

**CORONARY ARTERY DISEASE PREDICTION USING MACHINE
LEARNING ALGORITHM**

BY

**BIVUTI VUSHAN BHADRA
ID: 221-25-114**

This Report Presented in Partial Fulfilment of the Requirements for the
Degree of Master of Science in Computer Science and Engineering
(Major in Data Science)

Supervised By

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



**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2023**

APPROVAL

This Thesis titled “Coronary artery disease prediction using machine learning algorithm”, submitted by Bivuti Vushan Bhadra, ID No: 221-25-114 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 M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17-01-2023.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan, PhD
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Ms. Nazmun Nessa Moon
Associate Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

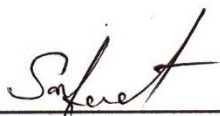
Internal Examiner



Dr. Fizar Ahmed
Associate Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Safaet Hossain
Associate Professor & Head

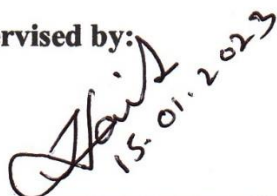
Department of Computer Science and Engineering
City University

External Examiner

DECLARATION

I hereby declare that, this thesis paper has been done by me under the supervision of **Dr. Sheak Rashed Haider Noori**, Professor and Associate Head, **Department of Computer Science and Engineering**, Daffodil International University. I also declare that neither this paper nor any part of this paper has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



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

Submitted by:



Bivuti Vushan Bhadra
ID: 221-25-114
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for me to complete the thesis paper successfully.

I really am grateful and wish my profound indebtedness to **Dr. Sheak Rashed Haider Noori**, Department of Computer Science and Engineering, Daffodil International University, Dhaka. The deep knowledge & keen interest of my supervisor in the field of “*Machine Learning*” has enthused me greatly to carry out this thesis paper. His endless patience, scholarly guidance, continual encouragement, energetic supervision, constructive criticism, valuable advice, and patience in reading many inferior drafts and correcting them at all stages have made it possible for me to complete this paper.

I would like to express my heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor and Head, Department of Computer Science and Engineering, for his kind help in completing my thesis paper. I am also grateful to other faculty members and the staff of the Computer Science and Engineering department of Daffodil International University for their kind support and help.

I would like to thank my entire course mate at Daffodil International University, who took part in many discussions and help me to complete this paper.

Finally, I must acknowledge, with due respect, the constant support, encouragement, and patience of my parents.

ABSTRACT

New technologies such as machine learning and big Data analytics has proven to be a promising solution for the biomedical community, health problems and patients care. It is accurate and therefore useful for early disease prognosis Interpretation of medical information. Disease control strategies Recognizing early symptoms can lead to further improvement Illness. This initial prediction also helps Disease symptoms and proper symptom control Treatment of illness. Machine learning methods can be used Prognosis of chronic diseases like kidney and heart Diseases form classification models. In this paper We propose a comprehensive preprocessing method to predict coronary arteries disease (CAD). This method involves zero substitution Standardization, resampling, normalization, classification, Prediction. This study aims to predict the risk of CAD use machine learning algorithms such as Random Forest, Decision Tree, Naïve Bayes, Logistic Regression and K-Nearest Neighbors. Comparative studies Among these algorithms, prediction accuracy is based on to be done. Additionally, generation are using k-fold cross validation. This algorithm has been tested dataset containing 1190 records and 12 features where Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbor, Naïve Bayes, and Support Vector Machine achieved 96% accuracy, 93%, 84% and 71% respectively. Therefore, using our in the preprocessing step, random forest classification gives more information more accurate results compared to other machine learning algorithms.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	i
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
CHAPTER	
CHAPTER 1: INTRODUCTION	1-6
1.1 Introduction	1
1.2 Motivation	3
1.3 Rationale of the Study	3
1.4 Research Questions	4
1.5 Expected Output	5
1.6 Report Layout	5
CHAPTER 2: BACKGROUND	7-13
2.1 Preliminaries	7
2.2 Related Works	8
2.3 Comparative Analysis and Summary	10
2.4 Scope of the Problem	12
2.5 Challenges	12
CHAPTER 3: RESEARCH METHODOLOGY	14-30
3.1 Introduction	14
3.2 Research Subject and Instrumentation	15
3.3 Data Collection Procedure	19
3.4 Statistical Analysis	23
3.5 Proposed Methodology	29
3.6 Implementation Requirements	30

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	31-40
4.1 Introduction	31
4.2 Experimental Results & Analysis	31
4.3 Descriptive Analysis	34
4.4 Discussion	39
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	41-42
5.1 Impact on Society	41
5.2 Impact on Environment	41
5.3 Ethical Aspects	42
5.4 Sustainability Plan	42
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	43-44
6.1 Summary of the Study	43
6.2 Conclusions	43
6.3 Recommendations	44
6.4 Implication for Further Study	44
REFERENCES	45-46
APPENDIX	47

LIST OF FIGURES

LIST OF FIGURES	PAGE NO
Figure 3.1: Flowchart of proposed work	15
Figure 3.2: Machine learning algorithm model	16
Figure 3.3: Co-Relations of Dataset	21
Figure 3.4: Remove outlier of dataset	22
Figure 3.5: Statistical Views of Coronary Artery Disease	24
Figure 3.6: Statistical Views of Coronary Artery Disease	24
Figure 3.7: Statistical view Gender & Age Wise	25
Figure 3.8: Statistical view Gender & Age Wise	25
Figure: 3.9 Statistical view Gender & Age Wise (normal patient)	26
Figure 3.10: Statistical view Gender & Age Wise (disease patient)	26
Figure 3.11: Statistical view of Chest Pain Type	27
Figure 3.12 Statistical view of Chest Pain Type heart patient	27
Figure 3.13 Features of histogram	28
Figure 3.14: Flowchart of proposed work	29
Figure 4.1: Accuracy chart	34
Figure 4.2: Confusion Matrix of Random Forest	35
Figure 4.3: Confusion Matrix of Decision Tree	35
Figure 4.4: Confusion Matrix of K-Nearest Neighbors	36
Figure 4.5: Confusion Matrix of Logistic Regression	36
Figure 4.6: Confusion Matrix of Naïve Bayes	37
Figure 4.7: Confusion Matrix of Support Vector Machine	37
Figure 4.8: ROC-AUC curve	38
Figure 4.9: Result Evaluation Score (After K-Fold Evaluation)	39

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Result Comparison	12
Table 3.1: Attributes of the dataset and their interpretation	20
Table 3.2: Descriptive Statistical Analysis	23
Table 4.1: Performance Statistics	33
Table 4.2: ROC-AUC Curve	38
Table 4.3: Result Evaluation Score (After K-Fold Evaluation)	39

CHAPTER 1

INTRODUCTION

1.1 Introduction

Machine learning is a strong research field for researchers. Different machine learning techniques are widely used in different fields. Marketing, health and medical disorders, weather forecasting, and socioeconomic activity analysis are used by machine learning. Many diseases in the health sector can be detected by using machine learning algorithms. Heart disease is a major global health problem in the 21st century. The high incidence of heart disease has a significant impact on the health and socioeconomic growth of the country. Likewise, the risk of rapid growth is high. Especially poor are faced with it. The global health organizations' focus on heart disease and perceptions. It is difficult that developed countries to focus and powerful integrative values to reduce heart disease. Hospitals store information about patients with heart diseases in their systems. By analyzing this information, you can identify various patterns that can help you predict. By applying data mining techniques to all this information, we can discover a large quantity of information and apply that to predict disruption. Many people are affected by heart diseases. Many people in Bangladesh don't mind the chaos. For this reason, the number of people affected by this disorder is increasing by the minute. You can control whether you can recognize or predict whether an individual is or will be affected. People should be extra careful not to be affected. You can use predictive to anticipate interruptions. Some of the possible approaches are classification, regression, and indexing. Many people believe that categorization is the preferred approach. Once the study is complete, it will certainly help predict coronary artery disease. People are aware of both illness and their condition. The leading cause of death worldwide, cardiovascular disease accounts for approximately one-third of all deaths annually. Coronary artery disease is the leading cause of death from all cardiovascular diseases [1] [2].

Heart disease, known as CAD, narrows the arteries of the heart, reducing the quantity of blood which can flow through them, reducing the flow of oxygen to the heart and increasing the work of the heart. The heart is working difficulty to pump blood through

the narrow passage. An artery is a complete block, causing heart attack. This obstacle of an artery usually occurs because of fatty deposits in the arteries that cause the arteries to slow down [3]. Lack of exercise on her was the main cause of the disease, and regular has been shown to greatly help prevent and rehabilitate his CAD [4].

Hence there is great interest in the diagnosis and treatment and prevention of coronary artery disease. To do this, machine learning algorithms are widely employed to build a model which can help identify the illness by creating a classification algorithm that determines if people have CAD based on data in medical sector. I tried to create a heart disease registry. To manage a machine learning model to determine if a person has the coronary artery disease, fourteen widely used feature data were employed. Ongoing technological advancements have enabled researchers to develop novel strategies based on artificial intelligence and machine learning. Big data is being generated more frequently due to the rise of health problems and this data is being used to create automated computerized systems that can employ machine learning algorithms to predict disease. The dataset appears to be able to perform efficiently for various challenges. Along with other chronic health issues including obesity, smoking, and diabetes, cardiovascular disease is regarded as a high-risk factor. Population aging is associated with CVD especially in developed countries suggesting that the elderly are more capable to CVD [5]. Cardiovascular disorders such as myocardial infarction, chronic heart failure and cardiac arrest have been recorded according to a 2014 World Health Organization report [6]. 17 million persons passed away. Coronary artery disease can also be detected by heart sounds [7]. Since most coronary artery diseases can't be find out by traditional electrocardiogram (ECG) techniques, several prohibitive sensors or instruments, as like phonocardiogram (PCG), electromyography (EGM), etc. have been developed. In this study, we fine-tune the parameters of three various classification algorithms and then evaluate and contrast them. First, we choose the most topical features for classification using five different feature selection methods. The primary objective of this research is to use classification of machine-learning algorithms to compute states and warn when someone calls for disturbance, and to use other machine learning algorithms to detect which is the best machine learning algorithm. By evaluating the following situations, machine learning algorithms are used.

1.2 Motivation

The main destination of this work is to offer coronary artery disease predictive model for predicting coronary artery disease. This study seeks to determine the best classification system for identifying patients with coronary artery disease. To accomplish this task, three of his classification algorithms—Naive Bayes, Decision Trees, and Random Forests—were used in a comparative study and analysis. However, CAD prediction is a challenging task that requires the highest level of accuracy and often uses machine learning techniques. Three algorithms are evaluated using different CAD prediction evaluation levels and techniques. This creates better opportunities for researchers and clinicians. Coronary artery disease is one of the global worst diseases, but early detection can reduce its rate. It is very important to control it. Angiography is the most common procedure used to treat coronary artery disease. Angiography is costly and has drawbacks. Due to the difficulty of detection due to risk factors, modern techniques have been developed to detect the presence of heart disease.

1.3 Rationale of the Study

Favorable invention of coronary artery disease prediction could reduce death. However, it is always detected in the final stages of the illness and following a death. An elemental problem, therefore, is the inability to predict or understand coronary artery disease early. It is not assumed that coronary artery disease can be detected immediately, it can be medicated quickly with accurate treatment, but in reality, numerous tests are required to confirm the presence of heart disease, which is very tedious. This review can make people's lives easier and save time. In this context, machine learning methods are an excellent system for predicting heart disease. If various data mining techniques can be used to predict coronary disease at an early stage.

1.4 Research Questions

Heart disease is one of the many diseases that affect us it is a serious disease as many of us notice that most people die from heart disease or like similar diseases related to the heart [8, 9]. Maximum medical professionals observe the multiplicity of heart patients often do not live and die from heart attack. When I thought about turning the idea into a reality, the following questions came to my mind:

- **How can we predict coronary artery disease more accurately?**

In recent years, there have been methods involving machine learning (ML) for disease detection and diagnosis. ML approaches typically involve "training" an algorithm using a controller. Create a dataset with known disease states (presence or absence of disease) and apply this trained algorithm to a variable dataset for predicting a patient's disease state that has yet to be determined. As a larger group, the ML algorithm is well trained as a predictor of disease states. More specifically, disease ML-powered predictions enable clinicians to improve detection, diagnosis, classification and risk stratification. and ultimately, patient management, potentially reducing the need for clinical intervention.

- **Why is the main focus of this review on predicting coronary artery disease (CAD)?**

Coronary artery disease is probably the most worrying problem in the world. Things get worse with time. It progresses through several stages and the final one is when the heart stops working completely. During this period, a person becomes irrational and acts in the manner described below. With effective therapy and compliance with several rules and standards, the initial expected probability of heart disease is reduced. For this reason, the focus of this review was on coronary artery disease.

- **Why are machine learning methods reliable?**

Machine learning is one of the most famous methods used for basically any kind of forecasting to use a large amount of data to prepare a model and predict any outcome. The use of clinical datasets and ML methods is expected cad result without any problems in the running scenario of eras of modernity are being experienced throughout

the world. Presumably about 10 years a time when the use of artificial intelligence was less popular, this nice name was just numerical logic. Yet currently a large part of planetary innovation across the planet depends on AI, It might become more durable with greater accuracy and subsequent, proper practice in this area. However, it is trustworthy throughout this time.

1.5 Expected Output

Many people are unconsciously conscious of their health. Because of this, they are undergoing from different types of ailments that in the long term, death is a real possibility due to coronary artery disease. Heart disease can be accurately predicted using algorithms. Despite the fact that coronary artery disease is now a serious concern to the community of this predictive method will help identify risk factors and regulate if a patient is at risk of coronary artery disease. Detecting predictive risk levels for coronary artery disease, when monitored, can help you be aware of your heart health as it reduces mortality from cardiovascular disease, smoking, and more. There are already many studies that predict coronary artery disease based on several traits, so it is not possible to provide accurate coronary artery disease predictive rates, but this study did It analyzes many traits such as chest pain, blood pressure, age factors, stress, medication, gender, regular exercise use and more. This allows for a more accurate display of coronary artery disease detection.

1.6 Report Layout

This report proposes some models such as decision tree, logistic regression, naive bayes, KNN, and random forest machine learning classification algorithms for forecasting coronary artery disease (CAD) risk levels based on selected symptoms by evaluating a coronary artery disease dataset. The technology's predictions will help people understand their heart condition and if anything goes wrong, they'll go to the doctor as soon as possible because mortality is down. This report is based on six parts. An important introduction to this research project is provided in Chapter 1. Objectives of the study, justification of findings, relevant research questions, expected results, general management information, and financial implications of the study are all

included in this chapter. Provides an overview of conditions associated with coronary artery disease.

Chapter 2 provides background information on this study. Machine learning systems, categorization data, and associated tasks have been built based on the findings of this research study. This part describes the comparative analysis, elaborates the problem statement and lists the observed defects.

Chapter 3 which provides narrative details of the planned system and methodology for this research study. State-of-the-art algorithmic details are described for each algorithm used.

The complete results analysis for the results of each step is presented in chapter 4. The best algorithm ends up with the best accuracy score by doing the confusion matrix. At the conclusion of this chapter, misclassification, AUC-ROC Curve, Cross-Validation, mean absolute error, and mean squared error are explored.

The future scope of this study is presented in Chapter 5 with a brief explanation of how it is an extension of this effort. This chapter discusses how this research study will affect society, the environment, and sustainability.

Finally, conclusions and next task in future is presented in chapter 6. The key findings of this research are briefly described in the presentation that concludes this research report.

CHAPTER 2

BACKGROUND

2.1 Preliminaries

This part provides a detailed discussion of the background of the research study titled *An Analytical Prediction System Using Machine Learning for Coronary Artery Disease*. Its classification can be accomplished by using supervised machine learning and classification techniques. A machine learning system is independent and capable of continuously integrating data for the purpose of making decisions. Machine learning techniques come in many shapes and sizes, making it possible to make analytical observations learned from past events and use other techniques to create an ever-improving system. Supervised machine learning methods are widely used in this research. Supervised machine learning methods examine previously labeled data to make future predictions. The learning method runs tests on a well-known training dataset to generate an inferred function that predicts the output value. The learning algorithm can detect and fix any errors by comparing its output with its target. This thesis deals with early phase classification of cardiovascular disease using various supervised learning methods. In data analysis, a process called classification uses large amounts of data to build multiclass feature models. This is sometimes referred to as the class model classification model and is the most renowned and frequently applied machine learning technique currently accessible. Specific class predictions can be made by discrete, randomly directed predictors, when trained under supervision. For example, a classification can be created to determine. If a picture contains a picture of a frog or a picture of a fish. The result would be expected as "frog" or "fish". A classifier cannot be used to obtain an intermediate value. For example, one can establish a classification. To increase accuracy, a classification learning method can be used. In category learning, there are two categories. The first type is known as training leader, while the second is known as test data. The model is constructed utilizing training data. When test data is employed to ensure that the model is accurate. A stepwise approach can be used to describe the classification process. Regarding the structural anatomy of the heart, our heart mechanism, the connections between blood vessels that supply all parts of our body with blood, revolves around the heart. It carries nutrients to all human

organs, keeping them healthy and functioning effectively. When a fatty substance called plaque builds up in your arteries, it causes coronary artery disease. Plaque hardens over time and narrows the artery. When plaque clogs an artery, it acts like a clogged drain tube, reducing blood flow. When the heart doesn't pump enough oxygen-rich blood to our vital organs and muscles, we get tired and weak in our legs. Weight gain, swollen ankles, legs and belly.

2.2 Related Works

Researchers have used various machine learning techniques like as organization rules, grouping and clustering to develop models to predict coronary artery disease. Machine learning is widely used in the prediction process. Predictive methods are very important in medical field. Therefore, it is very useful in these areas. A large accumulation of recorded data can help predict. Many diseases can be predicted early and efficiently using various machine learning techniques. Researchers are using machine learning techniques for more serious diseases like coronary artery disease and cancer. We also show that using these techniques to predict breast cancer yields more accurate results in predicting the disease. A major concern in medicine is the prognosis of coronary artery disease. Coronary artery disease can be easily predicted using a variety of "machine learning algorithm" techniques such as SVM, DT, NB, KNN and RF (Random Forest). Different studies have been conducted and are increasingly being conducted to predict coronary artery disease to obtain more better results. They was applying various machine learning techniques to CAD evaluation and prediction with different accuracy results. Different authors take different approaches Create a classification model. Data collection is also possible measuring his ADL (activities of daily living) using a wearable [10]. The proposed framework implemented these data by cohort segmentation and monitors Unsupervised algorithms for batch-level machine learning and processing. Different type ML Algorithms such as KNN, SVM, and random forest are applied. Based on two kind of datasets and comparisons accuracy and model building time.

Shiva Kazempour Dehkordi¹ and Hedih Sajedi proposed a model which based on prescriptive machine learning approach [10]. They proposed an algorithm called Skate

Improvement System Accuracy. Boosting and bagging are analogous to skating ensemble techniques. They experimented with different labels and compared four classification methods, including DT, NB, KNN and Naive Bayes. The most specific classifier was blocked, they showed. Unlike other classification algorithms and techniques, An accuracy of 73.17% was provided by this classification algorithm.it is a relatively low performing method.

For example, Jan et al. ensemble discovered five different classifications by applying machine learning mining techniques using two benchmark datasets (Cleveland and Hungarian) collected from the UCI repository in 2018. They found algorithms like RF, neural network, NB and classification. Using Regression Analysis and Support Vector Machine (SVM) [11].

In their study, they found that the smallest performing algorithms were the regression method, with RF giving very high values. 98.136 Curacy. 2011, Jyoti Soni et al. Apply its DT with a genetic algorithm to improve classification performance. He also has two other algorithms, namely clustering by NB and classification [12].

They found the accuracy which is 99.2% of the proposed model. In 2017, Hend Mansoor et al. and others analyzed the performance of its LR and RF classification algorithms to estimate risk exposure in CAD patients [13]. She obtained an accuracy of 89% in the LR model and 88% in the RF model.

Austin et al. compared the performance of traditional classification trees with regression trees [14] in 2013. Conventional LR has shown best results in evaluating the possible presence of HD.

Le et al. In 2018, they used three classification methods that listed 58 features in a dataset collected from the UCI Machine Learning Repository [15]. They say that a support vector machine (SVM) with a linear kernel performed well with an accuracy of 89.93%. Taraweeh and Embark create12-feature hybrid approach and compared its performance with ANN, DT, KNN, SVM, and NB [16].

Which accuracy was 89.2. It has the best performance compared to other adaptive. Particle swarm optimization used by Muthukaruppan et al [17]. use an evolutionary fuzzy expert system to classify CAD and obtain high result accuracy [18].

On the Cleveland dataset, make a prediction. [17] and [18] to construct a fuzzy rule base, we use decision trees. But none of these previous studies used multiples. Methods for selecting training features using features 12 frequently used and classified The algorithm applied in each instance was not identical. One of the five different classification methods discussed in this work. The results in this white paper show just how true this is. Selection of model parameters and given properties Increasing all evaluation criteria for previous attempts with SVM, KNN and Naive Bayes. This From the overview in [16], Performance of this algorithm in previous work.

In this study, the model showed Optimal accuracy for fast CAD predictions. Coincidentally Random Forests and Decision Trees gave better results than confusion matrices.

2.3 Comparative Analysis and Summary

Many people are unaware because of their physical condition. Because of this, individuals have different types of illnesses, including coronary artery disease, which is popular and ultimately leads to death. Since monitoring lowers the death rate from coronary artery disease, you will be more aware of your cardiovascular health. By displaying your estimated risk of developing heart disease, you can simply raise awareness of dietary modifications, lifestyle modifications, quitting smoking, hospitalization, etc. The death rate can be decreased. It is impossible to predict with any degree of accuracy the likelihood of developing heart disease as there have been many studies on the subject. The condition is associated with several risk factors, including drug use, high blood pressure, chest discomfort, age, sex, and cholesterol and gender risk factors. In the modern world, the majority of hospitals take some action to manage their patient or medical information. This procedure gathers a lot of data, which is then presented as numbers, text, graphs, and graphics. But regrettably, the medical industry rarely employs these data in decision-making. In order to predict coronary artery

disease fast, we applied classification approaches in this scope and demonstrated the analysis of various machine learning algorithms. In our research, Naive Bayes, Random Forest Classification, Decision Tree, and Support Vector Machine, Logistic Regression and KNN are employed as algorithms. For clinical specialists with specialized demands or for accurate analysis of cardiac disease, these algorithms may be highly helpful. We present a set of current data on cardiovascular disease in this study. Then, based on accuracy and execution, we choose our ideal algorithm utilizing a variety of execution metrics. It will be simple to assist the specialists in predicting coronary artery disease using this strategy. providing a more precise identification of coronary artery disease after analysis. In my paper, Random Forest showed the best results on both fronts with 96% accuracy. Other algorithms such as Naïve Bayes, Decision Tree, Logistic Regression, K Nearest Neighbor have accuracies of 85%, 93%, 84%, 71%. Examine the effects of class imbalance in your data Performs with multi-layer perceptron performance. The performance of the multi-layer perceptron algorithm was evaluated using different learning rates. Also compare based on the execution time and accuracy criteria, the study was conducted. The aim was to recognize the importance of traits to classification results. Preprocessing and normalization are Executed on the dataset. Then measure the correlation. A correlation matrix between features was obtained. further away, Classification was done in his three stages: beginning, Feature selection; second AUC (Area Under Curve), ROC, outlier detection and removal, missing value checking, correlation-based Performance comparison of five algorithms: Random Forest, Decision Tree, Naive Bayes, and KNN. Original data submitted. Third, classifier performance was compared based on sensitivity, precision, and specificity AUC.

In this study [23], they used three models for prediction (CAD) where the KNN classification model gave the best outcome with an accuracy of 84%.

In our paper, using six creative models using 12 attributes, 1190 data. it was found that the Random Forest (RF) algorithm gave very good accuracy for (CAD) Prediction. In this part, the output of the model used in this article will be compared to models from other researchers. Everyone will receive a summary of the model's performance in this document as a result. Ultimately, the random forest technique was used to train the best model that was discovered. When this model is compared to other models, it is clear

that the accuracy score in this paper's ML model is higher than that of the other researchers' models shown in the following table-

TABLE 2.1: RESULT COMPARISON

Algorithm	Previous work [23]	Proposed work
	Accuracy	Accuracy
Naïve Bayes	84%	85%
SVM	83%	73%
KNN	80%	71%
Random Forest	-	95%
Decision Tree	-	93%
Logistic Regression	-	84%

2.4 Scope of the Problem

A classification-focused machine learning algorithm determines whether to forecast coronary artery disease. Performance is improved when machine learning techniques are used. The main cause of death in is coronary artery disease. Simply because of bad daily habits and diets, many people have perished from heart disease. We made the decision to investigate the prognosis of coronary artery disease in order to lower the death rate using our technology. In order for people to be aware of their heart's health, we presented a method that offers them a prediction value of artery disease risk. Therefore, the focus of current paper is coronary artery disease.

2.5 Challenges

There is a dark aspect to everything. Therefore, there were many difficulties in the research and investigation, implementation of our system. At times it was very difficult to deal with, God, we have overcome these difficulties. These next difficulties made us Harder Studies –

- Collecting data from different hospitals.
- When choosing an algorithm.
- The use of machine learning classification algorithms.

- For the implementation of the proposed system.
- In selecting external factors for heart disease.
- Many datasets are available online, but selecting an effective dataset is a difficult task for us.
- A large dataset was selected, which was difficult to process in subsequent work steps.
- It was difficult to evaluate the functionality of the dataset features because we collected the dataset online.
- The dataset contains 14 medical attributes, so the preprocessing process was not trivial for large datasets.
- It took a lot of time to choose effective machine learning algorithms because I had to acquire enough knowledge to implement these algorithms.
- It takes a lot of time to get a good knowledge on this topic as there are many previous documents and research studies on this topic.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Decision Trees, Naive Bayes, Random Forests, and Support Vector Machines are among the learning algorithms used in this study's five machines to predict the risk of coronary artery disease. The purpose of this study is to identify primary contributors to coronary artery disease using a computerized prediction technique. It helps individuals with heart problems live longer lives. We use a variety of machine learning methods, some of which are discussed in this work, to examine the data set in order to accomplish this goal. Techniques for classification are particularly useful for forecasting serious disorders. Therefore, we employ this data mining technique to reach our objective. Here, we go over the entire process of data analysis utilizing step-by-step categorization approaches, as well as how some specific algorithms are put into practice. In order to accomplish this analysis, the top five supervised machine learning algorithms were used. In its most basic form, an algorithm is a list of sequential instructions that instructs computer programs how to transform a collection of input data into useful information. Facts are statistics, and valuable data is any information that may be used by humans, machines, or algorithms. Similar operations are carried out by machine learning algorithms, with a little math thrown in for good measure. Mathematical transformations are not the same for all machine learning algorithms. The most important machine learning algorithms are discussed in this paper, each of which incorporates crucial algorithmic operations across system topologies. We initially needed to gather a dataset before we could apply these methods. Regarding the extrinsic characteristics of coronary artery disease signs. We previously knew that these five machine-learning algorithms will offer more accuracy from reading academic publications. Data preparation is introduced in raw data transformation, Information in a setting that is open for any use. The process to be followed to achieve a reliable algorithm for predicting artery disease in patients is illustrated in the proposed system. After collecting the dataset, it performs pre-processing before using it to create and test several machine learning algorithms. After that, more important methods including accuracy, misclassification, Jaccard score, cross validation and confusion matrix were

conducted to identify the top algorithms. Once all the required elements were exhausted and discussed, the best algorithm was selected after scoring all the algorithms.

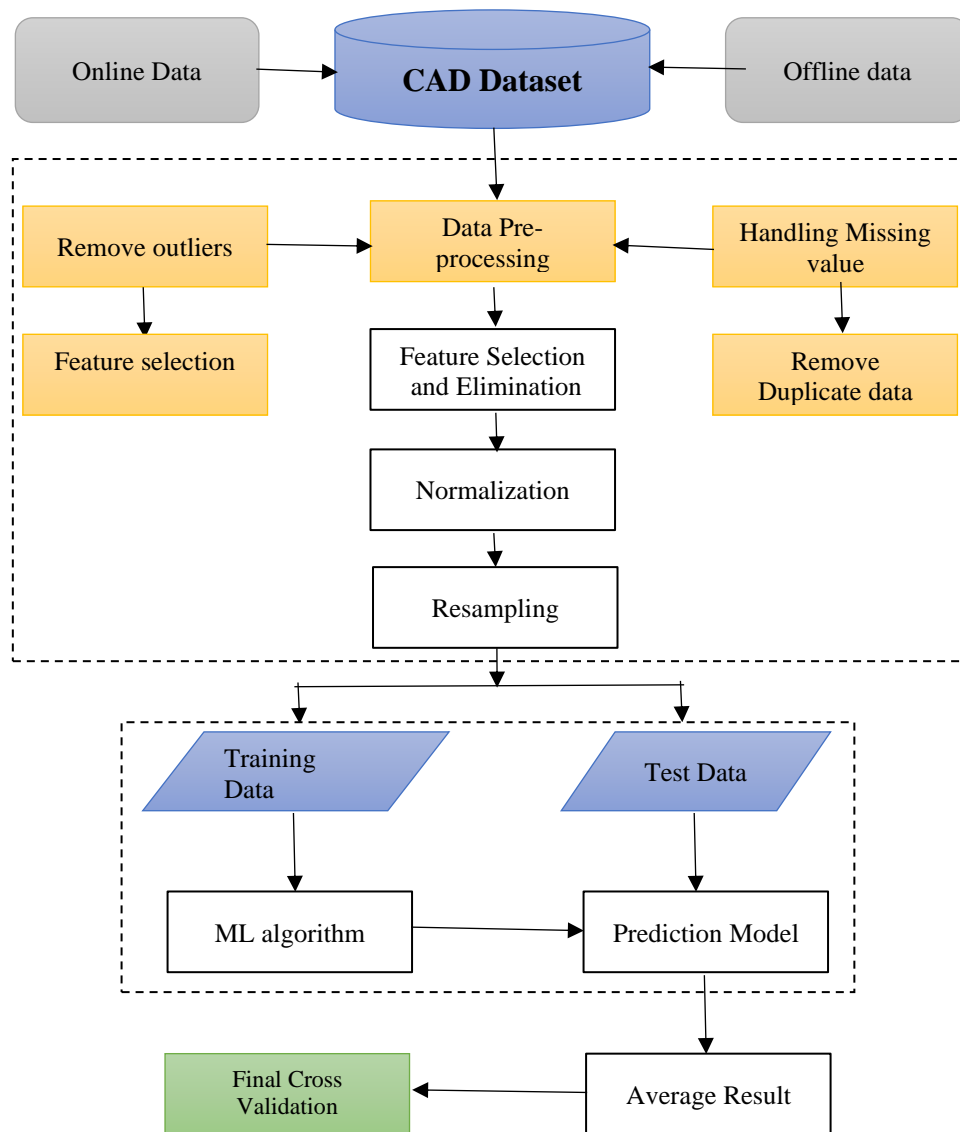


Figure 3.1: Flowchart of proposed work

3.2 Research Subject and Instrumentation

A variety of cardiac conditions, from abnormal heart rhythms to organ failure, are covered by the term "coronary artery disease." Many people face heart disease because of unhealthy lifestyle and diet, such as smoking, eating too much fat, and not exercising enough. Therefore, In the present work, we proposed a machine learning classification algorithm-based system that can help people learn about their daily routines and dietary

habits while assessing the risk of coronary artery disease. Although it will not completely eliminate coronary artery disease, it will reduce mortality.

At this time, data analysis relies heavily on machine learning algorithms for detection and prediction. We'll test various algorithms on the data we've acquired to discover which one complements our model the best. I am using use a range of machine learning algorithms. These include Naïve Bayes, Decision Tree, Random Forest, Logistic Regression, K Nearest Neighbor.

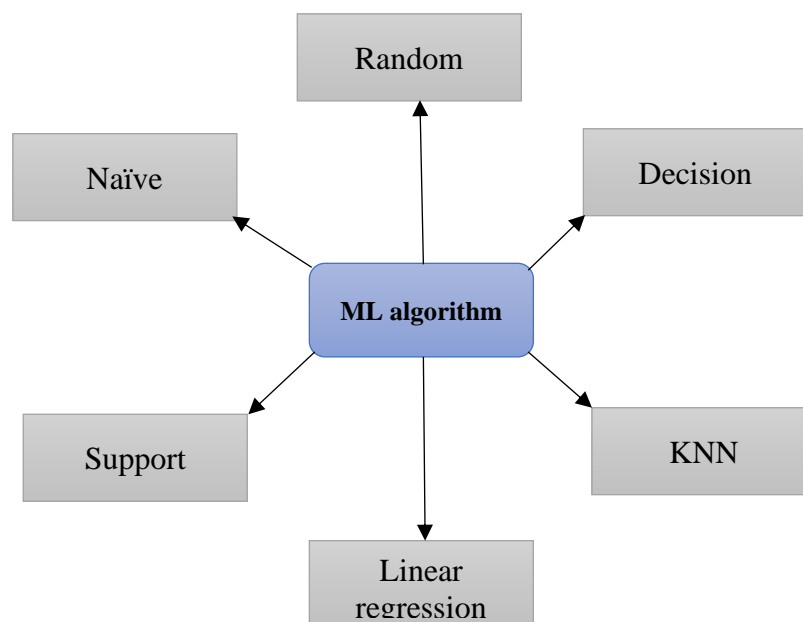


Figure 3.2: Machine learning algorithm model

3.2.1 Random Forest (RF)

Supervised machine learning is used in Random Forest (RF). It was a chance forest, as the name of the algorithm implies, that produced the decision tree. Basically, use it as a strategy. Adaptability that combines multiple learning models to improve your end results. Run Slack Create, manipulate and combine different decision trees to get more complex results. For this reason, it is one of the best machine learning algorithms. Get over there to get it over here. Split the nodes to get a random subset of features. Top features that affect your construction model the most. By providing random thresholds for each feature in the random forest, the results are significantly improved. Attribute

scoring also uses this method. Close to the base Relative feature relevance is determined by how much the feature pollutes the model. To outsiders, RF is powerful.

3.2.2 Decision Tree (DT)

Among the most basic algorithms that is also one of the most efficient and practical is the decision tree (DT). First chance knot, second decision knot, and end node are the three knots in an enveloping tree. A node where a choice is made based on outcomes is known as a decision node, and a random node will display potential outcomes for a given node. The path's final outcome is returned by a tree at the end node. The root node of a decision tree is the first node and is divided into many branches or nodes. section of Using probability as a guide, this root node is executed. Each link indicates a decision rule that was made at the node, and each node extracts information about the features of the data. On the Gini index and the law of entropy, trees are represented or drawn. One of the best prediction models is this one since it is so straight forward and easy to comprehend.

3.2.3 Logistic Regression

A classification technique called as logistic regression in statistics may forecast a binary result of 0 or 1. As a result of this method's prediction that artery disease detection will be either positive or negative for coronary artery disease, the results are rather easy to understand. Additionally, this approach makes predictions and implementation simple. These models have the ability to address a wide range of challenging issues by incorporating a collection of capabilities rather than a single feature. The variable's value rises as the Y-axis travels from 0 to 1. In actuality, the sigmoid function's maximum and lowest values are represented by these two integers. The data should ideally be divided into two groups. More about this source text required for additional translation information, Send feedback, Side panels.

3.2.4 Gaussian Naïve Bayes

Gaussian A variation of the Naive Bayes method which accepts continuous data and Gaussian distributions is called Naive Bayes. Naive Bayes is one of the most used supervised algorithms for classification engines, and it is based on the Bayes theorem. The categorization scheme is straightforward yet extremely efficient. In continuous data, it is general to assume that each class value is distributed with a normal (also known as Gaussian) distribution.

3.2.5 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a supervised classifier Predict the target class based on the algorithm. Label the training data for the model to indicate how similar the particular data is to other available data. It's understandable As, a property (characteristic) of this data, its purpose Labels must be predicted and features compared Existing data (except target class). Similar to for each class this data determines which class they belong to Per. KNN uses a method that compares unclassified data Calculates distance between, using categorical data Data point properties (Euclidean distance, using Manhattan) distance, etc. The model first gathers unclassified data. After that, data determines the distance between each feature. These data are characterized as sensitive data. This is done Select K short distance. Then the class is calculated. It appears most frequently among these K observations.

3.2.6 Support Vector Machine Classifier

Support vector machines are a related supervised learning model. A learning algorithm that analyzes the data used for classification and regression analysis. The SVM model is the example is mapped as a point in space, so the example of each category is separated by a clear gap. as wide as possible. Then the new example is mapped to this are predicted to belong to the category based on which side of the gap will it fall on.

3.3 Data Collection Procedure

Data must be collected to complete the investigation. Online and offline data were collected. We have collected relevant data of general population and cardiology patients from several hospitals in the city of Dhaka. This is known as offline data collecting. On the one hand, I've created a Google Form that I'm using to send links to get data from different people. The following questionnaire was administered and responses were recorded in questionnaire form. Therefore, we have collected all the necessary information. Some could not provide any information about the incident. We gather information from their medical history and family members. Data from earlier investigations is what we gather. For this reason, it was necessary to collect many research papers, journals, websites and reports and consolidate their data to collect the information needed to implement the concept. To assess heart disease, the system had to collect real-world data. Data was provided by UCI for model training.

To collect the medical data, I am created a survey form containing 3 common questions and 12 “yes/no” questions. We have disseminated these research materials to the general public as well as heart patients in various hospitals to get the desired results.

1. "What is your age?"
2. "What is your gender?"
3. "What type of person are you?"
4. "Are you a smoker?"
5. "Do you have diabetes?"
6. " Have you ever undergone high blood pressure treatment? " do you have?"
7. 'Are you taking any medicine?'
8. 'Are you stressed?'
9. 'Do you have heartburn?'
10. 'Are you dizzy or light-headed?'
11. , Do you feel out of breath? '
12. , "Do you usually work out?"
13. , "Are you experiencing any chest discomfort in the middle?"
14. Is there any discomfort or pressure in your chest?"

In all there were 1190 samples of data. Twelve predictors of future behavior were included in each group. There are 1190 entries total of 76 attributes; For them, the 12 most important characteristic are examined. The datasets are combined into a CSV file, which reduces the number of rows an amount that can be used in various machine learning methods. Machine learning algorithms require large amounts of data to make certain predictions. The raw data set contains three robust and 12 columns with few missing values in few rows and columns. The data set contains some missing values and random noise.

The description of all attributes is summarized below:

TABLE 3.1: ATTRIBUTES OF THE DATASET AND THEIR INTERPRETATION

Attribute	Description	Type
Age	Patient's age in completed years	<i>Numeric</i>
Sex	Patient's Gender (male represented as 1 and female as 0)	<i>Numeric</i>
Cp	1. Typical angina, 2. Atypical angina, 3. Non- angina pain and 4. asymptomatic	<i>Numeric</i>
Trestbps	Level of blood pressure at resting model (in mm/Hg at the time of admitting in the hospital)	<i>Numeric</i>
Chol	Serum cholesterol in mg/dl	<i>Numeric</i>
FBS	Blood sugar levels on fasting>120 mg/dl; represented as 1 in case of true, and 0 in case of false	<i>Numeric</i>
Resting	Result of electrocardiogram while at rest are represented as value 0, Abnormality in ST-T wave as Value 1, (which may include inversions of Tweve and/or depression or elevation of ST of > 0.05 Mv) and any probability or certainty of LV hypertrophy by Estes' criteria as Value 2	<i>Numeric</i>
Thali	The accomplishment of the maximum rate of heart	<i>Numeric</i>
Exang	Angina induced by exercise. (0 depicting 'no' and 1 depicting 'yes')	<i>Numeric</i>
Old peak	Exercise -include ST depression in comparison with the state of rest	<i>Numeric</i>
Slop	ST segment measured in terms of the slope during peak exercise depicted in three values: 1. unsloping, 2. Flat and 3. Downslope	<i>Numeric</i>
Target	Heart disease diagnosis represented in 5 values,with 0 indicating total absence and 1 to 4 representing the presence in different degrees.	<i>Numeric</i>

The first step in building a predictive model is to prepare the data for analysis. This support increases model performance when data is converted into an understandable format.

3.3.1 Data Preprocessing

Data preprocessing is an important step in organizing data for model building. Data preparation involves several important steps, including function selection, data cleaning, and data transformation. Remove outliers, missing numbers, and noisy values from data using data cleaning and transformation techniques so that it can be easily used to build a model. We started by cleaning the data, which was a tedious process. The layer that converts text data to numeric data should be encrypted if the data set includes a null value. Using the median and imputation, the missing value problem is solved. If the data set has noise values while utilizing a box, we can tell that the numerical data contains some noise. We used normalization to finish the data transformation. We often get datasets that have a lot of features. We might not have all the data we require to create a machine learning model that can generate the required predictions from the dataset. Thus, careful feature selection is crucial when building a machine learning model.

Correlation: The statistical concept of correlation, which is often used to describe how nearly linear a relationship between two variables is. High correlation properties are more linearly dependent and therefore affect the dependent variable virtually equally. So, we can exclude one of the two features when there is sufficient correlation between the two features.

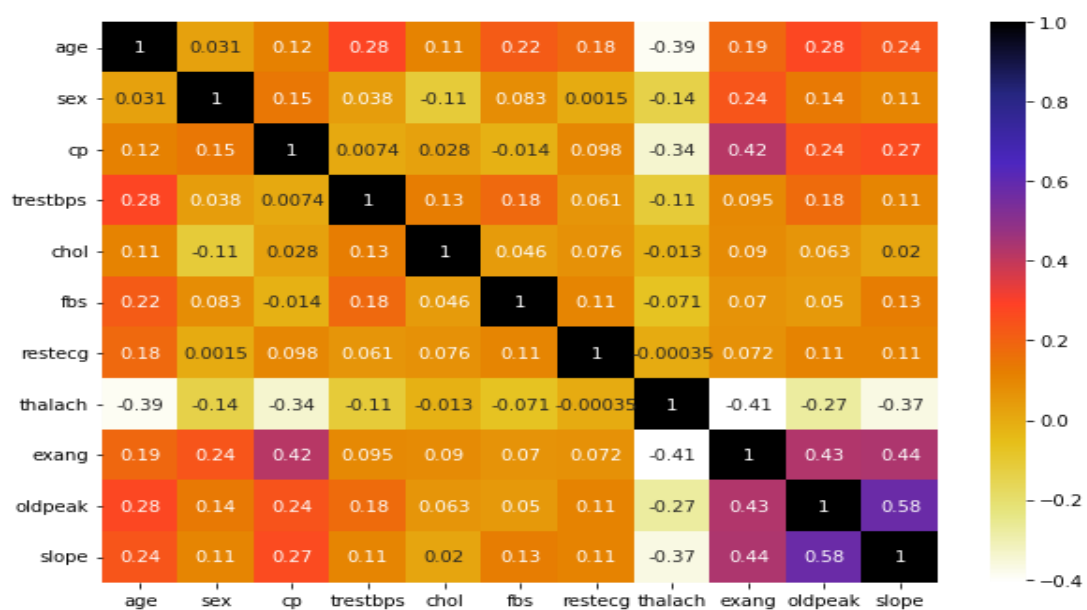


Figure 3.3: Co-Relations of Dataset

Outlier: An observation that differs from the majority of the data in a set is referred to as an outlier. Before eliminating our result function, which had dependent columns, we used outlier quantile detection to remove noisy numbers.

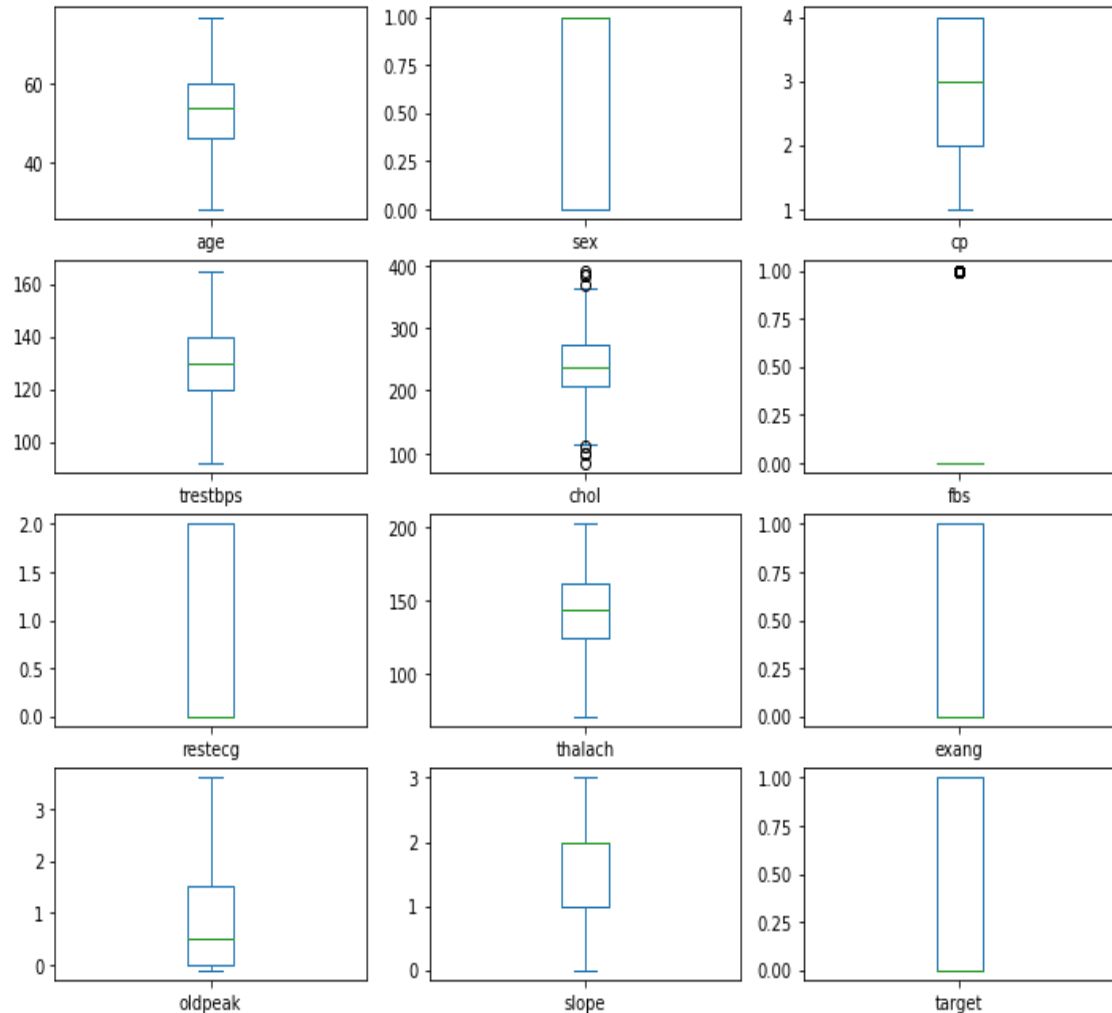


Figure 3.4: Remove outlier of dataset

As can be seen from the figure, outliers are excluded for age, sex, Trest_BPS, Chol, FBS, Restecg, CP, Thalach, Exang, Oldpeak and slope characteristics. However before removing the outliers I will check the dataset properties and their values. I'll start by using the formulas below to find the upper bound and lower bound using q_1 , q_2 , and IQR –

$$\text{Upper Limit} = q_3 + (1.5 * \text{IQR}) \tag{1}$$

$$\text{Lower Limit} = q_1 - (1.5 * \text{IQR}) \tag{2}$$

3.4 Statistical Analysis

We test and analyze this data in various ways after data collection. The first step is to pre-process the data. The dataset gathered for this research contains 1190, of which 74.3% male and 25.7% are female, respectively. The dataset should then be transformed to extract irregular data. We then selected an external factor and applied the selected algorithm to the collected dataset. A questionnaire was distributed to collect the dataset and the percentage results for all questions are shown in table 3.4.1. The table provides the questionnaire developed for this study 3.2.

TABLE 3.2: DESCRIPTIVE STATISTICAL ANALYSIS

	count	mean	std	min	25%	50%	75%	max
age	1190.0	53.720168	9.358203	28.0	47.0	54.0	60.00	77.0
sex	1190.0	0.763866	0.424884	0.0	1.0	1.0	1.00	1.0
cp	1190.0	3.232773	0.935480	1.0	3.0	4.0	4.00	4.0
trestbps	1190.0	132.153782	18.368823	0.0	120.0	130.0	140.00	200.0
chol	1190.0	210.363866	101.420489	0.0	188.0	229.0	269.75	603.0
fbs	1190.0	0.213445	0.409912	0.0	0.0	0.0	0.00	1.0
restecg	1190.0	0.698319	0.870359	0.0	0.0	0.0	2.00	2.0
thalach	1190.0	139.732773	25.517636	60.0	121.0	140.5	160.00	202.0
exang	1190.0	0.387395	0.487360	0.0	0.0	0.0	1.00	1.0
oldpeak	1190.0	0.922773	1.086337	-2.6	0.0	0.6	1.60	6.2
slope	1190.0	1.624370	0.610459	0.0	1.0	2.0	2.00	3.0
target	1190.0	0.528571	0.499393	0.0	0.0	1.0	1.00	1.0

3.4.1 Statistical Views of Coronary Artery Disease

These graphs illustrate how common heart disease is among healthy and affected individuals. 55.1% of people have heart disease, whereas 44.9% are healthy, according to the figure on the first. 517 healthy persons and 422 people with heart disease are shown in the second. The number of healthy individuals is 0, whereas those with heart disease are represented by the number 1.

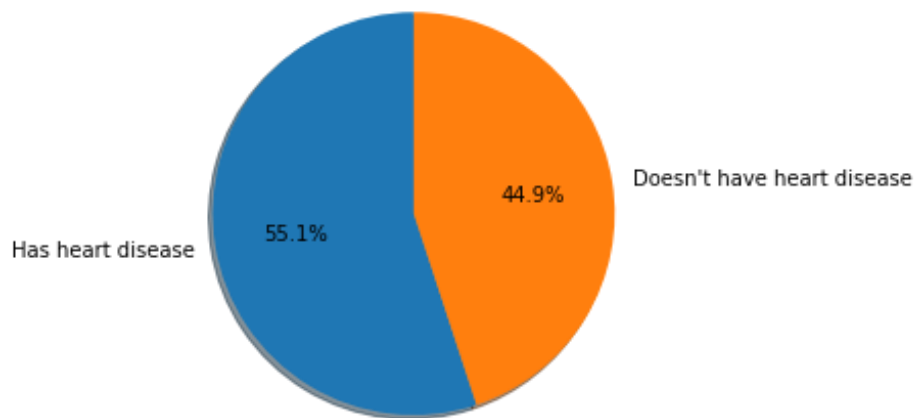


Figure 3.5: Statistical Views of Coronary Artery Disease

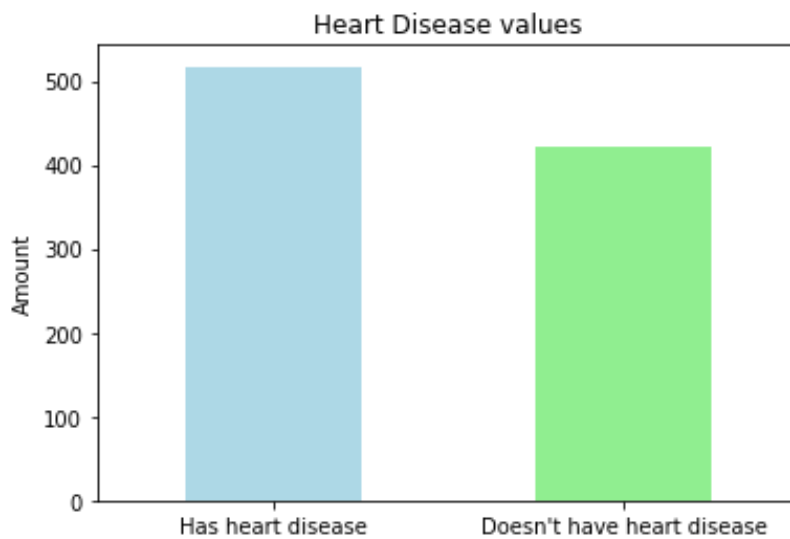


Figure 3.6: Statistical Views of Coronary Artery Disease

3.4.2 Statistical view Gender & Age Wise

The graph below's leftmost number reveals that the percentage of men in this dataset is significantly higher than that of females. The population consists of 26 percent females and 74 percent males. The age-wise distribution is represented by the figure on the right. Age is shown on the X axis while density is shown on the Y axis. The age of patients is around average 58.

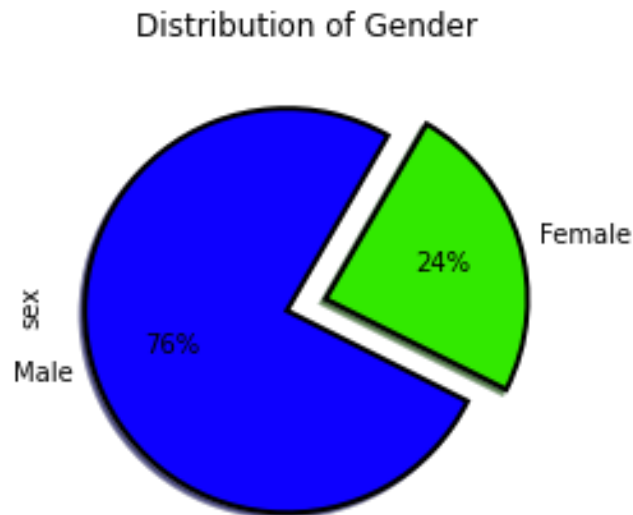


Figure 3.7: Statistical view Gender & Age Wise



Figure 3.8: Statistical view Gender & Age Wise

3.4.3 Gender and Age Wise Statistical view of Normal Patients & Heart Disease Patients

The graph below shows that male patients had more heart disease than female patients despite the average age of heart disease patients being between 57 and 60 years. The age and gender distribution of healthy people is shown in the top row, while that of heart disease patients is shown in the bottom row.

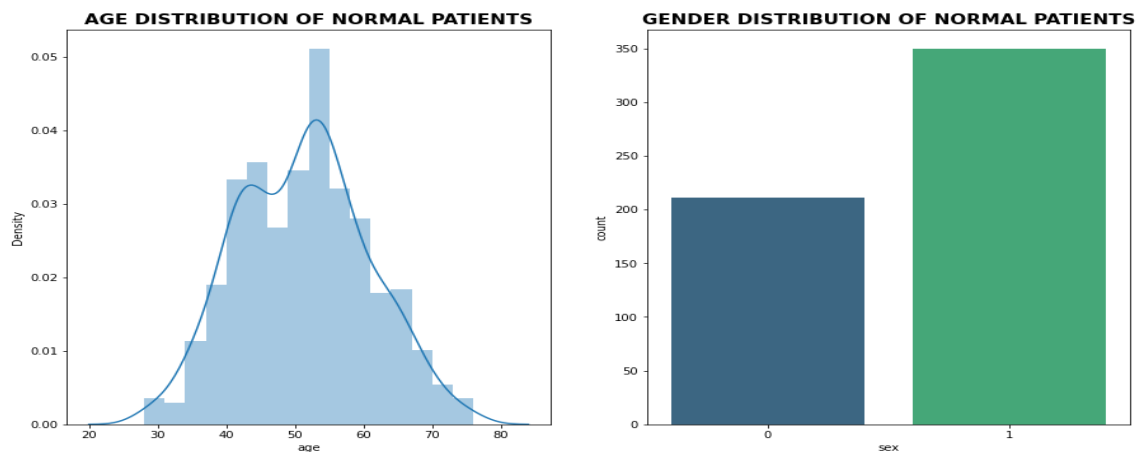


Figure 3.9: Statistical view Gender & Age Wise (normal patient)

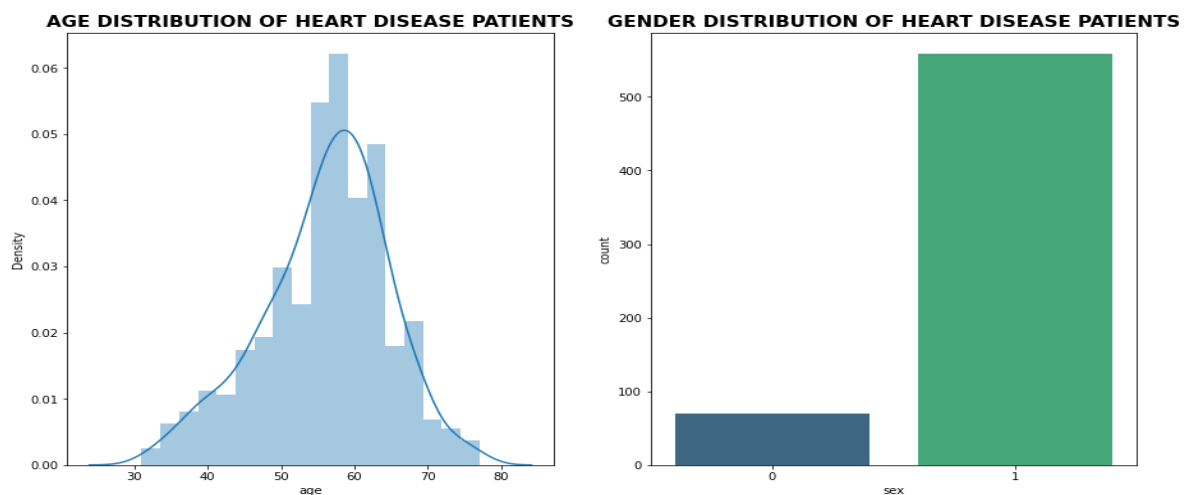


Figure 3.10: Statistical view Gender & Age Wise (disease patient)

3.4.4 Statistical view of Chest Pain Type

Chest discomfort is divided into four groups. The four categories of angina include asymptomatic, atypical angina, typical angina, and non-angina pain.

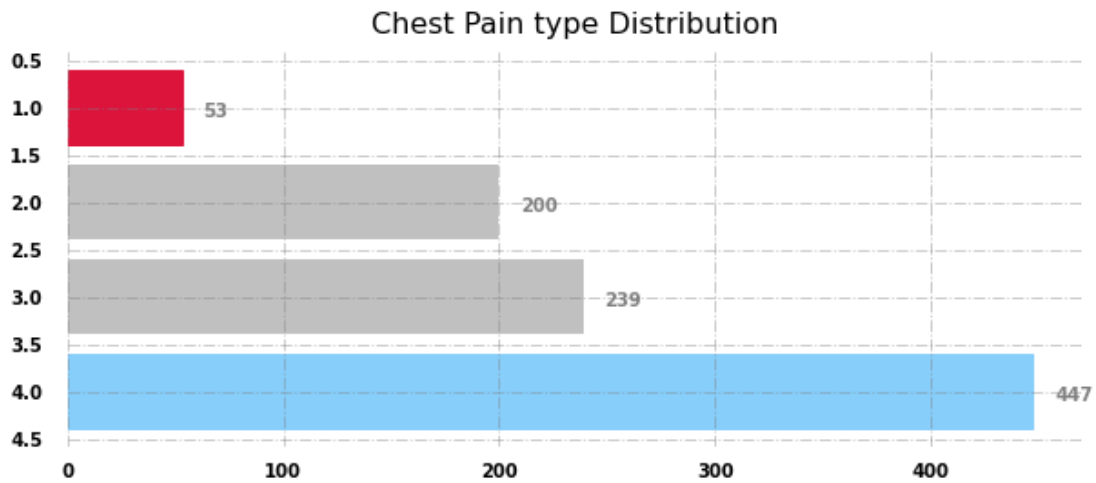


Figure 3.11: Statistical view of Chest Pain Type

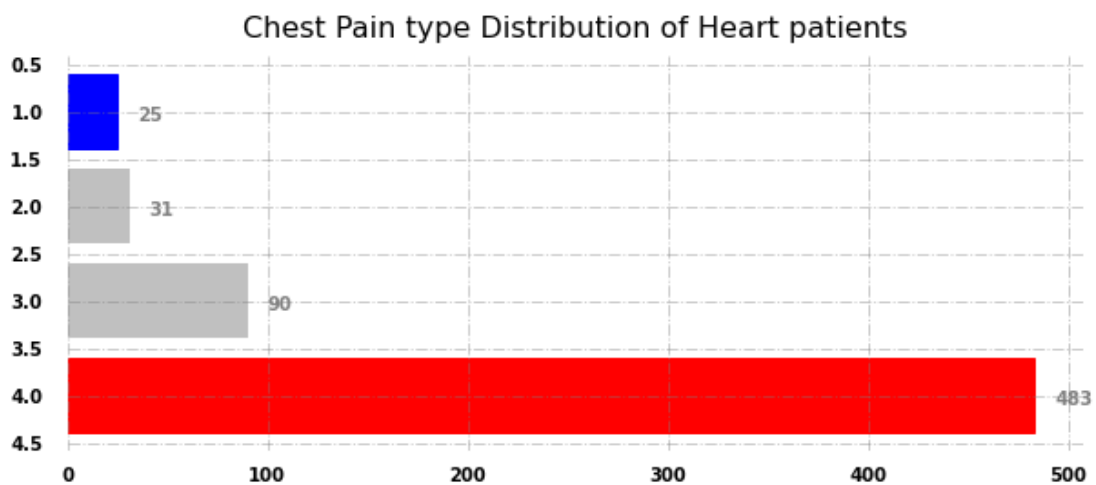


Figure 3.12: Statistical view of Chest Pain Type heart patient

3.4.5 Histogram

A histogram is a type of graph that uses bars with shifting stature to show data. The bars of a histogram separate the numbers into ranges. As the bars get higher, more information falls inside the run. A histogram speaks to the framing and scatter of continuous test data. At general, a histogram can be used anytime to compare the distribution of certain numerical data over different intervals. Examples of histograms can make complex concepts and patterns associated with a lot of data easier for the audience to observe and understand. It can assist in the decision-making process in numerous departments of a business or organization. This figure displays the amount and distribution of the continuous sample data from the dataset that was utilized for this study.

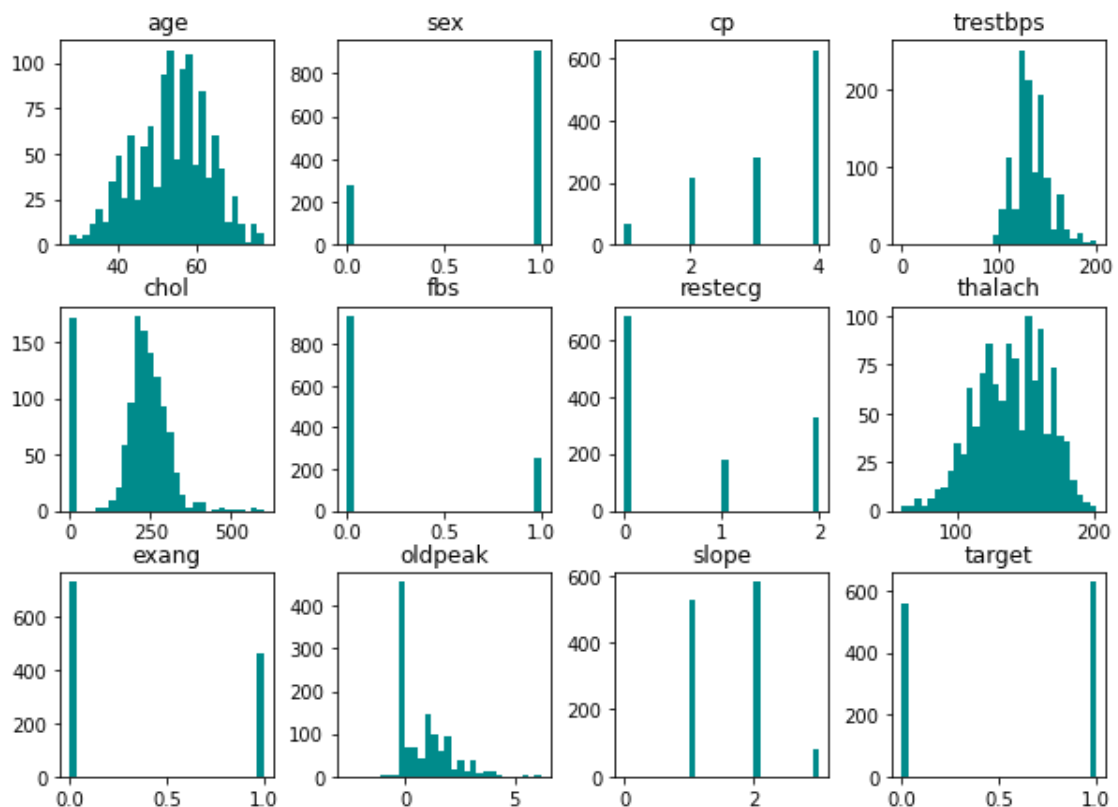


Figure 3.13: Features of histogram

Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slop, ca, thal, and target are all shown in this histogram.

3.5 Proposed Methodology

The proposed method demonstrates the processes required to derive a reliable algorithm for a patient's coronary artery disease prediction.

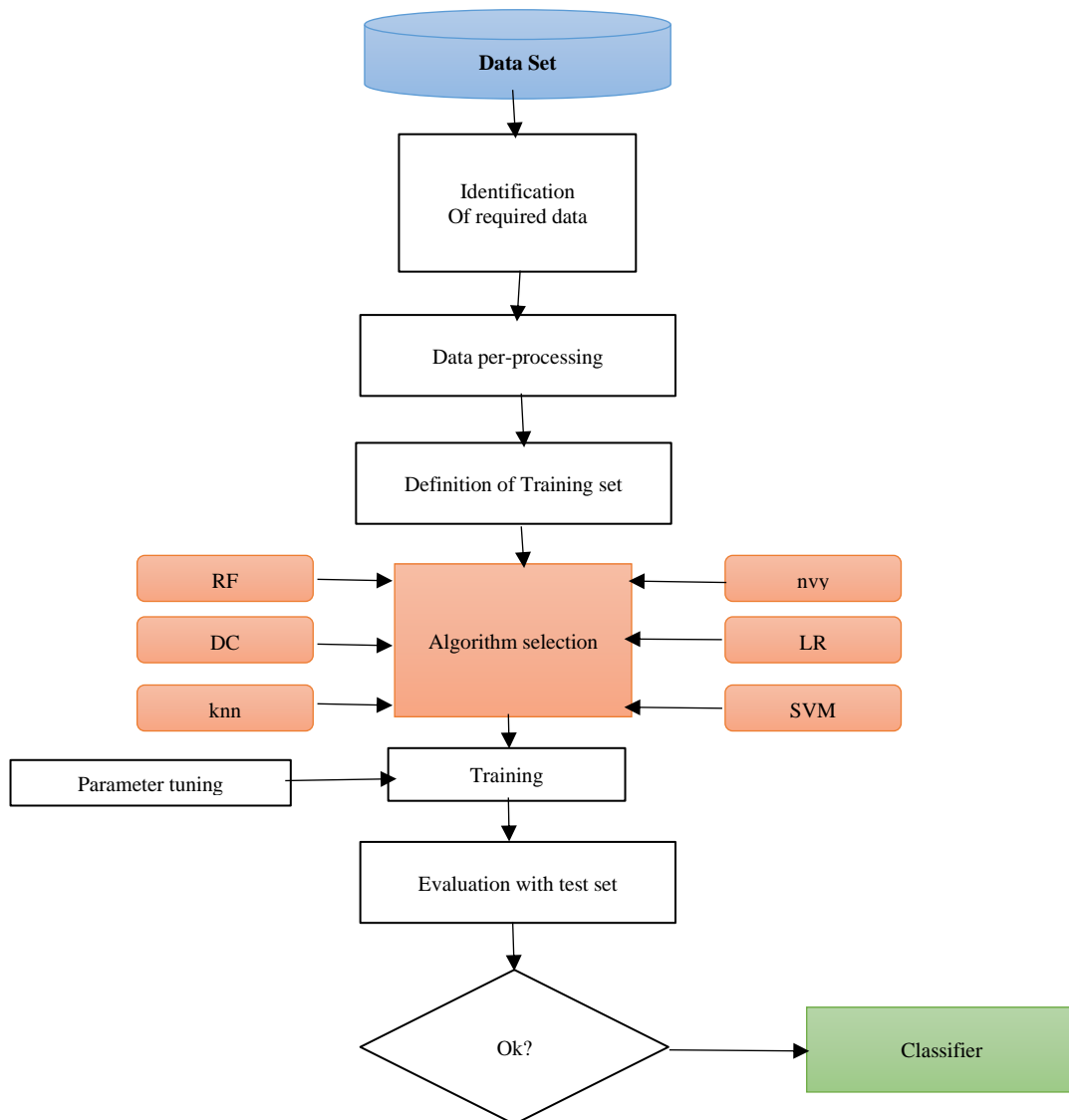


Figure 3.14: Flowchart of proposed work

The proposed system shows the steps taken to build a reliable method for a patient's coronary artery disease prediction. After collecting the dataset, we preprocessed the data before using it to build and observe multiple machine learning algorithms. Each ML computation is evaluated in two steps. The training set and test set together make up about 80% to 20% of the dataset. The model is used to fit the train data. During the training phase, the device can read the data structure model. If the selected model is

trained flawlessly, the machine provides fast feedback. Therefore, the test results must be represented. The test dataset enables this model to investigate reasonable hypotheses. The applicable dataset is split in an 80:20 ratio across the training and testing periods of this study. To run this model normally, the Python library is used. Other important components were then run to determine the best algorithm, such as: Accuracy, misclassification, Jaccard score, cross-validation, confusion matrix. After all necessary components were completed and described, the best algorithm was selected and all algorithms were evaluated.

3.6 Implementation Requirements

- Language - Python Version of 3.7.0
- Web application IDE - Google Colaboratory
- Library - Pandas analysis of data
- Library - Malplotlib visualization of data
- Library - Skit Learn of Machine Learning
- Microsoft office 2016
- Fundamental package for computing: NumPy
- Basic familiarity with computer programming
- A working understanding of Python

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

The author of this essay author invented a method to create a common set of items based on user symptoms. It helps identify the risk level for coronary artery disease based on external factors. The results help doctors predict the risk level of coronary artery disease patients. We have used six machine learning (ML) algorithms and used Random Forest (RF) as the best performer to achieve high accuracy. Artificial thought processes enable practitioners of a rich mental art and heart problems in this study, we present an overview of several models based on mapping and methodology and examine their performance Use targeted inspection methods to account for errors. Design templates are based on two or three supervised learning techniques. SVM, Decision Trees, Random Forest, Naive Bayes. Total counts performed will continue to be a fair indicator of disease progression. This classifier is also provided to determine the predictive value accurately as before.

4.2 Experimental Results & Analysis

The main objective of our proposed system is to foresee the possibility of coronary artery disease. Data mining and a variety of machine learning algorithms are increasingly being used to predict the severity of coronary artery disease. To apply these data mining techniques, data must be obtained and that data must be preprocessed with great precision. A total of 12 features were collected. Characteristics or symptoms in this study included age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope and target of breath. Next, we calculated accuracy coronary artery disease risk levels using a machine learning technique. Five machine learning classification techniques were used, and we obtained varying degrees of accuracy.

A matrix called confusion matrix is used to evaluate the performance of a model. A group of four terms Complexity matrix It is decided to appoint. This is the performance matrix:

True Positive (TP): When a result is The model successfully predicted the positive class.

True Negative (TN): A result when the class is negative correctly predicted by the model.

False Positive (FP): A result if the class is positive Misconceptions by the model.

False Negative (FN): A result when the class is negative Misconceptions by the model.

Accuracy: The percentage of test set tuples that were correctly classified by the classifier into that test set is the definition of a correct classifier on a given test set. It is easy to understand the evaluation of accuracy from an equation (i). Proportion of number of correct predictions the total number of instances given by the model.

$$Accuracy = \frac{(TP+TN)}{(TP+FT+FN+TN)} \quad (3)$$

Precision: You can think of precision as a way of evaluating the accuracy of an object (i.e, what percentage of tuples labeled as positive). The percentage of recovery cases that is truly significant is what it means in other words. The formula from which the accuracy is calculated is given by an equation (ii). In this work measures the proportion of individuals predicted to be at risk of developing CHD and had a risk of developing CHD.

$$Precision = \frac{(TP)}{(FP+TP)} \quad (4)$$

Recall/Sensitivity: A dataset's completeness, or the quantity of positive tuples found, is used by machine learning to assess its quality. A relevant instance is defined as the percentage of relevant instances that are located among all relevant instances discovered. Recall is calculated using the mathematical formula in equation (iii). Recall, in this task, measures proportion Individuals at risk of developing CAD and The algorithm predicted the risk of development CAD.

$$Recall = \frac{(TP)}{(FP+FN)} \quad (5)$$

F1 Score: The F measurement is also known as the weighted harmonic mean which is a technique for evaluating the precision and recall (for precision and recall) of a test. Equation (iv) contains the measurement-related mathematical formula F1- Score. F1 Score is the harmonic mean of precision and recall.

$$F1\ Score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \quad (6)$$

In machine learning, a categorization report is a statistic which used to assess the system's overall effectiveness. It is used with a training dataset to demonstrate the accuracy, recall, F1 score and support of a trained classification model. This metric represents the performance metrics of a classification-based machine learning model. This table displays the precision, recall, F1 score and accuracy. This gives a more realistic picture of the total performance of the trained model. All the metrics provided in the research study must be understood in order to understand the trained model.

Here is the Result Comparison given below (Accuracy Score, Precision Score, Recall Score, F1- Score) classification reports produced by the machine learning model.

TABLE 4.1: PERFORMANCE STATISTICS

Algorithm	Precision	Recall	F1-Score	Accuracy (%)
Random Forest	92.21 %	94.67 %	93.42 %	94.68 %
Decision Tree	86.25 %	92.0 %	89.03 %	90.96 %
Logistic Regression	79.73 %	78.67 %	79.19 %	83.51 %
Naive-Bayes Classifier	78.75 %	84.0 %	81.29 %	84.57 %
K-Nearest Neighbor	64.0 %	64.0 %	64.0 %	71.28 %
Support Vector Machine	69.84 %	58.67 %	63.77 %	73.4 %

I trained six different models. Results are produced by each model. To choose the best ML model for this study problem, I will compare the results of all models using precision score, recall score, F1-score and performance score as my metrics.

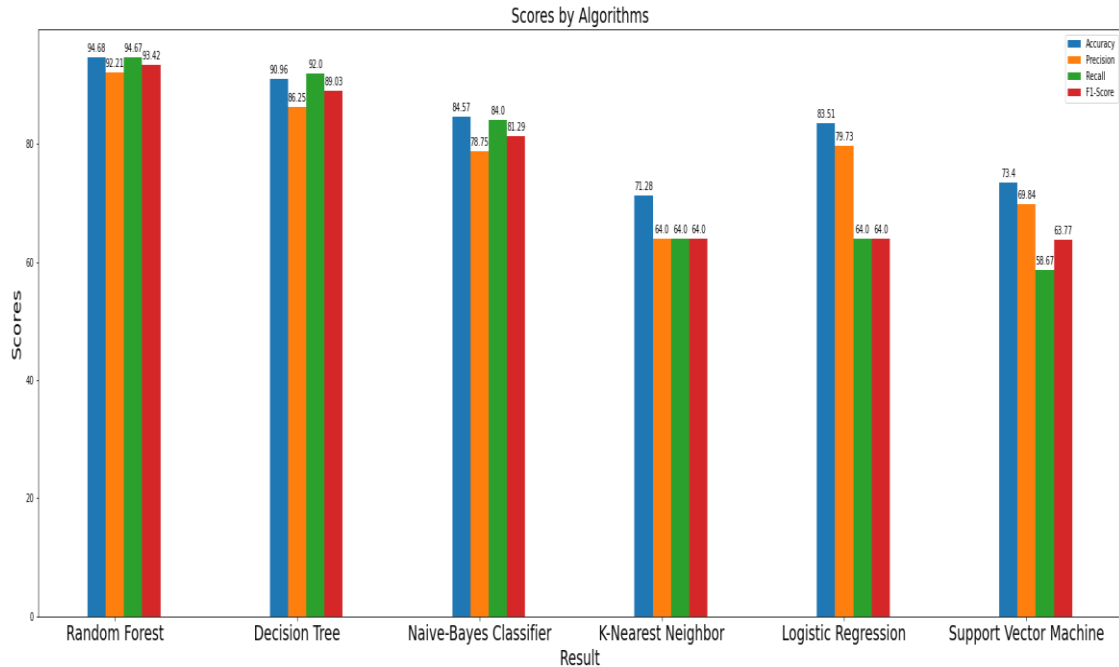


Figure 4.1: Accuracy chart

Figure 4.1 shows that among all methods, the RF algorithm has the highest accuracy 94.68, F1-score 93.42 and precision score 92.21. Overall, it is the best approach for this problem despite having a higher accuracy score than the KNN, LR, and SVM algorithms.

4.3 Descriptive Analysis

We evaluated the accuracy of several algorithms as well as their curve, SEM, precision, re-call and f1-score. Any product line needs to provide a model evaluation. In model evolution, some classifications must be evaluated. Classifiers are developed using test data collection for improved measurement.

The following is our confusion matrix for each machine learning method.

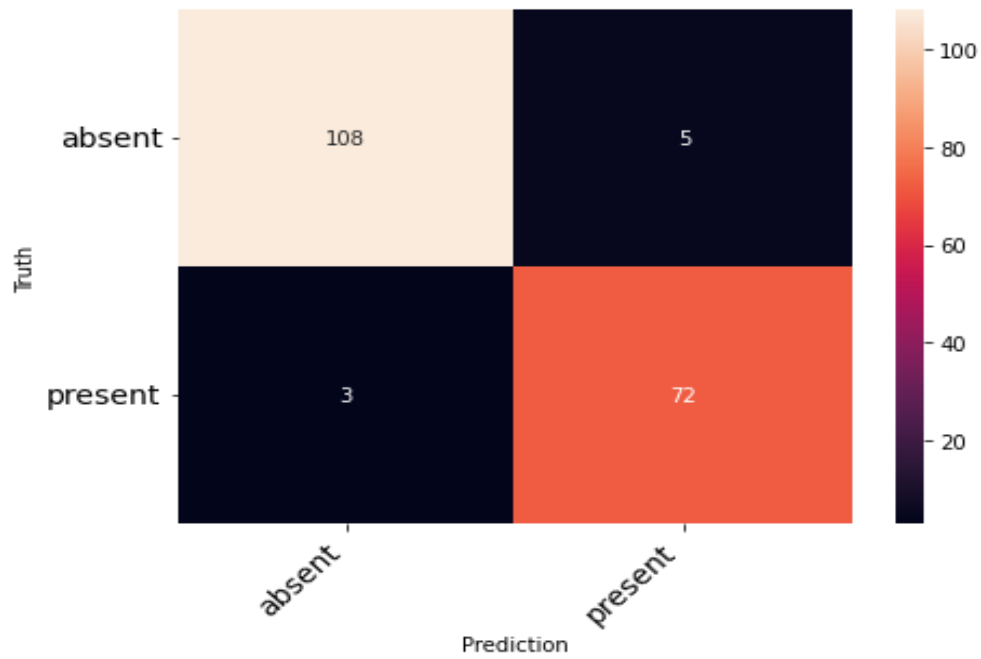


Figure 4.2: Confusion matrix of Random Forest

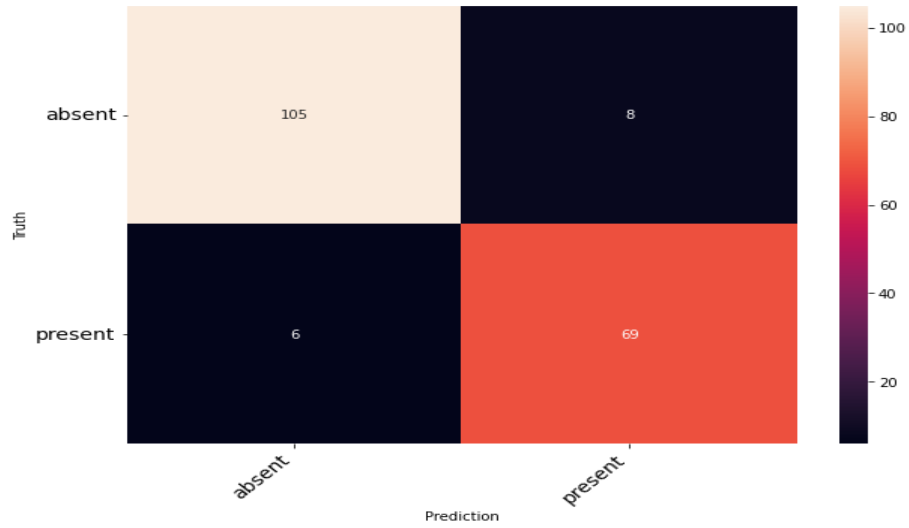


Figure 4.3: Confusion matrix of Decision Tree

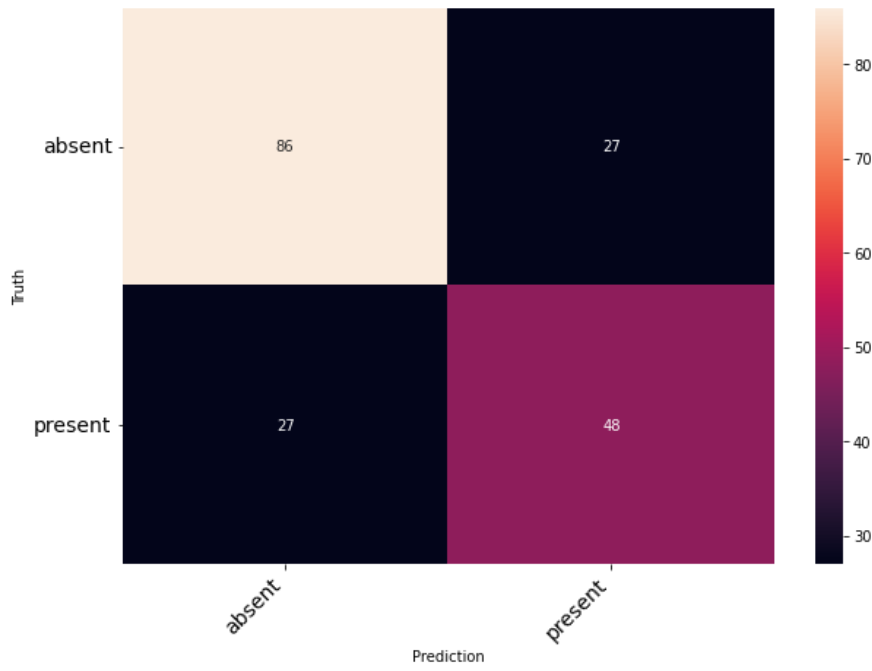


Figure 4.4: Confusion matrix of K-Nearest Neighbors

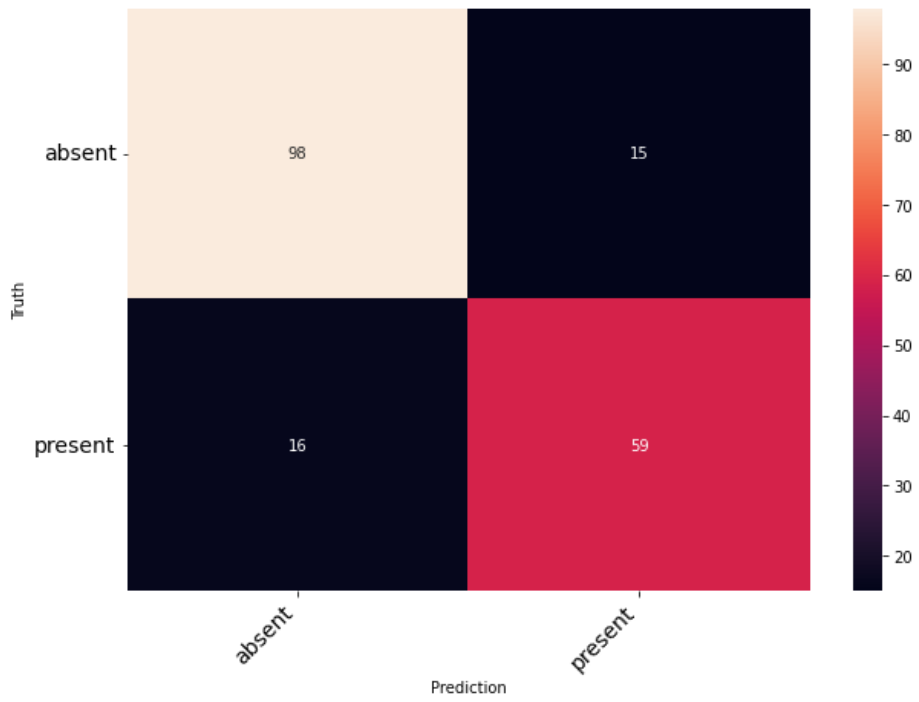


Figure 4.5: Confusion matrix of Logistic Regression

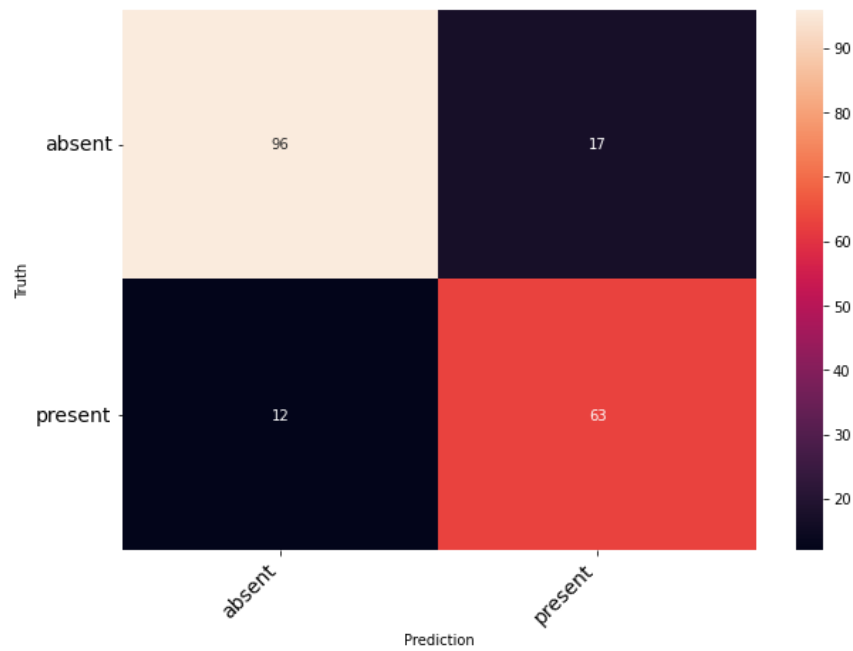


Figure 4.6: Confusion matrix of Naïve Bayes

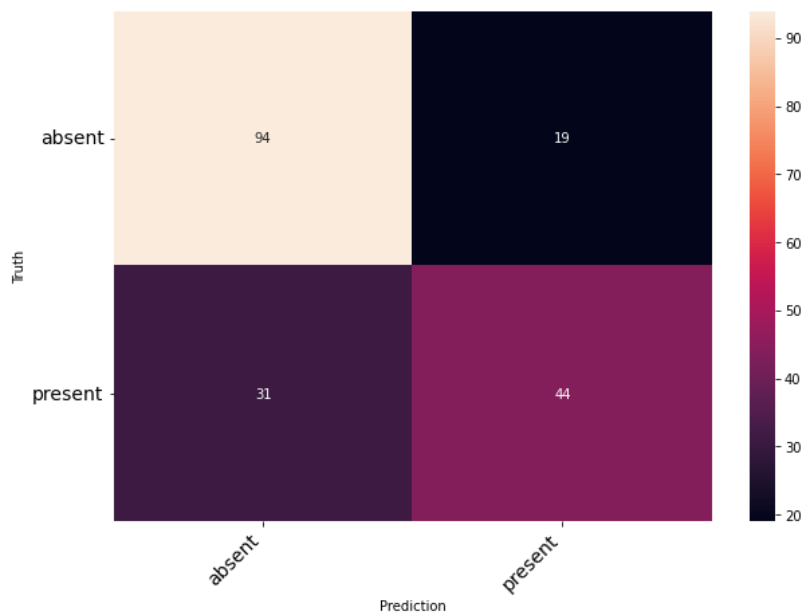


Figure 4.7: Confusion matrix of Support Vector Machine

AUC-ROC curve: An evaluation statistic for the performance of a classification model at different threshold levels is the AUC-ROC curve. AUC-ROC Curve is another tool I've used to assess models. The AUC value has a range from 0 to 100. An improved model is indicated by a higher AUC value.

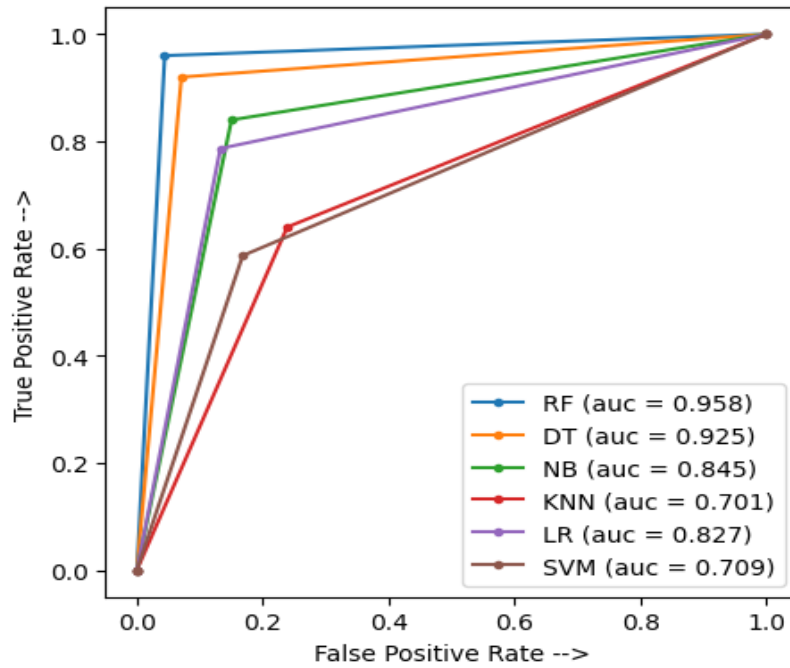


Figure 4.8: ROC-AUC curve

TABLE 4.2: ROC-AUC CURVE

Algorithm	AUC-value
Random Forest	0.958
Decision Tree	0.925
Naïve-bayes	0.845
K-Nearest Neighbor	0.701
Logistic Regression	0.827
Support Vector Machine 0.904	0.709

K-Fold Cross Validation: Cross-validation is a machine-learning assessment technique used to determine how effectively your machine-learning model can predict data outcomes that have not yet been observed. It is a popular option because it is easy to understand, performs well for a small data sample, and provides an estimate that is less biased.

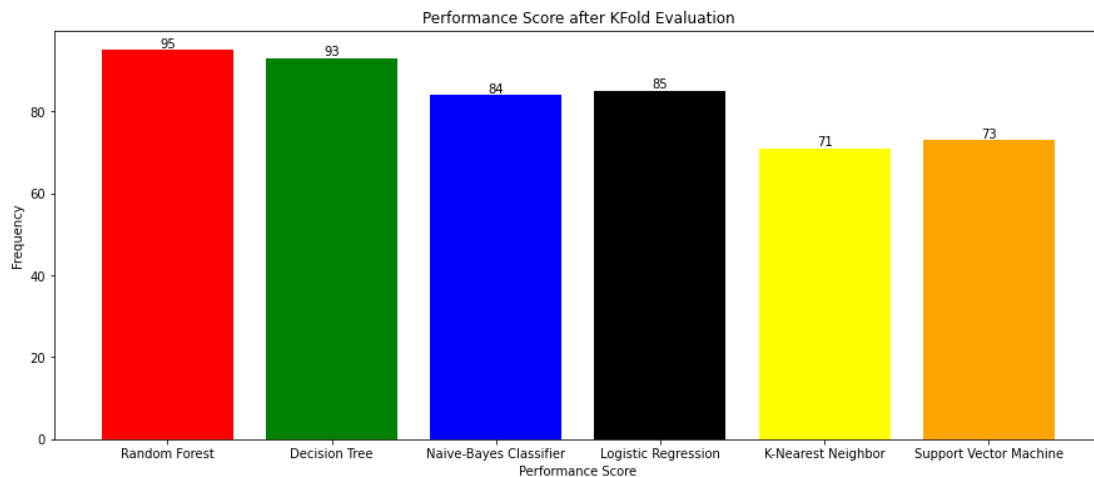


Figure 4.9: Result Evaluation Score (After K-Fold Evaluation)

K-fold cross-validation was used to evaluate the models, and the results showed that RF had the highest score among all techniques, with Decision Tree ranking second. Among the algorithms selected for model training, Logistic Regression gets the lowest performance rating.

TABLE 4.3: RESULT EVALUATION SCORE (AFTER K-FOLD EVALUATION)

Algorithm (Model)	Score (In Percent)
Random Forest	95%
Decision Tree	93%
Naïve Bayes	84%
K-Nearest Neighbor	85%
Logistic Regression	71%
Support Vector Machine	73%

4.4 Discussion

This thesis' major objective is to inspire others to take better care of themselves by presenting a prediction based on past performance. So that people can understand how their lifestyle and eating habits affect their risk of coronary artery disease. While they should take proper measures to maintain heart health. As we have collected information from many categories including seventy thousand, our thesis cannot provide any physical support or doctor's advice, through our accuracy and projections, people can

learn about the condition of their own hearts. Then, this data was carefully organized and put into an Excel spreadsheet. We used machine learning techniques with this data, including K Nearest Neighbor, Random Forest, Naive Bayes, Decision Tree, and SVM. In Tables 4.1 the absolute accuracy of the data is displayed. Among the machine learning techniques, we found that Random Forest and Decision Tree have the highest accuracy. Our best accuracy in deep learning algorithms is achieved with random forests.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Coronary artery disease is one of the major diseases in the world today. The majority of people in our society do not have adequate money to treat this illness because Bangladesh is a low-income nation. A significant contributor to early death is cardiovascular disease, along with other chronic conditions. Reason one. There are many people around us who have coronary artery disease but they don't know it because of the high cost of treatment. The system we developed predicts coronary artery disease at minimal cost and gives almost accurate results. Therefore, those with heart problems that go undiagnosed due to high medical costs can easily get their disease diagnosed at no additional cost. With the help of this system people can get quality services at affordable cost. Our system was developed using appropriate computer-based data and decision support technology, making it easy to achieve minimal clinical trial costs. As a result, the victims of our society are able to save precious lives and the death rate is decreasing day by day. There are many causes of this disease, including increased and decreased caloric intake, fatty foods, lack of exercise and improper diet. If we recognize the possibility of heart disease at an early age, we can be more careful and avoid harmful foods and habits, which are thought to have some influence on society.

5.2 Impact on Environment

The information used in this project is related to healthcare. This includes patient information. The system can analyze data and use machine learning techniques to make clinical decisions. This clinical decision-making, aided by computerized patient records, has the potential to improve healthcare systems by lowering clinical variability, enhancing patient safety and results, and eliminating medical errors. An environment that fosters cognition is created via machine learning, a powerful tool for data modeling and analysis. Enhance the standard of clinical judgments. This computerized background replaces the physician's practice of making clinical decisions based on

intuition and experience. In addition, this approach allows consumers to learn about risk factors for coronary artery disease, including air pollution, food workers, and water contamination with arsenic. In this way, you can raise awareness about environmental factors that contribute to coronary artery disease. We are working with medical data and trying to solve a problem in the healthcare sector, so the preferred environment in this case is the healthcare sector. Detecting coronary artery disease early, or even later, can help doctors. Doctors already use many types of diagnostic tests, and these tests can be added. Add another check to the test flow.

5.3 Ethical Aspects

It is an attempt, as individuals when we summarize a glimpse of the content, we tend to read it together, try to understand better and come up with a summary that captures the main points in a short time. Since PCs require human data and language barriers, changing the cross diagram is very problematic and not a trivial task.

- Do not disturb the patient.
- Protection of personal information
- Do not discriminate based on ancestry
- Equity in study planning.
- Accept accountability for study findings.

5.4 Sustainability Plan

The Custom Content Graph is a powerful review field that has a variety of business uses. In social events, a large amount of information in a concise form is useful for various downstream applications. Create news summaries, reports, news charts, and age-specific highlights. There are two known types of frame assumptions.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

The main motivation of this paper is to understand the extent of coronary artery disease and to be aware of their lifestyle, partly by reducing mortality. No medical therapy of any kind can be offered through our scheme. Think about it, eats a high-fat diet every day and does not need exercise. By applying our scheme, we can understand the degree of coronary artery disease risk and, consequently, the heart health that protects individuals from coronary artery disease. Data were collected from different hospitals and typically written down to apply the machine learning algorithm. Another way to collect information is the discharge summary for each patient. In this way, a total of 12 factors were collected from approximately 1190 data. This collected data was efficiently organized and organized in place in Excel. The use of this information is typically subject to various machine learning algorithms. 12 features such as age, gender, blood pressure, and smoking are extracted from the dataset to predict a patient's likelihood of developing coronary artery disease. These credits are reserved for RF, DT, KNN, SVM, and Naive Bayes grouping algorithms, as shown in Table 3. Here RF gave best results with best accuracy. The proposed algorithm outperforms other algorithms in accuracy in the literature review. From previous studies and Table 1, no study has a reference algorithm success rate higher than 86% for collected cardiac datasets. Considering the results review, it indicates that the proposed model provided meaningful results in the allocation of patients with probable coronary artery disease. The implementation of the proposed method is contrasted with existing algorithms and mastery of the proposed method.

6.2 Conclusions

One of the various illnesses that impact us is coronary artery disease. Indeed, it is a significant condition given that the majority of fatal diseases nowadays are heart disease and other conditions related to the heart [19, 20, 21]. It is the main cause of death on a

global scale. In accordance with the World Health Organization (WHO), 31% of all pedestrians worldwide died from coronary artery disease in 2012, killing 17.5 million people. 16 million persons under the age of 70 are anticipated to be affected by NCDs, and 82% of these individuals reside in lower middle-income nations. A stroke caused 6.7 million deaths, while coronary artery disease caused 7.4 million [22]. Therefore, the optimal treatment for this illness must be discovered. However, stopping it overnight is not simple.

6.3 Recommendations

Each classification algorithm was trained and evaluated in my experiment, which is associated with most papers, on a training dataset that includes both positive values and negative values. In addition, by gathering information from various devices and health-related sensors, clinical and medical facilities and providing more precise results for disease prediction and diagnosis, the work can be beneficial for chronic disease diagnosis and detection. In my opinion, there are many directions for future work in this area of study. I've only seen a few well-known supervised machine learning algorithms. More algorithms can be used to develop more accurate models for predicting various chronic diseases and performance can be improved. I also highlighted research trends and opportunities related to liver disease research and medical data analysis using ML-based techniques.

6.4 Implication for Further Study

In this study, we suggested a method that can understand the prognostic value of coronary artery disease. We are unable to provide medical support in relation to this research. However, we want to introduce a news section that will feature medicine names based on coronary artery disease in the future. Add another chapter that will create information for cardiologists so that individuals can contact cardiologists for consultation. Add a variety of sensors to assess diabetes, high blood pressure and cholesterol.

REFERENCES

- [1] Wong, Nathan D. "Epidemiological studies of CHD and the evolution of preventive cardiology." *Nature Reviews Cardiology* 11.5 (2014): 276-289.
- [2] Shi, Aimin, et al. "Epidemiological aspects of heart diseases." *Experimental and therapeutic medicine* 12.3 (2016): 1645-1650.
- [3] Hansson, Göran K. "Inflammation, atherosclerosis, and coronary artery disease." *New England journal of medicine* 352.16 (2005): 1685-1695.
- [4] Wise, Frances M. "Coronary heart disease: The benefits of exercise." *Australian family physician* 39.3 (2010): 129-133.
- [5] Kumar, Dinesh. *Automatic heart sound analysis for cardiovascular disease assessment*. Diss. Universidade de Coimbra (Portugal), 2014.
- [6] World Health Organization. *Global status report on noncommunicable diseases 2014*. No. WHO/NMH/NVI/15.1. World Health Organization, 2014.
- [7] Gjoreski, Martin, et al. "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers." *2017 International Conference on Intelligent Environments (IE)*. IEEE, 2017.
- [8] Babu, Sarath, et al. "Heart disease diagnosis using data mining technique." *2017 international conference of electronics, communication and aerospace technology (ICECA)*. Vol. 1. IEEE, 2017.
- [9] Banu, MA Nishara, and B. Gomathy. "Disease forecasting system using data mining methods." *2014 International conference on intelligent computing applications*. IEEE, 2014.
- [10] Purusothaman, G., and P. Krishnakumari. "A survey of data mining techniques on risk prediction: Heart disease." *Indian Journal of Science and Technology* 8.12 (2015): 1.
- [11] Burke, Lora E., et al. "Current science on consumer use of mobile health for cardiovascular disease prevention: a scientific statement from the American Heart Association." *Circulation* 132.12 (2015): 1157-1213.
- [12] Alagugowri, S., and T. Christopher. "Enhanced Heart Disease Analysis and Prediction System [EHDAPS] Using Data Mining." *International Journal of Emerging Trends in Science and Technology* 1 (2014): 1555-1560.
- [13] Lubna, Shamshad Rahman. "Predicting coronary heart disease through risk factor categories." *ASEE 2014 Zone I Conference*. 2014.

- [14] Pouriyeh, Seyedamin, et al. "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.
- [15] Masethe, H.D., Masethe, M.A.: Prediction of heart disease using classification algorithms. In: *World Congress on Engineering and Computer Science Vol II WCECS, San Francisco, USA, 22–24 Oct 2014*
- [16] Mahmoodabadi, Zahra, and Mohammad Saniee Abadeh. "CADICA: Diagnosis of coronary artery disease using the imperialist competitive algorithm." *Journal of Computing Science and Engineering* 8.2 (2014): 87-93.
- [17] Muthukaruppan, S., and Meng Joo Er. "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease." *Expert Systems with Applications* 39.14 (2012): 11657-11665.
- [18] Y. N. Devi and S. Anto, "An Evolutionary-Fuzzy Expert System for the Diagnosis of Coronary Artery Disease," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 3, no. 4, pp. 1478–1484, 2014.
- [19] Babu, Sarath, et al. "Heart disease diagnosis using data mining technique." *2017 international conference of electronics, communication and aerospace technology (ICECA)*. Vol. 1. IEEE, 2017.
- [20] Banu, MA Nishara, and B. Gomathy. "Disease forecasting system using data mining methods." *2014 International conference on intelligent computing applications*. IEEE, 2014.
- [21] Krishnaiah, V., "Diagnosis of heart disease patients using fuzzy classification technique." IEEE International Conference on Computer and Communications Technologies (ICCCT), 2014.
- [22] WHO (2015). Cardiovascular diseases. Retrieved August 28, 2015 from <http://www.who.int/mediacentre/factsheets/fs317/en/>. [Accessed on 19 July 2019].
- [23] Nassif, A. B., Mahdi, O., Nasir, Q., Talib, M. A., & Azzeh, M. (2018, November). Machine learning classifications of coronary artery disease. In *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). IEEE.

APPENDIX

This is my first study in the ML field. I knew very little about ML algorithms before this. So, throughout this study, I encountered various problems. At one point, getting enough datasets was also challenging. Here, I had to rely on a secondary dataset. Therefore, data collection was one of my most difficult tasks. Other difficulties came up when I dealt with the models, coded for data visualization, normalization and displayed the results. I managed to overcome the difficulties, though. I had to spend a lot of time studying several articles related to my credit risk analysis and learning about ML from scratch as I had no expertise in this field. As a result, I was able to effectively complete this study while also learning a significant amount and developing my skills.

Plagiarism Report

Bivuti_Vushan_Bhadra.pdf

ORIGINALITY REPORT

13%

SIMILARITY INDEX

10%

INTERNET SOURCES

6%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to Daffodil International University Student Paper	2%
3	Divya Krishnani, Anjali Kumari, Akash Dewangan, Aditya Singh, Nenavath Srinivas Naik. "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms", TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019 Publication	1%
4	ieeexplore.ieee.org Internet Source	1%
5	doctorpenguin.com Internet Source	1%
6	Md. Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni. "Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis	1%