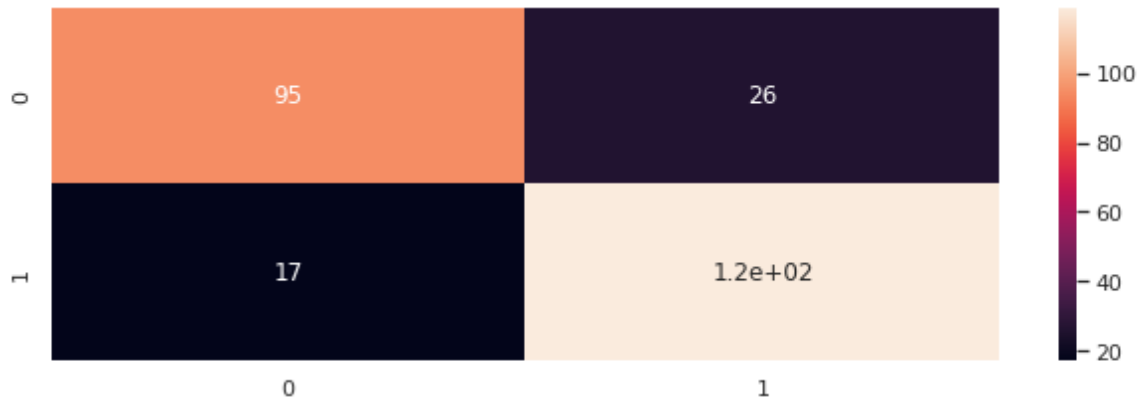
g)



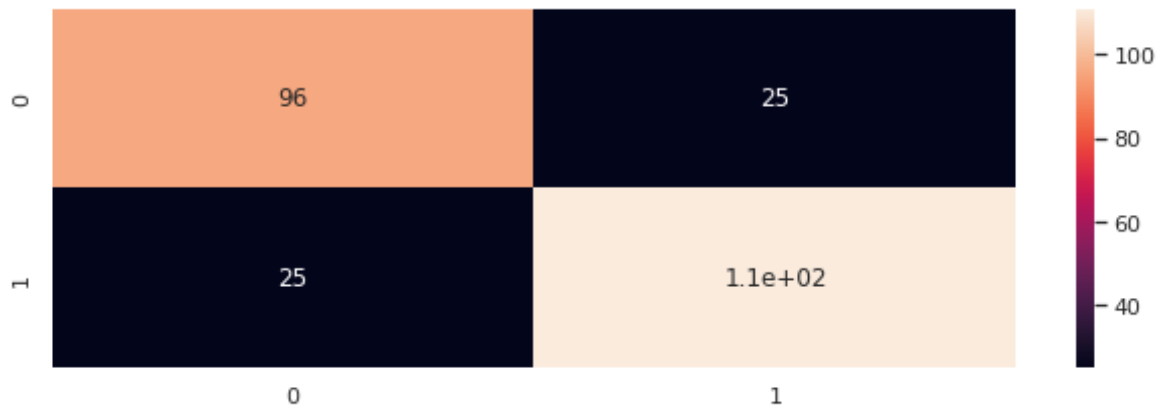
h)



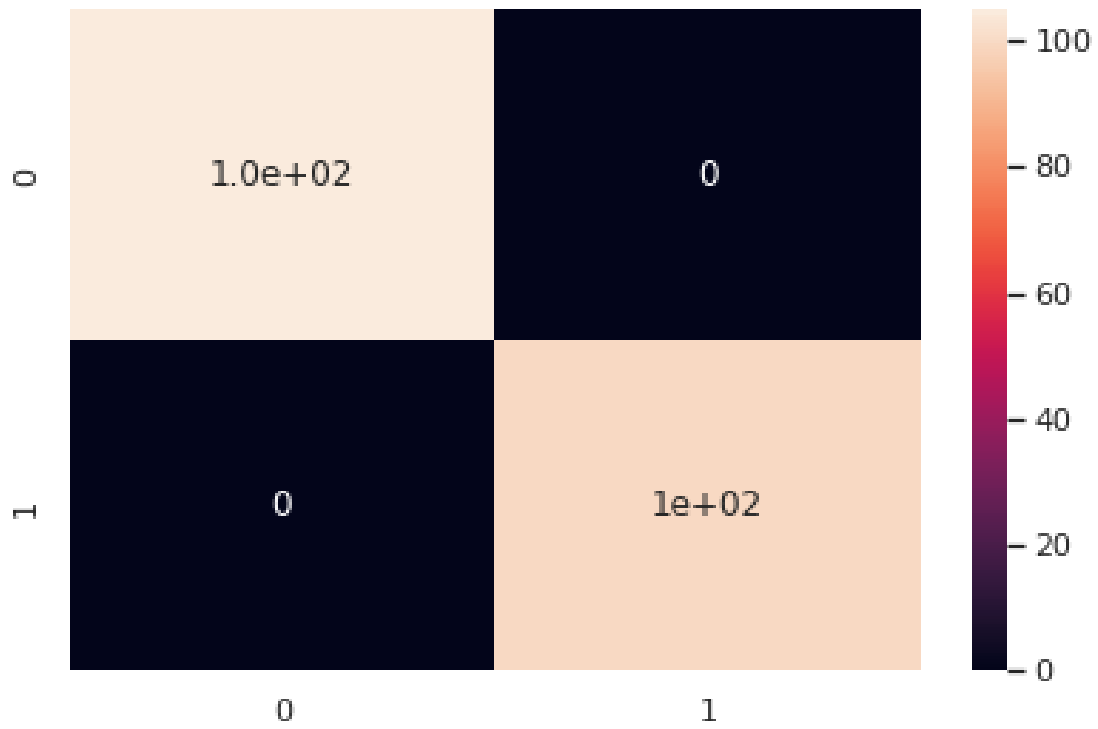
i)



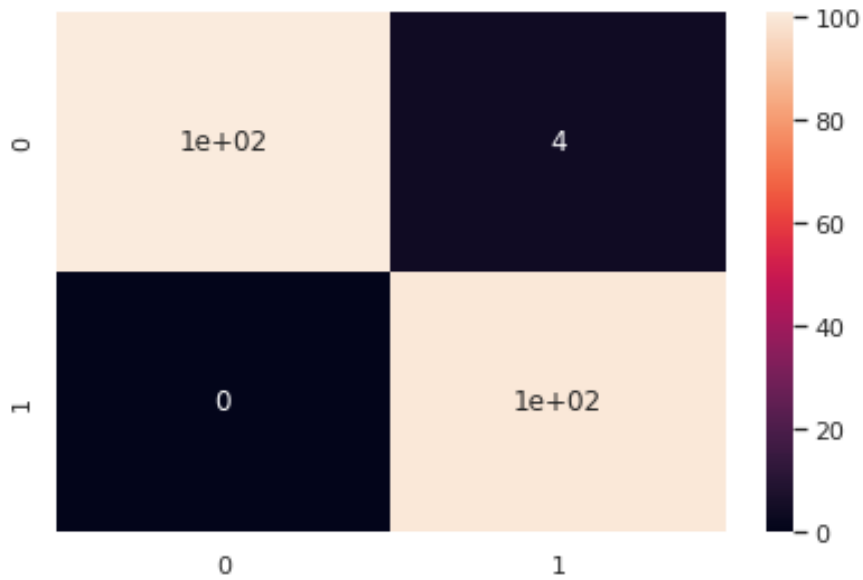
j)



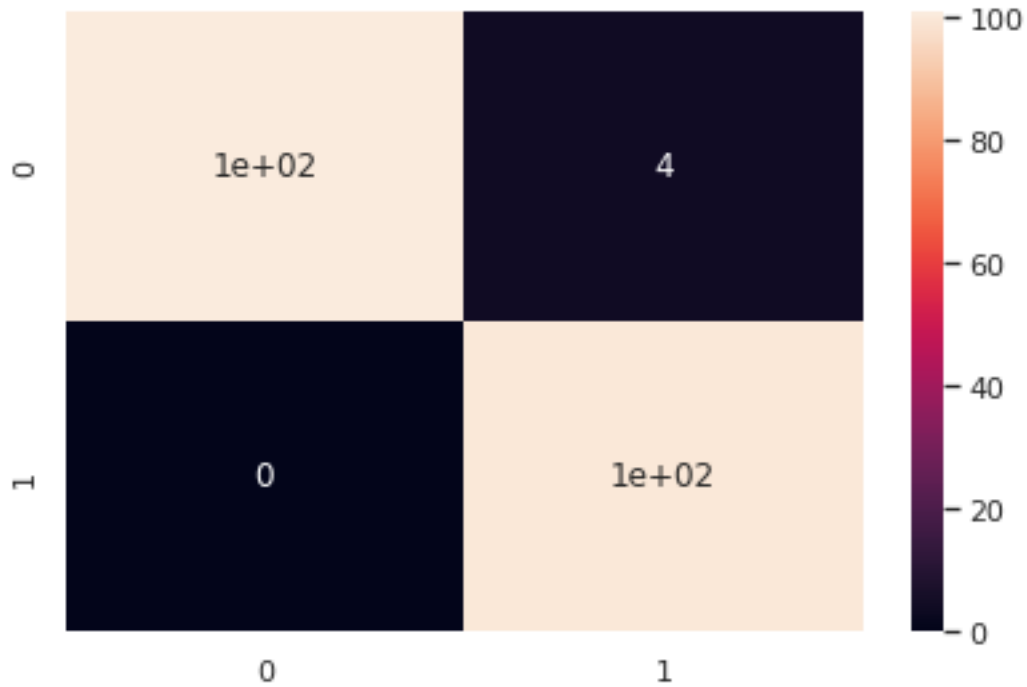
k)



l)



m)



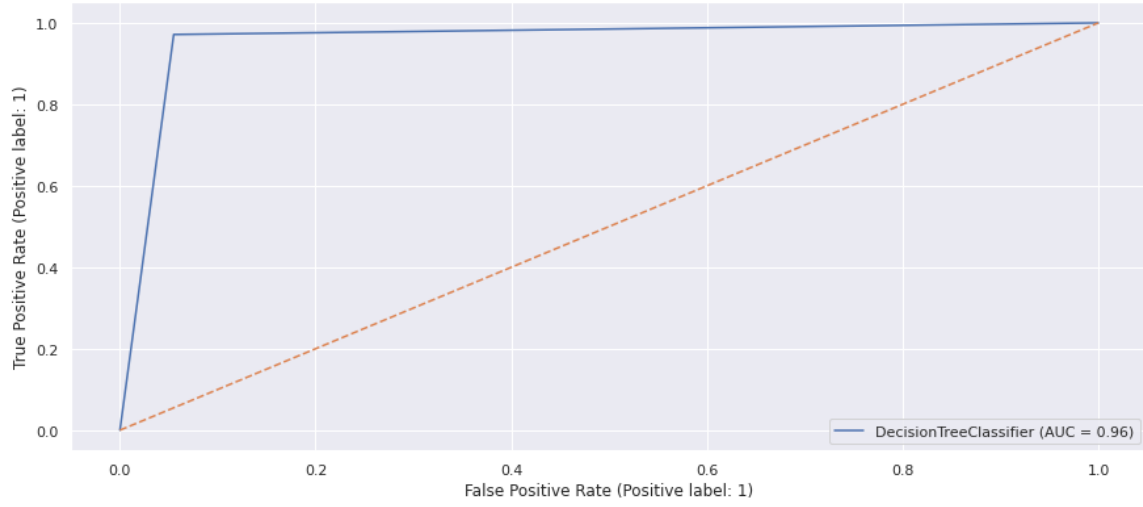
Confusion matrices for each of the classifiers are shown in Figure 4.1. a. Decision Tree, b. Random Forest, c. K-Nearest Neighbor, and d. Logistic Regression. c. K-Nearest Neighbor. e. Search CV Using Grid, f. Search CV Using Randomized g. XG booster, h. Support Vector Machine, i. Gaussian Naive Bayes, j. Bernoulli Naive Bayes, k. Univariate Selection, l. Model-Based Feature Selection, and m. Recursive Feature Elimination.

TABLE 4.1. Different ML algorithms' performance metrics using the default hyperparameter

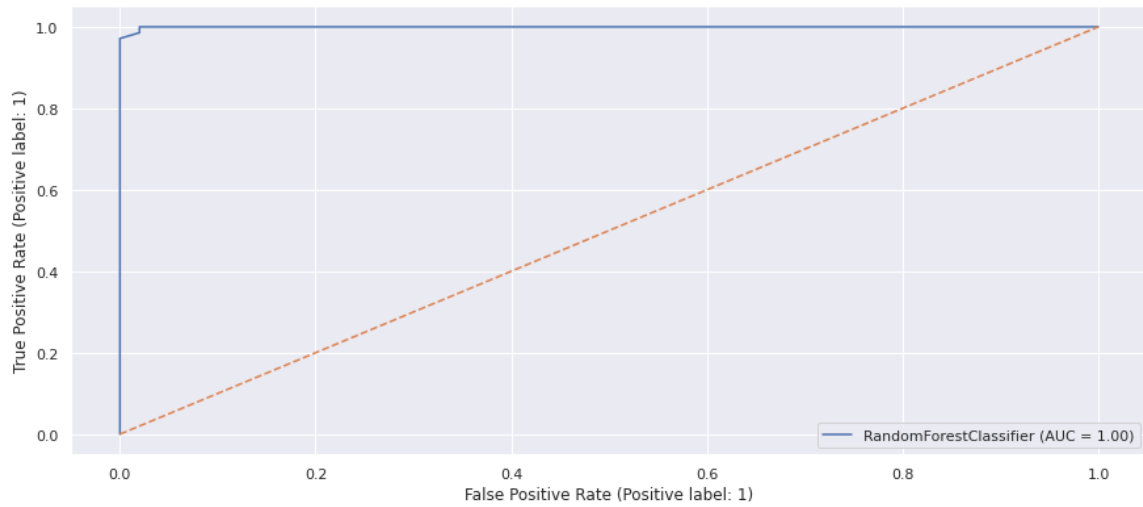
Name of the algorithm	Accuracy (%)	Precision	Sensitivity	Specificity	F1 Score	ROC-AUC
DT	97.07	0.962	0.980	0.960	0.971	0.97
RF	98.70	0.974	100.0	0.974	0.982	100.0
RCV	99.02	100.0	0.990	100.0	0.990	100.0
GC	99.02	100.0	0.990	100.0	0.990	100.0
KNN	68.5	0.677	0.699	0.670	0.688	0.84
LR	83.44	0.740	0.921	0.773	0.821	0.93
XGB	88.3	0.841	0.913	0.858	0.876	0.95
SVM	70.05	0.595	0.721	0.666	0.652	0.75
GNB	78.59	0.714	0.825	0.756	0.765	0.85
BNB	82.10	0.793	0.833	0.810	0.813	0.90
US	98.7	0.974	100.0	0.974	0.987	100.0
MBFS	98.7	0.974	100.0	0.974	0.987	100.0
RFE	98.7	0.974	100.0	0.974	0.987	0.99

As can be seen in Figure 4.2, a receiver operating characteristic curve (ROC) was generated for each of the models that were used in this study in order to do more research and analysis on the models that were developed. Figure 4.2 displays the results of calculating the area under each curve. The receiver operating characteristic (ROC) curve illustrates the classifier's ability to perform diagnostic analysis. The area value of the ROC curve is a measure of how well a model can detect a problem. The better off you are, the closer you are to one.

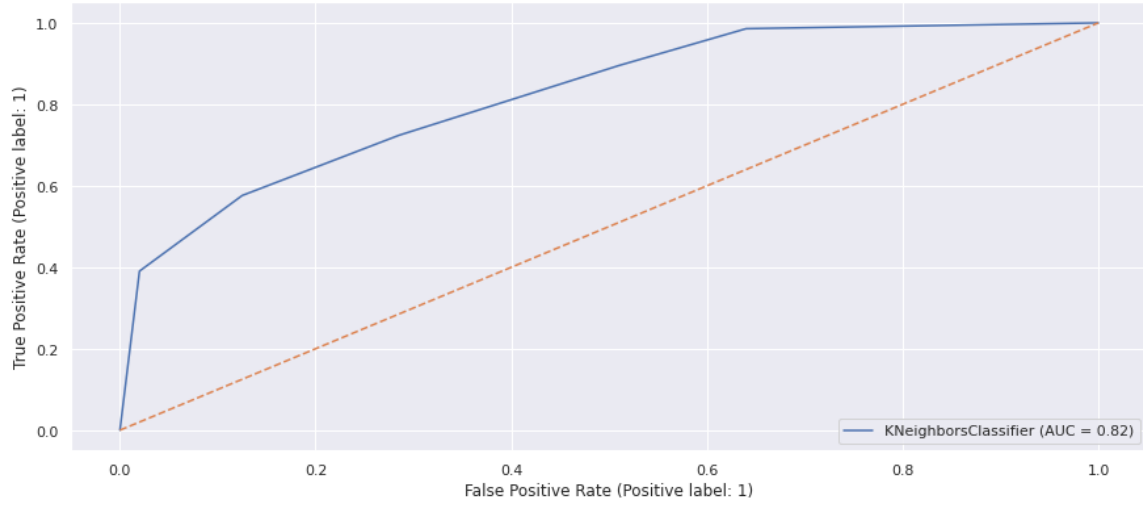
a)



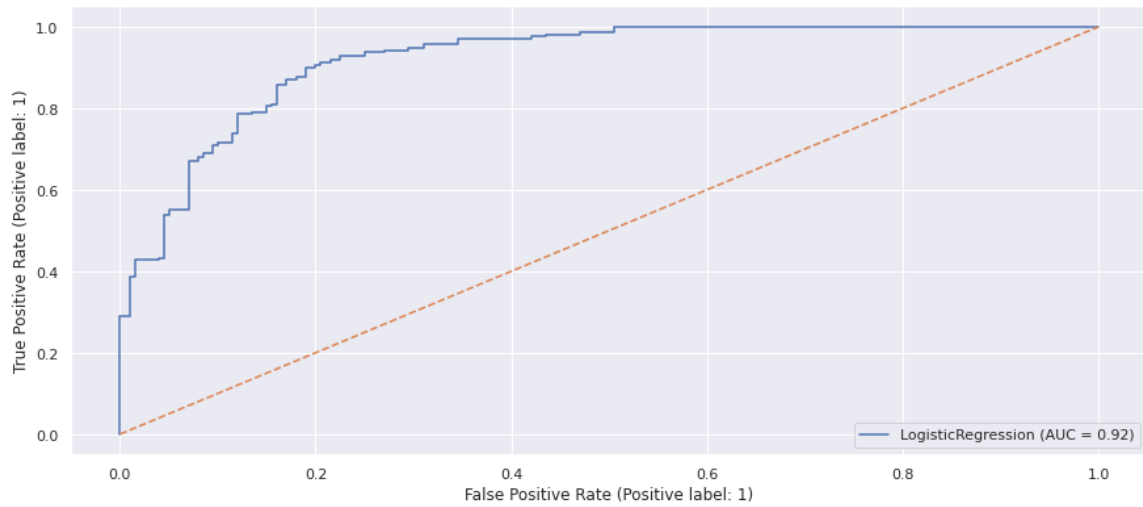
b)



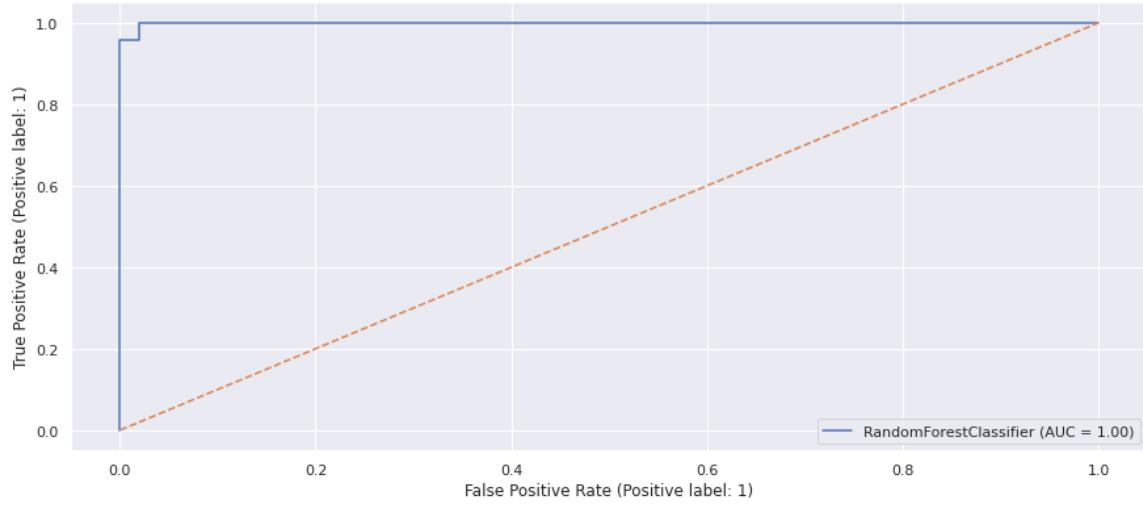
c)



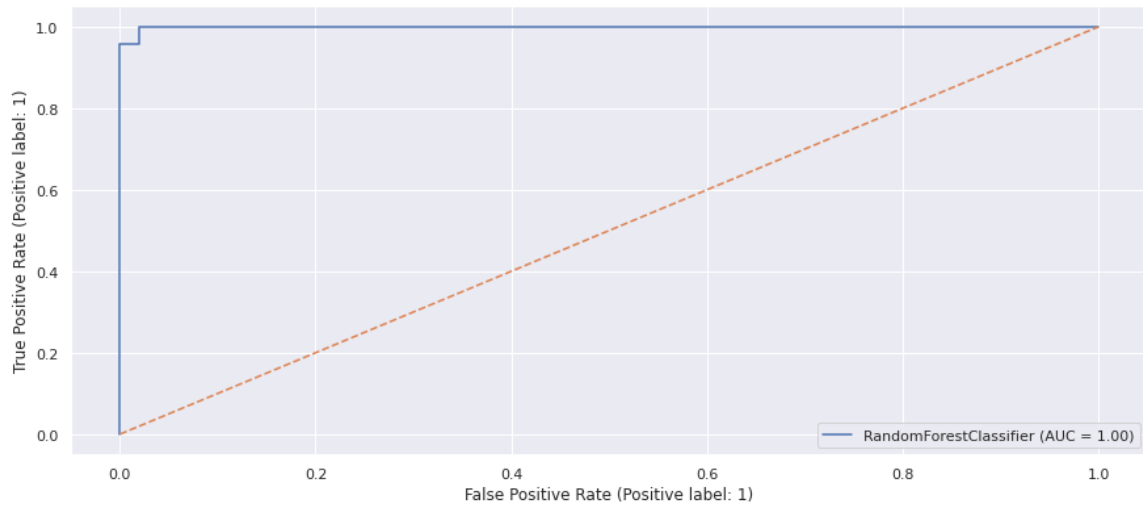
d)



e)



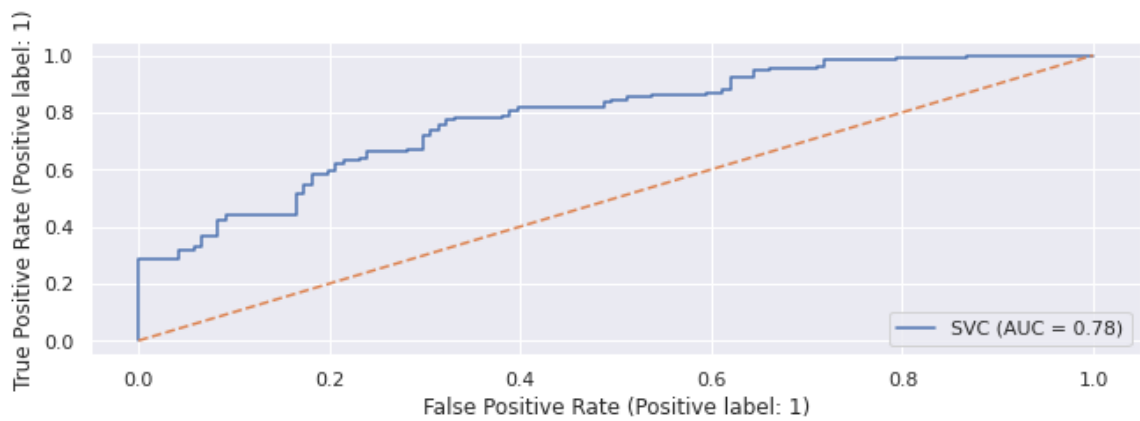
f)



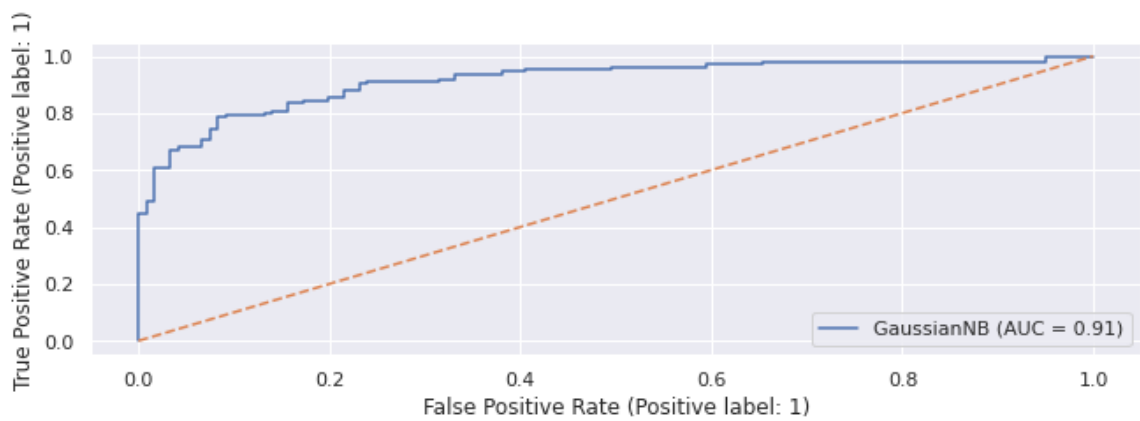
g)



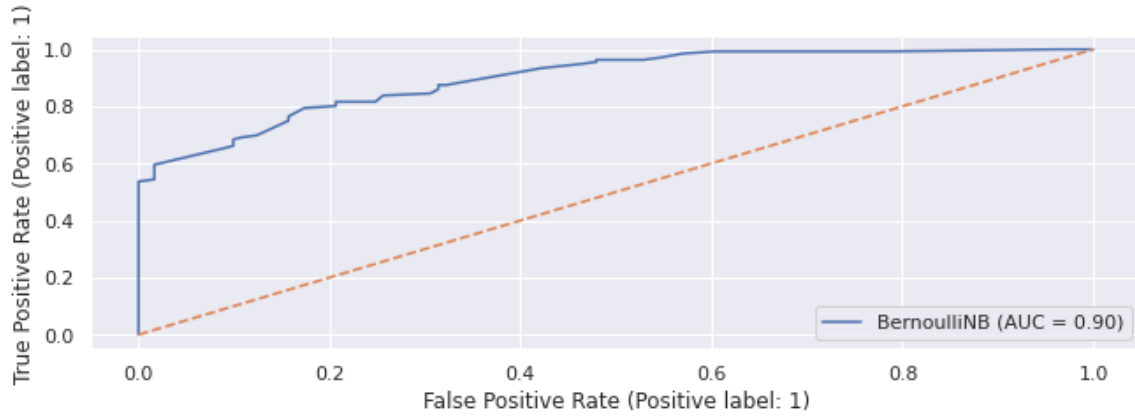
h)



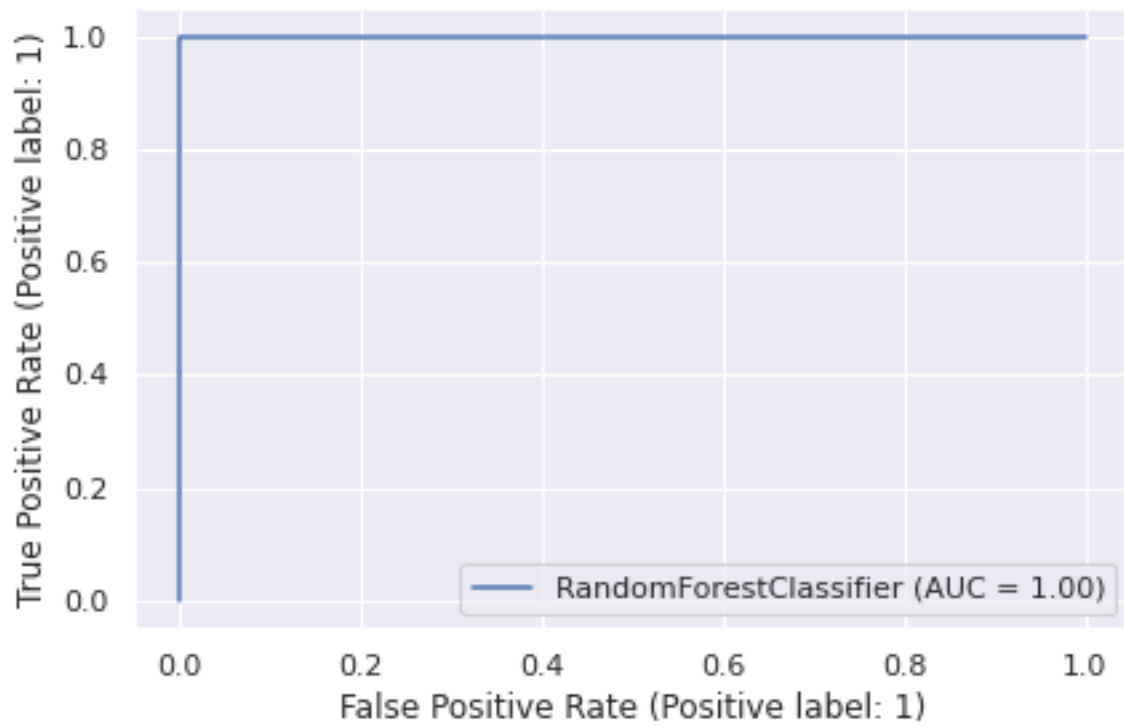
i)



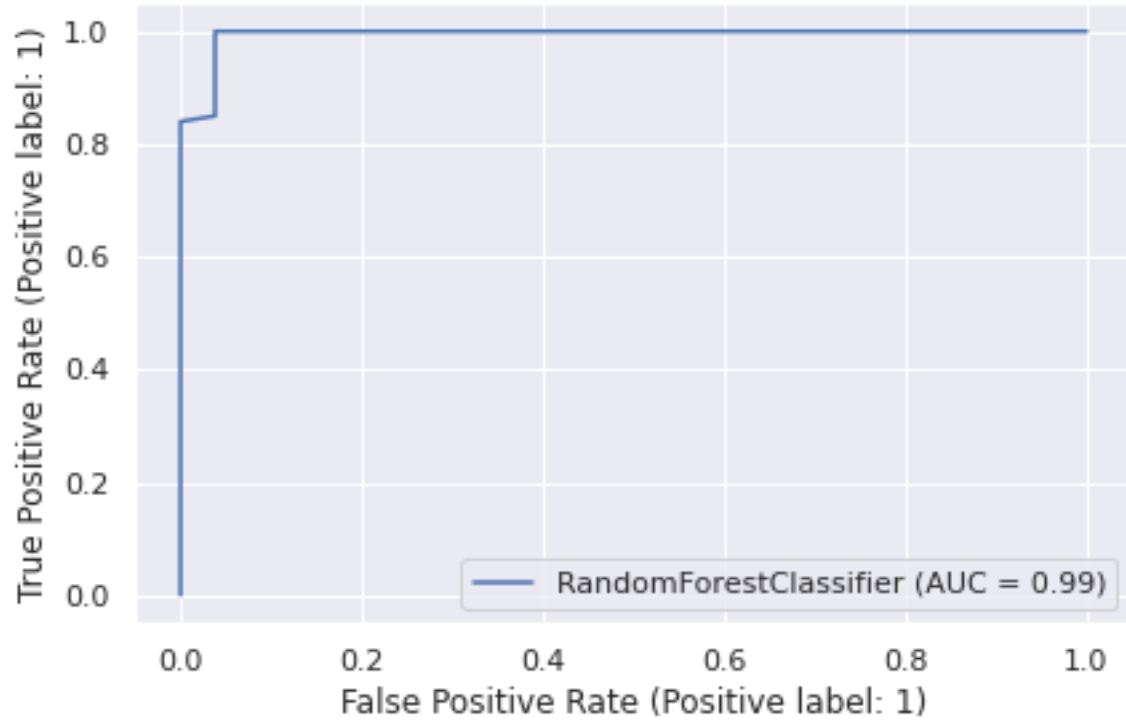
j)



k)



1)



m)

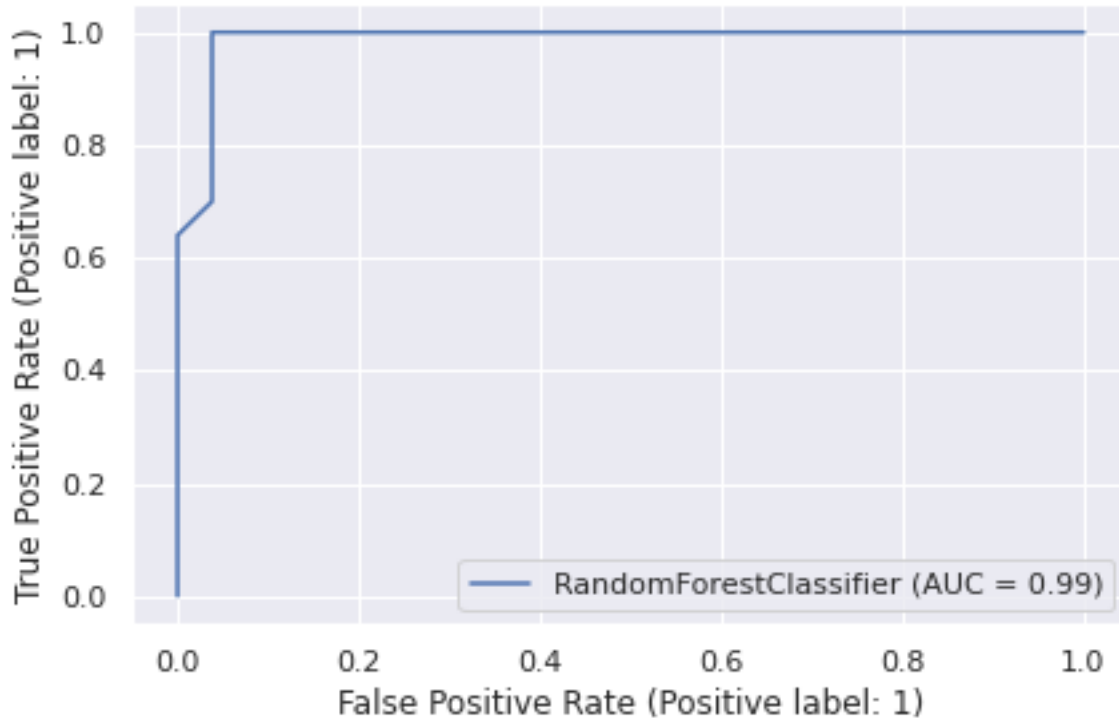


Figure 4.2. ROC curve for all models. a. Decision Tree, b. Random Forest, c. K-Nearest Neighbor, and d. Logistic Regression. c. K-Nearest Neighbor. e. CV with a Grid Search and f. CV with a Randomized Search g. XG booster, h. Support Vector Machine, i. Gaussian Naive Bayes, j. Bernoulli Naive Bayes, k. Univariate Selection, l. Model-Based Feature Selection, and m. Recursive Feature Elimination.

The final outcome of the hard vote ensemble technique demonstrated that the overall accuracy of this study was 98.70% obtained from Random forest Classification model , ensuring the predictive usefulness of bioinformatics to assist medical personnel in the rapid detection of cardiovascular illness.

CHAPTER 5

LIMITATION

Even when it is driven by facts, machine learning may often have the same biases as traditional methods. It is possible for these to become troublesome in high-stakes healthcare settings, despite the fact that they are often not problematic in corporate environments, which are typically concerned with accuracy. One of these pitfalls is known as selection bias, and it includes both sampling bias and observer selection bias. Even though there is often a reduction in selection bias in clinical trials that have a defined framework, it might be difficult to identify and resolve selection bias in datasets that are utilized in machine learning. Imagine a model of machine learning that was constructed making use of an electronic health record from a hospital. It is possible that this model will underestimate the frequency of a certain illness due to the fact that the patients from whom it learns typically undergo proper regular screening. The use of internet of things (IOT) technology may significantly mitigate the effects of these problems.

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Most people are quite concerned about developing heart disease since the heart is an essential organ. Machine learning algorithms trained on information about cardiovascular disease might be very useful in preventing this tragedy and so saving countless lives. Heart disease research will benefit from early discovery of any abnormalities, and so will healthcare. In this research, we employed a variety of machine learning techniques to make predictions on the development of heart disease. In this paper, we give a comprehensive comparative analysis, the findings of which reveal that the maximum accuracy (99%) was achieved by using hard and soft voting Randomized Search CV. In light of this, these methods could aid in the battle against heart disease by facilitating a more precise diagnosis and facilitating the provision of appropriate treatment.

6.2 Future Scope

Modeling in the computer, the creation of artificial data and patients, as well as mobile health solutions, are only two examples of the new fields that are made possible by AI and IoT. Human body dynamics are studied and simulated using computers in medical computational modeling. By combining several diagnostic data obtained from clinical modalities, it provides a platform for the virtual examination and therapeutic optimization, all while allowing for the building of a customized heart. Machine learning and computer modeling share the goal of predicting unknown consequences from available data, however machine learning is more likely to be data-driven whereas computer modeling is often deterministic. Recent research has shown the effectiveness of computational modeling using machine learning methodologies, namely in the areas of fluid dynamic simulations³⁶ and adverse medication responses. There will be no privacy or expense concerns when employing synthetic data, making it ideal for large-scale clinical research; it may also serve as a substitute for creating training data for computers.

Reference:

- [1] W.L. Duvall, Cardiovascular disease in women, *The Mount Sinai Journal of Medicine*, New York, 70 (2003) 293-305.
- [2] A.D. Callow, Cardiovascular disease 2005—the global picture, *Vascular pharmacology*, 45 (2006) 302-307.
- [3] R. Kones, Primary prevention of coronary heart disease: integration of new data, evolving views, revised goals, and role of rosuvastatin in management. A comprehensive survey, *Drug design, development and therapy*, 5 (2011) 325.
- [4] D. Shah, S. Patel, S.K. Bharti, Heart disease prediction using machine learning techniques, *SN Computer Science*, 1 (2020) 1-6.
- [5] R. Das, "A Comparison of Multiple Classification Methods for Diagnosis of Parkinson Disease." *Expert Systems with Applications*, 37 (2010) 1568–1572,
- [6] M. Hassoon, M.S. Kouhi, M. Zomorodi-Moghadam, M. Abdar, Rule optimization of boosted c5. 0 classification using genetic algorithm for liver disease prediction, 2017 international conference on computer and applications (icca), IEEE, 2017, pp. 299-305.
- [7] M. Abdar, N.Y. Yen, J.C.-S. Hung, Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees, *Journal of Medical and Biological Engineering*, 38 (2018) 953-965.
- [8] N. Shukla, M. Hagenbuchner, K.T. Win, J. Yang, Breast cancer data analysis for survivability studies and prediction, *Computer methods and programs in biomedicine*, 155 (2018) 199-208.
- [9] H. Yang, Y.-P.P. Chen, Data mining in lung cancer pathologic staging diagnosis: correlation between clinical and pathology information, *Expert systems with applications*, 42 (2015) 6168-6176.
- [10] M. Abdar, W. Książek, U.R. Acharya, R.-S. Tan, V. Makarenkov, P. Pławiak, A new machine learning technique for an accurate diagnosis of coronary artery disease, *Computer methods and programs in biomedicine*, 179 (2019) 104992.

- [11] T. Shu, B. Zhang, Y.Y. Tang, Effective heart disease detection based on quantitative computerized traditional chinese medicine using representation based classifiers, *Evidence-Based Complementary and Alternative Medicine*, 2017 (2017).
- [12] P. Pławiak, Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system, *Expert Systems with Applications*, 92 (2018) 334-349.
- [13] R. Alizadehsani, J. Habibi, Z.A. Sani, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, Diagnosis of coronary artery disease using data mining based on lab data and echo features, *Journal of Medical and Bioengineering*, 1 (2012).
- [14] A.H. Alkeshuosh, M.Z. Moghadam, I. Al Mansoori, M. Abdar, Using PSO algorithm for producing best rules in diagnosis of heart disease, 2017 international conference on computer and applications (ICCA), IEEE, 2017, pp. 306-311.
- [15] *Carpathian Journal of Electronic and Computer Engineering*, 8 (2015) 31. M. Abdar, "Using decision trees in data mining for identifying factors affecting of heart disease."
- [16] Comparing Data Mining Algorithms' Performance in Predicting Heart Diseases, *International Journal of Electrical & Computer Engineering* (2088-8708), 5 (M. Abdar, S.R.N. Kalhori, T. Sutikno, I.M.I. Subroto, and G. Arji) (2015).
- [17] L. Verma, S. Srivastava, P. Negi, A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data, *Journal of medical systems*, 40 (2016) 1-7.
- [18] Hinchcliff, T.M. Frech, T.A. Wood, C.-C. Huang, J. Lee, K. Aren, J.J. Ryan, B. Wilson, L. Beussink-Nelson, and M.L. Whitfield. Machine learning of the cardiac phenome and skin transcriptome to characterize heart disease in systemic sclerosis, *bioRxiv*, (2017) 213678, by M.E.
- [19] C. Zou and H. Deng, Fuzzy concept lattice for intelligent disease diagnosis, *IEEE Access*, 5 (2016) 236-242.
- [20] A. Lahsasna, R.N. Ainon, R. Zainuddin, A. Bulgiba, Design of a fuzzy-based decision support system for coronary heart disease diagnosis, *Journal of medical systems*, 36 (2012) 3293-3306.
- [21] Fuzzy soft expert system in prediction of coronary artery disease, N. Hassan, O.R. Sayed, A.M. Khalil, and M.A. Ghany, *International Journal of Fuzzy Systems*, 19, 1546–1559 (2017).

- [22] A.K. Paul, P.C. Shill, M. Rabin, R. Islam, K. Murase, Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease, *Applied Intelligence*, 48 (2018) 1739-1756.
- [23] S. Pouriye, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, J. Gutierrez, A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease, *2017 IEEE symposium on computers and communications (ISCC)*, IEEE, 2017, pp. 204-207.
- [24] D. Han, J.H. Lee, A. Rizvi, H. Gransar, L. Baskaran, J. Schulman-Marcus, F.Y. Lin, J.K. Min, Incremental role of resting myocardial computed tomography perfusion for predicting physiologically significant coronary artery disease: a machine learning approach, *Journal of Nuclear Cardiology*, 25 (2018) 223-233.
- [25] J.H. Tan, Y. Hagiwara, W. Pang, I. Lim, S.L. Oh, M. Adam, R. San Tan, M. Chen, U.R. Acharya, Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals, *Computers in biology and medicine*, 94 (2018) 19-26.
- [26] A unique ensemble approach for biomedical classification based on ant colony optimization was developed by L. Shi, L. Xi, X. Ma, M. Weng, and X. Hu.
- [27] H. Huang, J. Liu, Q. Zhu, R. Wang, G. Hu, A new hierarchical method for inter-patient heartbeat classification using random projections and RR intervals, *Biomedical engineering online*, 13 (2014) 1-26.
- [28] J. Patel, D. TejalUpadhyay, S. Patel, Heart disease prediction using machine learning and data mining technique, *Heart Disease*, 7 (2015) 129-137.
- [29] S.F. Weng, J. Reys, J. Kai, J.M. Garibaldi, N. Qureshi, Can machine-learning improve cardiovascular risk prediction using routine clinical data?, *PloS one*, 12 (2017) e0174944.
- [30] S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan, T. Zhu, Cardiovascular risk prediction method based on CFS subset evaluation and random forest classification framework, *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, IEEE, 2017, pp. 228-232.

- [31] E.Y. Boateng, D.A. Abaye, A review of the logistic regression model with emphasis on medical research, *Journal of data analysis and information processing*, 7 (2019) 190-207.
- [32] T. Yu and H. Zhu, Hyper-parameter optimization: A survey of methods and applications, arXiv preprint arXiv:2003.05689 (2020).
- [33] Algorithms for hyper-parameter optimization, *Advances in Neural Information Processing Systems*, 24 (J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl) (2011).

CARDIOVASCULAR DISEASE DETECTION USING MACHINE LEARNING ALGORITHMS

ORIGINALITY REPORT

19%

SIMILARITY INDEX

10%

INTERNET SOURCES

9%

PUBLICATIONS

12%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Jacksonville University Student Paper	8%
2	Submitted to University of Birmingham Student Paper	1%
3	Moloud Abdar, Wojciech Książek, U Rajendra Acharya, Ru-San Tan, Vladimir Makarenkov, Paweł Pławiak. "A new machine learning technique for an accurate diagnosis of coronary artery disease", Computer Methods and Programs in Biomedicine, 2019 Publication	1%
4	www.ijirset.com Internet Source	1%
5	"Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making", Springer Science and Business Media LLC, 2020 Publication	<1%
6	www.ncbi.nlm.nih.gov Internet Source	<1%

7	www.arxiv-vanity.com Internet Source	<1 %
8	www.ripublication.com Internet Source	<1 %
9	Submitted to Liverpool Hope Student Paper	<1 %
10	www.engineeringletters.com Internet Source	<1 %
11	Submitted to University of Bradford Student Paper	<1 %
12	hdl.handle.net Internet Source	<1 %
13	mdpi-res.com Internet Source	<1 %
14	ebin.pub Internet Source	<1 %
15	Submitted to University of Moratuwa Student Paper	<1 %
16	Submitted to Sim University Student Paper	<1 %
17	Shivangi Diwan, Gajendra Singh Thakur, Sunil K. Sahu, Mridu Sahu, N. K. Swamy. "Predicting Heart Diseases through Feature Selection and	<1 %

Ensemble Classifiers", Journal of Physics: Conference Series, 2022

Publication

18

Shixue Liang, Yuanxie Shen, Xiaodan Ren. "Comparative study of influential factors for punching shear resistance/failure of RC slab-column joints using machine-learning models", Structures, 2022

Publication

<1 %

19

Talha Anwar, Seemab Zakir. "Machine Learning Based Real-Time Diagnosis of Mental Stress Using Photoplethysmography", Journal of Biomimetics, Biomaterials and Biomedical Engineering, 2022

Publication

<1 %

20

people.fas.harvard.edu
Internet Source

<1 %

21

doctorpenguin.com
Internet Source

<1 %

22

forgos.uni-eszterhazy.hu
Internet Source

<1 %

23

rdw.rowan.edu
Internet Source

<1 %

24

"Intelligent Computing Methodologies", Springer Nature, 2014

Publication

<1 %

25 Aswin Kumar.K, S. Gowri, John Wilifred David J, Y. Bevish Jinila. "An Efficient Association Rule Mining from Distributed Medical Database for Predicting Heart Disease", 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022
Publication <1 %

26 nemertes.library.upatras.gr
Internet Source <1 %

27 Submitted to Eiffel Corporation
Student Paper <1 %

28 Hiroki Mizuochi, Masato Hayashi, Takeo Tadono. "Development of an Operational Algorithm for Automated Deforestation Mapping via the Bayesian Integration of Long-Term Optical and Microwave Satellite Data", Remote Sensing, 2019
Publication <1 %

29 www.coursehero.com
Internet Source <1 %

30 Lecture Notes in Computer Science, 2012.
Publication <1 %

31 Submitted to University of Surrey
Student Paper <1 %

32 bmcbioinformatics.biomedcentral.com

Internet Source

<1 %

33

ideaexchange.uakron.edu

Internet Source

<1 %

34

media.neliti.com

Internet Source

<1 %

35

www.researchgate.net

Internet Source

<1 %

36

Leondes, Cornelius T. "NON-PARAMETRIC PIXEL APPEARANCE PROBABILITY MODEL USING GRID QUANTIZATION FOR LOCAL IMAGE INFORMATION REPRESENTATION", *Medical Imaging Systems Technology*, 2005.

Publication

<1 %

37

Submitted to University of Stellenbosch, South Africa

Student Paper

<1 %

38

Lakshmi Prasanna Kothala, Prathiba Jonnala, Sitaramanjaneya Reddy Guntur. "Localization of mixed intracranial hemorrhages by using a ghost convolution-based YOLO network", *Biomedical Signal Processing and Control*, 2023

Publication

<1 %

39

Rahma Atallah, Amjed Al-Mousa. "Heart Disease Detection Using Machine Learning

<1 %

Majority Voting Ensemble Method", 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), 2019

Publication

40

Submitted to University of Wales Institute, Cardiff

Student Paper

<1 %

41

Zheng Liu, Rehan Sadiq, Balvant Rajani, Homayoun Najjaran. "Exploring the Relationship between Soil Properties and Deterioration of Metallic Pipes Using Predictive Data Mining Methods", Journal of Computing in Civil Engineering, 2010

Publication

<1 %

42

ingenieria.ute.edu.ec

Internet Source

<1 %

43

theses.gla.ac.uk

Internet Source

<1 %

44

A U Azmi, A F Hadi, D Anggraeni, A Riski. "Naive bayes methods for rainfall prediction classification in Banyuwangi", Journal of Physics: Conference Series, 2021

Publication

<1 %

45

Submitted to Jordan University of Science & Technology

Student Paper

<1 %

46 Pincheira, Jose. "Reinforced Concrete Design", Oxford University Press <1 %
Publication

47 airconline.com <1 %
Internet Source

48 de.slideshare.net <1 %
Internet Source

49 jurnalnasional.ump.ac.id <1 %
Internet Source

50 otp.tools.investis.com <1 %
Internet Source

51 www.jetir.org <1 %
Internet Source

52 Ali Al Bataineh, Sarah Manacek. "MLP-PSO Hybrid Algorithm for Heart Disease Prediction", Journal of Personalized Medicine, 2022 <1 %
Publication

53 Devansh Shah, Samir Patel, Santosh Kumar Bharti. "Heart Disease Prediction using Machine Learning Techniques", SN Computer Science, 2020 <1 %
Publication

54 Ernest Ng. "Implicit Stress Integration in Elastoplasticity of n-Phase Fiber-Reinforced <1 %

Composites", Mechanics of Advanced
Materials and Structures, 11/2007

Publication

55

IFIP Advances in Information and
Communication Technology, 2013.

Publication

<1 %

56

Sascha Caron, Jong Soo Kim, Krzysztof
Rolbiecki, Roberto Ruiz de Austri, Bob Stienen.
"The BSM-AI project: SUSY-AI-generalizing
LHC limits on supersymmetry with machine
learning", The European Physical Journal C,
2017

Publication

<1 %

57

coek.info

Internet Source

<1 %

58

dergipark.org.tr

Internet Source

<1 %

59

digitalcommons.du.edu

Internet Source

<1 %

60

etda.libraries.psu.edu

Internet Source

<1 %

61

researchcongress.tec.mx

Internet Source

<1 %

62

www.slideshare.net

Internet Source

<1 %

www2.mdpi.com

63

Internet Source

<1 %

64

D.R. Lovell, C.R. Dance, M. Niranjana, R.W. Prager, K.J. Dalton, R. Derom. "Feature selection using expected attainable discrimination", Pattern Recognition Letters, 1998

Publication

<1 %

65

Dineshkumar Muthuvel, Bellie Sivakumar, Amai Mahesha. "Future global concurrent droughts and their effects on maize yield", Science of The Total Environment, 2023

Publication

<1 %

66

ink.library.smu.edu.sg

Internet Source

<1 %

Exclude quotes Off

Exclude matches < 3 words

Exclude bibliography On