

**EFFICACIOUS CARDIOVASCULAR DISEASE ESTIMATION USING MACHINE
LEARNING
BY**

**Md. Ashrak Al Arif Shohas
ID: 182-15-2171**

**Mahedi Hasan Bijoy
ID: 182-15-2150
AND**

**Meherab Hossain
ID: 182-15-2143**

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

Supervised By

Mr. Ohidujjaman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

MR. Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
FEBRUARY 2022**

APPROVAL

Md. Ashrak Al Arif Shohas, ID No: 182-15-2171, Mahedi Hasan Bijoy, ID No: 182-15-2150, and Meherab Hossain, ID No: 182-15-2143, submitted this project titled "EFFICACIOUS HEART DISEASE PREDICTION USING MACHINE LEARNING" to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements. The presentation took place on February 26th, 2022.

BOARD OF EXAMINERS

(Name) Chairman Designation

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

(Name) Internal Examiner Designation

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

(Name) External Examiner Designation

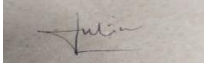
Department of-----Jahangirnagar

University

DECLARATION

We thus declare that we completed this study under the supervision of Mr. Ohidujjaman, Assistant Professor, Daffodil International University's Department of CSE. We further affirm that neither this project nor any part of it has been submitted for the granting of any degree or certificate to anybody else.


Supervised by:



Mr. Ohidujjaman

Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Md. Sabab Zulfiker

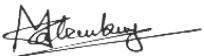
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md. Ashrak Al Arif Shohas

ID: 182-15-2171
Department of CSE
Daffodil International University



Mahedi Hasan Bijoy

ID: 182-15-2150
Department of CSE
Daffodil International University



Meherab Hossain

ID: 182-15-2143
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

By The grace of almighty Allah, we are successfully able to finish our final year Research-Based Project.

Mr. Ohidujjaman, Assistant Professor, Department of CSE, Daffodil International University, Dhaka, is our supervisor, and we owe him a great debt of gratitude. To complete this research and with our supervisor having extensive knowledge and a deep interest in the topic of "Data Mining." & His everlasting patience, knowledgeable direction, persistent encouragement, constant and vigorous supervision, helpful suggestions, and reading numerous poor versions and revising them at all stages allowed this project to be completed.

We would like to express our heartiest gratitude to Mr. Ohidujjaman, and Head of, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

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

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

ABSTRACT

We live in a modern era when our daily lives are undergoing numerous changes that have direct and indirect consequences on our health, which can be positive or negative. For this changing nature, where heart disease has grown more frequent, different forms of illnesses have substantially increased. Heart disease has been the most frequent cause of mortality throughout past years. The number of fatalities on heart among both men and women rises by the day. Changes in blood pressure, cholesterol, pulse rate, and other factors can contribute to cardiac disorders such as restricted or blocked blood arteries. Because most heart problems are identified at the very end, a precise forecast may lessen the tragedy associated with heart diseases. Because of this In this context, we use five machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, Support Vector Machine, and Naive Naves to three heart disease datasets combined to compare their performance in terms of attaining accurate prediction. The dataset comprises sixteen health characteristics that have been linked to heart disease. We also proposed combining these three datasets to produce a unique prediction that might discover a new accuracy point by offering a good forecast on the data of 5730 persons.

TABLE OF CONTENTS

| CONTENTS | PAGE |
|---|-------------|
| CHAPTER 1: INTRODUCTION | 1-4 |
| 1.1 Introduction | 1 |
| 1.2 Motivation | 2 |
| 1.3 Purpose of Study | 3 |
| 1.4 Research Questions | 3 |
| 1.5 Expected Output | 3 |
| 1.6 Report Layout | 4 |
| | |
| CHAPTER 2: BACKGROUND | 5-7 |
| 2.1 Introduction | 5 |
| 2.2 Related Works | 5 |
| 2.3 Research Summary | 6 |
| 2.4 Scope of the Problem | 6 |
| 2.5 Challenges | 6-7 |
| | |
| CHAPTER 3: METHODOLOGY OF RESEARCH | 7-22 |
| 3.1 Introduction | 7 |
| 3.2 Dataset | 8 |
| 3.3 Dataset Description & Preprocessing | 8 |
| 3.4 Co-Relations between features | 9 |
| 3.5 Classification Algorithms | 13-15 |
| 3.5.1 Logistic Regression | |
| 3.5.2 Decision Tree | |
| 3.5.3 Support Vector Machine | |
| 3.5.4 Naïve Bayes | |
| 3.5.5 Random Forest | |

| | |
|--|--------------|
| CHAPTER 4: EXPLORATORY RESULTS AND DISCUSSION | 16-24 |
| 4.1 Introduction | 16 |
| 4.2 Experimental Results | 17 |
| 4.3 Potential Future Improvement | 24 |

| | |
|---|-----------|
| CHAPTER 5: FUTURE IMPLICATION & CONCLUSION | 25 |
| 5.1 Study Synopsis | 25 |
| 5.2 Conclusion | 25 |
| 5.3 Evaluation | 25 |
| 5.4 Opportunities for Future Research | 25 |
| References | 26-27 |

| | |
|--|--------------|
| LIST OF FIGUREUREURES & TABLES | 15-31 |
| FIGURE1. The architecture of our proposed system | 15 |
| FIGURE 2: Heart Disease Frequency for Sex (Here are the images of 4 datasets) | 15-16 |
| FIGURE 3: Comparison Between Target Class Using Sex (heart_data & heart_data_2) | 16 |
| FIGURE 4: Comparison Between Target Class Using Sex (heart_data_3 & heart_data_4) | 16-17 |
| FIGURE 5: heart_data_4 results | |
| FIGURE 6: distribution curve for heart_data, heart_data_2, heart_data_3 | 18 |
| FIGURE 7: distribution curve of Dataset heart_data_4. | 18 |
| FIGURE 8: Co-Relations of Dataset heart_data_4 | 19 |
| FIGURE 9: Accuracy Score Of heart_data | 24 |
| FIGURE 10: heart_data_3 Accuracy Scores | 25 |
| FIGURE 11:heart_data_2 Accuracy Score | 26 |

| | |
|---|----|
| FIGURE 12:heart_data_4 Accuracy Score | 27 |
| Table 1.1: Report Layout | 04 |
| Table 4.1 Classification Report of Logistic Regression | 21 |
| Table 4.2 Confusion Matrix of Logistic Regression | 21 |
| Table 4.3 Classification Report of Naïve Bayes | 21 |
| Table 4.4 Confusion Matrix of Naive Bayes | 22 |
| Table 4.5 Classification Report of SVM | 22 |
| Table 4.6 Confusion Matrix of SVM | 22 |
| Table 4.7 Classification Report of Decision Tree | 23 |
| Table 4.8 Confusion Matrix of Decision Tree | 23 |
| Table 4.9 Classification Report of Support Vector Machine | 23 |
| Table 4.10 Confusion Matrix of Support Vector Machine | 24 |

CHAPTER 1

INTRODUCTION

1.1 Introduction

Human life is dependent on the heart's ability to operate properly. If the heart's capability is inadequate, it will have an impact on other organs of the human body, such as the cerebrum, kidneys, and so on. Failure of the heart may result in harm to other vital organs such as the kidneys and the brain. Heart disorders are classified into numerous groups. Congenital heart disease occurs when a newborn infant is born with heart problems. If, on the other hand, a person develops heart disease after reaching infancy, this is referred to as 'acquired heart disease.' Nowadays, the majority of heart disease case studies fall into the category of acquired heart problems. Cardiovascular disorders, heart attacks, coronary heart disease, and stroke are the most frequent heart-related ailments. Stroke is a kind of heart disease caused by the blockage of blood arteries by excessive pressure of blood [1][2][3].

In 2008, around 17.3 million people died exclusively as a result of heart disease. Nearly 23.6 million deaths by Heart disease will happen before 2030, according to the World Health Organization[5]. 85 percent of these deaths were responsible for a heart attack or stroke. More than three-quarters of all CVD deaths occur in low- and middle-income countries. CVDs accounted for 38% of the 17 million premature deaths caused by noncommunicable diseases in 2019. [4]

Coronary artery disease (CAD), a kind of heart disease, is becoming an increasingly serious medical and public health issue, and it is the top cause of death in Bangladesh and other nations. The underlying pathophysiology is unknown. Statistics demonstrate that risk factors for heart disease have a multiplicative effect[6]. Every day, many heart-related data are generated. By examining this data, we can determine which stage we are now in and whether or not we will be influenced soon. A plethora of procedures for identifying heart problems have been created as technology has advanced. However, forecast accuracy has not yet reached the desired level. Because our technology will anticipate heart illness for the user, our initiative intends to raise public awareness about heart disease. In this work, we will look at comparative examinations of several machine learning algorithms used in predicting heart disease using three distinct datasets.

1.2 Motivation

The prevalence of heart disease is rising at an alarming rate. People's hectic lifestyles in our period, with all the fast food at lunch breaks and then returning to sitting and working, have pushed us over the brink. Along with this, individuals nowadays are less active and do not get enough exercise. For the majority of them, leisure consists of watching another movie in bed or doing anything technologically related. Physical activity has been substantially curtailed. These conditions contributed to an alarmingly high prevalence of heart disease. The prevalence of heart disease in a developing country like ours has the same impact. Since

1990, the yearly death rate per 100,000 persons from cardiovascular illnesses in Bangladesh has climbed by 128.9 percent, or 5.6 percent per year [7].

Heart disease prediction is a tough and dangerous endeavor. Because it is directly related to people's health, precision is essential. It can be devastating if not precisely predicted. As a result, this study compares several data mining strategies for predicting it. It displays a comparison of the various approaches. To compare the strategies, cross-validation error is employed. We picked Decision Tree, Random Forest, Multiple Linear Regression, Support Vector Machine, and Naive Bayes because they are the most effective widely used techniques in determining diseases.

The internet has given us access to the whole planet. For health care objectives, a variety of websites and web applications are employed. We are aware that the Internet has grown in popularity in our nation, therefore we decided to establish a website that would save our people time while providing the anticipated results. As a result, we were inspired and sought to merge a complete heart disease prediction system with machine learning, with an emphasis on acquiring actual data or risk factors that may cause heart disease. If we can raise the awareness of at least 10% of the population regarding their heart health, we will have accomplished our most important goal for the project.

1.3 Purpose Of The Study

Heart disease is now regarded as one of the most dangerous illnesses. Many individuals die as a result of heart disease. It is regarded as the most appealing sickness in terms of getting associated with the heart. So the basic issue is that we can't recognize or comprehend cardiac sickness in its early stages. In this case, the Machine Learning approach provides an excellent solution for detecting coronary artery disease. We picked heart illness as our study topic after doing extensive investigation and analysis. The study topic has been chosen to reduce the number of people who die as a result of heart disease. Finally, the paper has been working on this to make a better recommendation that would assist us in lessening the number of deceased individuals in our current day.

1.4 Research Question

Several dangerous illnesses have already been discovered in humans. Although each illness has a preventive remedy, it is not attainable for everyone owing to unconsciousness. Everyone wants to live a joyful life when the only impediment is an illness. If the illness is still in its early stages, any type of disease prevention is conceivable. As a consequence, we developed a prediction method that assists in determining the illness stage and gives us the outcome of whether or not he or she is impacted. Among all diseases, heart disease is regarded as one of the most serious. Many people have perished as a result of this sickness. Finally, for our gratification, we chose it as our research topic.

Heart disease is one of several disorders that have a major impact on our lives. It is a severe condition since we frequently hear that heart disease is the leading cause of death and that other similar illnesses are heart-related [8][9][10]. When we considered putting our theory into action, the following questions arose:

- How can we estimate heart disease risk with more accuracy?
- How can individuals be made aware of their heart health?
- How can we minimize sudden death from heart disease?
- How can people be made aware of their eating habits?

1.5 Expected Output

Many people are completely unaware of their health. As a result, people are suffering from a variety of ailments, the most prevalent of which is heart disease, and in the long run, death is knocking. People will be more aware of their heart health if they use our system to monitor their heart state at all times. As a consequence, the death rate from heart disease will be reduced. We cannot give any medical assistance through our initiative, but we can raise awareness about changing one's lifestyle, eating habits,

quitting smoking, and so on by displaying the risk level of heart disease. There has been a lot of research done about heart disease risk prediction based on a few numbers of features, which is why they don't give us an accurate rate of risk of heart disease, but in this research, we analyze a lot of features such as smoking, family history of heart disease, cholesterol, blood pressure, chest pain, age factor, gender, stress, regular exercise, taking the drug or not, and so on, and as a result, we can show the risk of heart disease with higher accuracy.

We want to develop the research utilizing machine learning techniques following the research topics. We anticipated a benchmark to fulfill our aim of achieving the following outcomes:

- Analyze the datasets Heart Disease of the patient should be classified with a high degree of accuracy.
- Determine the relationship between various features in the dataset and the development of Heart Disease.
- Selection of key characteristics for detecting heart disease condition
- Contrast several machine learning classification approaches.

As we said earlier that our primary dataset will be the one combination of three datasets which has six classes and having diabetes as the target class makes it more interesting to know about the people who have diabetes & who also have heart disease.

1.6 Report Layout

We Planned to structure our report with these five portions.

| | |
|----------------|--|
| First portion | In the first portion of the project report, we looked at the review, the rationale for the project, our goal, and the results. |
| Second portion | The second segment delves further into our background research on heart disease and literature review. We have also enrolled in many studies in this sector. |
| Third Portion | The third section discusses the research approach that we used. We also briefly discussed the classifier algorithms that were employed in this study. |
| Fourth Portion | The fourth chapter includes a detailed summary of our findings as well as comparative classifier accuracy studies. |
| Fifth Portion | Finally, the fifth portion evaluates the rundown, future scope of the investigation, and discusses other areas for concentration in the comparable field. |

Table 1.1: Reporting 5 portions of Layout

CHAPTER 2

BACKGROUND

2.1 Introduction

Is heart disease a term established to describe a wide range of heart-related healthcare? Machine learning classification techniques make forecasting cardiovascular disease a bit simpler. The biggest issue in medical science is heart disease. Using the classification technique in 'Machine Learning,' it is easier to predict the risk of heart disease. In 'Machine Learning,' training and testing data from different classes were categorized using a variety of approaches. We've previously discussed why we chose this project; we want to assist individuals to acquire their heart disease prediction results through our system, and we want to raise awareness about heart disease. People can benefit greatly from testing their risk factors for heart disease regularly and being aware of any type of heart disease before it affects them. In this chapter, we will go through all of the tasks that must be completed before proceeding. We would want to discuss any connected papers, comparative studies, the scope of the problem, and challenges in this section.

2.2 Related Works

Heart disease is currently a standout amongst the most dangerous diseases all over the planet. Studies have played an important role in the advanced therapeutic framework. Numerous writers have contributed to the field of cardiac disease prediction systems by applying various data mining approaches and machine learning algorithms. The objective is to improve accuracy and make the system more successful over time so that it can anticipate the likelihood of a heart attack.

In 2017, Assistant Professor Sanjay Kumar Sen demonstrated a feasible Heart Disease prediction article using machine learning. They calculated using Naive Bayes, SVM, Decision Tree, and K-Nearest Neighbor Machine Learning. The highest exactness (84.1584) was provided by the Support Vector Machine [12].

Mydhili and Sujata Joshi compare the performance of three classification algorithms: particular decision tree, Naive Bayes. On a certain dataset, it determined the best prediction method in terms of accuracy and error rate [14].

We already said that there are some relevant works that we discovered and specified some of them. The Efficient Heart Disease Prediction System is a heart disease risk level prediction online tool that is intended to raise public understanding of heart disease terminology.

- Make a forecast of human heart disease.
- Collect statistics and disseminate information regarding cardiovascular disease.

It is a web program that seeks to forecast human heart disease, although it is not completely reliable in its output results.

Disadvantages of the current app

- It must be viewed using a web browser.
- It is unable to deliver alerts to the user.
- It does not have a lot of information regarding cardiovascular illness. Only limited stereoscopic prediction systems are accessible; no user feedback mechanism is offered.
- The security problem is unsolvable.

In our suggested application, we give as many features as possible to users to assist them and increase system efficiency.

2.3 Research Summary

According to prior research and analysis, there has been a small number of studies in this sector. In their own right, the studies have been fairly effective. Many additional illnesses have been studied using this sort of automated classification issue. Researchers have gone through many different paths, from exploring various algorithms to re-optimizing the present method to get better outcomes. The observable component is that, despite the high accuracy, we have yet to see any genuine application of these procedures. Probably, the thought of contacting a computerized diagnosis system for a condition isn't as appealing to the general people as consulting a doctor. However, with more accuracy and some testing, a completely automated diagnosis would most likely be as common as contacting a doctor.

2.4 Scope of the Problem

The major cause of mortality is heart disease. Many people died of heart disease as a result of their poor daily routines and eating habits. That is why we decided to research heart disease risk prediction to minimize the death rate using our technology. We presented a method that provides people with the predictive value of their heart disease risk so that they are aware of their heart health. As a result, the current study focuses on heart disease.

2.5 Challenges

One of the most difficult aspects of forecasting accuracy is data collecting. It is not feasible to forecast without data, and it cannot predict. Then comes preprocessing, which is a new problem. After preprocessing, our data set contains no null values, which allows us to make a solid forecast. Following that, feature scaling aids in putting all feature values on the same value scale. As a result, a new algorithm was applied to the suggested design. Finally, a technique for obtaining accurate anticipated value has been established. Several issues arose as a result of the working technique.

CHAPTER 3

METHODOLOGY OF RESEARCH

3.1 Introduction

We already knew from reading research articles that these five machine learning approaches provide greater accuracy. Data cleaning is a method that cleans information by removing missing information, copying information, and resolving information anomalies. As a result, information quality improves, increasing the usefulness of the information. Data transformation refers to the process of moving information or data from one organization to another. When a source configuration is expected to move over into the needed organization for a specific reason, this is often done. It is essentially the tentative or experimental translation of numeric or alphabetic advanced data into a revised ordered and rearranged structure. The basic goal of information reduction is to reduce several measurements of information into useful data.

Machine learning, abbreviated as ML, is a branch of computer science, specifically artificial intelligence, that uses statistical approaches to enable computers to learn using data without being explicitly taught (coded just once) [17].

Machine Learning (ML) is described as a technology that processes and analyzes vast amounts of data in a database to find relevant patterns/trends and new relationships between characteristics to achieve various intended goals. There are two kinds of machine learning algorithms. [18]

1. Supervised learning
2. Unsupervised learning

Supervised learning infers outputs from labeled training data consisting of a collection of training instances, whereas unsupervised learning infers outputs from unlabeled training data. Supervised learning is concerned with classification problems, whereas unsupervised learning is concerned with clustering problems. Another technique to machine learning is reinforcement learning, which refers to making judgments depending on their surroundings. Machine learning algorithms are divided into two stages. [19]

1. Training phase
2. Test phase

In the training phase, a machine learning algorithm is used to train the system using the provided data, and in the testing phase, new data is fed into the previously formed system for output. The system processes the new input and adjusts its models accordingly.

The most recent explosion of medical data has occurred as a result of machine

automation and the use of digital technology in illness diagnosis and treatment. Machine learning has been used to uncover and extract new patterns and beneficial information in medical advances. Although the adoption of automated illness categorization is still not widespread and acceptable in the medical community, it remains a study topic with huge promise for data scientists and researchers worldwide. As a result, we seek to investigate the notion of machine learning to create a machine-learned system to assess cardiac disease.

3.2 Dataset

For our study, we used three data sets that were already prepared or gathered. The dataset's claimants have made the dataset available in the UCI Repository, from which we acquired access [20][21][22]. The following are the contact details for the creators of this dataset.

3.3 Dataset Description & Preprocessing

Let us consider our first dataset as “**heart_data**”, we collected this data from the Kaggle website and It comes from a running heart study from Framingham, Massachusetts people. It has 4241 data. The classification's main target is to determine if a patient has a 10-year risk of future coronary heart disease which is known as CHD. The dataset contains information about the patients. There are almost four thousand records and fifteen qualities in all Variables.

Each trait has the potential to be a risk factor. There are risk variables that are demographic, behavioral, and medical.

Now let us consider the second dataset as “**heart_data_2**”, The collection contains 1190 patient records from the United States, the United Kingdom, Switzerland, and Hungary. It consists of 11 characteristics and 1 target variable.

For the third dataset Let us consider “**heart_data_3**”, Davide Chicco and Giuseppe Jurman data set: Machine learning can predict survival in heart failure patients based solely on serum creatinine and ejection fraction.

Lastly, let us consider the combined dataset as “**heart_data_4**”, In this dataset, we have taken 6 features that are similar in those three datasets, and among those three datasets combined, we made the diabetes dataset the target dataset in this dataset.

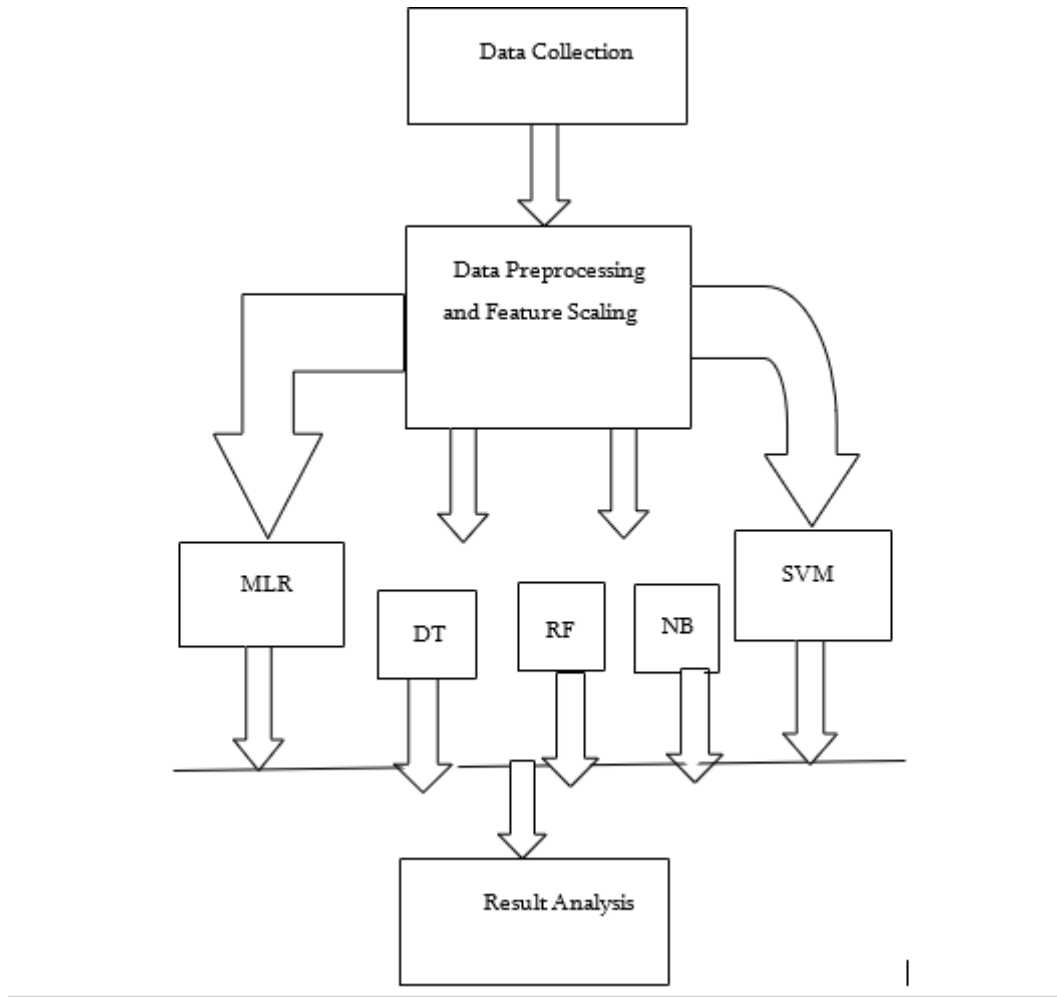


FIGURE. 1. The architecture of our proposed system.

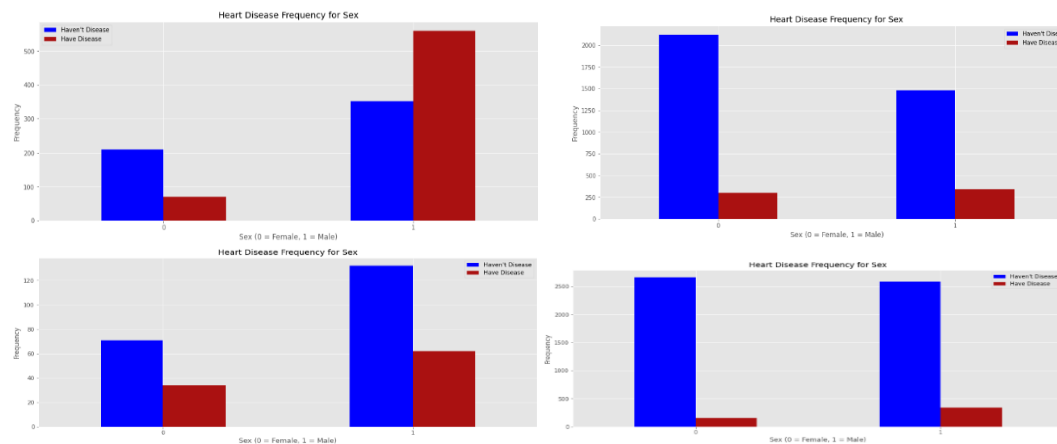


FIGURE 2: Heart Disease Frequency for Sex (Here are the images of 4 datasets)

In those four images, we find out the frequency of sex for Heart Disease where we can see that our primary data “heart_data_4” has more female patients who have heart disease. Here red color indicates those who have heart disease and those who have not the disease are marked in Blue. Here 0 means the female sex and 1 means the male sex.

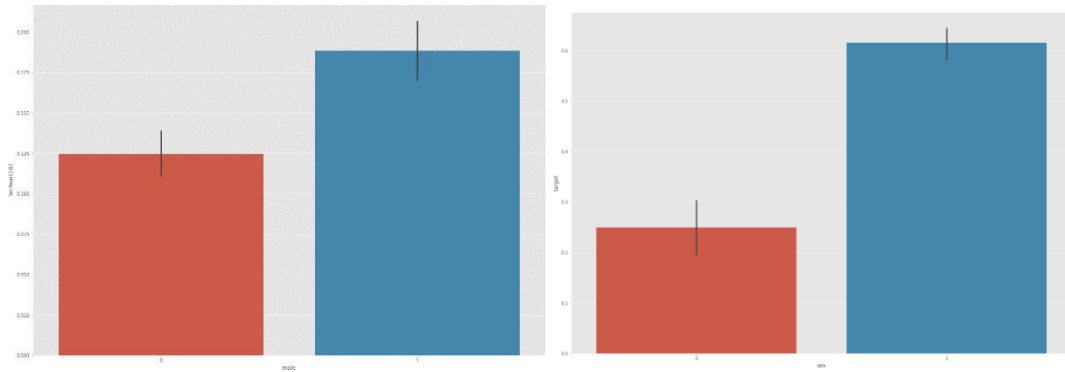


FIGURE 3: Comparison Between Target Class Using Sex (heart_data & heart_data_2)

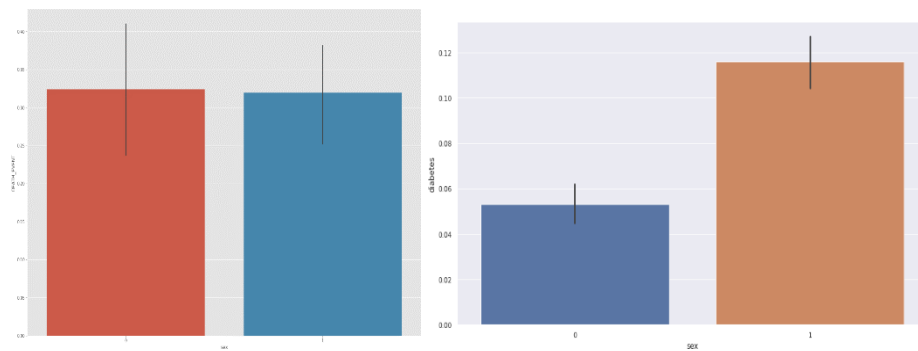


FIGURE 4: Comparison Between Target Class Using Sex (heart_data_3 & heart_data_4)

Here 0 means the female sex and 1 means the male sex.

We can mainly Focus on the heart_data_4 Result because that dataset is the combination of heart_data, heart_data_2 & heart_data_3. But the target class of that dataset is diabetes which is why we separately show the images.

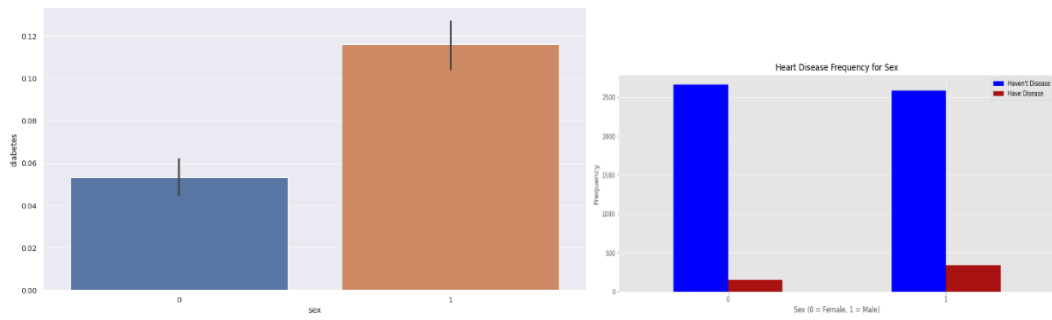


FIGURE 5: heart_data_4 results

3.4 Co-Relations between features



FIGURE 6: Distribution curve for heart_data, heart_data_2, heart_data_3

From these curves we see that 52-53-year-old people are most in the heart_data_2, 41-42-year-old people are most in the heart_data, 59-60-year-old people are most in the heart_data_3.

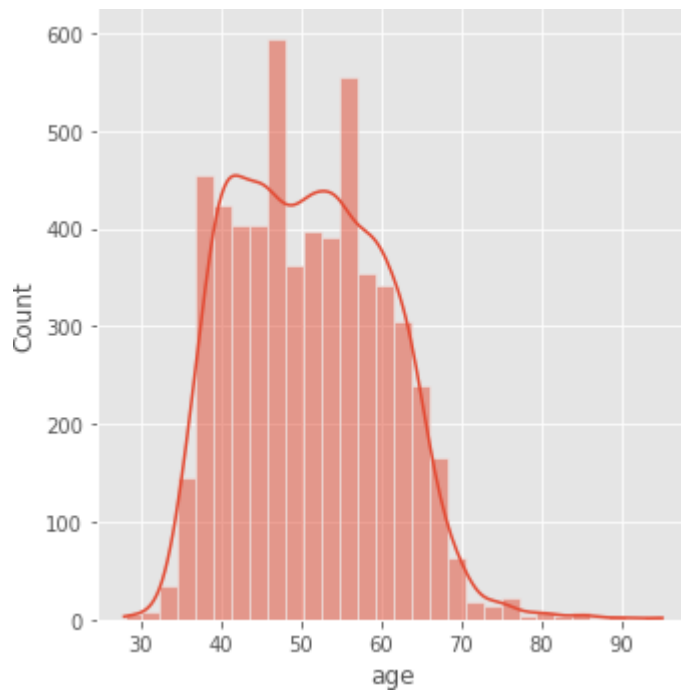


FIGURE 7: Distribution curve of Dataset heart_data_4

48-49-year-old people are most in the heart_data_4 As this data is the combination data of those three datasets so it is clear that the Maximum Person at the age of 48-49 faces heart diseases.

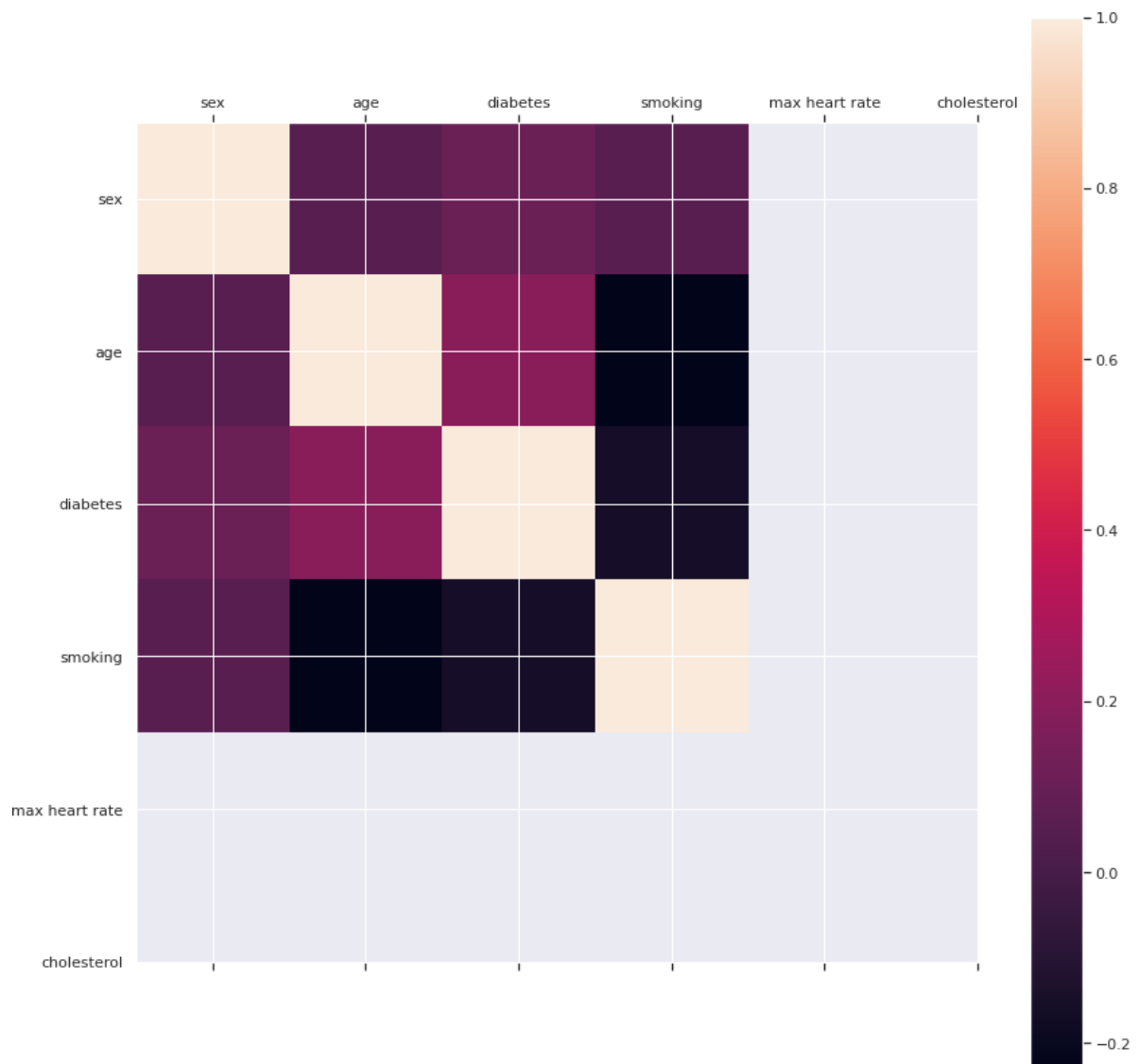


FIGURE 8: Co-Relations of Dataset heart_data_4

3.5 Classification Algorithms

Machine learning computations have recently grown in popularity. Machine Learning Algorithms enable PCs to learn from data through the use of quantitative approaches. In this method, a comparable computation may be linked to datasets from other places without requiring any changes to its internal constructions. Here are several types of machine erudition computations, but we have used approximately of them in our context. The algorithm is described in full below:

3.5.1 Logistic Regression

It addresses the classification concerns. It is used to forecast two-fold outcomes for a given set of free elements. The outcome of the reliant variable is discrete. Loads or coefficient respects are used to link information regarding (x) straightly to section yield regard (y). The yield regard displayed is a composite quality (0 or 1) rather than numeric regard, which distinguishes it from straight backslide.

3.5.2 Decision Tree

Decision trees are an important type of computation for showing predictive machine learning. Traditional decision tree computations have been known for a long time, and modern variants such as arbitrary timberland are among the most cutting-edge processes accessible. CART, which stands for Classification and Regression Trees, is a more recent moniker for the modest decision tree computation. This computation continues to be a recursive approach until the typical results are discovered. It provides excellent precision and versatility. The following equations provide access to this formula.

3.5.3 Support Vector Machine (SVM)

Given validated planning data (oversaw taking in) at the end of the day, the algorithm produces a flawless hyperplane that arranges new points of reference.

A line that separates the space of information variables is called a hyperplane. In SVM, a hyperplane is chosen to best separate the foci in the information variable space by their class, which is either class 0 or class 1. This may be seen as a line in two dimensions, and we should expect this line to isolate the majority of our information focuses. $B_0 + (B_1 * X_1) + (B_2 * X_2) = 0$ (3)

The learning computation determines the coefficients (B1 and B2) that determine the inclination of the line and the capture (B0) from equation 3, while X1 and X2 are the two information elements.

3.5.4 Naive Bayes

Naive Bayes is a classification technique based on the Bayes Theorem that calculates a likelihood by counting the occurrence of characteristics and the mix of qualities in verifiable data. The Bayes hypothesis calculates the chance of an event occurring

given the likelihood of a previous event. $P(B \text{ in the presence of } A) = P(A \text{ and } B)/P(A)$

The benefit of the Naive Bayes method is that it requires less data preparation to evaluate the parameters essential for classification.

3.5.5 Random Forest

Random Forests is a classification and regression approach that uses ensemble learning. It generates several Decision trees during training and outputs the class that is the mode of the outcome classes by individual trees. It also seeks to mitigate the concerns of high variation and high bias by identifying an average natural equilibrium between the two extremes.

CHAPTER 4

EXPLORATORY RESULTS AND DISCUSSION

4.1 Introduction

The outcomes of the guided experiment will be discussed in the following chapters. We will investigate and compare the accuracy and performance of several classifiers. We will display the findings in graphs as well as tables.

4.2 Experimental Results

Many machine learning methods (Random Forest Classifier, for example) are used in our work. Regression logic. Classifiers such as Naive Bayes, Decision Trees, Support Vector Machines, Logistic Regression). Among our four datasets, we can say the last dataset which we made by combining the three datasets are considered as the primary dataset for this experimental Result part, we divide the dataset into two sets using the Hold out technique Each data set:

1. Training Set (80% of the data) – 4583 data.
2. Test Set (20% of the data) – 1146 data.

The models are trained using a training set and after building a model, it is tested on a test set.

The following sections extensively discuss the results from our study for all datasets. Lastly, we will show the Classification Report & Confusion Matrix of the heart_data_4.

Now Here are some FIGURE that we obtained from google coal rotary where we tested our datasets and get the predictions.

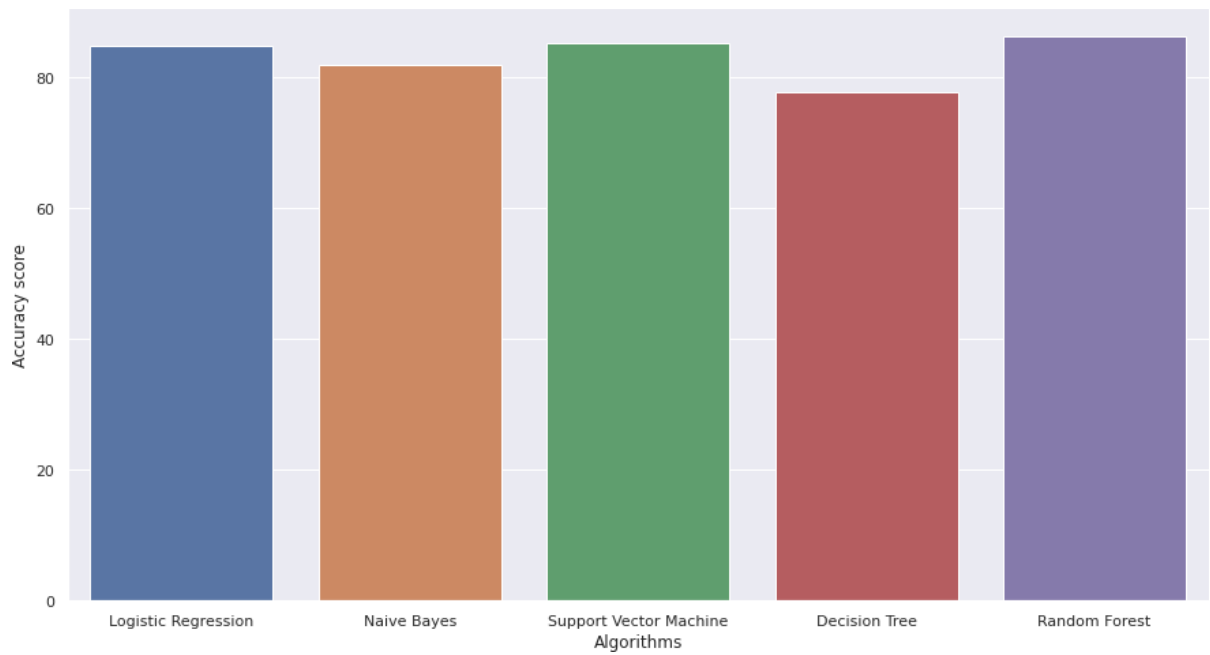


FIGURE 9: Accuracy Score Of heart_data

Logistic Regression earned an accuracy score of 84.79 percent.

The accuracy attained using Naive Bayes is 81.96 percent.

The Support Vector Machine earned an accuracy score of 85.14 percent.

The Decision Tree earned an accuracy score of 77.59 percent.

Random Forest obtained an accuracy score of 86.2 percent.

From these five classifiers, Random forest gave the highest accuracy.

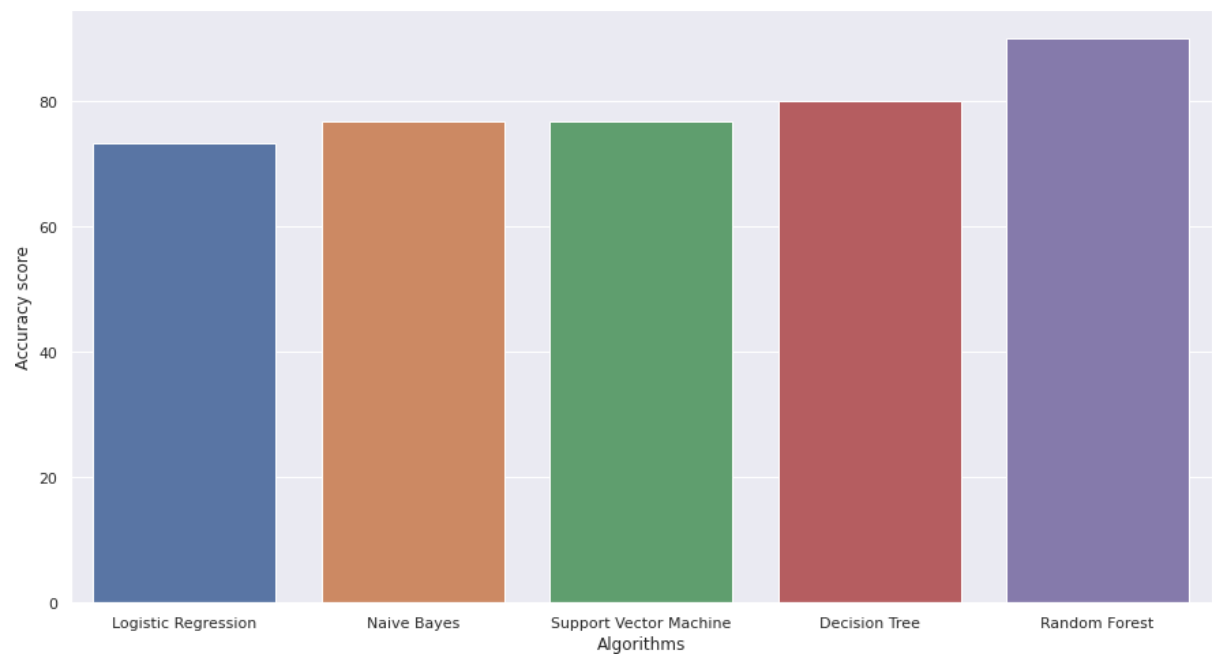


FIGURE 10: heart_data_3 Accuracy Scores

Logistic Regression earned an accuracy score of 73.33 percent.

The accuracy attained using Naive Bayes is 76.67 percent.

The Support Vector Machine earned an accuracy score of 76.67 percent.

The Decision Tree earned an accuracy score of 80.0 percent.

Random Forest received a 90.0 percent accuracy score.

From these five classifiers, Random forest gave the highest accuracy.

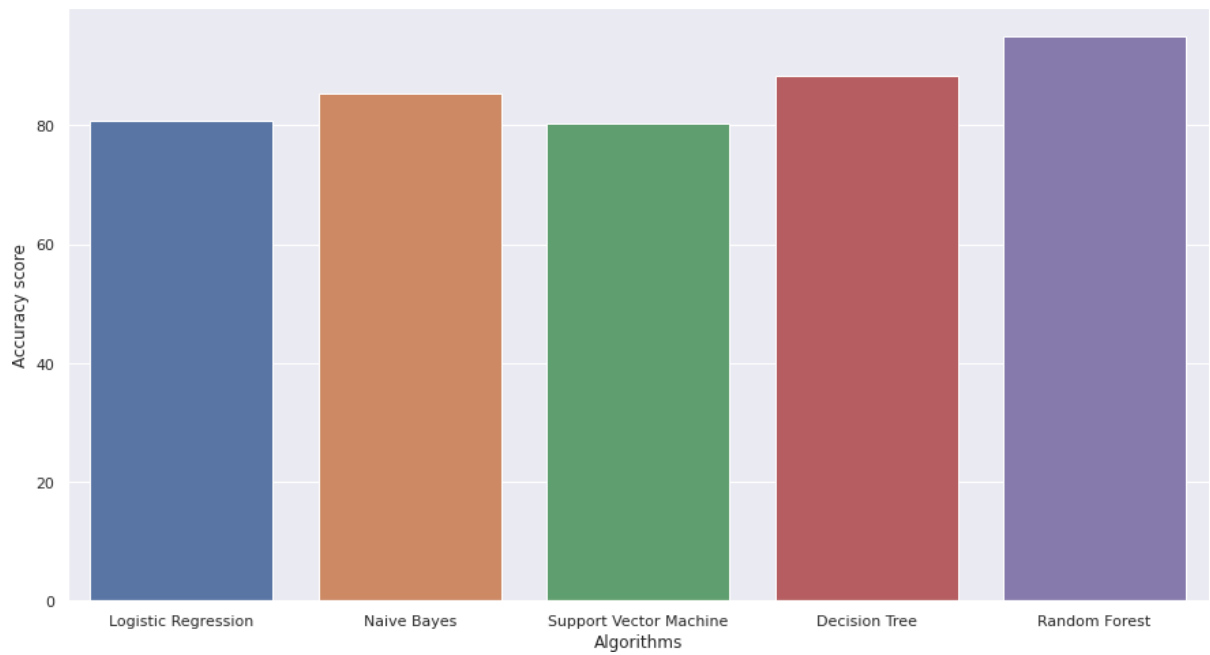


FIGURE 11:heart_data_2 Accuracy Score

Logistic Regression earned an accuracy score of 80.67 percent.

The accuracy attained with Naive Bayes is 85.29 percent.

The Support Vector Machine earned an accuracy score of 80.25 percent.

The Decision Tree earned an accuracy score of 88.24 percent.

Random Forest earned an accuracy score of 94.96 percent.

From these five classifiers, Random forest gave the highest accuracy.

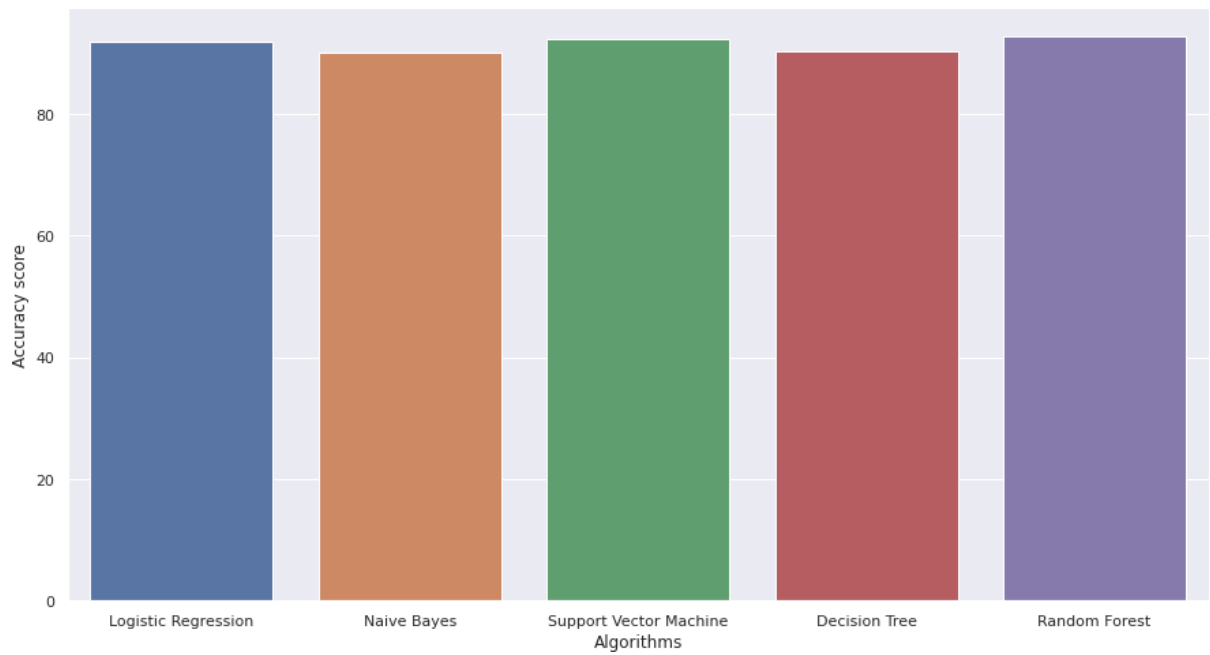


FIGURE 12:heart_data_4 Accuracy Score

Logistic Regression earned an accuracy score of 91.8 percent.

The accuracy attained with Naive Bayes is 90.05 percent.

The Support Vector Machine earned an accuracy score of 92.23 percent.

The Decision Tree earned an accuracy score of 90.4 percent.

Random Forest earned an accuracy score of 92.76 percent.

From these five classifiers, Random forest gave the highest accuracy.

Logistic Regression

The Classification Report, Confusion Matrix, Cross-Validation (Average Accuracy), and Standard Deviation are generated when Logistic Regression is applied to the dataset.

| | Precision | Recall | F1 score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.99 | 0.93 | 0.96 | 1127 |
| 1 | 0.08 | 0.37 | 0.13 | 19 |
| Accuracy | | | 0.92 | 1146 |
| Macro Avg | 0.53 | 0.65 | 0.54 | 1146 |
| Weighted Avg | 0.97 | 0.92 | 0.94 | 1146 |

Table 4.1 Classification Report of Logistic Regression

| | T | F |
|---|------|----|
| T | 1045 | 82 |
| F | 12 | 7 |

Table 4.2 Confusion Matrix of Logistic Regression

Naive Bayes

We obtained the following test accuracy score, classification report, confusion matrix, cross-validation (average accuracy), and standard deviation after applying Naive Bayes on the dataset.

| | Precision | Recall | F1 score | Support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.94 | 0.95 | 0.95 | 1043 |
| 1 | 0.44 | 0.38 | 0.41 | 103 |
| Accuracy | | | 0.90 | 1146 |
| Macro Avg | 0.69 | 0.67 | 0.68 | 1146 |
| Weighted Avg | 0.89 | 0.90 | 0.90 | 1146 |

Table 4.3 Classification Report of Naïve Bayes

| | | |
|----------|------------|-----------|
| | T | F |
| T | 993 | 50 |
| F | 64 | 39 |

Table 4.4 Confusion Matrix of Naive Bayes

Support Vector Machine

We obtained the following test accuracy score, classification report, confusion matrix, cross-validation (average accuracy), and standard deviation after applying Support Vector Machine on the dataset.

| | Precision | Recall | F1 score | Support |
|--------------|-----------|--------|----------|---------|
| 0 | 1.00 | 0.92 | 0.96 | 1046 |
| 1 | 0.00 | 0.00 | 0.00 | 0 |
| Accuracy | | | 0.92 | 1146 |
| Macro Avg | 0.50 | 0.46 | 0.48 | 1146 |
| Weighted Avg | 1.00 | 0.92 | 0.96 | 1146 |

Table 4.5 Classification Report of SVM

| | | |
|----------|-------------|-----------|
| | T | F |
| T | 1057 | 89 |
| F | 0 | 0 |

Table 4.6 Confusion Matrix of SVM

Decision Tree

We obtained the following test accuracy score, classification report, confusion matrix, cross-validation (average accuracy), and standard deviation after applying a Decision tree on the dataset.

| | Precision | Recall | F1 score | Support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.94 | 0.95 | 0.94 | 1040 |
| 1 | 0.45 | 0.37 | 0.41 | 106 |
| Accuracy | | | 0.90 | 1146 |
| Macro Avg | 0.69 | 0.66 | 0.68 | 1146 |
| Weighted Avg | 0.89 | 0.90 | 0.89 | 1146 |

Table 4.7 Classification Report of Decision Tree

| | T | F |
|---|-----|----|
| T | 986 | 49 |
| F | 71 | 40 |

Table 4.8 Confusion Matrix of Decision Tree

Random Forest

We obtained the following Test Accuracy Score, Classification Report, Confusion Matrix, Cross-Validation (Average Accuracy), and Standard Deviation after applying Random Forrest to the dataset.

| | Precision | Recall | F1 score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 1.00 | 0.92 | 0.96 | 1146 |
| 1 | 0.00 | 0.00 | 0.00 | 0 |
| Accuracy | | | 0.92 | 1146 |
| Macro Avg | 0.50 | 0.46 | 0.48 | 1146 |
| Weighted Avg | 1.00 | 0.92 | 0.96 | 1146 |

Table 4.9 Classification Report of Random Forrest

| | T | F |
|---|------|----|
| T | 1026 | 57 |
| F | 31 | 32 |

Table 4.10 Confusion Matrix of Random Forrest

4.3 Potential Future Improvement

The findings show that an automated system may be utilized to diagnose Heart Disease Prediction in clinical settings. This might be useful in clinical diagnosis in Bangladesh. It can be said a more efficient algorithm can be found with more training data. If Bangladesh's medical sector can collect more data with proper certifications then it could create some valuable opportunities for researchers like us.

CHAPTER 5

CONCLUSION AND FUTURE IMPLICATION

5.1 Study Synopsis

This section summarizes the results of our study. FIGURE 12 shows the accuracy obtained from the different models of the combined dataset heart_data_4. It can be concluded that our proposed model gives the most desirable accuracy of 92.76% where our targeted class was diabetes.

5.2 Conclusions

We discovered that we utilized 5 different machine learning algorithms, the most accurate of which was our suggested primary model heart_data_4. Although the other algorithms and other datasets produced results that were quite similar to and accurate in comparison to our suggested model heart_data_4 as for the combination which makes that dataset somewhat unique from the other three datasets.

5.3 Evaluation

Considering the study, finding additional clinical data that is correctly arranged will provide us with more precision.

5.4 Opportunities for Future Research

Comparable strategies can be used for a variety of different ailments, and additional data on other clinical health problems should be gathered to conduct similar investigations. More research in the field is needed to achieve clinical accuracy and dependability.

With authorization and supervision from authorized authorities, a web-based or Android-based application can be built for public use of the platform.

References

- [1] T. Dent, Predicting the risk of coronary heart disease, PHG foundation publisher, 2010.
- [2] "Global status report on non communicable diseases," World Health Organization, 2014.
- [3] "<https://medlineplus.gov/encyclopedia.html>," [Online]. Available: <https://www.nlm.nih.gov/>.
- [4] "Fact sheets-Cardiovascular diseases (CVDs)," 11 June 2021. [Online]. Available: www.who.int.
- [5] K. A. Pronab Ghosh, "Comparative Study on Different Machine Learning Algorithms for Achieving Accurate Prediction for Heart Diseases," Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.
- [6] M. M. MD.MASUD RANA, "HEART DISEASE PREDICTION," Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.
- [7] "Prediction System for Heart Disease Using Naive Bayes," International Journal of Advanced Computer and Mathematical Sciences, vol. 3, pp. 290-294, 2012. Shadab Adam Pattekari and Asma Parveen, "Prediction System for Heart Disease Using Naive Bayes," International Journal of Advanced Computer and Mathematical Sciences, vol. 3, pp. 290-294, 2012.

- [8] Sarath Babu, "Data mining approach for heart disease detection." IEEE, 2017. International Conference on Electronics, Communication, and Aerospace Technology (ICECA).Vol. 1.
- [9] "Disease forecasting system utilizing data mining approaches," in Intelligent Computing Applications (ICICA), 2014 International Conference on. IEEE, 2014. Banu, MA Bishara, and B. Gomathy.
- [10] "Diagnosis of heart disease patients using fuzzy classification approach," by V. Krishnaiah. 2014 International Conference on Computer and Communications Technologies (ICCCT), IEEE, 2014.
- [11] By Jaymin Patel, Prof Tejal Upadhyay, and Dr. Samir Patel, heart disease may be predicted using machine learning and data mining techniques.
- [12] Asst. Professor, Computer Science & Engineering, Orissa Engineering College, Bhubaneswar, Odisha - India. Predicting and Diagnosing Heart Disease Using Machine Learning Algorithms.
- [13] "Prediction of Heart Disease Using Classification Based Data Mining Techniques," by Sujata Joshi and Mydhili K.Nair, Springer India, volume 2, 2015.
- [14] Wikipedia, https://en.wikipedia.org/wiki/Machine_learning. [Online]. Last Accessed: 1 April 2019.
- [15] Coursera, <https://www.coursera.org/learn/machine-learning>. [Online]. Last Accessed: 129 April, 2019.
- [16] "Introduction to machine learning," [Online]. Available: <http://www.zarantech.com/blog/an-introduction-to-machine-learning-why-it-matters.%5BOnline>.
- [17] "Cardiovascular Study Dataset," [Online]. Available: <https://www.kaggle.com/christofel04/cardiovascular-study-dataset-predict-heart-disea?select=train.csv>.

- [19] "Heart Disease Dataset (Comprehensive)[statlog + cleveland + hungary dataset],"
[Online]. Available: <https://www.kaggle.com/sid321axn/heart-statlog-cleveland-Hungary-final>.
- [20] "Heart Failure Prediction[12 clinical features for predicting death events],"
[Online]. Available: <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>.