EVALUATING MACHINE LEARNING ALGORITHMS FOR HEART DISEASE PREDICTION: AN EXPLORATORY STUDY

BY

MD. SOFIQUL ISLAM
ID: 191-15-2508
AND

SHAH MOAZZAM HOSSEN SHIBLEE
ID: 191-15-2507

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

Supervised By

Amatul Bushra
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Mahfujur Rahman
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY, 2023

APPROVAL

This Project titled **"Evaluating Machine Learning Algorithms For Heart Disease Prediction: An Exploratory Study"**, submitted by Md.Sofiqul Islam, ID No: 191-15-2508 and Shah Moazzam Hossen Shiblee, ID No: 191-15-2507 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 01 February 2023.

BOARD OF EXAMINERS

Chairman

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

Internal Examiner

Tania Khat

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

Internal Examiner

Lamia 1.2.23

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

External Examiner

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

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Amatul Bushra, 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:

21.01.23

Name: Amatul Bushra
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Name: Md. Mahfujur Rahman
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:

Sofiqul

Name: Md. Sofiqul Islam
ID: 191-15-2508
Department of CSE
Daffodil International University

Shiblee

Name: Shah Moazzam Hossen Shiblee
ID: 191-15-2507
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 **Amatul Bushra**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Amatul Bushra, Md. Mahfujur Rahman, and 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

One of the significant motives of death globally is a heart-attack. Heart-attack detection early on can save lives. In this exploration, we mention a predictive machine learning algorithm of heart attack. The algorithm is trained on a dataset of worldwide patients which is taken from Kaggle. The dataset includes features such as gender, smoking status, age, cholesterol level, systolic pressure, and familial heart-disease history. Heart-attack prediction is a difficult problem due to the complex nature of the data and the lack of understanding of the underlying causes of the disease. However, predictive models that can be used to identify people at high risk can be created using the machine learning algorithm. This study employed a machine learning system to forecast cardiac attacks in a sizable population. A set of clinical and demographic variables served as the training ground for the algorithm. In the test dataset, the results demonstrated that the algorithm was capable of accurately predicting heart attacks. Cardiovascular failure is the main source of death around the world. Heart attack detection early can save lives. The algorithm can be used to predict heart attack in future patients. Heart attack prediction is a challenging problem in the machine learning field. This report's objective is to develop a machine learning system algorithm that is able to accurately forecast the occurrence of heart attacks. The dataset used for this purpose contains information on various risk factors such as age, gender, smoking habits, medical history, etc. The machine learning algorithm is trained on this dataset and is able to accurately predict the occurrence of heart attacks. The Decision Tree algorithm is able to achieve an accuracy of 99.70% in predicting heart attack. In future, we will do more research about heart attack & will make an android app so that people can easily detect their disease. It will help everyone to predict their disease.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-6**
1.1 Introduction	1-2
1.2 Motivation	2-4
1.3 Rationale of the Study	4
1.4 Research Questions	4-5
1.5 Expected Output	5
1.6 Project Management and Finance	5-6
1.7 Report Layout	6
CHAPTER 2: BACKGROUND	**7-14**
2.1 Preliminaries	7
2.2 Related Works	7-11
2.3 Comparative Analysis and Summary	11-13
2.4 Scope of the Problem	14
2.5 Challenges	14
CHAPTER 3: RESEARCH METHODOLOGY	**14-28**

3.1 Research Subject and Instrumentation	15
3.2 Data Collection	15-16
3.3 Statistical Analysis	16-17
3.4 Proposed Methodology	17-18
3.4.1 Dataset collection	18
3.4.2Dataset labeling	19
3.4.3 Data preprocessing	19-20
3.4.4 performance measures	20-21
3.5 Implementation Requirements	21-28
3.5.1 Decision Tree	21-23
3.5.2 Random Forest	23-24
3.5.3 Logistic Regression	24-25
3.5.4 K-nearest neighbor	25-26
3.5.5 Naive Bayes	26-28
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	**29-32**
4.1 Experimental Setup	29
4.2 Experimental Results & Analysis	29-31
4.3 Discussion	31-32
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	33-34
5.1 Impact on Society	33
5.2 Impact on Environment	33-34

5.3 Ethical Aspects	34
5.4 Sustainability Plan	34
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	**35-36**
6.1Summary of the Study	35-36
6.2Conclusions	36
6.3Implication for Further Study	36
REFERENCES	37-38

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.2.1 The collection of datasets	16
Figure 3.3.1 Accuracy of all algorithms	17
Figure 3.4.1 System Design	18
Figure 3.5.1 Decision Tree	22
Figure 3.5.2 Random Forest	23
Figure 3.5.3 Logistic Regression	24
Figure 3.5.4 K-nearest neighbor	26
Figure 3.5.5 Naive Bayes	27
Figure 4.2.1: Report on precision recall and f1 score	30
Figure 4.2.2 Structure of Decision Tree	31
Figure4.3.1 confusion matrix	32

LIST OF TABLES

TABLES	PAGE NO
Table 2.3.1. Comparative analysis with previous work	11-13
Table 4.2.1 test Result	29
Table 4.2.2 Model performance Table	30

CHAPTER 1

Introduction

1.1 Introduction

In Bangladesh, Cardiovascular disease is one of the primary causes of mortality. Every year, thousands of people die from heart attacks. Most of these deaths could be prevented if people knew they were at risk and took steps to reduce their risk. Machine learning algorithms have the potential to predict who is at risk of a heart attack. Doctors may be able to determine who is most likely to have a heart attack with the assistance of these predictions. This research paper aims to use machine learning algorithms to figure out who in Bangladesh is at risk for a heart attack. The paper will firstly review the literature on machine learning algorithms and heart disease. It will then describe the data that will be used in the study. Finally, it will discuss the results of the study and the implications for Bangladesh. Heart failure is one of the top causes of mortality globally. Heart disease can be detected and diagnosed earlier, which can increase the likelihood of successful treatment and improve patients' quality of life. AI calculations can possibly give precise forecasts of coronary illness. Machine learning is a subfield of artificial intelligence. Algorithms that allow computers to continuously improve their performance in response to data are the focus of this field. In order for a computer to learn, it must examine its prior knowledge in order to uncover helpful patterns and regularities, even those that a person may overlook. Creating models, like patterns and rules, automatically from data is a primary goal of machine learning research. Data warehouses and subjects like "Inductive argument, analytics, pattern classification, data mining, and conceptual computer science" are closely connected to machine learning. Algorithms for machine learning have been used in the past to diagnose cardiac problems. Sadly, the specificity, sensitivity, and accuracy were extremely low. The art of machine learning involves controlling a system without using explicit computation. They are used to examine the analytical setup in large-scale, varied data sets, such as those pertaining to heart ailments. They are utilized in the identification of arrangements (patterns) that enable forecasting and control mechanisms for analysis and medicine. This study, we will utilize machine learning to develop algorithms that can predict heart attacks. The model will be trained on a data set of patient medical records.

We will then evaluate the accuracy of the model and compare it to other prediction models. Early prediction of heart attack can save everyone in Bangladesh & worldwide. The design and development of algorithms that are capable of learning from data and making predictions is the subject of the machine learning subfield of artificial intelligence. In this exploration, we propose an AI calculation for the expectation of respiratory failure. At first, we have tried many kinds of machine learning algorithm to make a model. We have tried Naïve Bayes, K-Nearest Neighbor, Decision Tree, Random Forest, Logistic Regression, &Logistic Regression. Among all, decision tree performed the high accuracy for us. This accuracy was 99.70%. The algorithm is trained on a dataset of patients with heart disease and is tested on a separate dataset. The results show that the algorithm is able to accurately predict heart attacks. The following is a summary of this work's contributions:

- For the purpose of diagnosing heart disease, a pre-trained machine learning model-based framework is proposed. The following are the portions of this report:
- A compare analysis with other researcher research's work
- Identifying the best performance of each model and then evaluating it on the test set again to ensure the model is more generalizable.

The methods used in this report are described in Section 2. In part 3, the quantitative results and discussion are presented, and in chapter 4, the method used for comparison with previous studies is shown. In chapter 5, the study's conclusion is finally presented.

1.2 Motivation

Cardiovascular illness has been one of the most ordinary causes of death in the medical sector. The number of persons suffering from cardiovascular illnesses is growing annually as a result of an improvement in people's living conditions and an increase in life pressure. Acute cardiovascular disorders and chronic cardiovascular diseases make up the majority of cardiovascular diseases. The computer technique is crucial in the treatment of cardiovascular disorders since traditional wet-lab studies used to diagnose cardiovascular problems frequently prove to be ineffective and time-consuming. Cardiovascular disease, often known as heart disease, is regarded as a lethal condition that is becoming worse faster in our contemporary society. In 2030, the World Health Organization (WHO) anticipates 24.5 additional deaths. In addition, the main causes of this fatal disease, which is now widely prevalent among individuals of all ages, family history, high cholesterol, smoking, include obesity, high blood pressure, and drinking. However, given that cardiovascular illness is accompanied by a variety of symptoms, a rapid and precise diagnosis of the condition seems to be fairly difficult for medical specialists. As a result, a sizable quantity of data is being gathered internationally by the healthcare sector in order to learn more about cardiac disorders and uncover information that will help specialists better comprehend the condition and guarantee that patients receive effective therapy. However, gathering data requires extensive screening and processing in order to efficiently extract information. These analyses on huge datasets, nevertheless, were previously impractical using conventional statistics. As a result, machine learning (ML) has become the most effective technology available today for processing data and using that data to advance the healthcare industry. The human body's most important organ is the heart due to its critical function in blood pumping. The mortality rate associated with cardiac illnesses can be significantly decreased by using machine learning to forecast heart health and anticipate disease. The causes of heart disease might vary widely. The evolution of lifestyle variables such as smoking, physical activity, eating habits, diabetes, and obesity, as well as biochemical elements like glycemia or blood pressure. Because of this, it is necessary to document key cardiac behavior specific to each form of heart illness and to develop a system that aids clinicians in establishing accurate and effective diagnoses. In reality, a medical diagnosis is a categorization mission in which a doctor attempts to locate the flaw by examining the values of several qualities. So, a wide range of clinical sciences, study of fundamental sciences, implementation science, public health sciences, and policy and system studies are required to fill these gaps in order to effectively control and manage the heart disease as well as prevent and respond to future epidemics. Our research Objectives are is given below:

- To determine the outcome of heart disease.
- Quality of life.

- To investigate the correlated issue

1.3 Rationale of the Study

The broad term for illnesses that damage the heart or blood vessels is cardiovascular disease (CVD). Unhealthy eating, inactivity, smoking, and dangerous drinking are the level equivalent risk factors for coronary heart disease and stroke.. Individuals might encounter elevated blood pressure, elevated blood sugar, elevated blood lipids, and obesity or stoutness because of conduct risk factors. Typically, it is linked to an increased risk of blood clots and the accumulation of fatty deposits in the arteries, which is known as atherosclerosis. Even though computer science and medicine appear to be distinct fields, they have worked together for a few decades. One aspect of this collaboration is data mining, a relatively new multidisciplinary field that can extract useful information from large data sets. Despite this, it was rarely used in cardiology-related studies. We assume that some expensive, time-consuming, sometimes dangerous medical examinations, complex, uncomfortable, time-consuming can be replaced by data mining tools. The prediction and diagnosis of heart related illness necessitate greater precision, accuracy and perfection, because even a minor error can result in fatigue or death; the number of heart-related deaths is growing exponentially every day. A prediction system for disease awareness is absolutely necessary to address the issue. Artificial intelligence is a subfield known as machine learning offers prestigious assistance in predicting any kind of event by drawing on lessons learned from natural occurrences. We are motivated to address these difficulties because of this.

1.4 Research Question

- What is this database's most important features?
- How does this research's algorithm function?
- How can you anticipate when heart disease will be detected early?

- If a person develops heart disease, what is the success rate?

1.5 Expected output

We are trying to detect-heart disease. Through our project, we can find out whether heart infected and if it is infected, we can also take out the percentage of it that is infected. Thinking about the future, we have worked on such a program so that in the future we can detect heart disease very early and get good quality treatment for it. We worked with a programming language called machine learning in heart disease detection project. In this project, we took 14 attributes. And we also used different five algorithm. So, we hope our system will accurately and consistently identify diseases.

1.6 Project Management and Finance

The supervision of a project's financial aspects, such as its cost, revenue, and profit, is referred to as project financial management, or project accounting. In order to accomplish this, it integrates billing, estimation, budgeting, funding, and project expense management. In order to enhance patient experience and satisfaction, reduce costs, and improve patient care, today's healthcare leaders are constantly improving and developing their processes. Over the course of the past ten years, there has been a significant rise in the use of project management principles in healthcare. These principles have become increasingly significant to facilities due to the fact that they aid in cost control, risk mitigation, and overall project outcomes. There are countless electronic framework executions within health care. Project management has therefore emerged as one of the most sought-after business skills. The management of a project's financial aspects, such as its cost, revenue, and profit, is referred to as project financial management, or project accounting. In order to accomplish this, it integrates billing, estimation, budgeting, funding, and project expense management. Of this large number of parts of monetary task, the executive's, viable undertaking planning is by a long shot the most significant. From there, managing that budget over the course of a project is the challenge, with the goal of ensuring that the work is finished within the allotted budget. To better understand emerging patterns of roles and responsibilities, data from the survey of project management roles and responsibilities was acquired. Project management includes scheduling and planning meetings, facilitating the entire study, and entering data into databases. Study progress was recorded via timely release of meeting agendas and minutes outlining progress and action items. These resources accumulated throughout time to create an archive that is centrally accessible via the Madcaps communication platform. To produce high-quality data that can influence disease prevention and control as well as policy requirements, taking Competent project management is essential for executing health-related research that involves taking into account the local and nationwide political and social situations. This occurs from the fact that efficient project management enables the administration and adherence of the appropriate tools, methodologies, tactics, and predictable approaches. By incorporating project management abilities into efforts to establish health care systems in nations with low and middle incomes, the population's health may be boosted.

Finance: bill-paying and account-balancing, money-management, grant administration, and budget creation.

1.7 Report Layout

- Describe of background study
- Process of Research Methodology
- Find out Experimental Result and Discussion
- Discuss to Summary, Conclusion and Future Analysis
- Reference

CHAPTER 2
Background study

2.1 Preliminaries

One of the most difficult works in the world is taking care of people's health. One of the main reasons for a rise in mortality is cardiovascular disease. Cardiovascular diseases (CVDs) account for a significant portion of the global death rate. The World Health Organization (WHO) estimates that cardiovascular disease-related deaths accounted for 17.7 million deaths in 2015, or 31% of all deaths worldwide. Countries with low or middle incomes, like Bangladesh, where 80% of these deaths take place, have populations that are most affected. From 2011 to 2025, CVD accounts for nearly half of the estimated $7.28 trillion in cumulative economic losses from all noncommunicable diseases in these nations. As a result, cardiovascular disease (CVD) is regarded as a major public health concern worldwide. The prevalence and mortality rates of noncommunicable chronic diseases have skyrocketed in Bangladesh over the past few decades. As a result of its rapid economic growth over the past few decades, Bangladesh has experienced rapid urbanization, and it has recently emerged as a developing nation. Changes in eating habits, such as raise the possibility of a further rise in the burden of chronic diseases, inconsistent meal times, increasing the demand for and availability of processed food, and decreased physical activity, as a result of this growth and urbanization.

2.2 Related Works

Practicing clinical medicine in the healthcare sector is likely to change as machine learning technology advances. Frequently used cardiovascular datasets are categorized using modern machine learning model such as Naive Bayes, M5P tree, Random Tree, J48, and JRIP, REP Tree, Linear Regression.

Nadakinamani, R.G. et al. [1] The Random Tree model did a great job, achieving a maximum accuracy of 100%. The accuracy obtained for the Random Tree, M5P, Linear Regression, REP Tree has a score of 99.81%, 75.75, 74.32%, and 88.44%.

Yu, W. et al.[2] utilized Naive Bayes, the method with the highest accuracy in this instance. Additionally, an accuracy of 83% was achieved when SVM was applied to the Dataset from the Health and Nutrition Examination Survey. SVM, with the most elevated exactness as against other previously mentioned general procedures and helping, stood apart as the best technique for expectation.

Ahmed, H. et al. [3] talked about a real-time system that uses streams of medical information that demonstrate a patient's present state of health to predict heart-disease. To choose significant features from the dataset, two different types of feature selection methods are utilized: Relief and the selection of univariate features. Four machine learning algorithms were compared the Random Forest Classifier, Decision Tree, Logistic Regression Classifier, Support Vector Machine, and Random Forest Classifier. While the training data were utilized 90% of the time, the testing data were utilized 10% of the time. SVM has the lowest accuracy, while random forest has the highest (94.9%). In a similar vein, DT and LR have 88.40 percent and 89.9%, respectively.

Hertzog, M.A. et al. [4] used 14 distinct attributes to create three clusters using the clustering method. Logistic regression, SVM, and Random Forest etcetera can be utilized for this reason. This was done with Nave Byes, and the results are very accurate.

Kwon, K. et al. [5] were regarding the clustering method, Logistic Regression has proven to be the most accurate, and 5] worked. Among different endeavors, a UI-based Cardiovascular checking framework was created utilizing versatile passages and observing servers. For the purpose of establishing distinct patterns for various patient subgroups, logistic regression was utilized.

Sharma, H. et al. [6] could predict 67% of this disease. Algorithms for machine learning and deep learning open up new possibilities for accurate prediction of heart attacks. This disease can be diagnosed using a variety of algorithms. Naive Bayes, Neural Network, Decision Tree, and KNN algorithms they used. They have a 95% success rate in Neural Network.

Jindal, H. et al. [7] used Logistic Regression, KNN, and the Random Forest algorithm in this regard. Heart disease is becoming more common by the day. This type of disease foretells the future. This disease is difficult to diagnose. As a result, it must be done precisely and efficiently. Using Logistic Regression and KNN, heart disease patients can be predicted. Machine Learning is a broad and varied field. Machine learning classifiers are used to predict and find accuracy. The most efficient algorithm is the KNN Algorithm. In this paper, KNN provided an accuracy of 88.52%. Logistic Regression also provided an accuracy of 80.50%. The algorithms in this paper have an average accuracy of 87.5%.

Takci, H. et al. [8] combined feature selection and machine learning algorithms in this study. Cardiologists use traditional clinical methods for this task, but Machine learning-based computer-vision diagnosis systems are also utilized. They utilized numerous machine-learning techniques as well as the best feature selection algorithms. With an accuracy of 84.81 percent, this pair was the most accurate. In this study, twelve classifiers from various categories and four feature selection algorithms from two distinct categories were utilized for the purpose of heart attack prediction. They used 10% of the data for testing and 90% for training. Model accuracy was 84.81% using naive Bayes and SVM-linear methods.

To compare the outcomes in this study, Bharti, R. et al. [9] employed a variety of deep learning and machine learning techniques. The UCI Machine Learning Heart Disease dataset was examined. The dataset that will be used in the investigation has 14 key features. Decision tree accuracy was 82.3%, Logistic regression accuracy was 83.3%, SVM accuracy was 83.2%, Random Forest accuracy was 80.3%, Decision tree accuracy was 82.3%, and K neighbors accuracy was 84.8%.

Marimuthu, M. et al. [10] applied Nave Bayes, Fuzzy Logic, artificial Neural Network (ANN), Support Vector Machine, Decision Tree, and K-Nearest Neighbor (KNN) are some machine learning and data mining techniques used to predict heart disease (SVM).

Yahaya, L. et al. [11] performed classification algorithms such as the Artificial Neural Network (ANN), Decision Tree (DT), and Nave Bayes (NB) are used. The heart disease dataset used in this study only contains 14 features and 303 instances. utilizing geographically diverse data sources to enhance disease prediction accuracy.

Khourdifi, Y. et al. [12] classified data using various classification algorithms such as SVM, Random Forest, Multilayer Perception, KNN, and Nave Bayes. Artificial Neural Network optimized using (ACO) and (PSO) approaches. The different machine learning algorithms are compared using various indicators of performance such as precision, f1-score, accuracy, recall, and so on. The optimized model that was proposed by FCBF, ACO, and PSO had a maximum classification accuracy of 99.65 percent with KNN, and 99.6 percent with RF.

Louridi, N. [13] increased the accuracy of cardiovascular disease prediction. It assisted in recognizing the patient's cardiac condition and helped a doctor determine whether or not the patient had cardiovascular disease. Using the accuracy, precision, and recall performance measures to demonstrate their findings, they compared the effectiveness of a number of algorithmic learning techniques. The models used were Naive Bayes (NB), Support vector machine (SVM), with K-nearest neighbors (KNN). SVM was employed to get the 86.8% score accuracy.

Asif, M. [14] studied to compare and contrast how various machine learning algorithms predicted cardiovascular disease. The model was trained and evaluated using data from of the UCI. The performance of the default DHP, GSCV, and random search cross-validation was examined for twelve distinct machine learning algorithms RSCV methods performed. Additionally, the GSCV and RSCV calculation times were established. Both soft and hard voting ensemble classifiers have an EVCS and EVCH accuracy of 92 percent.

Dinesh, K. G. [15] offered many machine learning methods for predicting the degree of uncertainty in cardiovascular disorders based on various factors. Support vector machine, naive bayes classifier, and random forest, Gradient Boosting, and logistic regression are the models that are used to forecast the illnesses. Logistic regression has the highest accuracy (86.5%).

Amin, M. S. [16] presented a cloud-based solution. Using Arduino, a genuine monitoring system was created that allows for the detection of several metrics, such as temperature, blood pressure, and heartbeat, every ten seconds. SVM demonstrates its effectiveness in this investigation with accuracy of above 95%.

2.3 Comparative Analysis and Summary

Physicians may be able to make more accurate diagnoses thanks to advances in medical data collection. In addition, computational biomedical systems have the potential to speed up the decision-making process and enhance the accuracy of predictions for a variety of diseases, including cancer, skin conditions, diabetes, kidney conditions, and heart conditions. Cardiovascular diseases have been identified as having the highest mortality rate among these conditions in the majority of countries WHO, Electrocardiogram (ECG), angiography, cardiac magnetic resonance imaging (MRI), stress tests, echocardiogram (heart ultrasound), are all commonly used by doctors to identify cardiovascular issues. However, the entire community cannot afford the relatively high diagnostic and treatment costs associated with cardiovascular disease. Data mining techniques make it possible to quickly determine if a more likely to be a patient to develop heart disease early on, which lowers the cost of diagnosis and treatment. Table 1. demonstrates the analysis of previous work in comparison.

Table 2.3.1. Comparative analysis with previous work

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Nadakinamani, R.G. et al. [1]	J48, and JRIP, M5P Tree, REP Tree, Naive Bayes, Random Tree, Linear Regression.	Random Tree (100%)
2.	Yu, W. et al. [2]	Naïve Bayes, SVM	SVM (83%)
3.	Ahmed, H. et al. [3]	SVM, Decision Tree, Logistic Regression, Random Forest	Random Forest (94.9%)

4.	Hertzog, M.A. et al. [4]	Logistic regression, Nave Bayes, SVM, and Random Forest	______
5.	Kwon, K. et al. [5]	Logistic regression	______
6.	Sharma, H. et al. [6]	Naive Bayes, Decision Tree, KNN, Neural Network	Neural Network (95%)
7.	Jindal, H. et al. [7]	Logistic Regression, KNN, Random Forest	KNN (88.52%)
8.	Takci, H. et al. [8]	SVM-linear, naive Bayes	84.81%

9.	Bharti, R. et al. [9]	K neighbors, Random Forest, Logistic Regression, Decision Tree, SVM, and Deep Learning	Deep learning (94.2%)
10.	Marimuthu, M. et al. [10]	ANN, SVM, Decision Tree, KNN, Fuzzy Logic, Nave Bayes,	_______
11.	Yahaya, L. et al. [11]	Decision Tree, Artificial Neural Network, Naive Bayes	________
12.	Khourdifi, Y. et al. [12]	KNN, SVM, Nave Bayes, Random Forest, Multilayer Perception	KNN (99.65%)
13.	Louridi, N. et al. [13]	Support vector machine (SVM), K-nearest neighbors, and Naive Bayes (NB)	SVM (86.8%)
14.	Asif, M. et al. [14]	grid search cross-validation (GSCV), random search cross-validation, and hyperparameter (DHP) (RSCV)	92%
15.	Dinesh, K. G. et al. [15]	Logistic regression, Random Forest, Support Vector Machine, Naive Bayes classifier, Gradient Boosting,	Logistic regression (86.5%)
16.	Amin, M. S. et al. [16]	SVM	95%

2.4 Scope of the Problem

We've read a lot of research papers on this research. They worked on applying different deep and machine learning methods, cardiac illness. The data collection is a major problem to develop any Machine Learning project. This paper we had some trouble collecting data but we couldn't collect much data.

2.5 Challenges

- **Data collection**
- **Model selection**
- **Model train**

Chapter 3
Research Methodology

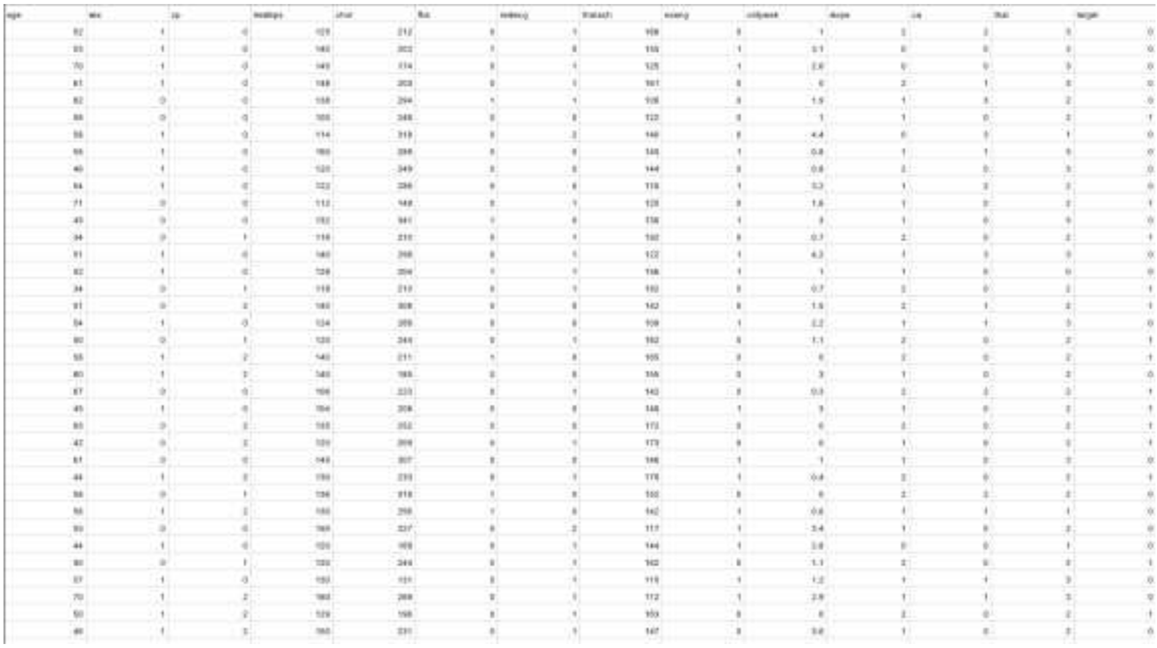
The goal of this study is to employ a mathematical prediction approach to determine the source of heart disease. It can help save the lives of those suffering from heart disease. To accomplish this purpose, we study the data set using several machine learning methods described in this work. Classification approaches can help forecast serious illnesses. So, to attain our purpose, we employ data mining techniques. In this section, we examine the whole data analysis approach utilizing step-by-step classification techniques, as well as the execution of chosen algorithms.

3.1 Research Subject and Instrumentation

We attempted to provide a clear picture of our study approach here. We obtained our dataset from the internet. We worked on the Windows operating system. Google's Colab notebooks tools were employed. Collaboratory, or "Colab" for short, is a Google research tool that may be programmed in Python. Colab is the best tool for learning, data analysis, and machine learning since it enables anybody to create and run arbitrary Python script in a browser. Colab is a no-setup sharing Jupyter service that offers free access to computing tools like GPUs.

3.2 Data Collection Procedure

Using machine learning, we identified persons who had heart disease. That is why we needed to compile a dataset. Patient datasets are often filled with data from them. Due to timing restrictions, we were unable to gather the data ourselves. The dataset was obtained via Kaggle. That is why we chose solid and high-quality datasets in Kaggle, with fewer missing and null values. As a consequence, accuracy will improve. Diagram 3.2.1 shows the collection of dataset.



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1
42	0	2	120	209	0	1	173	0	0	1	0	2	1
61	0	0	145	307	0	0	146	1	1	1	0	3	0
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
58	0	1	136	319	1	0	152	0	0	2	2	2	0
56	1	2	130	256	1	0	142	1	0.6	1	1	1	0
55	0	0	180	327	0	2	117	1	3.4	1	0	2	0
44	1	0	120	169	0	1	144	1	2.8	0	0	1	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
70	1	2	160	269	0	1	112	1	2.9	1	1	3	0
50	1	2	129	196	0	1	163	0	0	2	0	2	1
46	1	2	150	231	0	1	147	0	3.6	1	0	2	0

Figure 3.2.1 the collection of dataset

We read the data and then removed the missing numbers from our dataset. Our dataset has a size of (1025,14). The dataset we obtained has 1025 data points, 14 of which include information. The information are Age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, maximum heart rate reached, exercise-induced angina, old peak = ST depression caused by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, and resting electrocardiographic results (values 0, 1, 2)

3.3 Statistical Analysis

We obtained our dataset from the internet. Data from China is difficult to obtain online. We had to do a lot of research online for this. Finally, we were able to obtain our requested dataset via Kaggle. There are 14 characteristics in the dataset. Various strategies can be used to comprehend the dataset. The statistical analysis aids in making the dataset more understandable. We used 5 algorithms in all, with the Decision Tree method achieving the greatest accuracy of 99.70%. Figure 3.3.1 shows all accuracy

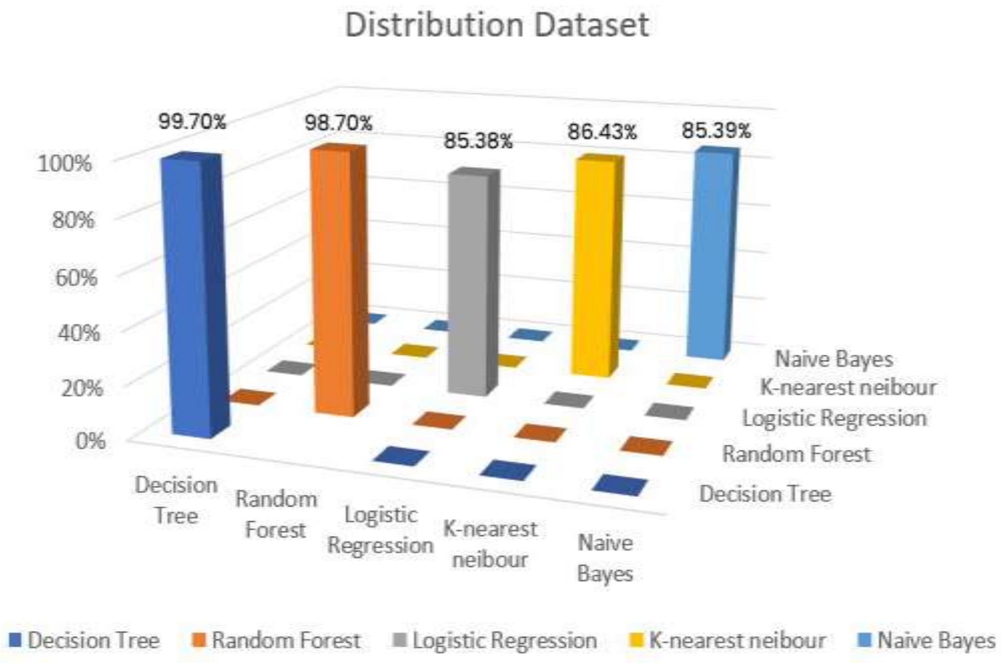


Figure 3.3.1 accuracy of all algorithms

3.4 Proposed Methodology/Applied Mechanism

Heart disease has taken a bad turn in recent years. So we attempted to detect it using a machine-learning approach. We detected it using multiple methods. Among them are the most accurate algorithms are K-nearest neighbor, Decision Tree, Random Forest, and Naive Bayes, and Logistic Regression. To complete the procedure, we took various steps. The stages are depicted in an attractive flowchart format.

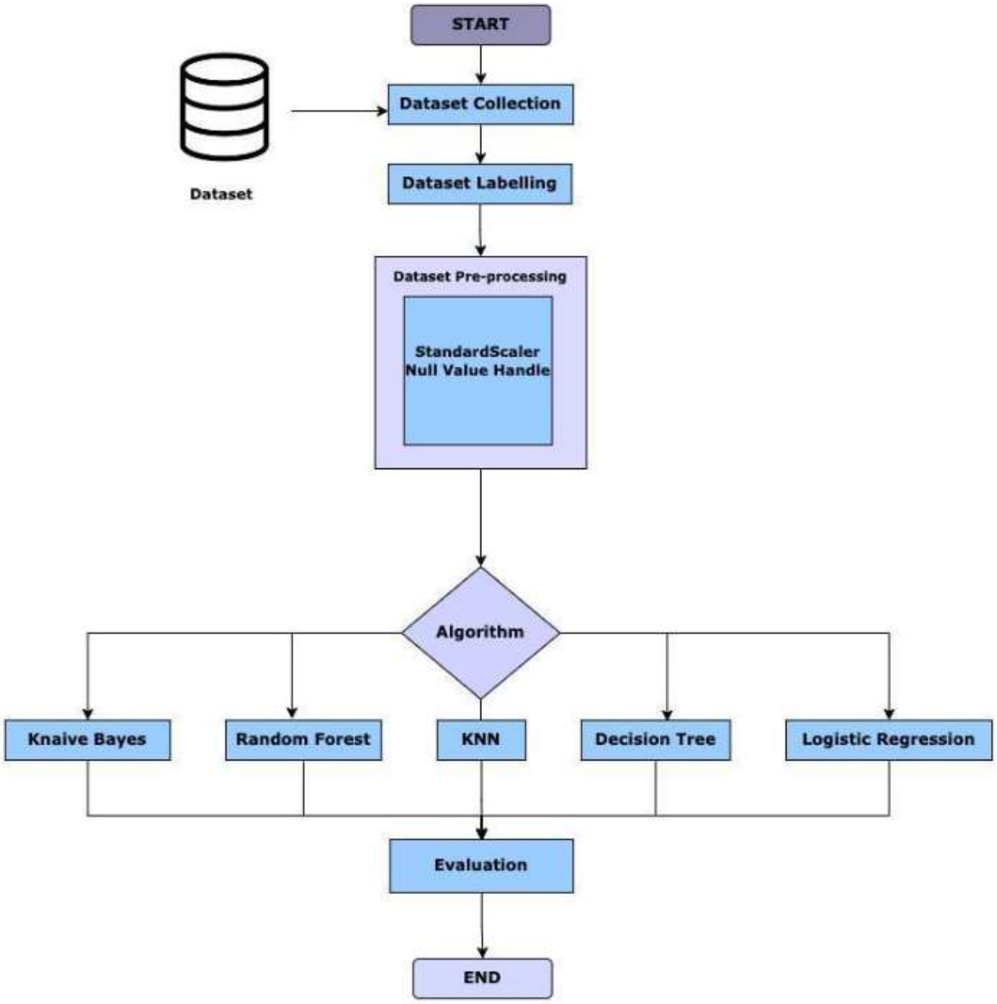


Figure 3.4.1 System Design

This diagram depicts the entire machine learning process. Data collection, dataset labeling, dataset preprocessing, method application (Logistic-Regression, Decision-Tree, Naive-Bayes, K-nearest neighbor, and Random-Forest), model evaluation, and conclusion are the procedures. The procedures are as follows:

3.4.1 Dataset collection:

The collection of information and data for a job is known as dataset collection. A dataset for machine learning is a group of data used to train the algorithm. A dataset serves as an example for the algorithm for machine learning to learn how to generate predictions.

3.4.2Dataset labeling

It is insufficient to provide a massive quantity of original data to a model of machine learning and expected it to understand it. Because this will provide a sudden outcome, it is required to pre-process the information, and Data Labelling is among the stages of pre-processing records. We offer some recognition to raw data (which could be a picture, voice, or word) and add certain tags to it throughout the data procedure. Those tags indicate whatever class of item the information relates to, allowing the ML model to learn from it and generate the best accurate estimate.

3.4.3 Data preprocessing

Preparing original data for machine learning methods is known as data preprocessing. It is the first and most crucial phase in creating a machine learning model.

When developing a pattern recognition project, we cannot always run across clear and prepared data. Furthermore, data must be cleaned up and categorized before being used in any way. So, we use the data preparation task for everything.

Actual data normally contains noise, and incompleteness, and could be in an unsuitable format that cannot be utilized properly for models of machine learning. In order to clean up the data and get it ready for a classification model, which increases the accuracy and efficiency of machine learning algorithms, pre-processing is a vital task.

It consists of the following steps:

- Getting the dataset
- Adding libraries
- Dataset importation
- Locating Missing Data
- Categorical Data Encoding
- dividing the dataset between testing sets and training sets
- Scaling of features

For our dataset, we used certain preprocessing procedures. For preparing our dataset, we used two steps. We used two kinds of pre-processing techniques: null value management and regular scaling.

- Null Values handling: We know that missing values occur at random. Missing values can be handled by eliminating the rows or columns that contain null values.

A column can be removed if many over half of a rows in it are null. Rows having missing value inside one or even more columns can be removed as well.

- Scaling Rules: Another scaling is standardization approach where the values are centralized around the means with a standardized deviation. As a result, the property's mean changes to zero, and the resulting dispersion has a variance of one.

3.4.4 Performance measures

Our information is used to determine the overall Recall, accuracy, and F1 score, which are used to assess architectural efficiency and precision. TP means for true-positive and FP means for false-positive in this study. FN means for false-negative, whereas TN stands for true-negative. According to our dataset, the highest result is around 99.70 %

Positive Absolute Accuracy Precision is measured by dividing total of positive outcomes by the entire amount of outcomes. The entire amount of actual positive outcomes is divided by the predicted positive class of results to compute recall. The harmonized mean, or F1 score, of memory and accuracy strikes a balance.

Recall, accuracy, and F1 scores are used to assess categorization task performance.

- **Precision:** Since it assesses the accuracy of a favorable prediction made by the algorithm, precision is one metric of a machine learning model's performance. By dividing the total number of positive predictions by the number of genuine positives, precision is calculated.
- **Recall:** The number of true positives based on number of positive cases in the data collection is referred to as recall (true positives plus false negatives). Recall is an important metric of the model's ability to recognize the class labels.
- **F1 scores:** The F1 score is a single assessment metric that attempts to compensate for and optimize accuracy as well as recall. It is defined as the arithmetic mean of accuracy and recall. F1 scores are used by data scientists to base on pre - defined predictive accuracy during model iteration stages by combining accuracy and recall into a single statistic. This allows teams to conduct hundreds of tests at the same time and objectively identify top-performing models. A simulation will achieve a strong F1 score if it has a high accuracy and recall. A model, however, will have a poor highest accuracy if one element is low, and when the other is 100%.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

TP = True positive

TN = True negative

FP = False positive

FN = False negative

.

3.5 Implementation Requirements

We used the following models in our project: Naive Bayes, K-nearest Neighbor, Logistic Regression, Decision Tree, and Random Forest. Among them, we received the highest Decision Tree score of 99.70%.

3.5.1 Decision Tree

Decision Tree is a Directed Learning method that may be used for classification and regression problems, although it is most frequently used for categorization. This extractor has a tree-like structure, where leaf nodes represent the results, support vectors represent dataset properties, and branching represents the rule base. The Decision Node and the Leaf Node are two network nodes that connect a decision tree. Tree structures indicate the outcomes of those choices and have no further branches, whereas decision nodes are used to make decisions and have many branches. It is a visual representation for gathering all possible responses to a problem based on predetermined criteria. The judgements or tests are based on the qualities of the engages the productive. Since it starts with the root node and evolves on succeeding branches to form a tree-like design, it is dubbed a decision tree.

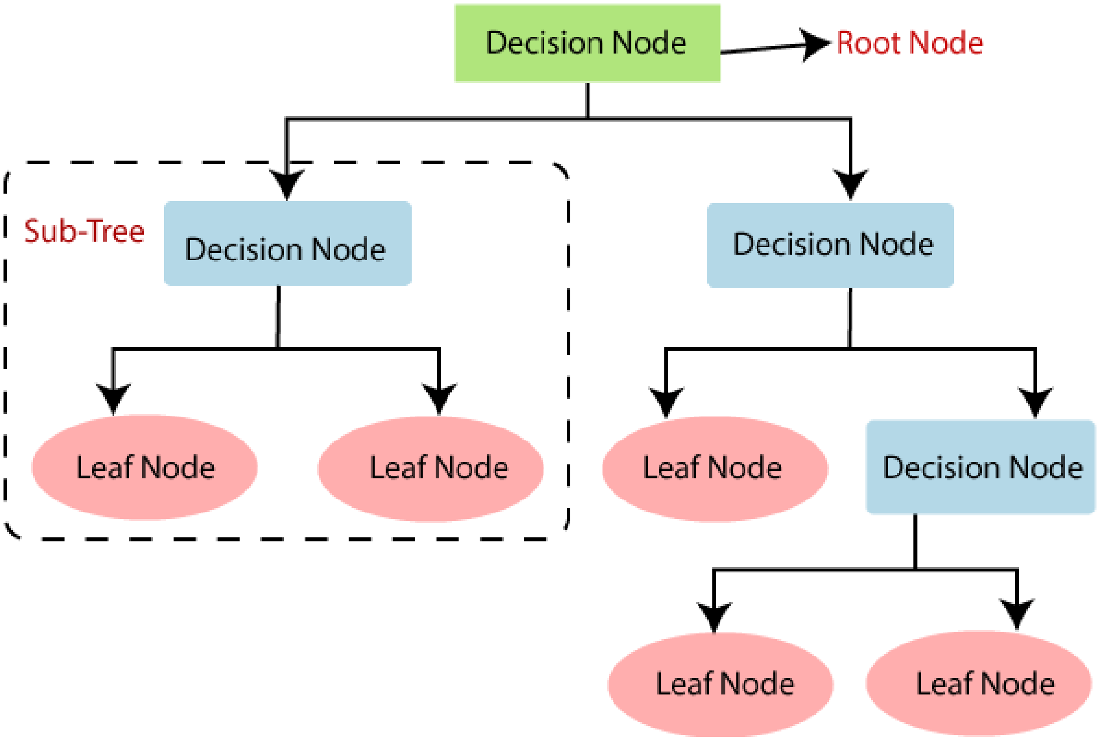


Figure 3.5.1 Decision Tree

The most important thing to keep in mind while creating a machine learning approach is to choose the best method for the dataset and problem because there are many different machine learning techniques. The following are the two justifications for using the decision tree:

- Decision trees are designed to mimic human decision-making processes, which makes them easy to understand.
- The decision tree's logic is easily understood because of its tree-like structure.

Terminologies for Decision Trees:

- Node in the Central core
- Node of the Leaf
- Splitting
- Subtrees and branches
- Pruning
- Node of Parent/Child

This method runs at the root node of a decision tree like this and moves its way up to determine the category of a certain dataset. This method compares the contents of the root property with the features of the record (real dataset) and, based on the similarity, branches to the next node.

Using the other sub-nodes, the algorithm verifies the parameter value before moving on to the following node. It keeps doing this until it gets close to the tree's root system.

3.5.2 Random Forest

The supervised learning strategy uses the well-known machine learning algorithm Random Forest. It may be used to solve machine learning classification and regression problems. It is based on the idea of supervised techniques, which is a technique that combines many categories to address a complex problem and improve accuracy. As the name suggests, "Random Forest" is a classification that includes a choice of decision trees on services for low of the supplied dataset & picks the mean to improve the anticipated accuracy of that dataset. The rf gathers the predictions from each decision tree and anticipates the ultimate output based on the simple majority of predictions rather than relying just on one decision tree.

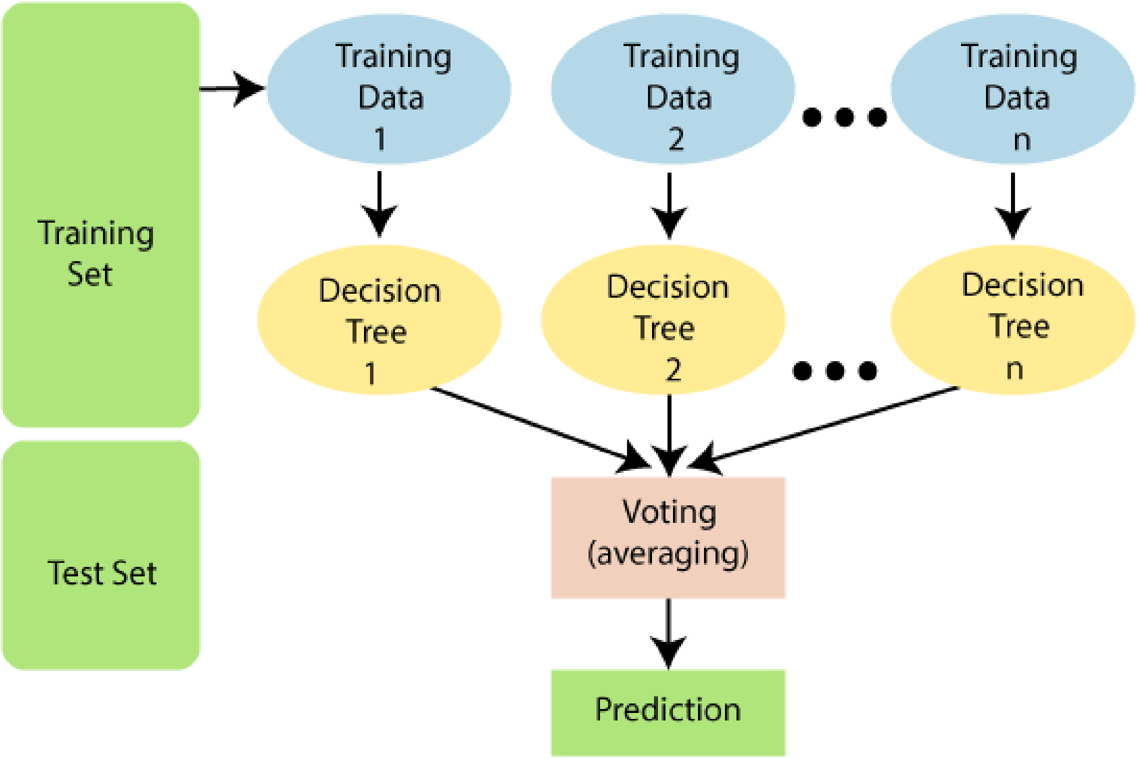


Figure 3.5.2 Random Forest

Some decision tree algorithms may decide the right output while others might or might not since the random forest blends a number of trees to anticipate the category of the dataset.

But when all of the trees are considered together, they accurately forecast the outcome. As a result, the following are two requirements for a stronger Classification algorithm:

- There must be some real numbers in the dataset's reports an experimental so that the algorithm can anticipate correct results as opposed to speculations.
- Each tree's estimates must have very weak correlation.

The following are some justifications for using the Random Forest algorithm:

- In comparison to other algorithms, it needs less training time.
- It accurately predicts output and performs well even with large datasets.
- When a sizable portion of the data is missing, it could nevertheless maintain accuracy.

Random Forest has the following advantages:

- can handle both regression and classification challenges in Random Forest.
- can manage large datasets having complexity.
- It becomes better prediction performance and prevents the fitment issue.

3.5.3 Logistic Regression

The Directed Learning methodology includes the popular machine learning technique of logistic regression. Forecasting the categorized response variable from a collection of independent factors is done using it. Logistic regression is used to forecast the outcome of a dependent variable with categories. The conclusion must thus be categorical or discrete. It can be True or False, 0 or 1, False or True, and so on. However, rather than showing exact values like 0 or 1, it instead displays probability values that lie between 0 and 1.

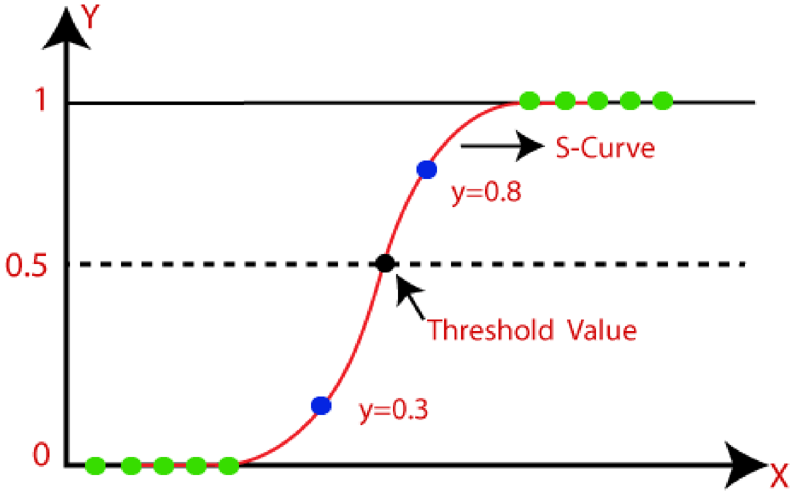


Figure 3.5.3 Logistic Regression

Logistic regression and regression analysis are quite similar, with the exception of how they are used. Regression analysis is used to resolve issues, as opposed to using regression analysis to resolve classification issues. We fit a "S" curved logistic function that forecasts two peak values instead of fitting a regression (0 or 1). The logistic function curve indicates the potential for anything, including whether or not the cells are cancerous, whether or not a mouse is overweight based on its weight, and so on. Because it can provide probabilities and categorize new data using both discrete and continuous datasets, logistic regression is a crucial machine learning technique.

$$\text{y}= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

$$\frac{\mathbf{y}}{\mathbf{1-y}}\textbf{; 0 for y= 0, and infinity for y=1}$$

$$\boldsymbol{log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n}$$

Depending on the groups, there are three different types of logistic regression:

- Binomial: The response variable in a binomial regression analysis can only be one of two types, such as 0 or 1, Pass or Fail, and so on.
- Multinomial: The response variable in a multinomial regression analysis may be one of three or more unordered sorts, such as "cats," "dogs," or "sheep."
- When using ordinal logistic regression, three or more ordered classes of response variable, including such "low", "Medium", and "High".

3.5.4 K-nearest neighbor

In fact, the Guided Learning strategy is used in the fundamental Machine Learning technique known as K-Nearest Neighbor. By assuming similarities between the new particular instance and the existing instances, this technique places the new case in the category that is most similar to the existing cases.. It saves all accessible data and classifies fresh data points based on their similarity. This suggests that when fresh data appears, it may be quickly classified using the K-NN technique into an appropriate suite category. The K-NN approach may be used for both classification and regression problems, however it is usually used for classification problems.

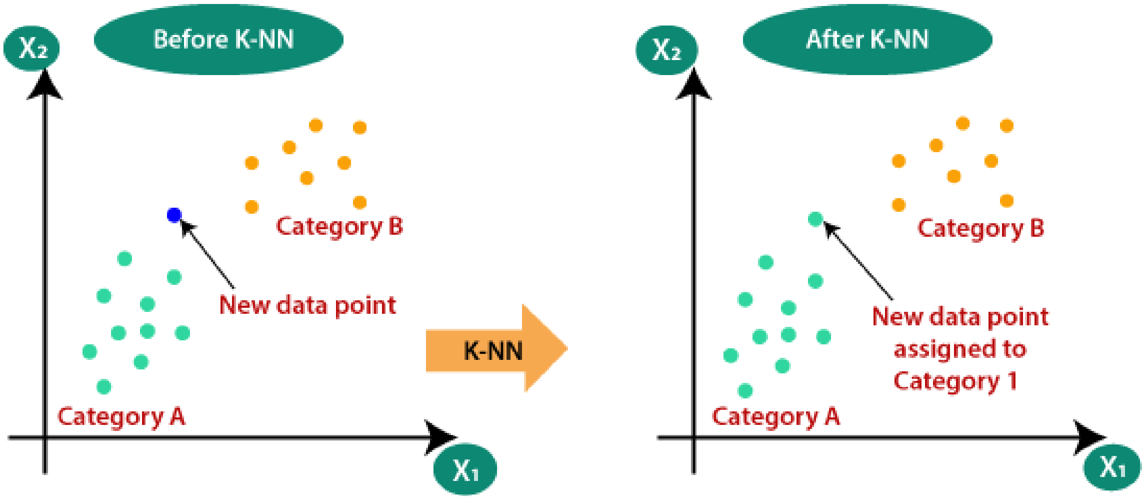


Figure 3.5.4 K-nearest neighbor

Since K-NN is a semi-method, it does not assume anything about the underlying data. It is sometimes referred to as a slow learning strategy since it stores the dataset before acting on it during classification rather than learning immediately from the training set. This technique merely saves the dataset throughout the training phase, and when new data is received, it assigns it to a group that is quite similar to the new data.

The following are the benefits of the KNN Algorithm:

- It is simple to implement..
- Noisy training data do not affect it.
- If the training set is actually large, it could be more effective.

The following are the disadvantages of the KNN Algorithm:

- The values of K must constantly be determined, which can occasionally be challenging.
- Since the distance between data sets for every training set is determined, the computation cost is high.

3.5.5 Naive Bayes

Because it thinks that the existence of one feature has nothing to do with the existence of other traits, it is known as naive. For instance, if fruit is categorized according to its color, shape, and flavor, then a fruit that is red, spherical, and tasty is considered to be an apple.

As a consequence, each trait, by itself, contributes to classifying it as an apple. The Bayes' Theorem serves as the foundation for Bayes.

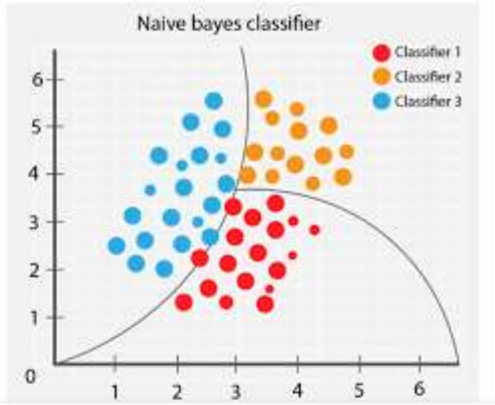


Figure 3.5.5 Naive Bayes

The Nave Bayes method is a supervised training approach that addresses classification problems by applying the Bayes hypothesis. It is mostly used for classification jobs that need a sizable training sample. The Nave Bayes Classifier is a straightforward and effective classification algorithm that facilitates the creation of machine learning models with a high rate of prediction accuracy.

Because it is a classification algorithm, it infers things based on how likely they are to be. Some popular Nave Bayes Algorithm uses include spam filtering, sentiment analysis, and article categorization.

Bayes' formula, often referred as Bayes' rule or Bayes' rule, is a mathematical formula used to determine the likelihood of a claim given past knowledge. It is determined by conditional probability.

The Bayes theorem's formula is as follows:

$P(A|B)= P(B|A)\ P(A) / P(B)$

Where,

$P(A|B)$ represents Posterior Probability, $P(B|A)$ represents Likelihood Probability, $P(A)$ represents Prior Probability, and $P(B)$ represents Marginal Probability.

Benefits of the Nave Bayes Classifier:

- Nave Bayes is a quick and easy machine learning (ML) method for forecasting a class of datasets.
- Both binary and multi-class classifications are possible with it.

- It performs better in Multi-class projections when compared to the other Algorithms.
- The most popular approach for solving text classification problems is this one.

Chapter 4

Experimental Results and Discussion

4.1 Experimental Setup

Using real data, we investigate the consequences of the application's final stage in this paper. Using our estimates, we are able to achieve a fairly exact performance. In our analysis, we employed around 1025 pieces of data. In each dataset, we extracted 14 characteristics. This implies that we gathered data from 1025 people and extracted their 14 favorite pieces of information. The data was then trained using machine learning in Google Collab.

4.2 Experimental Results & Analysis

All model has been thoroughly analyzed here. Previous to that, we needed to develop the model and select the models with the highest accuracy. In addition, we standardized the data to obtain precise and uniform accuracy.

Table 4.2.1 test Result

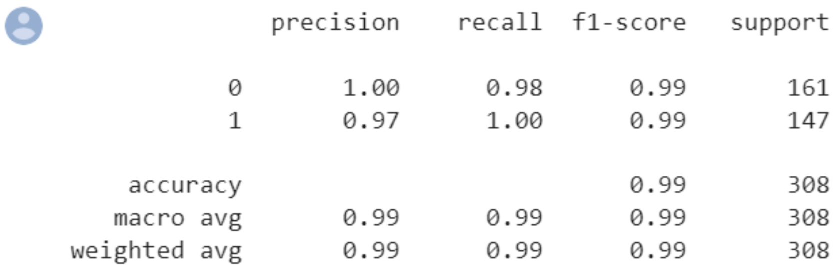
	Algorithm Name	Accuracy Score
1	Decision Tree	99.70%
2	Random Forest	98.70%
3	K-nearest neighbor	86.43%
4	Naive Bayes	85.39%
5	Logistic Regression	85.38%

We can observe that we have the greatest accuracy in the Decision Tree algorithm. We achieved 99.70% accuracy with the Decision Tree Algorithm. It is almost perfectly accurate. Random Forest performed worse than this Logistic Regression Algorithm. That is 98.70% and 85.38%, respectively. The algorithms' precision, recall, and F1-score are listed below.

Table 4.2.2 Model performance Table

	Precision	Recall	F1-score
0	1.00	0.98	0.99
1	0.97	1.00	0.99

The ability of a model to locate all pertinent examples in a data set. The number of true-positives divided by the total number of positive cases plus the total number of false-negatives is how we mathematically define recall. The ability of a classification model to find just pertinent data points is known as recall and precision. The number of false-positives divided by the sum of the true-positives plus the number of false-positives is how precision is mathematically defined.



```
              precision    recall  f1-score   support

           0       1.00      0.98      0.99       161
           1       0.97      1.00      0.99       147

    accuracy                           0.99       308
   macro avg       0.99      0.99      0.99       308
weighted avg       0.99      0.99      0.99       308
```

Figure 4.2.1: Report on precision recall and f1 score

One well-known machine learning method is the Decision Tree approach. In order to solve the machine learning problem, Decision Tree transforms the data into a tree structure. Each leaf node in the tree represents a class label, whereas each internal node represents an attribute.

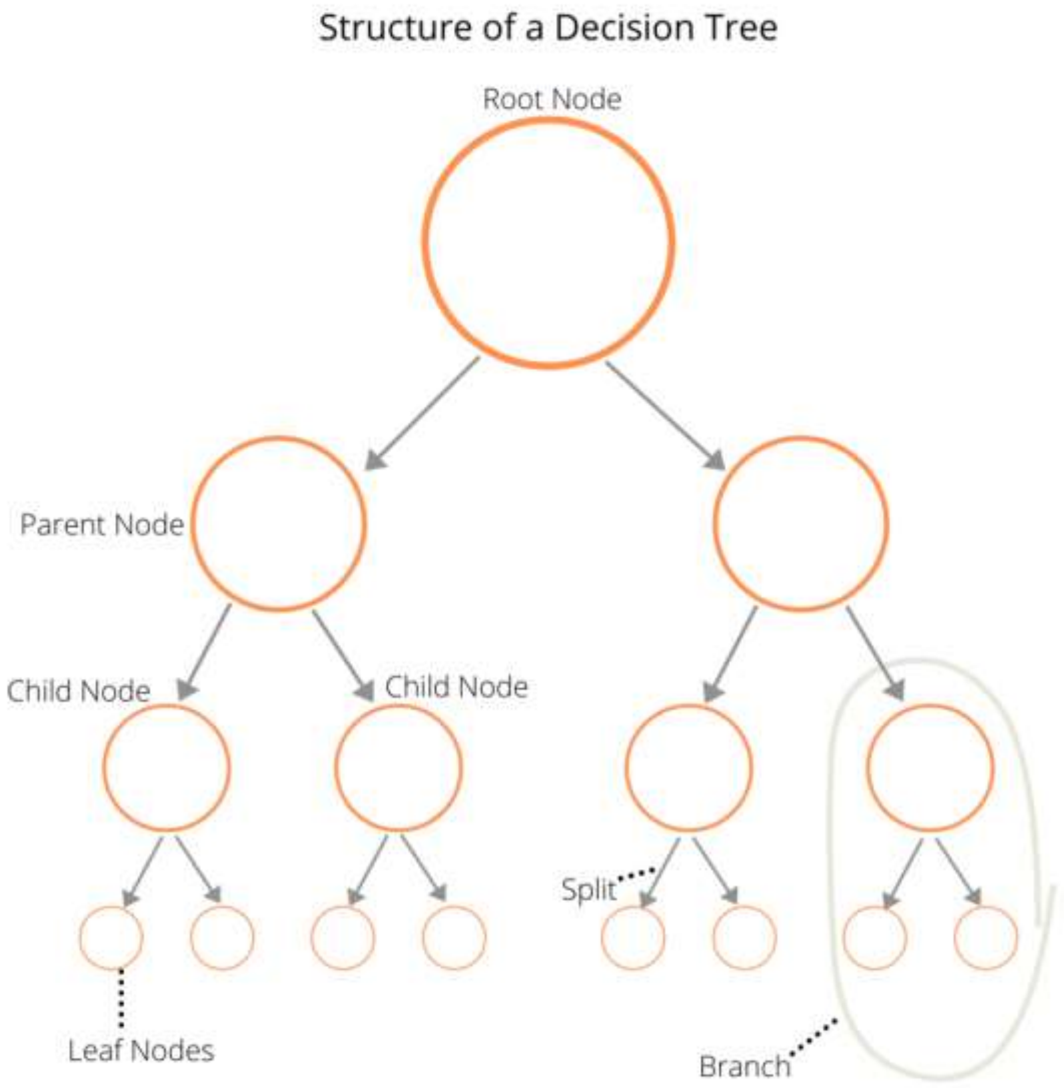


Figure 4.2.2 Structure of Decision Tree

Compared to other approaches, decision trees need less labor to preprocess data. No data normalization is necessary for a decision tree. It is not required to scale the data. The process of creating a decision tree is not significantly impacted by random mistakes in the data. To technical and stakeholder groups, this approach is very clear and easy to explain.

4.3 Discussion

While training the data, we divided the entire dataset into two halves. The first is the train dataset, while the second is the test dataset. We save 70% of the data for training and 30% for testing. We used five different algorithms and chose the best one to ensure that our forecasts are correct.

A classification algorithm's performance is evaluated using a N x N matrix called the confusion matrix, where N stands for the number of class labels. The matrix compares the actual target values to the predictions made by the machine learning model. Figure4.3 depicts a graphic depiction of the confusion matrix for easier comprehension. Figure 4.3.1 depicts a graphic depiction of the confusion matrix for easier comprehension.

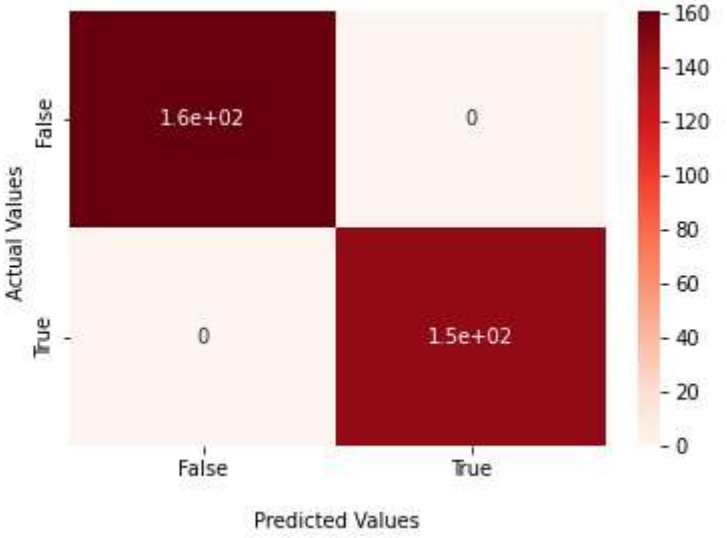


Figure4.3.1 confusion matrix

Chapter 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Bangladesh is a low-income country, and the majority of the people there do not have the money to consult a doctor. Furthermore, the number of specialists in Bangladesh is quite low. As a result, despite the desire, it is often impossible to meet with the doctor. Meanwhile, heart disease, like other ailments, is a major concern for the people of Bengal. A huge portion of the population in this country is now suffering from heart disease and is unable to see a doctor owing to financial constraints and a physician shortage. However, the approach we developed nearly always correctly determines if a patient has the cardiac disease. Because our system is totally electronic, making decisions does not take long. Heart issues are detectable. When several symptoms are noticed, it is possible to conclude that there is a cardiac disease. And because we study these symptoms and make diagnoses using machines, much like a specialized doctor, our technique is cost-effective. It does not cost anything to test or otherwise. Every year, many individuals in our nation die as a result of heart disease, as do other fasts. Those who have been unable to meet the dater owing to a lack of funds or time can now readily do heart disease identification using our method. They will obtain medical treatment without paying the little expense of clinical examinations, thereby improving society and lowering mortality.

5.2 Impact on the Environment

We used healthcare-related data in our research. It contained patient data. The system can evaluate data and apply data mining methods to make healthcare decisions. Clinical choices backed by computer-based patient data have the potential to enhance the medical method by minimizing medical mistakes, enhancing safety and outcomes for patients, and eliminating unwanted practice variances. Data mining is a powerful data modeling and evaluation method that has helped to build a knowledge-rich environment. This has the potential to enhance the quality of professional judgments. This computerized situation will take the role of practices in which doctors make therapeutic choices based on instinct and experience. Furthermore, individuals will be aware of the risk factors associated with arsenic contamination in water, restaurant workers, air pollution, and so on. As a result, we can raise our knowledge of the environmental elements that cause heart disease.

5.3 Ethical Aspects

While working on this project, we must keep several ethical considerations in mind. Because we are dealing with a lot of people's personal information. As a result, it is our moral obligation to protect their privacy. We should also bear in mind:

- Not to endanger any patients.
- Private information security,
- Not to demonstrate genetic prejudice.
- Objectivity in study design.
- Take accountability for the study result.

5.4 Sustainability Plan

In our investigation, we reached nearly perfect accuracy. The hurdles to seeing a doctor in our nation are really high. We don't have any money and there aren't any physicians available. Our automated technology enables rapid diagnosis without the use of clinical testing. So, if we can offer our project to individuals, it will be really beneficial to them.

CHAPTER 6
Summary, conclusion, recommendation and implication for future research

6.1 Summary of the Study

The human body's most necessary organ is the heart due to its critical function in blood pumping. The mortality rate associated with cardiac illnesses can be significantly decreased by using machine learning to forecast heart health and anticipate disease [4]. The causes of heart disease might vary widely. The evolution of lifestyle variables such as smoking, physical activity, eating habits, diabetes, and obesity, as well as biochemical elements like glycemia or blood pressure. Because of this, it is necessary to document key cardiac behavior specific to each form of heart illness and to develop a system that aids clinicians in establishing accurate and effective diagnoses. In reality, a medical diagnosis is a categorization mission in which a doctor attempts to locate the flaw by examining the values of several qualities. It can successfully predict some of the most dangerous diseases in at-risk patients, this machine learning technique is the most effective in recent times. As a result of ML, healthcare professionals are able to provide better feedback, direction, and support for maintaining health because they have a better understanding of the people, they care for day-to-day patterns and requirements. In addition to facilitating patient flow, AI in hospitals can assist in the creation of pharmaceuticals, the storage and analysis of patient data, and even the diagnosis of diseases like cancer. There are also demerits like incomplete data or small size of the data that can impact the analysis. Patients with heart failure are becoming more prevalent every day. A system that can be used to create or classify data rules is required to get out of this dangerous situation and reduce the likelihood of heart disease. As a result, this study of machine learning techniques discusses, proposes, and implements a machine learning algorithm that combines five different techniques. When compared to prior research, this study has demonstrated a considerable improvement and high level of accuracy. In machine learning, preprocessing is a vital step that promotes improved outcomes. In order to increase their accuracy, this paper compared machine learning algorithms with several performance criteria. In our method, missing data from the preprocessing stage is replaced with the mean value. The dataset was expanded using the proper methods. Using the Decision Tree model, best accuracy of 99.70% was attained. Moreover, others algorithms have such as Random Forest (98.70%), Logistic Regression (85.38%), KNN (86.43%), Naive Bayes (85.39%) accuracy.

6.2 Conclusions

The study found that the machine learning algorithm was able to 99.70% accurately predict heart attacks in Bangladesh with a high level of precision. The algorithm was be able to identify risk factors for heart attacks and accurately predict them. This study provides valuable insights into the machine learning use for heart attack identify and may contribute to increasing predictions' accuracy in the future. There are a few limitations to this research paper. Firstly, the study was conducted in Bangladesh, which may not be representative of the general population. Secondly, self-reported data were used in the study, which could lead to recall bias. Thirdly, the study did not include a control group, which makes it difficult to draw causal conclusions. Finally, the study was relatively small, which may limit its generalizability. The next step in this study would be to implement the algorithm for machine learning to a larger dataset in order to improve the prediction's accuracy. Additionally, the algorithm could be tweaked and improved based on the results of this study. Finally, it would be interesting to see if this algorithm could be used to predict other cardiovascular diseases such as stroke.

6.3 Implication for Further Study

We will work with more data sets and use more algorithms to determine accuracy and try to improve our performance. In future, we will do more research about heart attack & will make an android app so that people can easily detect their disease. It will help everyone to predict their disease.

Reference:

[1] Nadakinamani, R.G., Reyana, A., Kautish, S., Vibith, A.S., Gupta, Y., Abdelwahab, S.F. and Mohamed, A.W., 2022. Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. *Computational Intelligence and Neuroscience*, *2022*.

[2] Yu, W., Liu, T., Valdez, R., Gwinn, M. and Khoury, M.J., 2010. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10(1), pp.1-7.

[3] Ahmed, H., Younis, E.M., Hendawi, A. and Ali, A.A., 2020. Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation Computer Systems, 111, pp.714-722.

[4] Hertzog, M.A., Pozehl, B. and Duncan, K., 2010. Cluster analysis of symptom occurrence to identify subgroups of heart failure patients: a pilot study. Journal of Cardiovascular Nursing, 25(4), pp.273-283.

[5] Kwon, K., Hwang, H., Kang, H., Woo, K.G. and Shim, K., 2013, January. A remote cardiac monitoring system for preventive care. In 2013 IEEE International Conference on Consumer Electronics (ICCE) (pp. 197-200). IEEE.

[6] Sharma, H. and Rizvi, M.A., 2017. Prediction of heart disease using machine learning algorithms: A survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), pp.99-104.

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

[8] Takci, H., 2018. Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering and Computer Sciences, 26(1), pp.1-10.

[9] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., 2021. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021.

[10] Marimuthu, M., Abinaya, M., Hariesh, K.S., Madhankumar, K. and Pavithra, V., 2018. A review on heart disease prediction using machine learning and data analytics approach. International Journal of Computer Applications, 181(18), pp.20-25.

[11] Yahaya, Lamido, N. David Oye, and Etemi Joshua Garba. "A comprehensive review on heart disease prediction using data mining and machine learning techniques." American Journal of Artificial Intelligence 4, no. 1 (2020): 20-29.

[12] Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. International Journal of Intelligent Engineering and Systems, 12(1), 242-252.

[13] Louridi, N., Amar, M. and El Ouahidi, B., 2019, October. Identification of cardiovascular diseases using machine learning. In *2019 7th mediterranean congress of telecommunications (CMT)* (pp. 1-6). IEEE.

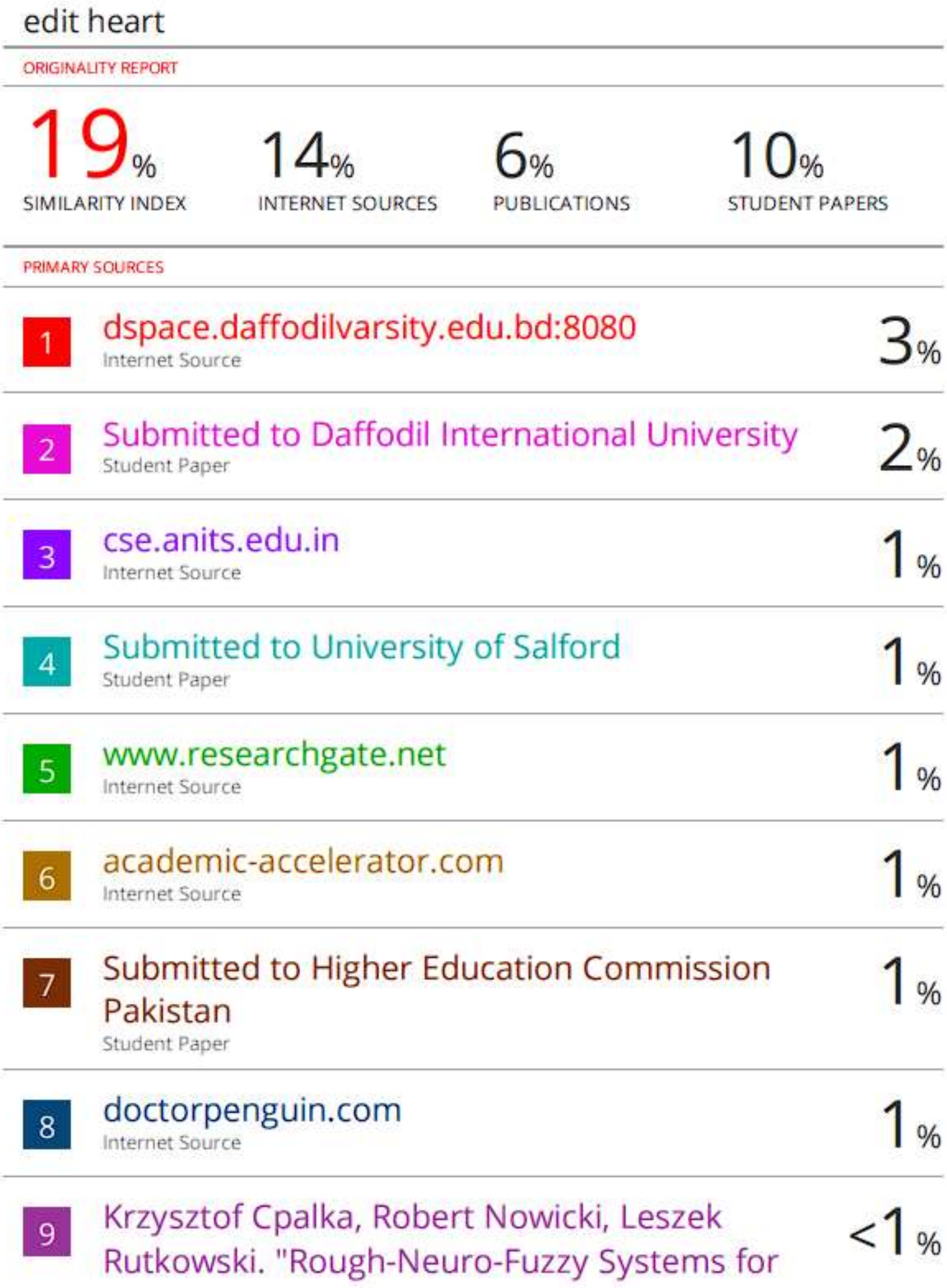
[14] Asif, M., Nishat, M.M., Faisal, F., Dip, R.R., Udoy, M.H., Shikder, M. and Ahsan, R., 2021. Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease. *Engineering Letters*, *29*(2).

[15] Dinesh, K.G., Arumugaraj, K., Santhosh, K.D. and Mareeswari, V., 2018, March. Prediction of cardiovascular disease using machine learning algorithms. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)* (pp. 1-7). IEEE.

[16] Amin, M.S., Chiam, Y.K. and Varathan, K.D., 2019. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, pp.82-93.

[17] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Plagiarism Report



edit heart

ORIGINALITY REPORT

19%	14%	6%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Submitted to Daffodil International University Student Paper	2%
3	cse.anits.edu.in Internet Source	1%
4	Submitted to University of Salford Student Paper	1%
5	www.researchgate.net Internet Source	1%
6	academic-accelerator.com Internet Source	1%
7	Submitted to Higher Education Commission Pakistan Student Paper	1%
8	doctorpenguin.com Internet Source	1%
9	Krzysztof Cpalka, Robert Nowicki, Leszek Rutkowski. "Rough-Neuro-Fuzzy Systems for	<1%