

**AN APPROACH FOR HEART DISEASE PREDICTION USING SUPERVISED
MACHINE LEARNING ALGORITHMS**

BY

Koli Mondol Mira

ID: 221-25-133

This report is presented in partial compliance with the Qualifications
Requirements for Computer Science and Engineering.

Supervised By

Professor Dr. Md. Ismail Jabiullah

Professor

Department of CSE

Daffodil International University

Co-supervised By

Md. Sadekur Rahman

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY


DHAKA, BANGLADESH

JANUARY 17, 2023

APPROVAL

This Thesis titled “An Approach for Heart Disease Prediction Using Supervised Machine Learning Algorithms”, submitted by Koli Mondol Mira, ID No: 221-25-133 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


17/1/23

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

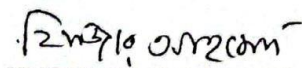
Associate Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Dr. Fizar Ahmed

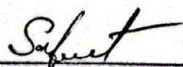
Associate Professor

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

We therefore make declaration that this work has been done by us under the watchful eye of, **Professor Dr. Md. Ismail Jabiullah**, Professor in the Department of CSE Daffodil International University. We also announce that neither this project nor any part of this project has been relocated to be awarded any degree or diploma.

Supervised by:



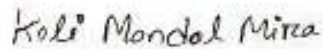
Professor Dr. Md. Ismail Jabiullah
Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Koli Mondol Mira
ID: 221-25-133
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First of all, we express our deepest gratitude and gratitude to Almighty God for His divine blessing enabling us to successfully complete our final year project.

We are very grateful and wish our deepest debt to **Professor Dr. Md. Ismail Jabiullah**, Professor, Department of CSE Daffodil International University, Dhaka. In-depth knowledge and in-depth interest of our manager in the field of “Web Application” to undertake this project. His unwavering patience, expert guidance, constant encouragement, unwavering supervision and enthusiasm, constructive criticism, valuable advice, a lot of low draft learning and correction at all stages made it possible to complete the project.

We would like to express our deepest gratitude to our Parents, our Family, and the Head of the CSE Department “**Professor Dr. Touhid Bhuiyan**”, for his kind assistance in completing our project and for the other members of the faculty and staff of the CSE department of Daffodil International University.

We would like to thank the entire study of our partner at Daffodil International University, who participated in this discussion while completing the course work.

Finally, we must respectfully acknowledge the support and support of our parents' patients.

ABSTRACT

The academic community has recently become interested in categorizing medical datasets using machine learning, despite the fact that it is a challenging task. The conclusion of the procedures is aided by the use of numerous machine learning algorithms to a collection of data. The use of machine learning to predict disease has been the subject of numerous studies in the past. However, there are several chances for development. This work aims to examine alternative machine learning-based models for diabetes prediction using pre-processing techniques, classical classifiers, and ensemble classifiers. The major cause of death globally during the past few decades has been heart disease, sometimes referred to as cardiovascular disease. It includes a variety of disorders that have an impact on the heart. One of the hardest difficulties in the medical industry right now is the prognosis of heart disease. There are numerous risk factors associated with heart disease, and it is urgent to find accurate, trustworthy, and practical methods to make an early diagnosis and achieve fast disease management. Machine learning approaches have advanced the health industry by several researches as a result of current technological advancements. The purpose of this study is to develop an ML model for heart disease prediction using the relevant factors. For this research project, we have collected a dataset from "kaggle" that consists of 13 different parameters connected to heart disease. Machine learning methods such Random Forest, Logistic Regression, Naive Bayes, and Decision Tree have been used in the model's design. With the aid of conventional machine learning techniques, we also attempted to identify correlations between the various features present in the dataset with the purpose of effectively predicting the risk of heart disease. The results demonstrate that Random Forest provides better prediction accuracy in less time than other ML approaches.

TABLE OF CONTENTS

CONTENTS	PAGE
APPROVAL	i
DECLARATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
TABLE OF CONTENTS	v-vii

CHAPTER	PAGE
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1-2
1.2 Motivation	2
1.3 Objective	3
1.4 Research Questions	3
1.5 Report Layout	3
CHAPTER 2: BACKGROUND STUDY	4-6
2.1 Introduction	4
2.2 Review Works	4
2.3 A Comparison of the Review Works	5
2.4 Summary of Related Work	6

CHAPTER 3: RESEARCH METHODOLOGY	7-15
3.1 Introduction	7
3.2 Description of the Proposed Method	7
3.3 Diagram of the Proposed Method	8
3.4 Proposed Algorithm	9
3.5 Pre Processing	10
3.5.1 Categorical data encoding	10
3.5.2 Missing Value Imputation	10
3.5.3 Feature Selection Technique	10
3.6 Classification Algorithms	10
3.6.1 Logistic Regression (LR)	11
3.6.2 K-Nearest Neighbor	11
3.6.3 Decision Tree (DT)	12
3.6.4 Random Forest	12
3.7 Research Instrumentation	13
3.8 Dataset Description	13-14
3.9 Performance Evaluation Measure	14-16
CHAPTER 4: EXPERIMENTAL RESULTS & DISCUSSION	17-22
4.1 Introduction	17
4.2 Feature Analysis	17-20
4.3 Experimental Results	20-22
CHAPTER 5: CONCLUSION	23-24
5.1 Conclusion	23
5.2 Future Work	24
REFERENCE	25

LIST OF FIGURES

NAME OF FIGURES	PAGE
Figure 3.1: Proposed method	8
Figure 3.2: Logistic Regression	11
Figure 3.3: Decision Tree	12
Figure 3.4: Confusion Matrix	15
Figure 3.5 : Performance evaluation measures formula	16
Figure 4.1: cardio Distribution by age attribute	18
Figure 4.2: smoke	18
Figure 4.3: gender	18
Figure 4.4: Cholesterol	19
Figure 4.5: Gluc	19
Figure 4.6:active	20

LISTS OF TABLES

NAME OF TABLES	PAGE
Table 2.1: Summary of Related Work	6
Table 3.1: Details of datasets	13
Table 4.1: Experimental Results of Traditional Classifiers	21
Table 4.2: Experimental Accuracy of Different Algorithms	22

CHAPTER 1

INTRODUCTION

1.1 Introduction

A disorder that affects the heart or blood vessels is referred to as a heart disease. Since the heart is a vital organ in our bodies, its component functioning is essential to life. A person will pass away within a few minutes if their heart doesn't beat properly, which will also cause their brain and other organs to stop functioning. The prevalence of numerous heart-related disorders is rising as a result of changes in lifestyle, stress at work, and poor eating habits.

Even though we have great technology, we lack the information and awareness necessary to address the issue that is directly responsible for the majority of human deaths. People are not aware of the risk of cardiovascular diseases, as seen by the 17.9 million deaths annually from heart disease. Heart attacks and other heart diseases are the leading causes of death in various nations (such as Kazakhstan, Russia, Afghanistan, Tajikistan, and Mongolia). The COVID-19 virus, also known as the coronavirus, is more deadly and a leading cause of mortality for those with cardiovascular conditions, according to the World Health Organization. According to WHO data, people with any form of cardiac disease have a higher death rate. If the heart is healthy, good health and improved immunity can be guaranteed. Cardiac disease is predisposed to by a number of behavioral risk factors, including smoking, excessive alcohol and caffeine use, stress, and physical inactivity, in addition to physiological risk factors like obesity, hypertension, high blood cholesterol, and pre-existing heart diseases.

Technology must be developed to find a way to detect cardiovascular disease because it is the most serious of all diseases and can either be treated or prevented. In the research community, machine learning techniques have attracted a lot of interest. Since we have a large number of medical datasets, machine learning can assist us in extracting patterns and useful information from them. Machine learning has a wide range of applications, but in the medical industry, it is primarily used to forecast disease. Due to the fact that machine learning may speed up the diagnosis process while also improving accuracy and efficiency, it has attracted the attention of many researchers. Machine learning algorithms can be used to identify a variety of diseases,

however this study will concentrate on heart disease diagnosis. We may utilize machine learning to solve this problem by predicting if a person has cardiac disease or not.

The following sentences were used to conduct our research as we looked closely at these procedures for this paper.

- We pre-processed the dataset using a variety of techniques, including addressing unbalanced data and feature scaling.
- Using supervised machine learning algorithms (RF, LR, KNN and DT), we were able to predict the disease state with greater accuracy.
- We've figured out the outcomes of a lot of evaluation measures to compare with other people's performances.

The addition of five new ones has improved the remaining sections of the paper. We cited the literature review in Section 2 of the paper. Section 3 offered a dataset description. We provide an overview of our experimental strategy in section 4. Section 5 has a description of the results of our experiment. Our investigation comes to a conclusion in Section 6.

1.2 Motivation

In the current world, the prevalence of cardiovascular disease (CVD) is rising daily. The risk of death can be decreased by detecting any heart-related illnesses in their early stages. The identification of heart disease is one of the most crucial and challenging tasks in the medical field. To save lives, it needs to be detected swiftly, effectively, and correctly. Before diagnosing CVD, numerous tests are carried out, including auscultation, ECG, blood pressure, cholesterol, and blood sugar. Prioritizing the tests is crucial since they are frequently extremely time-consuming when a patient's health is potentially serious and treatment must be started right away. In order to take action to save death, an effective, precise, and early medical diagnosis of heart disease is essential. With the aid of machine learning, the system can quickly identify cardiac diseases by learning from past data sets. Because of this, researchers have been interested in predicting cardiac disease and have created a number of systems employing machine learning algorithms. As a result, in this study work, we use machine learning to score the diagnostic tests and identify some lifestyle factors that contribute to heart disease.

1.3 Objectives

- To determine the severity of heart disease.
- To investigate related issue.
- Life-Quality
- To raise awareness of heart diseases among the public

1.4 Research Questions

- What are the main characteristics of this database?
- How does the research's algorithm operate?
- How do you forecast heart diseases?
- If a person has heart disease or not, what would be the success rate?

1.5 Report Layout

- Background
- Research Methodology
- Experimental Result and Discussion
- Conclusion and Future Analysis
- Reference

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Through the analysis of all these variables and the use of a variety of approaches, including machine learning, we attempted to predict heart disease in this study.

This section includes a variety of the heart prediction research that has been done. Techniques for determining whether a person will develop heart disease or not can be very beneficial for both patients and the medical field. If we are aware of the dangers of having heart disease, we can raise awareness among the populace or advise them to take health precautions. As a result, some of them are reviewed in this section given that numerous researchers have employed various methodologies or models to predict heart disease.

2.2 Review Works

Here we have discussed five heart diseases detection related research work.

1. In [3], researchers investigated and analyzed various data mining techniques used for the prediction of heart illness.
2. A prototype for heart disease prediction called Intelligent Heart Disease Prediction Systems was created by the authors of [4].
3. For predicting cardiac illness, writers in [5] used artificial neural networks (ANN), k-nearest neighbor (KNN), decision trees, and Naive Bayes.
4. The authors of [6] offer different heart disease-related characteristics as well as a model built on supervised learning techniques such Naive Bayes, decision trees, K-nearest neighbors, and random forest algorithms.
5. Rine Nakanishi [7] et.al evaluated ML methods for improving the prediction rate of coronary heart disease (CHD).

2.3 A Comparison of the Review Works

The authors of [3] analyzed Naive Bayes, Neural Networks, and Decision Trees and came to the conclusion that the accuracy of the prediction may depend on the quantity of characteristics employed. The authors of [4] have examined a number of heart-related disease risk factors and used classification matrices to determine whether or not the prediction is accurate. This approach can lower the cost of teaching and learning for medical students. Three data mining modeling techniques, Decision Trees, Naive Bayes, and Neural Networks, are used by the author in [4]. Comparing Decision Trees to the other two models, it exhibits a high accuracy rate of (89%). In [5], the patient risk level is categorized using data mining classification techniques as Naive Bayes, KNN, Decision Tree Algorithm, Neural Network, etc. When using more attributes, the risk level can be predicted with a high degree of accuracy. For the prediction of the cardiac disease, they took into account various characteristics and risk levels, such as greater than 50%, less than 50%, and 0. Data is trained, classified, and predicted or evaluated using the KNN and ID3 algorithms, respectively. The author [6] uses the current dataset from the UCI heart disease patient repository's Cleveland database. The author [6] makes advantage of the already-existing dataset from the Cleveland database of patients with cardiac disease at UCI. The algorithms producing the best outcomes in this model include random forest, Naive Bayes, and K-nearest neighbor. According to the results, K-nearest Neighbor provides the highest accuracy score. The Author [7] applied various machine learning algorithms on 6814 patient records. Here, we can observe that every author is using a small dataset. This is why their finding is not precise enough. In my research, I'm employing supervised machine learning methods to work with a large dataset. To assess the precise performance of our existing model, we also apply other performance evaluation criteria in this case.

2.4 Summary of Related Work

Table 2.1: Summary of Related Work

Source Methodology	Objectives	Result
Naive Bayes, Neural Networks, and Decision Trees	They have discussed Naive Bayes, Neural Networks, and Decision Trees and come to the conclusion that the accuracy of the prediction may depend on the quantity of characteristics used.	They concentrated on data mining techniques for categorization utilized for data discovery.
ANN, Decision tree, Naïve Bayes	Using data mining techniques, this research has created a prototype Intelligent Heart Disease Prediction System (IHDPS).	It permits the establishment of crucial knowledge, such as patterns and correlations between medical aspects connected to heart disease.
Naïve Bayes, KNN, Decision Tree Algorithm, Neural Network. etc.	Estimating each person's risk level based on their age, gender, blood pressure, cholesterol, and pulse rate.	This article provides an overview of the many classification methods.
Random forest, Naive Bayes, and K-nearest neighbor.	They have discussed Naive Bayes, Random forest, and K-nearest neighbor algorithms accuracy.	This article provides an overview of three classification methods

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The objective of the prediction methodology is to create a model that can extrapolate a predicted class' characteristics from a combination of other data. The goal of data mining in this study is to create models for class prediction based on chosen attributes. The goal of the study is to identify which model provides the best proportion of accurate predictions for the diagnoses by applying data mining techniques to forecast potential heart attacks from the patient dataset.

3.2 Description of the Proposed Method

In this study, the relevant information is collected from the UCI machine learning repository. The “Kaggle” website was used to access the dataset. Thirteen attributes are used to describe the aspects of the 70,000 sample dataset. The data was then processed in a Jupyter notebook using open-source Python tools. The current approach entails a number of processes, including data collecting and pre-processing. Various algorithms, including Decision Tree, Random Forest, KNN, and Logistic Regression, are used in the current work. Random forest has the highest level of accuracy. The training dataset was applied to the first stage of model learning. In Google Colab, the dataset was imported for this purpose using the Read CSV operator. Additionally, the dataset was duplicated four times in order to connect it to ML models. Train was used to prevent the selection of comparable values during the learning and testing phases of the model. This operator calculates a learning and testing phase's statistical analysis and model performance. Every Model was connected to the same dataset and changed at the same time as part of maintenance to determine the exact quality and accuracy of the test.

3.3 Diagram of the Proposed Method

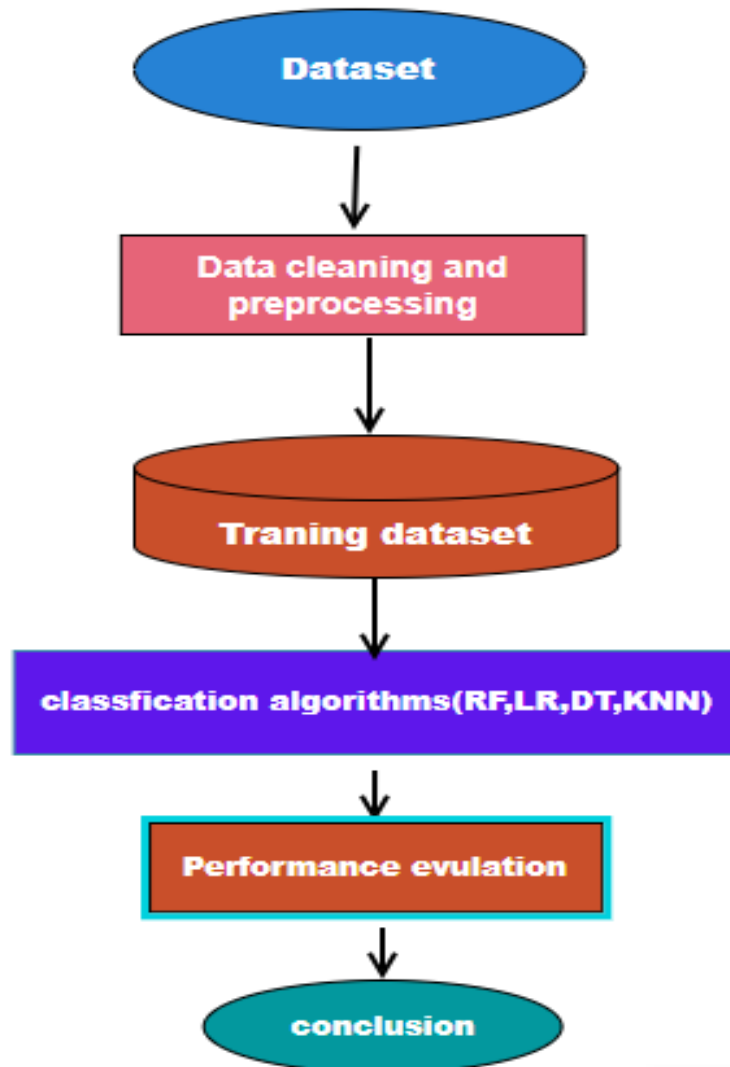


Figure 3.1: Proposed method

3.4 Proposed Algorithm

Step 1: Dataset Selection/Data Preprocessing {

Overview of the data

Identify and eliminate outliers

Outliers must be identified and removed.

Checked for missing data

}

Step 2: Model Evaluation {

Understanding the worth of data

Choosing Machine Learning Model

}

Step 3: Model Execution {

Data Import

Implementing all models concurrently

}

Step 4: Performance Evaluation {

Use the "Performance" operator to calculate accuracy.

Analyzing the outcome using the Confusion Matrix

}

Step 5: Compare the Results {

Comparing the precision of all models

}

3.5 Pre-processing

3.5.1 Categorical data encoding: The process of converting categorical data to a numerical value is known as categorical data encoding. We are aware that input and output variables for machine learning models must be numbers, hence categorical data in a dataset must first be converted to numbers before the model can be fitted. Our cardio-train dataset exclusively contains numerical data, with nominal data present in each column. Therefore, encoding the dataset was not necessary.

3.5.2 Missing Value Imputation:

The technique of replacing missing data with imputed values that were determined through research with other dataset data is known as "missing value imputation." Thankfully, there are no missing values in our datasets.

3.5.3 Feature Selection Technique: To identify significant features among those features, we conducted a new statistical analysis. We use univariate selection technique in our study work. `f_classif` score function are used here. We have greater final score accuracy thanks to the Chi-Square test. In order to make the model simpler and more effective, we deleted the less significant column after determining the score of each attribute.

3.6 Classification Algorithms

Data are categorized into the necessary number of classes using the supervision approach known as classification [9]. Discovering the causes of heart disorders and determining if a person has heart disease or not are the goals of this research. This makes it simple for us to forecast the heart patient.

In order to classify data, we have used four different techniques: Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and k-nearest neighbors (KNN). The best fitting algorithm for this part of the problem was then determined by comparing their performances based on various model evaluation measures.

3.6.1 Logistic Regression (LR): It's a Machine Learning (ML)-based approach where the class label contains two categories, yes/no like binary (0/1). Although they address a scenario in which they can be combined linearly through a logistic operation, logistic regression assumes that the seer are insufficient to resolve the response variable. When the seer does not meet the requirements to provide a more probabilistic evaluation of the feedback, LR excels in a select few aspects.

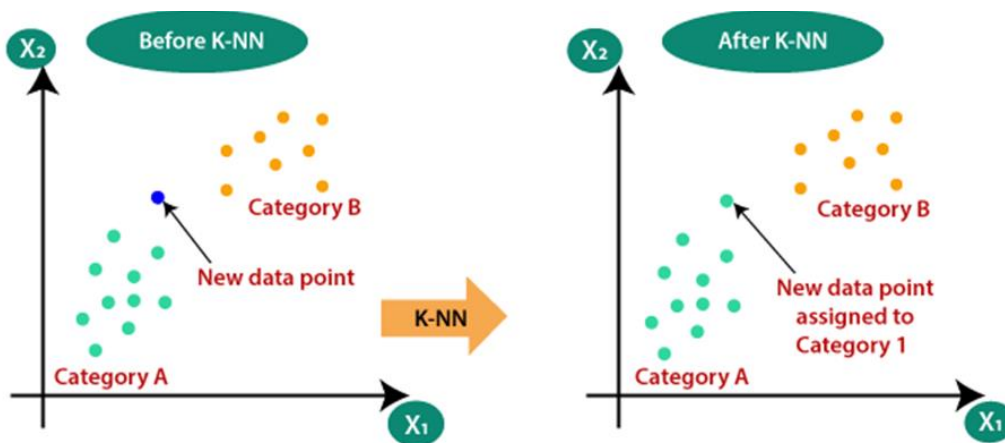


Figure 3.2: Logistic Regression

3.6.2 K-Nearest Neighbor: It is one of the simplest supervised learning-based machine learning algorithms. The new case is put in a part that is comparable to the existing section according to the KNN algorithm's estimation of the similarity between the new data and the existing cases. An additional data point is classified by this algorithm depending on the match and is stored along with all of the existing data. In other words, the K-NN algorithm can quickly classify new data into a better suite of categories when it arises. Although it can be used for classification and regression, the majority of the time, classification problems are the ones it is employed for.

3.6.3 Decision Tree (DT): It is a machine learning (ML) classifier that is supervised. In this model, the structure is shaped like a tree, with the leaf node representing the class label and the internal nodes representing features. It takes a completely different method of classification. In most cases, it develops a set of if-then rules that designate an observation to a particular branch of the tree. With ordinal outcome variables, it has more than two possible outcomes.

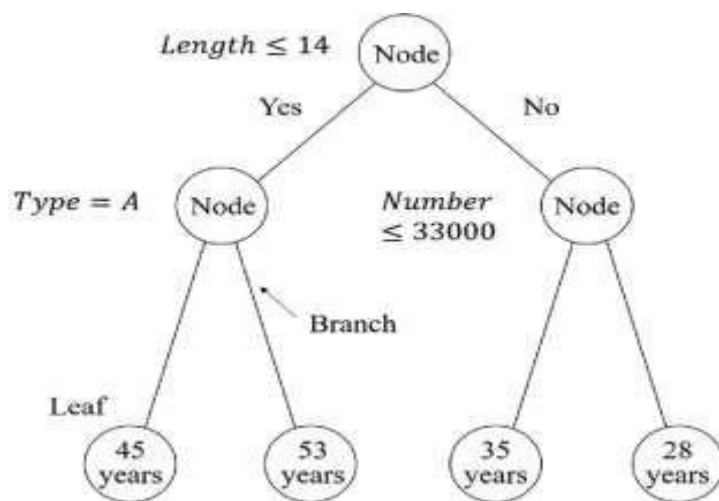


Figure 3.3: Decision Tree

3.6.4 Random Forest: A classifier that combines bagging and random feature selection is called a random forest. There is no need for preprocessing when using random forests. Probability estimate and prediction have both employed the Random Forest technique. For large data sets, random forest performs well. It is among the most precise classifiers. For a variety of data sets, particularly the data set on heart disease, it generates a classification that is incredibly accurate. We achieve the maximum accuracy in our analysis utilizing the Random Forest algorithm.

3.7 Research Instrumentation

Python is used across the platform Google Collaborator to implement all the experiments. Through the browser, it is possible to execute and write any Python code. With no setup required, Colab is a hosted Jupyter notebook. The system used for all studies has an Intel(R) Core(TM) i5-8250U CPU clocked at 2.70GHz, Windows 10 Pro 64-bit, and 8GB of RAM.

3.8 Dataset Description

The cardio train dataset, used in this study, was obtained from Kaggle Platform. The cardio train dataset, which was used in this study, was gathered through the Kaggle Platform. The dataset includes 70,000 patient records, 11 characteristics, and a target. For one patient, those 12 features are the most useful for predicting heart disease. Additionally, this dataset is essential for identifying the cause of heart diseases. The following table, Table 3.1, lists each entire description along with the number of attributes and values assigned to each attribute:

Table 3.1: Details of datasets

No	Attribute Description	Types of features
1	Age Objective Feature age int (days)	Objective Feature
2	Height Objective Feature height int (cm)	Objective Feature
3	Weight Objective Feature weight float (kg)	Objective Feature
4	Gender Objective Feature gender categorical code	Objective Feature
5	Systolic blood pressure Examination Feature ap_hi int	Examination Feature
6	Diastolic blood pressure Examination Feature ap_lo int	Examination Feature
7	Cholesterol Examination Feature cholesterol 1: normal, 2: above normal, 3: well above normal	Examination Feature

8	Glucose Examination Feature gluc 1: normal, 2: above normal, 3: well above normal	Examination Feature
9	Smoking Subjective Feature smoke binary	Subjective Feature
10	Alcohol intake Subjective Feature alco binary	Subjective Feature
11	Physical activity Subjective Feature active binary	Subjective Feature
12	Presence or absence of cardiovascular disease Target Variable cardio binary	Target Variable

3.9 Performance Evaluation Measure

There are a few performance evaluation metrics that we can use to determine the precise performance of our current model. These techniques evaluate overall performance based on unobserved data.

Confusion Matrix: The performance of an algorithm can be seen in a table called a confusion matrix. Two rows and two columns of the confusion matrix specify TP, FP, TN, and FN for two class problems.

To compare the number of accurate and inaccurate predictions provided by the model with the actual categorization of the heart disease data set, confusion matrices are employed. Below you can find the conventional classification matrix.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.4: Confusion Matrix

1. Specificity= $TN / (FP+TN)$
2. Sensitivity = $TP / (TP+FN)$
3. Disease prevalence= $(TP+FN) / (TP+FP+TN+FN)$
4. Positive predictive value (PPV): $TP / (TP+FP)$
5. Negative Predictive value (NPV): $TN / (FN+TN)$
6. Accuracy= $(TP+TN) / (TP+FP+TN+FN)$

Where TP=> Positive tuples that are correctly labeled by the classifier. TN=> Negative tuples that are correctly labeled by classifier. FN=> Positive tuples that are incorrectly labeled by classifier. FP=> Negative tuples that are incorrectly labeled by classifier. Positive predictive value (PPV) is defined as probability that the heart disease is present when the diagnosis test is positive.

Positive predictive value (NPV) is defined as probability that the heart disease is absent when the diagnosis test is negative

Accuracy: It refers to the proportion of test data predictions that were accurate. Where accuracy works, measurements can be accessed with actual measurements. It was founded on just one factor. Systematic errors are addressed with accuracy.

Precision: The percentage of accurately predicted positive observations is what it refers to. In actuality, precision identifies the real true portion of all the occasions when they would have anticipated true.

Recall: Actually, precision reveals the actual true portion of all the instances where they would have anticipated true.

F1 Score: It's basically the harmonic mean of the precision and recall.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 - score} = \frac{2 * P * R}{P + R} \quad (4)$$

Figure 3.5: Performance evaluation measures formula

CHAPTER 4

EXPERIMENTAL RESULTS & DISCUSSION

4.1 Introduction

In this area, our goals were to comprehend the critical causes of heart disease and to help people avoid contracting the disease. Therefore, the first study focused on identifying the contributing elements, and the second was to identify people who are more likely to be attacked by cardiac diseases. We did experiments with the Hold out technique. We partitioned samples into two distinct data sets using the Hold out method. Eighty percent of the data set is used to train and develop the classifier. The remaining 20% of the data set is for testing. We conducted experiments utilizing confusion matrix and several experiment methods.

4.2 Feature Analysis

To determine the important factors, we performed a uni-variate analysis. We have identified several significant findings and insights through the examination of the data collection. Figure 4.1 shows how cardiac disease is more prevalent in older people. The age range represented by the data set ranges from 30 to 65. Heart disease is most common in the elderly, and the risk of developing it increases between the ages of 56 and 60. While considering the smoke and gender figure, some important information has been discovered. Figure 4.4 depicts the relationship between heart disease and high cholesterol. From the Figure 4.4, we can see that who have high cholesterol, they have highly risk to attack by heart disease. People with high cholesterol are more likely to be attacked than others. Another critical factor is glucose level. We notice that, who have high level of glucose, they have heart diseases. Figure 4.5 depicts the likelihood of having heart disease in those with high glucose levels. Another important attribute is “active”. Figure 4.6 represents the risk of developing heart disease among inactive people. We've seen that persons who are physically active have a lower risk of developing heart disease.

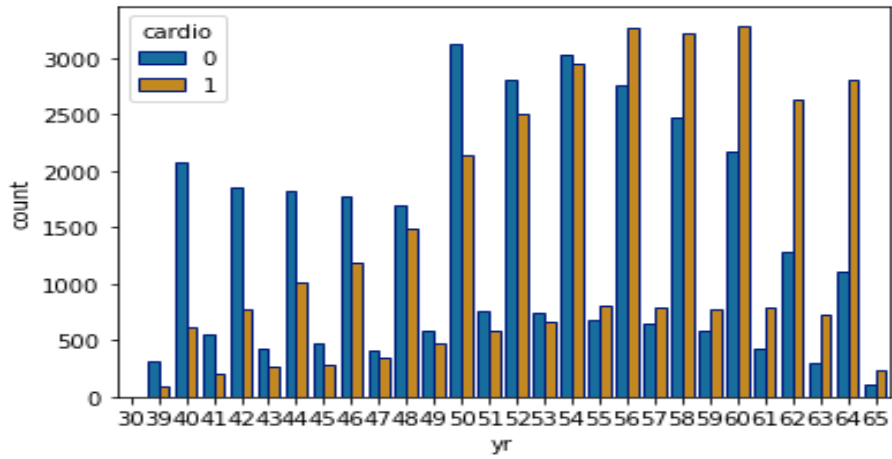


Figure 4.1: cardio Distribution by age attribute

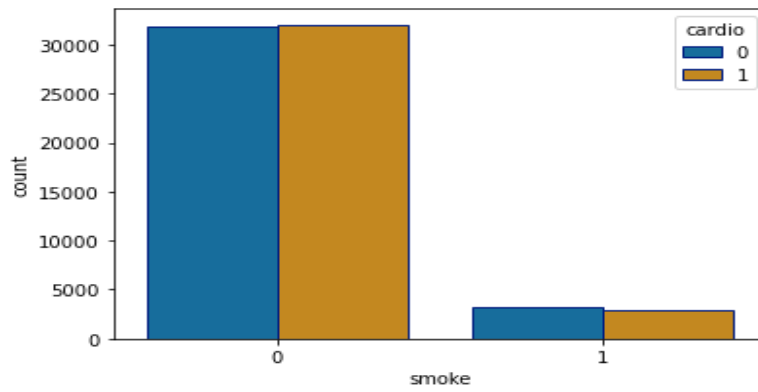


Figure 4.2: smoke

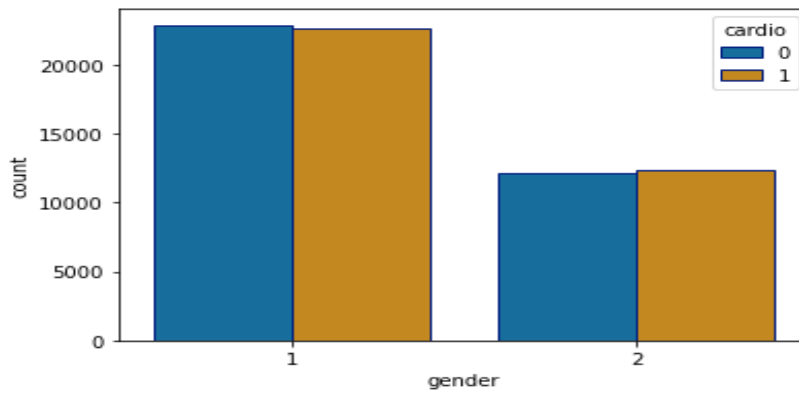


Figure 4.3: gender

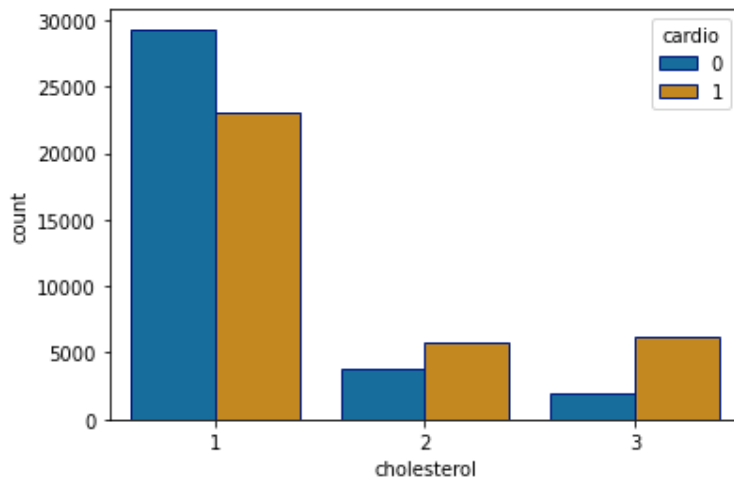


Figure 4.4: Cholesterol

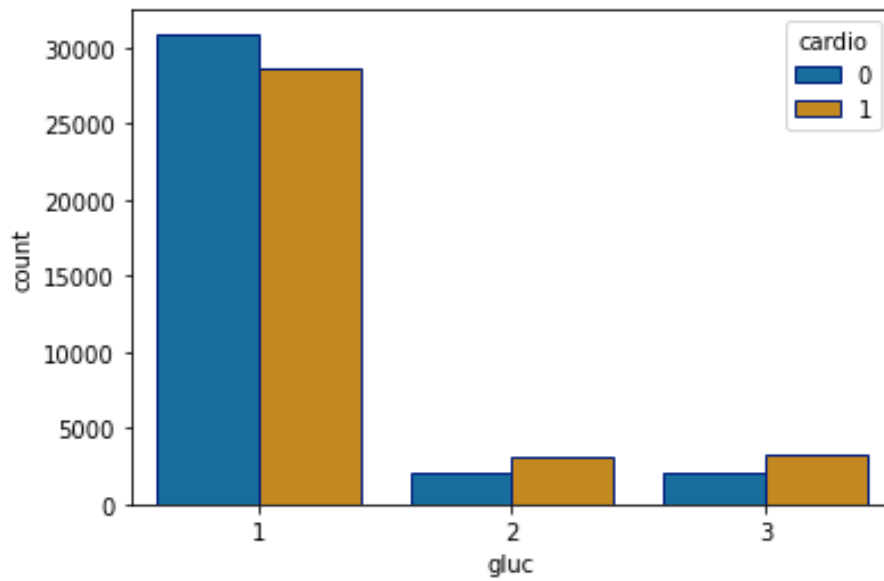


Figure 4.5: Gluc

```
In [74]: sea.countplot(x='active',hue='cardio',data=df,palette='colorbli
```

```
Out[74]: <AxesSubplot:xlabel='active', ylabel='count'>
```

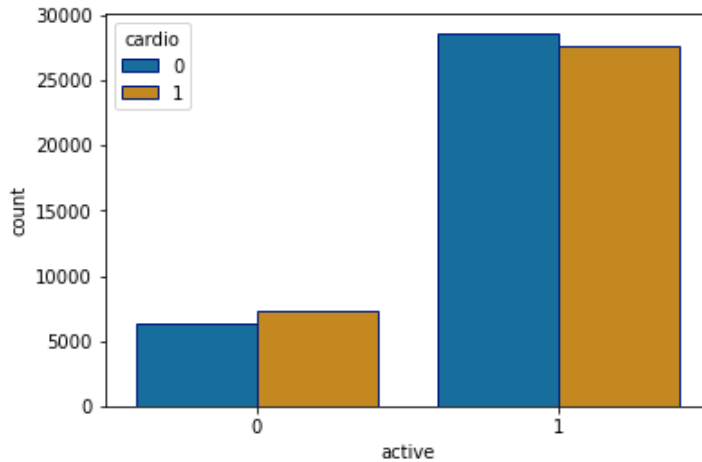


Figure 4.6: active

4.3 Experimental Results

Here, using the experimental findings of the evaluation metrics of the machine learning models of the diabetes dataset, we must do a comparative study. As a result, we have noted various performances while employing conventional classifiers.

Table 3, shows us Experimental results of supervised machine learning algorithms. A test for heart disease will likely come back negative when it is not there, according to specificity. When a diagnosis test is positive, the positive predictive value (PPV) represents the risk that a patient has cardiac disease. By using Random forest Algorithm, we find out 5103 TP value and TN value is 4987. Here success rate is 73%. Then using Logistic Regression algorithm, we find out TP value is 5118 and TN value is 4687. Here success rate is 71%. From the table 3, we can notice that we obtained high accurate result from Random Forest Algorithm. By using Decision Tree Algorithm, we find out 4442 TP value and TN value is 4450. Here success rate is 65%. Here we can say that we obtained low success rate from Decision Tree Algorithm.

Table 4.1: Experimental Results of Traditional Classifiers

Random Forest		
	True (0)	True (1)
Prediction (0)	5103	1814
Prediction (1)	2096	4987
Success Rate	73%	
Logistic Regression		
	True (0)	True (1)
Prediction (0)	5118	1799
Prediction (1)	2396	4687
Success Rate	71%	
K Nearest Neighbors		
	True (0)	True (1)
Prediction (0)	4924	1993
Prediction (1)	2397	4686
Success Rate	69%	
Decision Tree		
	True (0)	True (1)
Prediction (0)	4442	2475
Prediction (1)	2633	4450
Success Rate	65%	

Table 4, shows us classification report of 4 different algorithm. Here we calculated Accuracy, Precision, Recall and F1-score. First of all, considering the performances of traditional classifiers, the best accuracy has been obtained 73% using Random Forest classifier. For huge data sets, random forest performs well. As we are dealing with a vast dataset, the Random Forest Algorithm provides the best accuracy. By using Logistic Regression algorithm we obtained 71% accuracy. Its precision is 70.01%, F1 score is 70.5% and Recall is 70.2%.After that we find out 69% accuracy from k Nearest Neighbors classifier. Its calculative F1 score is 68.50%.Here, we obtained low accuracy from Decision Tree classifier algorithms. By using Decision Tree we obtained 65% accuracy. Its calculative F1 score is 65.50% and its precision is 63.50%.

Table 4.2: Experimental Accuracy of Different Algorithms

Algorithms	Accuracy	Precision	Recall	F1-score
Random Forest	73%	0.725	0.721	0.726
Logistic Regression	71%	0.701	0.702	0.705
K Nearest Neighbors	69%	0.685	0.6851	0.685
Decision Tree	65%	0.635	0.6356	0.645

Chapter 5

CONCLUSION

5.1 Conclusion

The classification of medical data is one of the most difficult tasks in medical informatics, according to some. However, it is one of the oldest works in the field of research. Thus, numerous approaches have been put out by others. Numerous areas still need to be improved. Through the use of supervised machine learning algorithms, we created an effective method for heart disease prediction in this study report. Heart disease forecasting is significantly aided by data mining. In order to extract the most crucial features from our dataset, we used feature selection technique. The dataset has been carefully reviewed by us. As a result, a variety of performances employing different algorithms have been noted. This research aims to make better use of datasets gathered from diverse medical databases by experimenting with multiple machine-learning methods to increase performance classifieds. It sheds light on the significance of machine learning in the healthcare industry and demonstrates how this technology may aid medical professionals by producing precise predictions. In this study, a strong machine learning algorithm is used to develop a reliable cardiac disease prediction system. This method makes use of old patient records from the past to forecast future cases at an early stage, saving lives. This study compares the effectiveness of various machine learning approaches and discusses which one provides the highest accuracy for a chosen dataset.

I'm using supervised machine learning techniques in my research to process a sizable dataset. We also use other performance evaluation criteria in this instance to determine the exact performance of our current model. They claim that the random forest method we presented (which we used for the heart data set) had the best accuracy, 73%. The mentioned experimental findings imply that our suggested method effectively achieves large degrees of dimensionality reduction and enhances accuracy with dominant characteristics. Our method performs better overall than previous methods. This inadvertently reduces the quantity of diagnostic tests required for the patient's diagnosis in order to forecast cardiac disease.

5.2 Future Works

The approach may be efficient and useful in the future for doctors and cardiac surgeons to quickly diagnose a patient's potential for a heart attack, according to the research. By including additional algorithms and using different datasets as our next step, we can finish our experiment. Our future goals should also include applying deep learning approaches and experimenting with other pre-processing methods. In the future, we are working with Bangladesh heart disease patient's purpose. Our future targets should also involve using an image processing approach for large data sets.

References

- [1] Obasi, Thankgod, and M. Omair Shafiq. "Towards Comparing and Using Machine Learning Techniques for Detecting and Predicting Heart Attack and Diseases." In *2019 IEEE International Conference on Big Data (Big Data)*, 2393–2402. IEEE, 2019.
- [2] Katarya, Rahul, and Polipireddy Srinivas. "Predicting Heart Disease at Early Stages Using Machine Learning: A Survey." In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 302–5. IEEE, 2020.
- [3] Gandhi, Monika, and Shailendra Narayan Singh. "Predictions in Heart Disease Using Techniques of Data Mining." In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, 520–25. Greater Noida, India: IEEE, 2015.
- [4] Palaniappan, Sellappan, and Rafiah Awang. "Intelligent Heart Disease Prediction System Using Data Mining Techniques." In *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 108–15. IEEE, 2008.
- [5] Thomas, J., and R. Theresa Princy. "Human Heart Disease Prediction System Using Data Mining Techniques." In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 1–5. IEEE, 2016.
- [6] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." *SN Computer Science* 1.6 (2020): 1-6.
- [7] Nakanishi, Rine, Damini Dey, Frederic Commandeur, Piotr Slomka, Julian Betancur, Heidi Gransar, Christopher Dailing, Kazuhiro Osawa, Daniel Berman, and Matthew Budoff. "Machine Learning in Predicting Coronary Heart Disease and Cardiovascular Disease Events: Results from the Multi-Ethnic Study of Atherosclerosis (Mesa)." *Journal of the American College of Cardiology* 71, no. 11S (2018): A1483–A1483.
- [8] Jabbar, M. A., B. L. Deekshatulu, and Priti Chandra. "Intelligent Heart Disease Prediction System Using Random Forest and Evolutionary Approach." *Journal of Network and Innovative Computing* 4, no. 2016 (2016): 175–84.
- [9] Dinesh, Kumar G., K. Arumugaraj, Kumar D. Santhosh, and V. Mareeswari. "Prediction of Cardiovascular Disease Using Machine Learning Algorithms." In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 1–7. IEEE, 2018.
- [10] Dwivedi, Ashok Kumar. "Performance Evaluation of Different Machine Learning Techniques for Prediction of Heart Disease." *Neural Computing and Applications* 29, no. 10 (2018): 685–93.

An Approach for Heart Disease Prediction Using Supervised Machine Learning Algorithms

ORIGINALITY REPORT

20%
SIMILARITY INDEX

17%
INTERNET SOURCES

4%
PUBLICATIONS

8%
STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	www.kaggle.com Internet Source	3%
3	www.researchgate.net Internet Source	2%
4	www.mdpi.com Internet Source	1%
5	Submitted to University of East London Student Paper	1%
6	Submitted to University of Bradford Student Paper	1%
7	Submitted to University of Sunderland Student Paper	1%
8	ijcseonline.org Internet Source	1%
9	www.coursehero.com Internet Source	<1%