Machine Learning Based Heart Disease Prediction

BY

SABBIR AHMED
ID: 172-15-9777

AND

FARHINA ALAM
ID: 172-15-9705

AND

NESHAT TASNIM MOITRI
ID: 172-15-9972

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

Supervised By

MD. TAREK HABIB
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2021

APPROVAL

This Project/internship titled **“Your Title”**, submitted by Name, ID No: Student ID 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 date.

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

Abdus Sattar

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Md. Jueal Mia

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

External Examiner

Farid

Dr. Dewan Md. Farid

Associate Professor

Department of Computer Science and Engineering

United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Tarek Habib, Senior Lecturer, 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:

THabib

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Co- Supervised by:

Department of CSE
Daffodil International University

Submitted by:

Sabbir Ahmed

Md. Sabbir Ahmed
ID: -172-15-9777
Department of CSE
Daffodil International University

Farhina Alam

Farhina Alam
ID: -172-15-9705
Department of CSE
Daffodil International University

Neshat Tasnim Moitri

Neshat Tasnim Moitri
ID: -172-15-9972
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 **Md. Tarek Habib**, **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 draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Touhid Bhuiyan, Professor 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

The heart is the most valuable muscular organ in living organisms. In our country, Now people are more sensible about their heart. A healthy heart is an interior to aggregate good health. The heart needs to care for living life wonderful. In the present situation, Due to the unhealthy environment and unhealthy lifestyle habits, people are faced with many heart disease problems. The enumeration of heart consequential disease needed more exactness, fullness, and right information. There are numerous cases where people are dying day by day because of heart problems. For reducing the heart-disease problem and knowing present heart status, we are proposed a solution to predict the heart condition in this paper. Our proposed solution would support people to keep their heart good and also plays the significant role in the medical field. We are using Machine Learning which is the branch of Artificial Intelligence. . In our work, we were collected important heart-related data and we computed the accuracy of MLA, and find out the result of the predicted heart disease problem. Here we applied various classification techniques and compared the accuracy. In Deep learning algorithms, Artificial Neural Network (ANN) archives overall 74.09% accuracy. In Machine learning algorithms, The SV machine archives 73.00% accuracy.

TABLE OF CONTENTS

CONTENTS PAGE

CHAPTER

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Heart is the central administration organ that serves the entire body. It is the most essential and valuable part of living life. It is a muscular organ that is kind of a tiny size of the closed fist. To maintain a healthy heart, people have to be aware of heart disease.

The Medical science field can be modernized by using the latest technologies. In the age of modernity, Technologies will increase the life expectancy of the overall population. A range of conditions that affected people is described by heart disease. The health system can be modernized with the new innovations, increasing the total population's life expectancy. Leading illnesses such as heart disease and cancer are responsible for a large number of deaths worldwide. Cancer is one of the most common causes of mortality in the globe. Every year, the death rate from cardiovascular disease rises at an unprecedented rate. According to a 2016 World Health Organization survey, cardiovascular disease was responsible for 35% of all deaths worldwide, with heart attacks and stroke accounting for 88% of all deaths.

In both developing and developed countries, the rising popularity of alcohol and tobacco directly leads to the risk of heart disease. Obesity rates are rising in developing countries such as the United States, England, Canada, and New Zealand, increasing the risk of heart disease. The phrase "cardiovascular disease" refers to situations in which blood vessels are narrowed or obstructed, resulting in a heart attack, chest discomfort (angina), or stroke. Other cardiac illnesses, such as those affecting the heart's muscle, valves, or rhythm, are frequently categorized as heart disease. CVDs claim the lives of 19.6 million people per year, accounting for about 32% of all deaths worldwide.

Nowadays, the healthcare industry generates a vast volume of data about patients, illness diagnoses, and so on. Given the global effects of cardiovascular diseases, a machine learning model for early detection becomes highly valuable. To deal with this growing massive problem, constant efforts using various technological advances are made.

In recent years, various bioengineering techniques have been developed to deal with the ever-increasing health problems. Continued research in this field is helping to increase the reduction rate.

In today's world, both the younger and older generations suffer from heart attacks. Heart attack cases are on the rise as a result of their eating habits and lifestyle. Physicians were unable to predict the patient's disease status in the early stages of the disease until they reached the final stage.

So, if doctors can use machine learning tools in these situations, there is a chance to predict heart disease status sooner, which might save more lives with proper care. The hospital administrator will store patient information in a database server, which contains a large amount of data.

Then, using machine learning algorithms, physicians can extract the data and use it to predict heart disease. A heart disease data collection is used in this scheme.

The main goal of this method is to estimate the patients' chances of developing heart disease in terms of percentage. Data mining classification methods are used to accomplish this. The classification method is used to divide the entire dataset into two categories: yes and no. Machine learning classification algorithms, such as Decision tree classification and Naive Bayes Classification models, are used to apply classification techniques to the dataset. These models are used to improve the classification technique's accuracy. This model is capable of both classification and prediction. The Python Programming Language is used to build these models.

1.2 Motivation

Heart disease is one of the world’s internecine problems because it is not easily seen externally. People will be attacked by heart failure at any time if their heart health has not been monitored before. Even in a very short time people reach the time limit of their death cycle. So detecting the heart disease problem is very important to everyone.

Even heart monitoring is an essential part of keeping the heart healthy.

In our research-based project, we propose a framework that allows users to identify heart disease automatically.

Our research-based project also has several clear motivations:

- To reduce the limitations of work and processing time needed for heart disease detection.
- Assists in cardio evaluation for people in remote areas.
- Identifies the stage of the heart condition
- Since our system will have a lot of previous results, it will assist doctors and also promote the education of undergraduate and postgraduate young physicians.
- It would be useful in the event of a heart attack in which there is no doctor available to treat the primary care.

1.3 Rationale of Study

There has been a lot of research done in the field of automated disease detection. In the world, heart disease is the leading cause of death for men, women, and citizens of the most racial and ethnic groups. In every 36 seconds, one person dies from cardiovascular disease. Now we can commonly see cardiovascular disease is the leading cause of death in middle-aged people. The type of disease and its manifestations are also evolving over time due to a variety of factors. Heart disease identification is also difficult due to the difficulty of obtaining data since changing of symptoms.

1.4 Outcome

Our research aims to predict heart disease automatically and stand against heart disease. It will help doctor & undergraduate and postgraduate young physicians. It will examine the heart disease issues and demonstrate the disease's accuracy.

1.5 Report layout

We have discussed about the introduction to the Heart Disease Identification, inspiration, the study's rationale, and the thesis's conclusion in this chapter. The report layout is then followed.

The history of our research subject will be discussed in Chapter 2.
The analysis methodologies used in our study will be discussed in Chapter 3.
We'll talk about classification and the model analysis in Chapter 4.

The obtained experimental results and discussion will be discussed in Chapter 5.
The conclusion and future work will be discussed in Chapter 6.

CHAPTER 2
BACKGROUND

2.1 Introduction

Heart disease has one of the highest mortality rate of both Bangladesh and abroad. According to WHO (World Health Organization) 2020 report Cardio-vascular disease is one of the number one cause of death all over the world. It is taking away around 17.9 million lives per year all over the world.

In “Bangladesh” the Heart Disease Deaths reached 118,287 or 15.23% of total yearly deaths (WHO last data published in 2018). So it is high time to recheck the death rate and which categories people are being effected by heart disease. We can use data mining knowledge and machine learning technologies to discover knowledge from any datasets. The discovered data can be used to solve any local people problem and any hospital or health care center can use this knowledge to improve their patient condition and they can improve their service towards an emergency patient of heart disease. This knowledge will also help the diabetic patient as we have data for sugar level.

The disease "heart disease," can be called by "cardiovascular disease" also because it refers to a group of illnesses that affect the cardiovascular system. Not only heart attacks, but a variety of conditions that affect the heart. Heart disorders include coronary heart disease, cardiomyopathy, and cardiovascular disease. The word "cardiovascular disease" refers to a huge amount of varieties of conditions. Conditions attacking the heart and the body because the heart is the only organ which is supplies blood to the whole body by pumping it, and helps us to breath. People often need to have a diagnosis, and this process need to be happen perfectly, any mistake can cause death of the patient. If the doctor is experienced enough then he is the perfect one to do that because this diagnosis is based on his experience. This leads to unfavorable outcomes and high medical costs for care given to patients.

A group of patients, an electronic medical diagnostic device is the neediest machine in this operation because it has made less the death rate.

The aim of this project is to help people by showing our accuracy rate by applying classification methods and predict heart disease risk. By reading our thesis study and knowing the machine learning techniques people can reduce the number of patients of heart disease by knowing the previous dataset analysis. People can know which lifestyle is good for them and how they can make themselves safe from heart disease.

2.2 Related Works

We can predict the risk of heart disease using "Machine Learning" techniques such as k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), Linear Regression (LR), Logistic Regression (LR), and Linear Discriminant Analysis (LDA). Various experiments have been conducted, and more are being conducted, for improving the accuracy rate of heart disease prediction. There are so many works in our topic they have used Machine Learning algorithms and Deep Learning algorithms like us.

The heart is the most valuable part of our body, it keeps us alive, it is responsible for pumping heart and blood circulation, heart is as important to the body as oxygen, so it is the most needed part for the human body. One of the main reasons why researchers are working on this is to ensure its safety. As a result, there are a lot of people working on it. There has many field like Artificial Intelligence, Data Mining, and Machine Learning that are contributing on this predicting disease field.

Table 2.1: Summary of Related Research Work

Work Done	**Problem Domain**	**Algorithm**	**Accuracy**
Singh et al.	Prediction	k-nearest neighbor	87%
Motarwar et al.	Prediction	Random Forest	95.05%
Krishnan et al.	Prediction	Decision tree classifier	91%

Haq et al.	prediction	Logistic regression	89%
Atallah et al.	Prediction	Hard Voting Ensemble Method	90%
Yaswanth et al.	Prognosis	Neural Networks	92.30%
Agrahara et al.	Prediction	Decision Tree	98.29%
Srinivas et al.	Prediction	Hybrid Linear Regression	89.13%

Srivastava et al.	Prediction	k-Nearest Neighbours	87%
Jindal at al.	Prediction	k-Nearest Neighbor	87.5%
Rajamhona et al.	Prediction	Multi-layer Perception with back propagation learning algorithm	94%
Javid et al.	Prediction	Hard Voting Ensemble Model	85.71%
Habib et al.	Detection	SVM	95.2%

In the paper of Singh & Kumar, they discovered that the precision of the K-NN is much more effective than other algorithms by using the machine learning method for testing and teaching. The use of algorithms Accuracy can be determined using the confusion matrix of each algorithm, where the number of counts of TP, TN, FP, and TP, TN, FP, and TP, TN, FP, and TP, TN, FP, and TP, TN, FP, and TP, TN. The value of FN has been measured using the equation of precision, and it has been determined that KNN is the highest among them, with an accuracy of 87 percent. They have got the other accuracy as follows SV machine: 83%; DT: 79%; LR: 78%; and k-NN: 87%.

In the paper of Motarwar, Duraphe et al. On their dataset, data visualization was used to simulate similarity or dependency between any of the listed variables. The one of the good attribute among their dataset were chosen using feature selection. For executing classification algorithms, this method provides the best data consistency. To improve the simple accuracy of each algorithm, additional enhancement techniques are used. With 80 percent of the results, 242 examples, the dataset was educated. 61 instances are expected from the remaining 20% results. The improved percentage for each technique. They have got the accuracy as follows Gaussian NB: 93.44%; Support Vector ; Machine: 90.16; Random Forest: 95.08%; Hoeffding Tree: 81.24%; Logistic Model Tree : 80.69%.

In the paper of Mr. Santhana Krishnan.J & Dr.Geetha.S two supervised data mining algorithms were added to the dataset to predict the likelihood of a patient developing heart disease, and the results were evaluated using the Nave Bayes Classifier and Decision tree classification models. These two algorithms are tested on the same dataset in order to determine which is the most accurate. The Decision tree model correctly predicted heart disease patients 91% of the time, while the Nave Bayes classifier correctly predicted heart disease patients 87% of the time.

In the research of Haq et al,A hybrid intelligent machine-learning-based predictive method was proposed in this research study for the diagnosis of disease of the heart. Three feature selection algorithms were used, including logistic regression, K-NN, ANN, SVM, NB, DT, and random forest. The essential features were chosen using relief, mRMR, and LASSO. The heart attack was properly classified by 0e ANN with Relief. 0e classier logistic regression 0e classier logistic regression 0e classier logistic regression 0e class Relief FS algorithm MCC was 89 percent on selected functions.

In the thesis of Rahma Atallah, Amjed Al-Mousa, they have got around 90% accuracy rate from their proposed model, which was higher among the other algorithms they have used. The first test was performed with the classifier's default parameters and they have got an accuracy of 80%. They have found the tailored parameters based on cross-validation after running a GridsearchCV, and the accuracy improved to 88 percent. Furthermore, the Logistic Regression classifier was the most recent model developed. The accuracy was 87 percent, and the accuracy remained the same after running GridsearchCV because the default parameters were the same as the configured parameters.

In the thesis of Raparthi Yaswanth, Dr.Y.Md.Riyazuddin they have used the most recent model produced was the Logistic Regression classifier. They have got the best accuracy of 87 percent, and since the default parameters were the same as the configured parameters, the accuracy remained the same after running GridsearchCV. This paper compares and contrasts different machine learning methods in order to determine which one has the better outcomes for effective heart disease prediction. Ensemble approaches and a mixture of artificial Neural Networks are used in the architectures and techniques, resulting in more precise performance.

In the thesis of Javid et al, they presented ML and DL ensemble models that combine many ML and DL models to provide optimum accuracy and performance. A robust mechanism for predicting the likelihood of developing heart disease. This Ensemble solution had an accuracy of 85.71 percent, which was higher than they predicted the accuracy of each individual model.

Md. Tarek Habib conducted an analysis using a machine learning classification technique to recognize papaya disease. They used color pictures of rotten papayas. Many of the images were resized to 300 x 300 pixels. Histogram and bicubic interpolation. For image processing, equalization was used. In their model, they used 129 photographs of imperfect and defect-free objects. They split their dataset into two sections, with two-thirds used for training and one-third used for research. They used a variety of machine learning classification methods. SVMs, C4.5, Nave Bayes, and Logistic regression are some of the methods used. BPN, CPN, and RIPPER are all acronyms for regression, KNN, Random Forest, BPN, CPN, and RIPPER. In their work, they had dealt with five different diseases. SVM has outperformed all of these methods. Of all classifiers, SVM has a 95.2 percent accuracy rate.

2.3 Research Summary

In Bangladesh there are 70% people are having many type of heart related problems. And they do not even conscious about their health. For this unconscious behavior towards their health they are having many type of physical problem. Peoples are always used to eat unhealthy foods, as a result they are having blood sugar, high level of blood sugar, high level of cholesterol etc. and this physical problems are taking them towards heart disease.

If people always monitor their food habit and health checkup, using our research, prediction and analysis they will be more conscious and benefited because this prediction and analysis will help them to know how risky their food habits are and how not to become a heart disease patients. We can’t help people by physical activity or medicine but we can provide them a support of awareness to change their lifestyle by showing the predictive risk of heart disease.

There are a lot of research papers on this topic of prediction of heart disease but those research does not show any accurate rate of heart disease but in our research we have taken data about whether they smoke too much, drinks a lot, diabetes level, and so on and the data are categories by male and female. After analysis the various categories data we are predicting the risk of heart disease with higher accuracy.

2.4 Scope of the Problem

Heart disease is a very common disease and cause of death not only in Bangladesh but also many other countries like us. Like in India heart disease has the high risk of mortality. So heart disease is the most common concerned health topic. Scope of our topic never going to be less. That is why we have chosen this topic so that we can help people to be aware of them by predicting the risk of heart disease. The research work of ours is to set a model by analyzing data and applying Machine Learning algorithms and Deep Learning algorithms. Our proposed model can predict the risk of heart disease. So people who are unable to go through a checkup can read our thesis and can be aware of their heart status. Nowadays because of COVID-19 getting a medical checkup has become so risky. In the villages the village people do not have that much facilities for their checkup and they also do not know about health care. So they will be able to gather much more knowledge about their heart health care and they will be aware when to take necessary steps from our data analysis and prediction. So our thesis will be helpful for all categories of people from village to town or city. We also know that heart patients are at high risk for affecting of COVID-19, so nowadays everyone is willing to know their risky sides so if anyone wants to know about their heart status and if they never went through any checkup they will be benefited from our thesis paper and prediction. So in the end we can say taking in mind the present situation and future situation also our thesis will be the most valuable and useful for everyone.

2.5 Challenges

Every light has its darkest side, if there is something great there are challenges too. We are also not different from that, we have faced difficulty and challenges too. As we know we are in a pandemic situation for COVID-19, so hospitals are the most risky placed for getting affected by it. For heart disease data collection our main place was the hospitals. There was the main challenge, all over the country there was lockdown and people was so scared to talk with unknown people and share their data. So, our main challenge were data collection because we were bound to travel anywhere. That is why we left no chance to accept online data collection. So, we started to find data online but there were challenges too. There was already a lot of work on our topic so our data should be accurate and unique. So, after so many challenges we have collected our wished dataset which is about seventy thousand data. After that we have analyzed our data. Then there comes one more challenge of selecting the algorithms of Deep Learning and Machine Learning. After so many ups and downs we have decided which algorithms are giving more preferable accuracy by applying them. After so many ups and downs we have finished our work successfully.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction:

The methodology is a means of collecting evidence or knowledge that incorporates many principles and theories. Methodology encompasses the data or knowledge that contains different hypotheses and principles. In this research, we are going to predict the risk of heart disease prediction by using some machine learning algorithm such as Random Forest, K-Nearest Neighbor, Decision Tree, Naïve Bayes, Logistic Regression, Support Vector, Nu-Support Vector, Linear Support Vector and also use some deep learning algorithm such as Neural Network, Artificial Neural Network and Recurrent Neural Network since it outperforms other algorithms in terms of accuracy. The Prediction Model is built on knowledge about people's health. We had to gather dataset from Kaggle to implement these machine learning algorithms. We also knew from reading research papers that these machine learning algorithms have greater precision. Algorithms used in the model to classify data. To choose the best algorithm for the model, we estimated and computed the accuracy, sensitivity, specificity, precision and roc-curve of each algorithm. We discovered that Re-current Neural Network has the highest precision and as ideally suited to our proposed model.

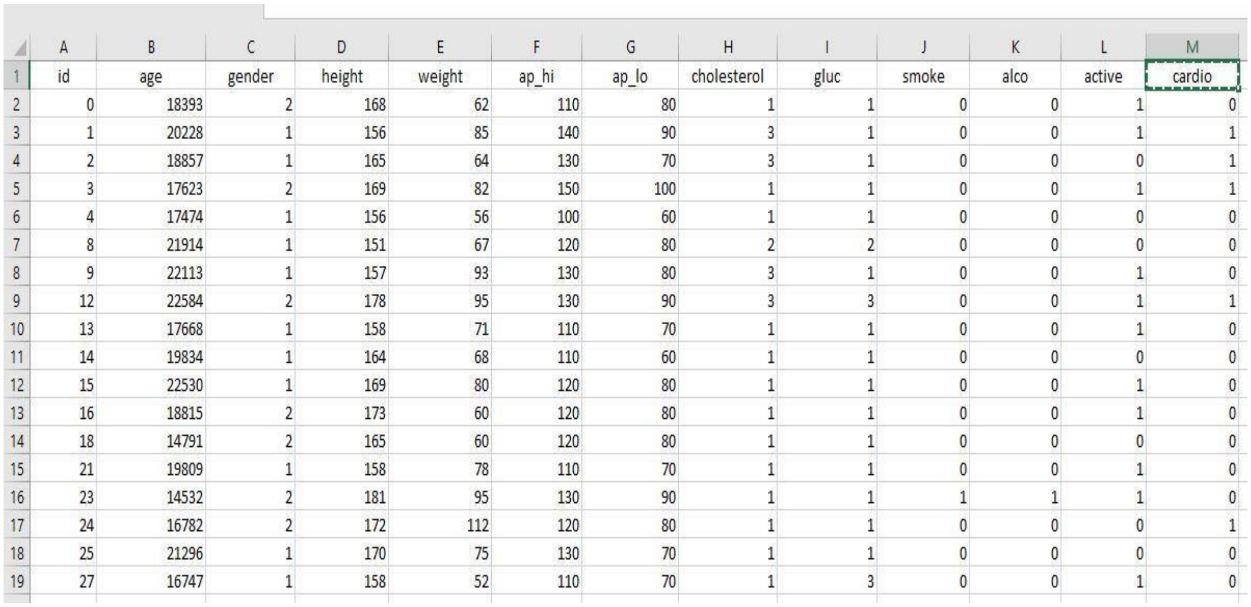
3.2 Data Source Procedure

An Organized Dataset of people was chosen with their history of heart attacks and other medical disorders in mind. A range of disorders that affect the heart are referred to as heart disease. According to the World Health Organization, cardiovascular diseases are the main cause of death in people in their middle years.We use a data base that contains the medical histories of seventeen thousand individual patients of various ages. This dataset provides us with much-needed details, such as the patient's age, gender, height, weight, Systolic blood pressure, Diastolic blood pressure, cholesterol, and so on, which aids us in determining whether the patient has a heart attack or not. This dataset includes 12 medical traits from seventy thousand patients that help us determine whether a patient is at risk of developing a heart attack or not, as well as identify patients that are at risk and those that are not. This Heart Disease dataset was obtained from the Kaggle Website. This dataset extracts the sequence that contributes to the diagnosis of patients at risk of developing heart disease. This dataset has seventy thousand rows and 13 columns, with each row representing a single record.

Table 3.1: contains a list of all attributes:

S.N0	Characteristic	Statement	Variety
1	age	age of the patients (in days)	Numerical
2	gender	gender of the patient (divided into two groups: 1 for female and 2 for male)	Titular
3	hgh	Hgh of the patients (in cm)	Numerical
4	wgh	Wgh of the patients	Numerical

		(in kg)	
5	ap_hi	ap_hi sort SBP	Numerical
6	ap_lo	ap_lo sort DBP	Numerical
7	cl	cl of patients (divided into 3 groups: 1 for normal, 2 for above normal and 3 for well above normal)	Titular
8	gluc	Glucose of patients (divided into 3 groups: 1 for normal, 2 for above normal and 3 for well above normal)	Titular
9	smk	smk of patients (divided into two groups: 0 for smoke addicted and 1 for without smoke addicted)	Titular
10	al	alcohol of patients (divided into two groups: 0 for alcohol addicted and 1 for without alcohol addicted)	Titular
11	active	active sort Binary feature	Titular
12	cardio	cardio of patients (divided into two groups: 0 for cardio addicted and 1 for without cardio addicted)	Titular



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
2	0	18393	2	168	62	110	80	1	1	0	0	1	0
3	1	20228	1	156	85	140	90	3	1	0	0	1	1
4	2	18857	1	165	64	130	70	3	1	0	0	0	1
5	3	17623	2	169	82	150	100	1	1	0	0	1	1
6	4	17474	1	156	56	100	60	1	1	0	0	0	0
7	8	21914	1	151	67	120	80	2	2	0	0	0	0
8	9	22113	1	157	93	130	80	3	1	0	0	1	0
9	12	22584	2	178	95	130	90	3	3	0	0	1	1
10	13	17668	1	158	71	110	70	1	1	0	0	1	0
11	14	19834	1	164	68	110	60	1	1	0	0	0	0
12	15	22530	1	169	80	120	80	1	1	0	0	1	0
13	16	18815	2	173	60	120	80	1	1	0	0	1	0
14	18	14791	2	165	60	120	80	1	1	0	0	0	0
15	21	19809	1	158	78	110	70	1	1	0	0	1	0
16	23	14532	2	181	95	130	90	1	1	1	1	1	0
17	24	16782	2	172	112	120	80	1	1	0	0	0	1
18	25	21296	1	170	75	130	70	1	1	0	0	0	0
19	27	16747	1	158	52	110	70	1	3	0	0	1	0

3.3 Research Subject and Instrumentation

MLA, data mining, and deep learning are currently very important for prediction and detection. We will run our gathered data through different algorithms to see which ones would work best with our model. We conduct a variety of ML algorithms and deep learning algorithms. They are Random Forest, K-NN, DT, NB, LR, SV, Nu-Support Vector, Linear Support Vector, NN, A-NN, and R-NN. We used “Python” as a PL and “Weka”, ”Google Colab” as DM tools and “Microsoft Excel” as our dataset in our research work.

3.4.1 Proposal Methodology

Our proposed methodology system is shown below in the Figure 3.1

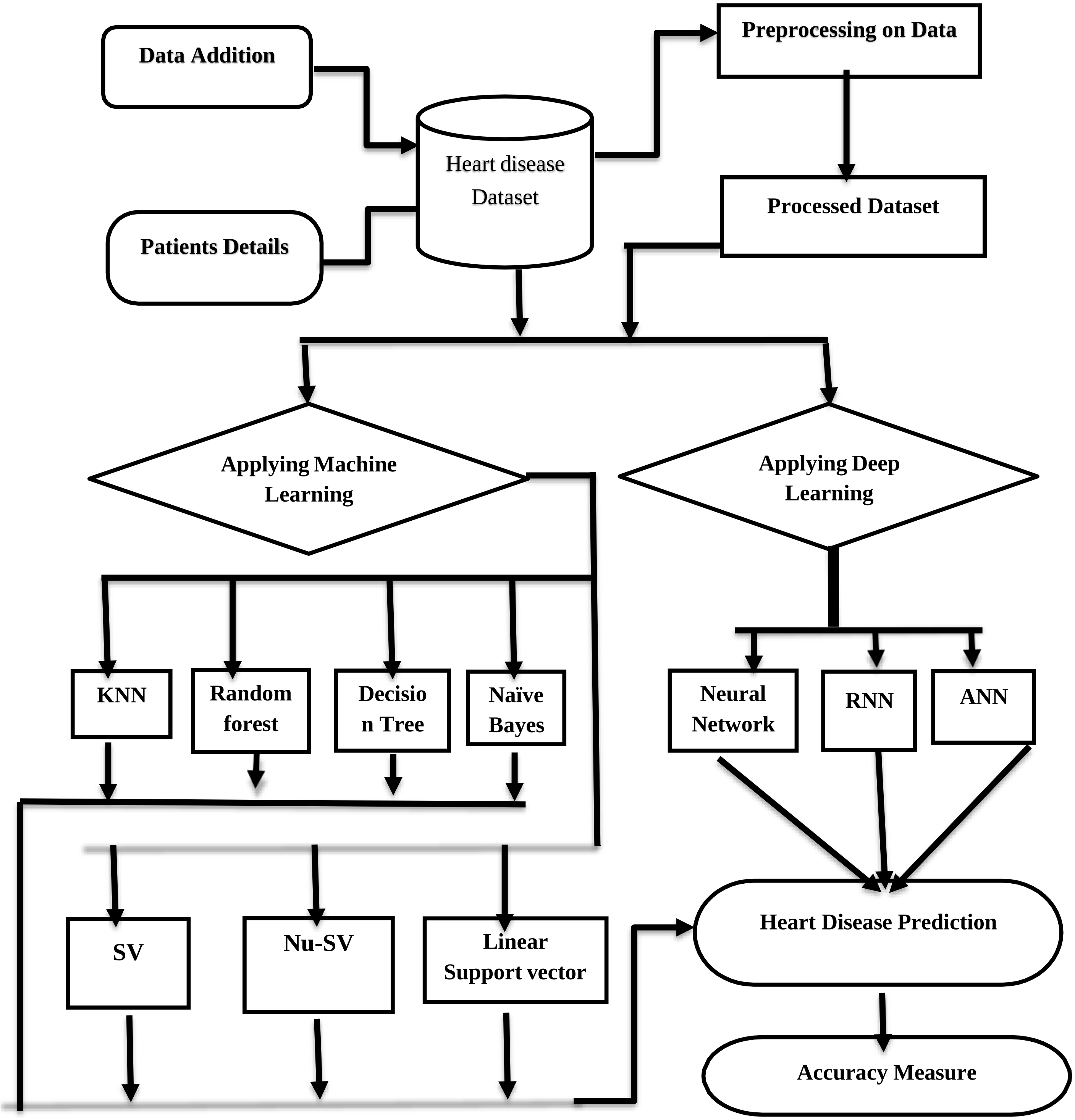


Figure 3.1: Our Proposal Methodology System

3.4.2 Data Preprocessing

We get some missing data, categorical data, numerical data, and nominal data after collecting the data. Then, using data analysis, we can type this data so that it is perfect for algorithms. After gathering data, data-analysis is the ability to convert it into a compatible format. Input or data is processed in a particular format to facilitate output.

Our data preprocessing system is display below in shape 3.2

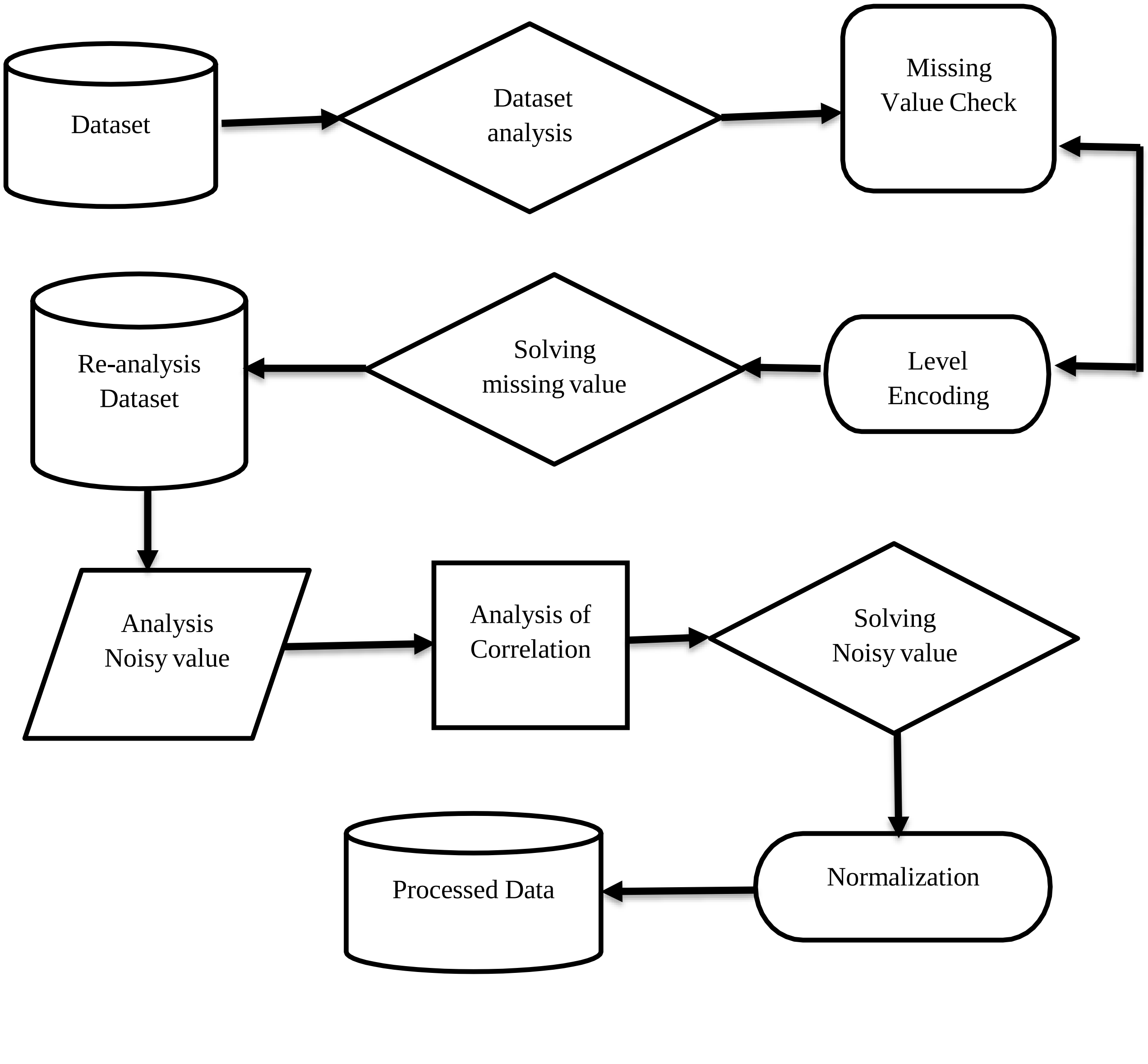


Figure 3.2 Steps of our data preprocessing system

Preprocessing data is an important phase in planning data for use in model construction. Data cleaning, data transformation, and function selection are all critical stages of data preprocessing. Data cleaning and transformation are techniques for removing outliers, missing values, and noisy values from data such that it can be conveniently used to construct a model. First, we began the laborious task of data cleaning. If the data set contains a null value, encrypt the level that transforms the text data to numerical data. We used imputer and median to solve the missing value dilemma. We can see that there are some noisy data in the numerical data if there is a noisy value in the data set using a box. We examine the correlation matrix as part of the data integration process. This matrix displays the ratio of each data connected to each data, with a positive value showing that the data is associated in a positive way, a negative value suggesting a negative relationship between the data, and a zero indicating that the data does not bind to itself. We use outlier quantile detection to delete noisy values before dropping our outcome function, which was the addicted column. In feature engineering, we create a unique histogram for each feature, which aids us in data reduction and visualization. We completed the data transformation by using normalization. As a result, we now have the stored data collection in our hands. The “Google Colab” , the “Anaconda navigator” and the “weka ”were used for the entire data processing operation.

3.5 Statistical Analysis

After gathering data, we interpret and process it in a number of stages. First and foremost, we must preprocess these results. The total number of participants in this survey is seventy thousand, with males accounting for 35.0 percent and females accounting for 65.1percent. The dataset is then transformed to retrieve incomplete and abnormal data. We gathered information from citizens of various Age, genders, smk, alc, cl, and so on.

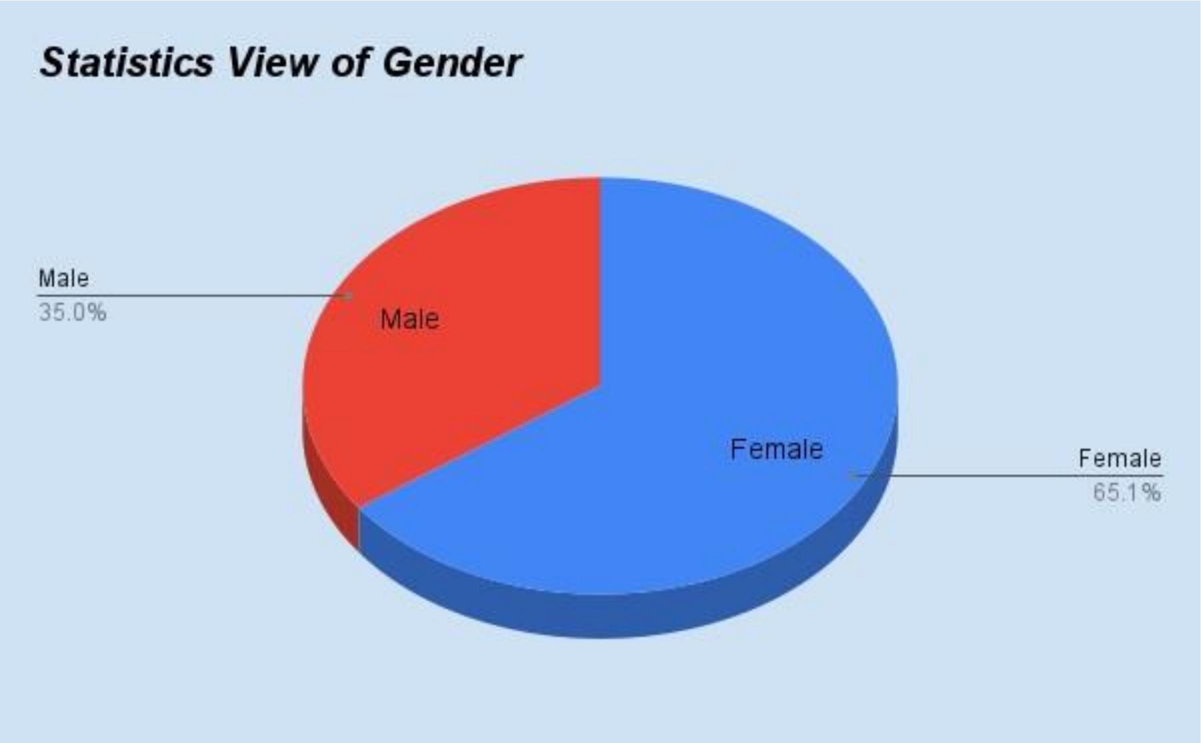


Figure 3.3: Statistics View of Gender

Figure 3.4 depicts data from citizens of various ages. This diagram illustrates how many individuals of various ages we have knowledge about. The majority of the information we gathered concerned the elderly people.

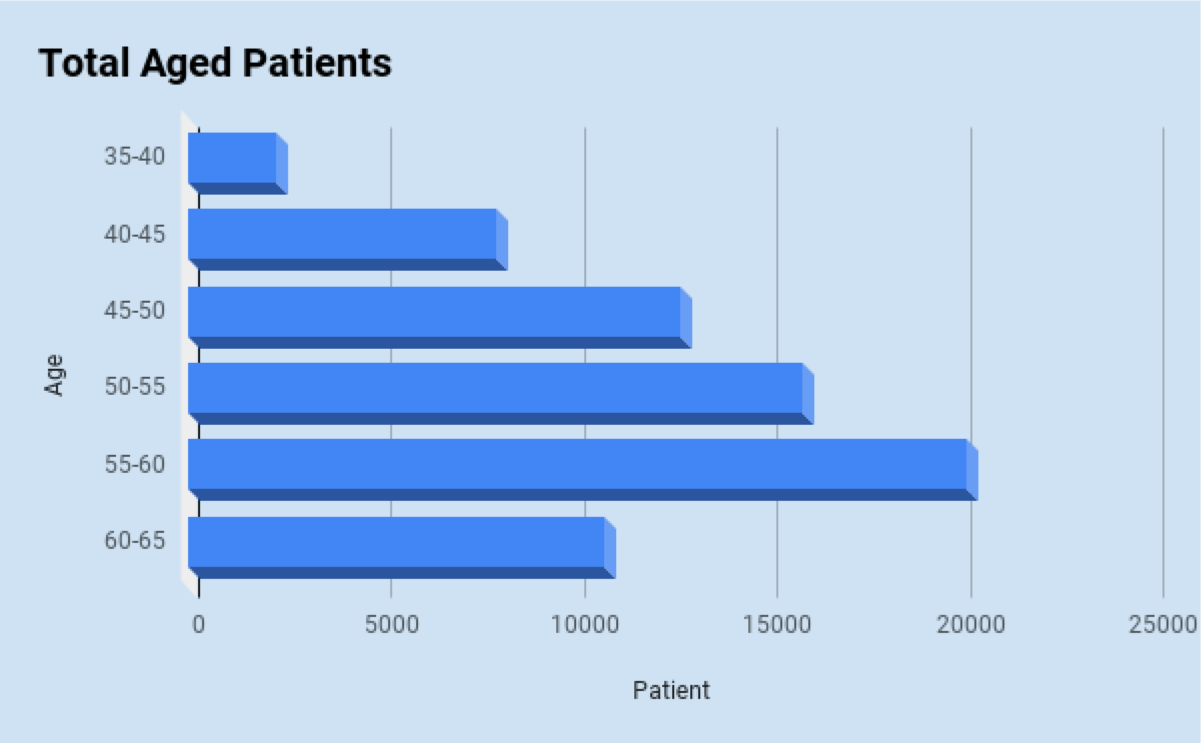


Figure 3.4: Statistics View of Aged Patients

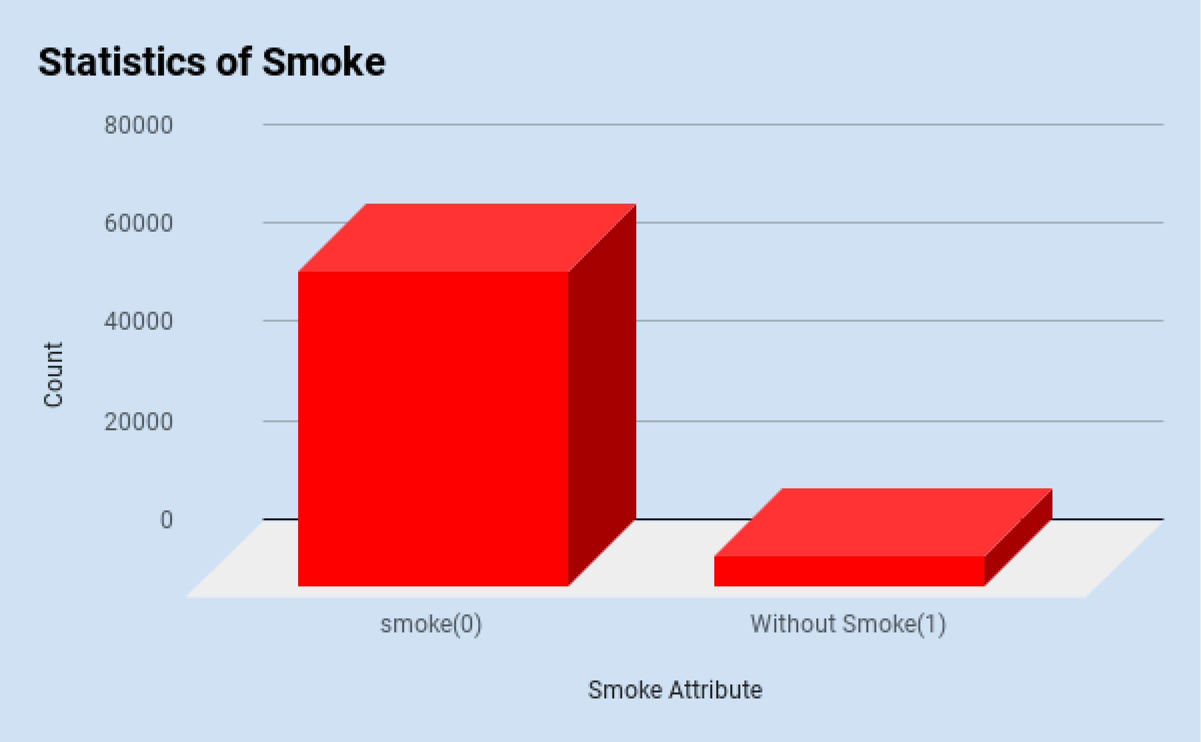


Figure 3.5: Statistics View of Smoke

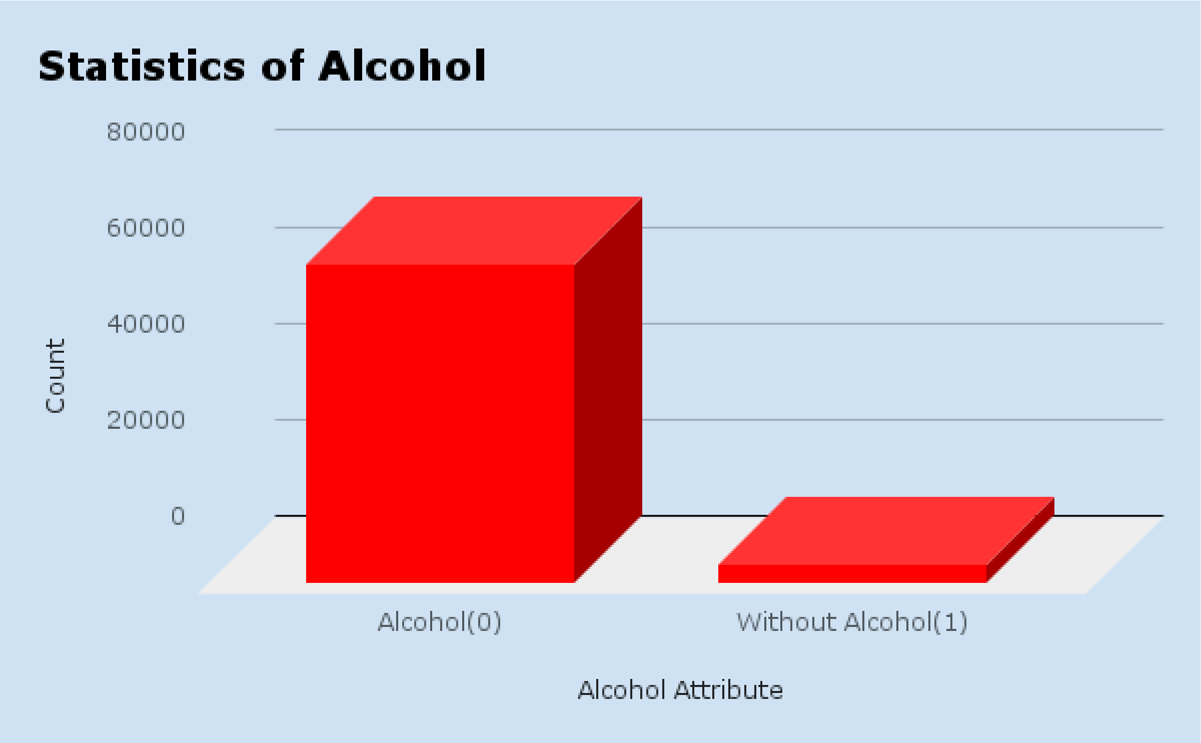


Figure 3.6: Statistics View of Alcohol

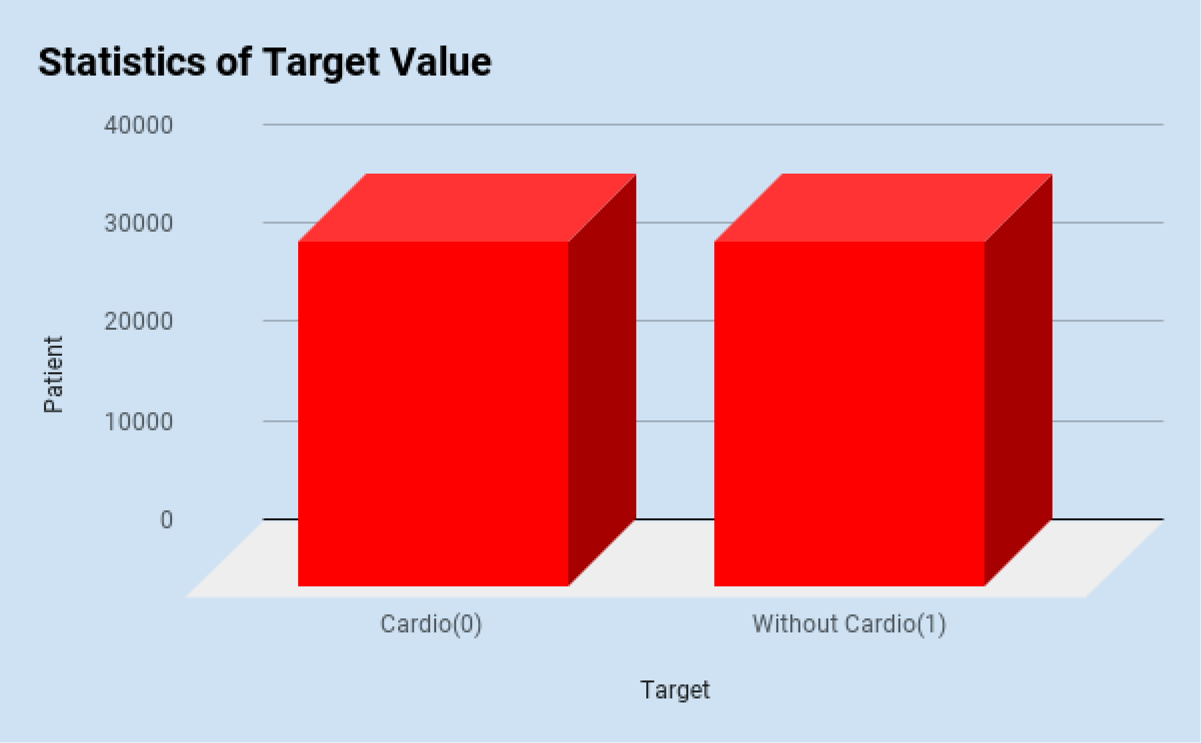


Figure 3.7: Statistics View of Target Class

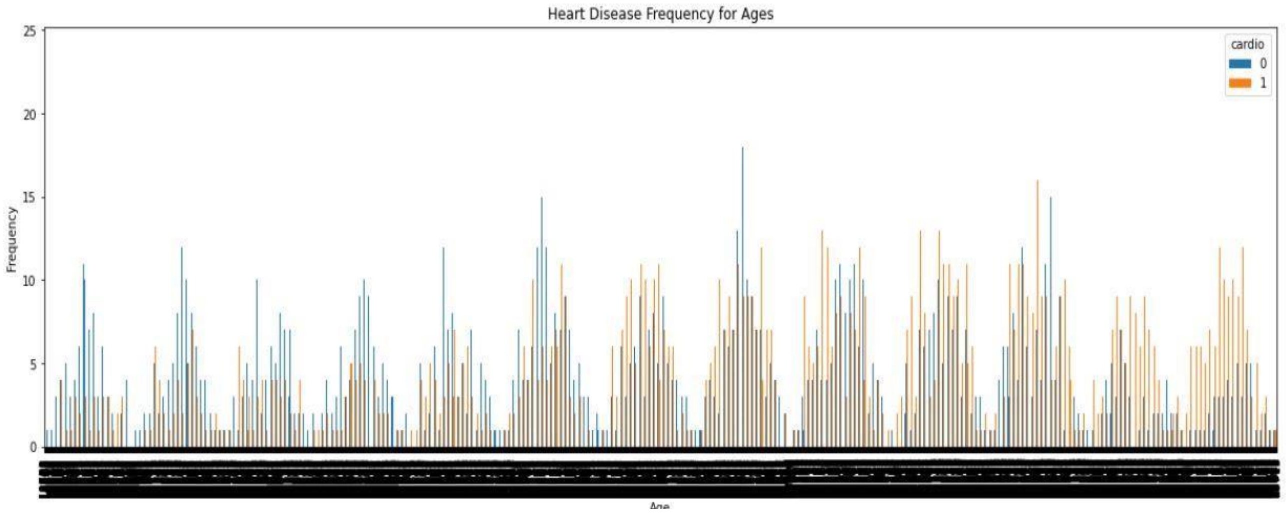


Figure 3.8: Heart Disease Frequency for Ages

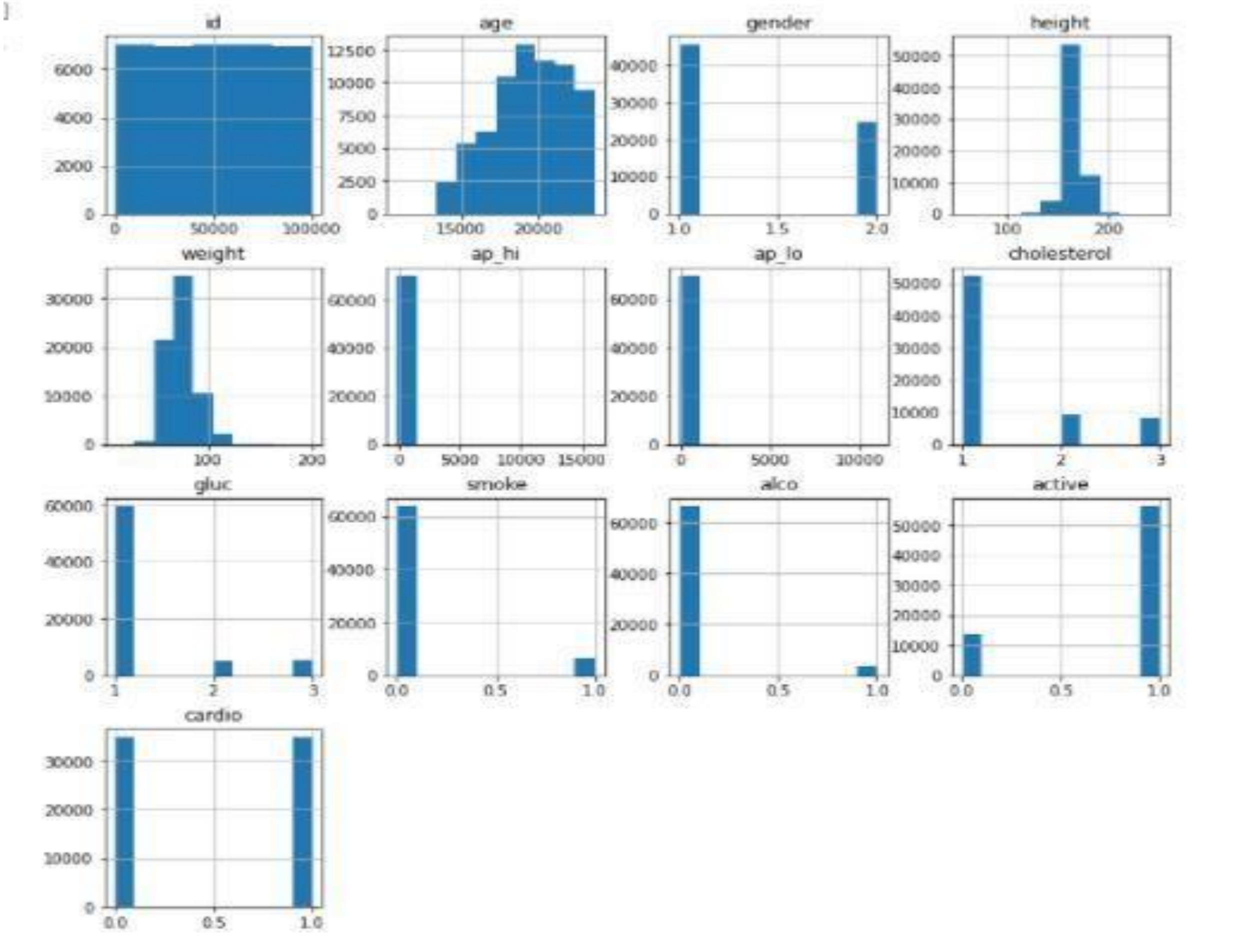


Figure 3.9: Statistical View on Database Attributes

3.6 Implementation Requirements

- Language: Python
- Open source web application: Google Colab Notebook
- Library: Pandas
- Library: num-py
- Library: sk-learn
- Library: Mal-plotlib
- Library: Scikit learn
- Library: ke-ras
- Library: TensorFlow
- Microsoft Excel
- Microsoft Word
- Basic knowledge of computing
- Data mining Tool: Weka

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

We covered the dataset, dataset methodology processes, data processing process and Requirements in the previous segment. The processed data is used in certain algorithms, the effects of which are described in this section. The algorithms are Random Forest, K-Nearest Neighbor, Decision Tree, Naïve Bayes, Logistic Regression, Support Vector, Nu-Support Vector, Linear Support Vector, Neural Network, Artificial Neural Network and Recurrent Neural Network classifier are all used and the results are evaluated to determine which algorithms are offer the best accuracy. We collected seventy thousand data of doth cardio addicted and non-cardio addicted patients along with 80 percent is used as training data set and 20 percent is used as testing data set. The name of our dataset csv file is "final cardio.csv". To achieve high precision, we use eight machine learning algorithms and three deep learning algorithms, with Recurrent Neural Network (RNN) being the best performer.

4.2 Experimental Results & Analysis

We contrasted eight machine-learning algorithms and three deep learning algorithms by measuring their accuracy, confusion matrix, precision, recall, F1 score, and support.

4.2.1 Experimental Evaluation

The primary goal of our proposed scheme is to estimate the likelihood of heart failure. Nowadays, various machine learning methods and data mining methods make predicting prediction levels easier. We must first collect information in order to implement these data mining techniques, and then we must carefully pre-process this information. There are a total of 13 characteristics that have been gathered. The prediction degree was thenestimated using MLA. After applying eight machine learning classification algorithms andthree deep learning algorithms achieved various degrees of accuracy.

Table 4.1: List of Machine Learning Algorithms accuracy

algorithms	accuracy (%)	precision (%)	recall (%)	f1 score (%)
K-NN	0.6414	0.64	0.64	0.64
Random Forest	0.716	0.72	0.72	0.72
Decision Tree	0.633	0.64	0.64	0.64
Naïve Bayes	0.513	0.53	0.52	0.47
Logistic Regression	0.7239	0.73	0.72	0.72
Support Vector	0.7269	0.73	0.73	0.73
Linear Support Vector	0.6707	0.67	0.67	0.67
Nu-Support Vector	0.6414	0.65	0.64	0.64

Table 4.2: List of cross-validation accuracy

Algorithms	Accuracy (%)
K Nearest Neighbor	0.6414
Random Forest	0.7158
Decision Tree	0.6336
Naïve Bayes	0.5139

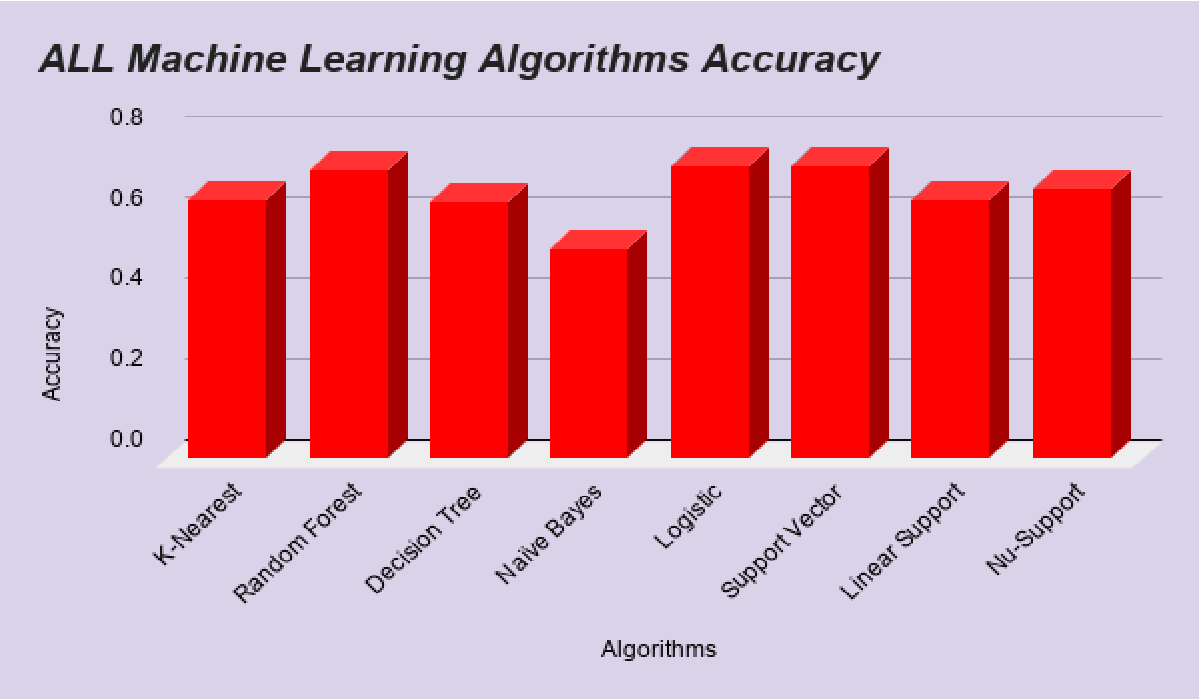


Figure 4.1: Accuracy for Machine Learning Algorithms

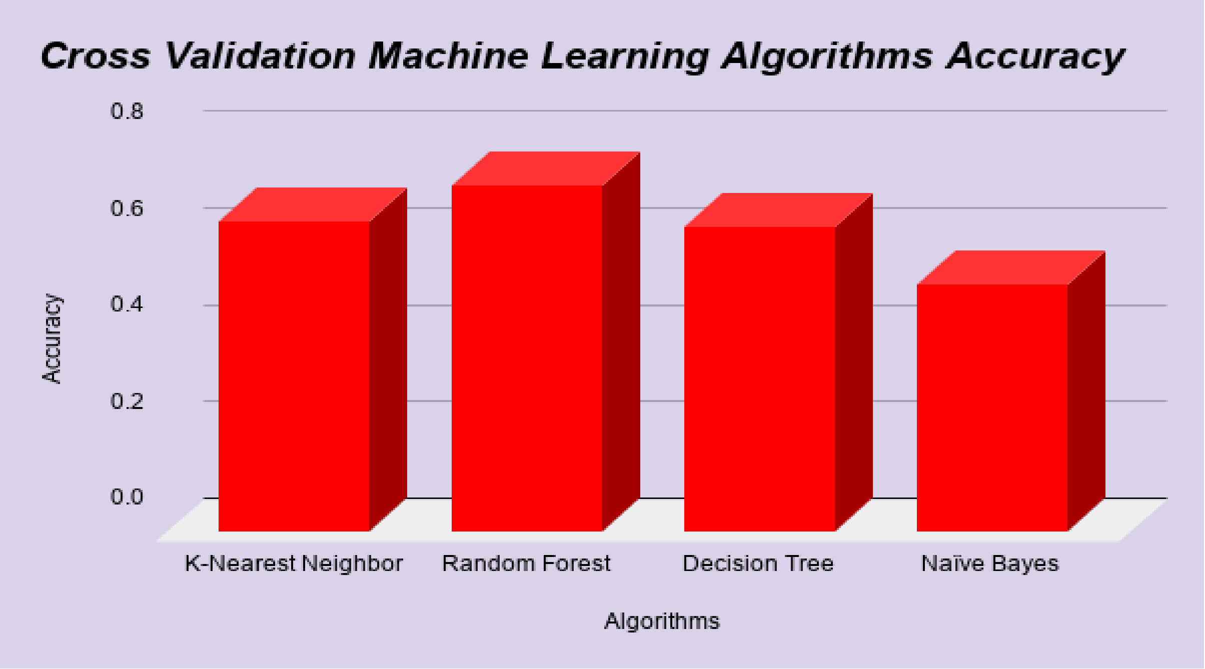


Figure 4.2: Cross-Validation Accuracy graph

Table 4.3: List of Deep Learning Algorithms accuracy

Algorithms	Accuracy (%)
Neural Network	73.72
Recurrent Neural Network	66.79
Artificial Neural Network	74.09

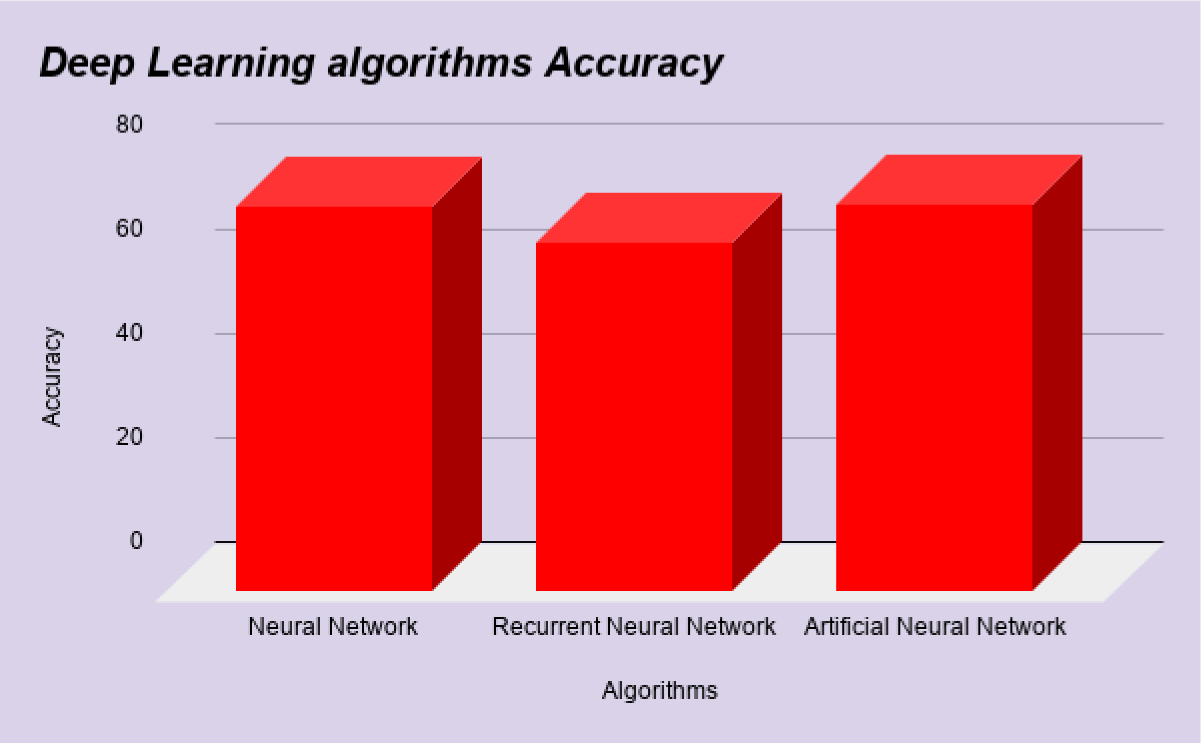


Figure 4.3: Accuracy for Deep Learning Algorithms

The K-NN algorithm is a straightforward supervised MLA. The K-NN algorithm will describe classification and regression problems. The K-NN algorithm remembers the training observation in order to classify the secret test results. The K-NN algorithm selects related items in a nearby neighborhood.

For prediction purposes, random forest generates a huge set of de-correlated trees. By introducing randomness into the tree-growing process, it decreases tree correlation. It uses split-variable randomization to generate results. At each tree break, the random forest has a smaller function search space.

A decision tree is a model that is built on trees. Using slicing laws, it divides the functions into smaller sections with identical answer values. The tree diagram is created using the divide-and-conquer technique. The decision tree needs no preprocessing and can effectively manage the categorical features without it.

Naïve Bayes is one of the earliest MLA. The Bayes theorem and basic statistics are used in this algorithm. The Naive bias model employs class probabilities and conditional probabilities. It adds attributes with a Gaussian distribution.

Logistic regression employed a logistic equation, which is known as a sigmoid function. An S-shaped curve takes real values and converts them to numbers ranging from 0 to 1.

A supervised MLA is an SVM. This method is also used for grouping and regression problems. Data structures are positioned in n-dimensional space, and the values of the features are shown at each coordinate. It produces the most homogeneous points in each subsection, which is why it is referred to as a hyperplane.

4.3 Descriptive Analysis

We determined not only the accuracy of many algorithms, but also their p-sion, re-call, f1-score, curve, and cm. Any product range must provide an evaluation of the model. Certain classifiers must be measured in the case of model evolution. For improved measurement, classifications are calculated using the test data collection.

K-Nearest Neighbor error rate diagram base on out data set.

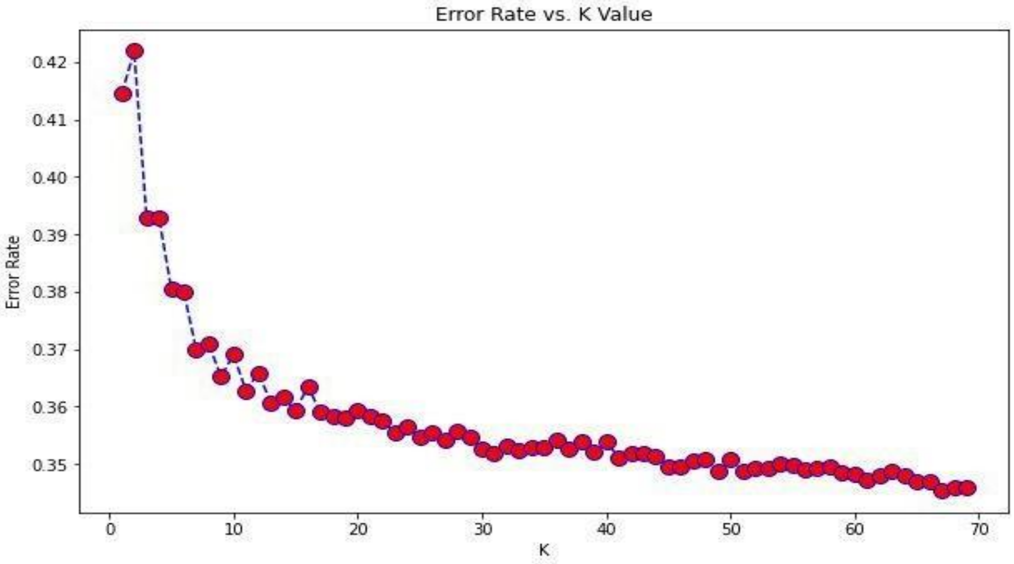


Figure 4.4: Error Rate K-Nearest Neighbor

Re-Current Neural Network model accuracy and model loss, 4.5 and 4.6 figure based on data set.

```
dict_keys(['loss', 'accuracy', 'val_loss', 'val_accuracy'])
```

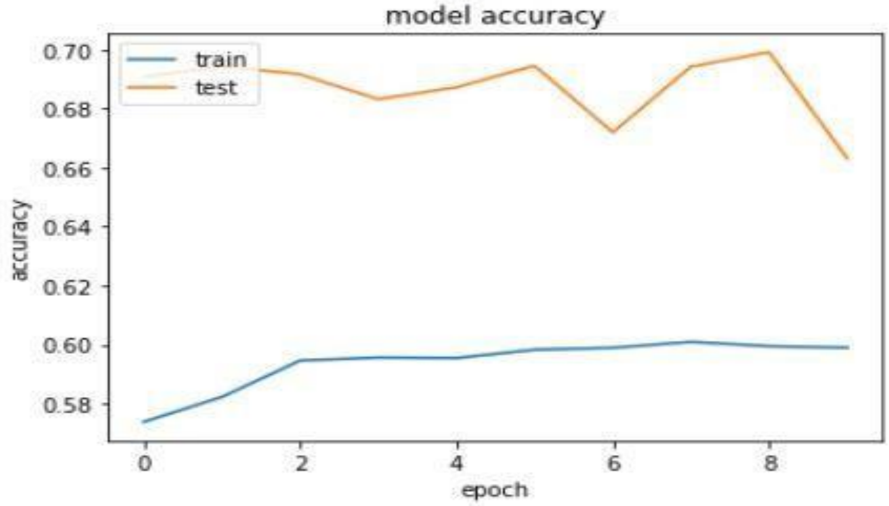


Figure 4.5: R-NN model accuracy

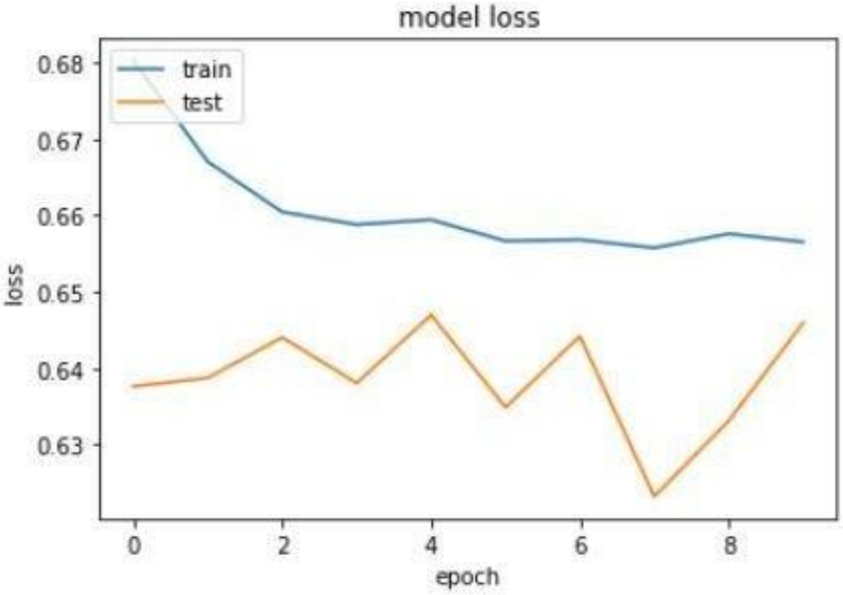
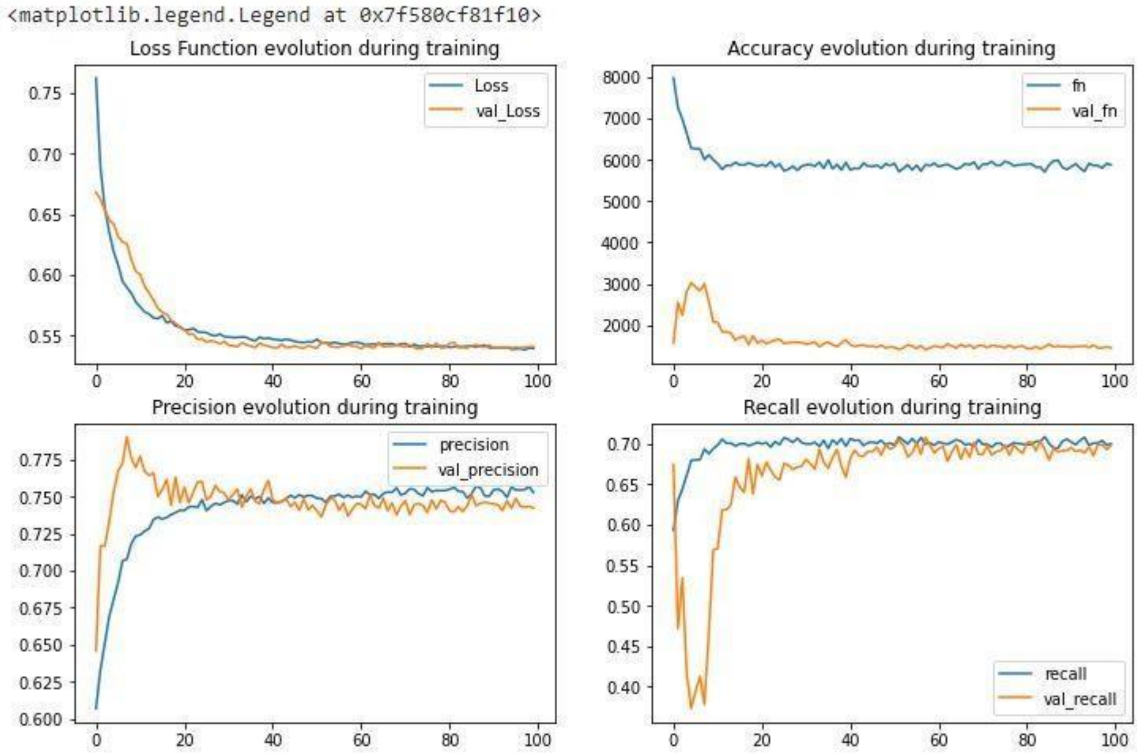


Figure 4.6: Recurrent NN model loss

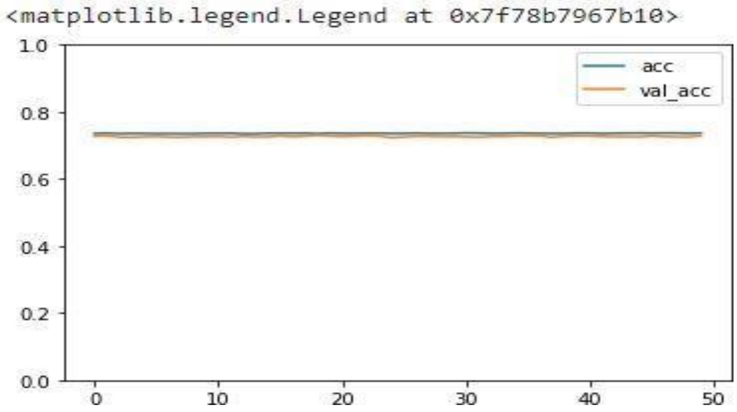
Artificial Neural Network accuracy evolution, loss function, precision evolution, and recall evolution 4.7 figure base on data set.



```
<matplotlib.legend.Legend at 0x7f580cf81f10>
```

Figure 4.7: Artificial Neural Network evolution

Neural Network accuracy evolution 4.8 figure base on data set.



```
<matplotlib.legend.Legend at 0x7f78b7967b10>
```

Figure 4.8: Neural Network evolution

Here is our confusion matrix for every machine learning algorithms has given below

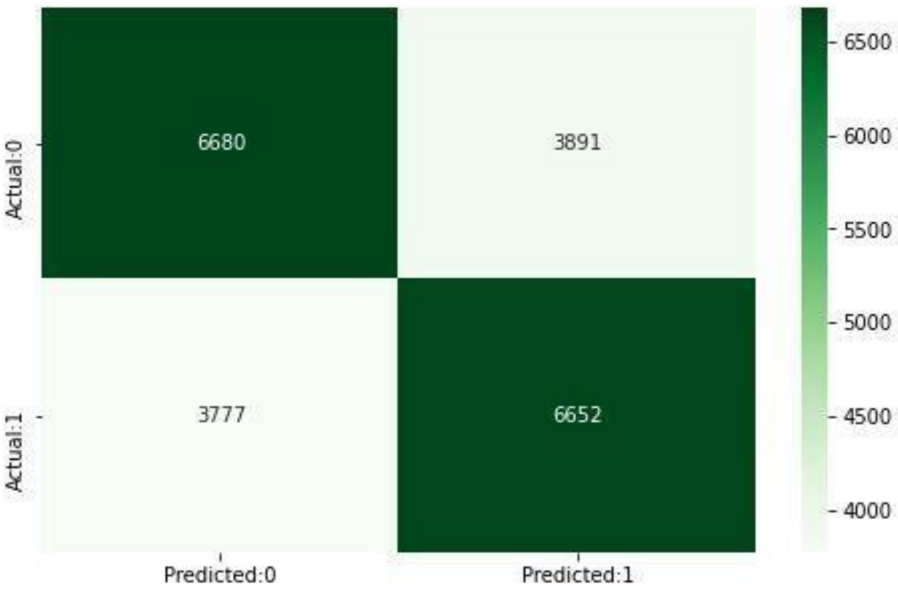


Figure 4.9: Decision Tree CM

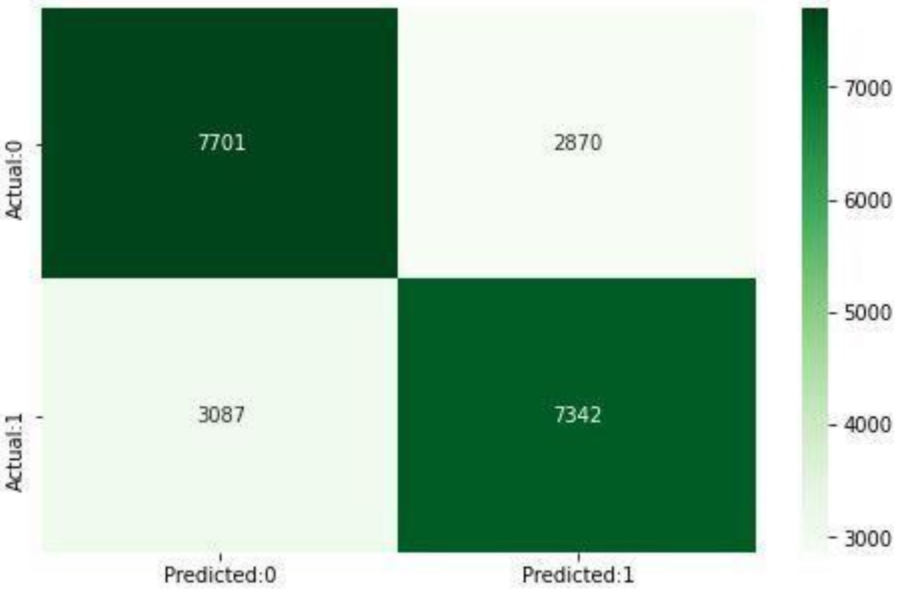


Figure 4.10: Random Forest CM

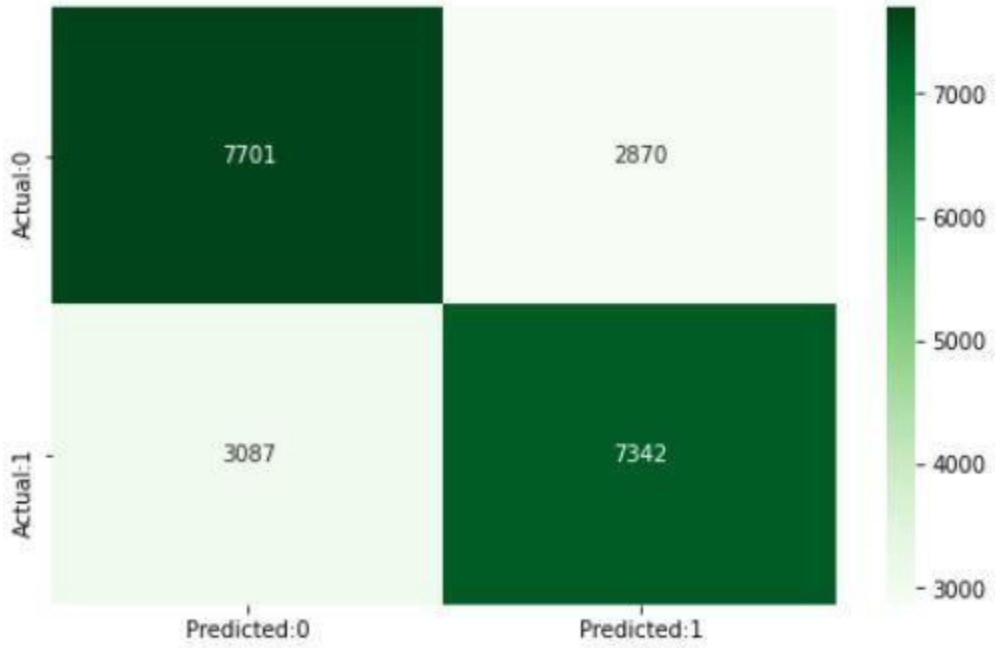


Figure 4.11: K Nearest Neighbor CM

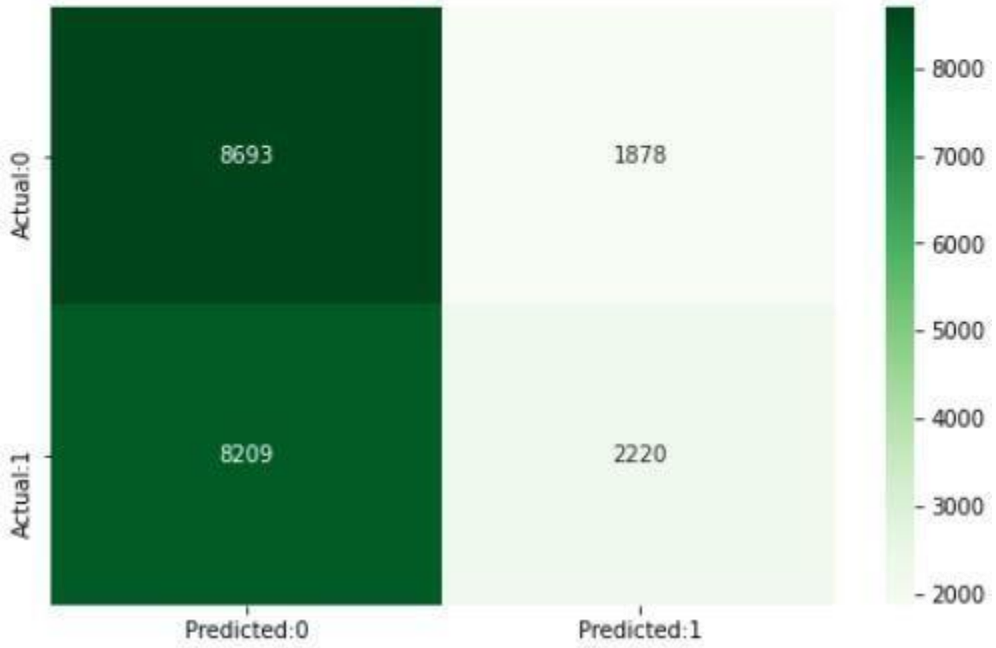


Figure 4.12: NB CM

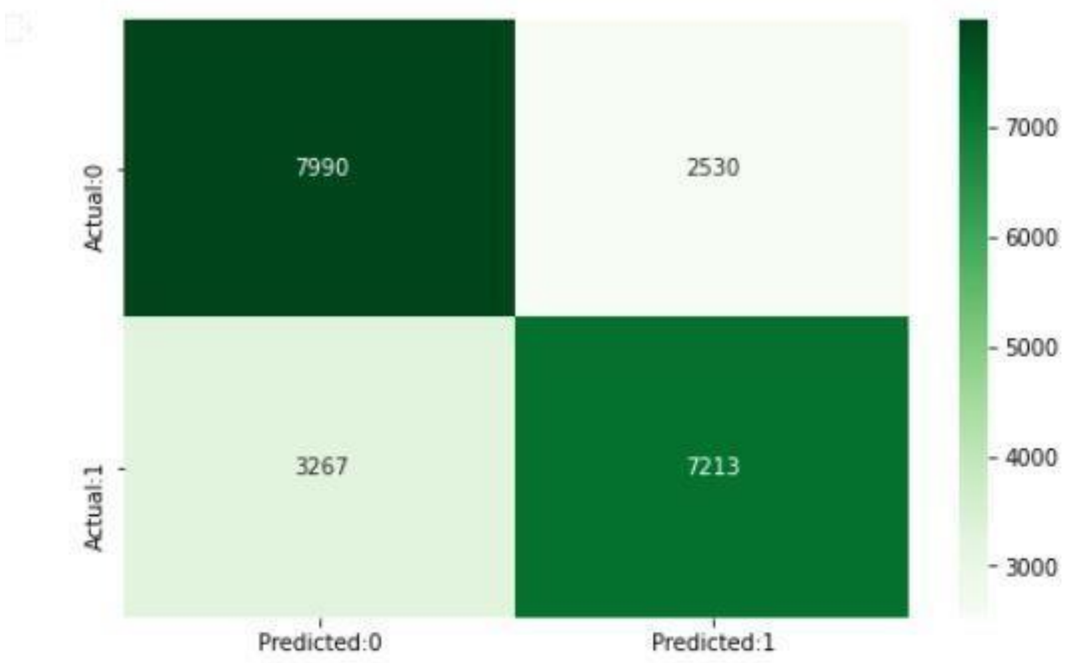


Figure 4.13: Logistic Regression CM

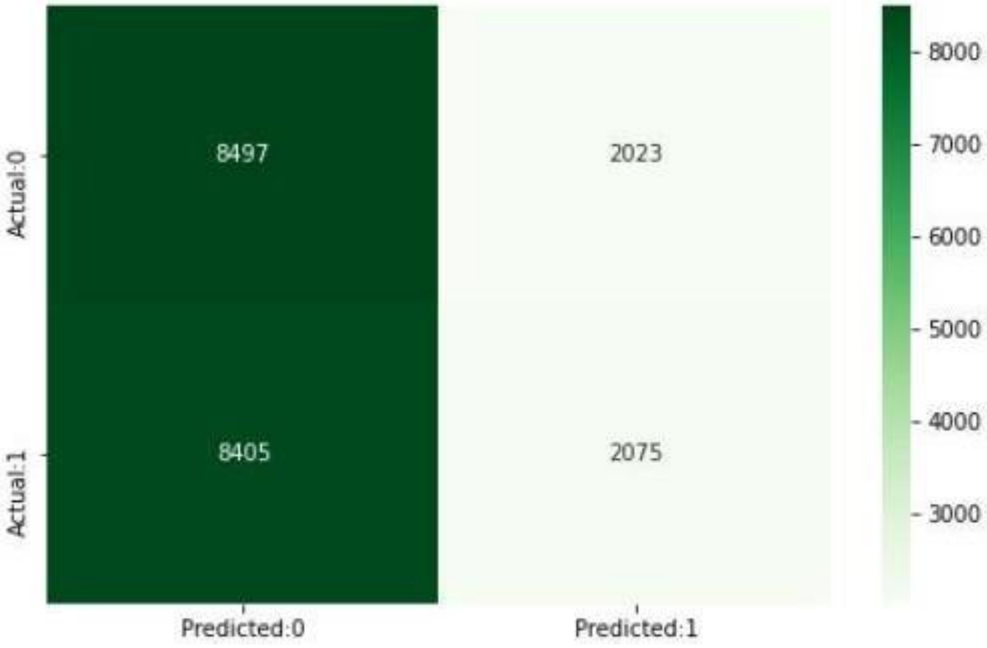


Figure 4.14: Support Vector Machine CM

4.4 Summary

The primary motive of this thesis is to motivate people about their health care by showing them a prediction of previous record. So that they can understand the risk of heart disease according to their lifestyle, food habit. When they should take necessary step for their healthy heart status. Though our thesis is unable to give any physical support or suggestions of any doctor, the people can understand about their heart status through our accuracy and prediction as we have taken many categories data about seventy thousand. This information was then expertly arranged and stored in an Excel spreadsheet. With this information we have applied Machine Learning algorithms and Deep Learning algorithms such as Neural Network, Recurrent Neural Network, Artificial Neural Network, K Nearest Neighbor, Random Forest, Naïve Bayes, Decision Tree. All the collected accuracy has been shown in Table 4.2 and Table 4.3. We have counted the best accuracy among Machine Learning algorithms in Logistic Regression and Neural Network. Among the algorithms of Deep Learning we have counted our best accuracy in Artificial Neural Network.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

There could be positive and negative impacts on society. People will be helped by our thesis in many ways. As heart disease is a serious cause of death, if people know more about this disease they will be more conscious. Our prediction will give anyone a general knowledge about heart disease risk. It will create social awareness and consciousness. Any health worker will be helped by our thesis so it will create an educational support for the health worker and the medical officers also. People don't need to go out searching for a dataset; they can monitor our data analysis and they can learn about heart disease. Usually people are not interested to share their personal data so our analysis will be helpful for those who want to learn about heart disease. It will help people to maintain social distance. Overall our work is a social work to us because any kind of people can be helped by our work. Any health center can maintain their policy and awareness learning from our work. Our thesis will create social awareness about heart disease. In the future we will make a website for any kind of person so that anyone can check their heart condition whether their heart is working healthy or not. It will definitely be a social work and every gender and any aged person will be helped with our website.

5.2 Impact on Environment

Our work is about heart disease prediction so there will be less effect on the environment. In this pandemic situation if anyone can learn from their house it will be helpful for our society and environment because people will be able to maintain social distance and from this they can save themselves from getting affected by the virus. This will have a positive effect on the environment.

5.3 Ethical Aspects

We have a plan to elaborate our work with making a website so that people can justify their health status. They will be able to check how healthy their heart is and how much they are at risk of heart disease. In that case there will be a chance of data piracy, people can be concerned about their personal data leak. In this concern we will not publicize our data, people will be able to see only their data which they have provided they will not be able to see others data. So anyone will be able to check their heart status according to their physical condition or physical health. We will not take anyone's name so there will not be left any chance of knowing whose data is this for us.

CHAPTER 6

CONCLUSION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Conclusion

Every day, the number of people suffering from heart failure rises. The proportion of heart failure patients has been increasing every day. The study's main aim is to increase heart disease prophecy accuracy. This paper compares and contrasts various machine learning techniques in order to decide which one produces the most reliable heart disease prediction outcomes. The performance of each algorithm was tested on the Cleveland dataset, and the results were compared in terms of accuracy. Random forest is also considered to be more suitable.

6.2 Implication for Further Study

Our long-term aim is to develop a smartphone application that enables patients to enter heart-related data and submit it to our expert system, which will interpret the information and provide recommendations to the patients. As a result, our system would help patients and they can easily detect if they have heart disease. The accuracy of our method is 74.09 percent, but this result falls short of our expectations. Our long-term aim is to improve accuracy by combining multiple models with a large number of various data sets.

References

[1] Motarwar, Pranav, Ankita Duraphe, G. Suganya, and M. Premalatha. "Cognitive Approach for Heart Disease Prediction using Machine Learning." In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1-5. IEEE, 2020.

[2] Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." *Mobile Information Systems* 2018 (2018)

[3] Almustafa, Khaled Mohamad. "Prediction of heart disease and classifiers' sensitivity analysis." *BMC bioinformatics* 21, no. 1 (2020): 1-18. [

4] Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, and Ruinan Sun. "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms." *Mobile Information Systems* 2018 (2018).

[5] Atallah, Rahma, and Amjed Al-Mousa. "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method." In *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, pp. 1-6. IEEE, 2019.

[6] Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." In *2020 international conference on electrical and electronics engineering (ICE3)*, pp. 452-457. IEEE, 2020.

[7] Muhammad, Yar, Muhammad Tahir, Maqsood Hayat, and Kil To Chong. "Early and accurate detection and diagnosis of heart disease using intelligent computational model." *Scientific reports* 10, no. 1 (2020): 1-17.

[8] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE Access* 7 (2019): 81542-81554.

[9] Samuel, Oluwarotimi Williams, Grace Mojisola Asogbon, Arun Kumar Sangaiah, Peng Fang, and Guanglin Li. "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction." *Expert Systems with Applications* 68 (2017): 163-172.

[10] Paul, Animesh Kumar, Pintu Chandra Shill, Md Rafiqul Islam Rabin, and Kazuyuki Murase. "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease." *Applied Intelligence* 48, no. 7 (2018): 1739-1756.

[11] Alizadehsani, Roohallah, Mohammad Javad Hosseini, Abbas Khosravi, Fahime Khozeimeh, Mohamad Roshanzamir, Nizal Sarrafzadegan, and Saeid Nahavandi. "Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries." *Computer methods and programs in biomedicine* 162 (2018): 119-127.

[12] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." *International Journal of Computer Applications* 47, no. 10 (2012): 44-48.

[13] Chaki, Dipankar, Amit Das, and M. I. Zaber. "A comparison of three discrete methods for classification of heart disease data." *Bangladesh Journal of Scientific and Industrial Research* 50, no. 4 (2015): 293-296.

[14] Saboji, Rashmi G. "A scalable solution for heart disease prediction using classification mining technique." In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 1780-1785. IEEE, 2017.

[15] Bui, Anh L., Tamara B. Horwich, and Gregg C. Fonarow. "Epidemiology and risk profile of heart failure." *Nature Reviews Cardiology* 8, no. 1 (2011): 30.

[16] 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*, pp. 108-115. IEEE, 2008.

[17] Rathnayakc, Bandarage Shehani Sanketha, and Gamage Upeksha Ganegoda. "Heart diseases prediction with data mining and neural network techniques." In *2018 3rd International Conference for Convergence in Technology (I2CT)*, pp. 1-6. IEEE, 2018.

[18] Malav, Amita, Kalyani Kadam, and Pooja Kamat. "Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy." *International Journal of Engineering and Technology* 9, no. 4 (2017): 3081-3085.

[19] Tarle, Balasaheb, and Sudarson Jena. "An artificial neural network based pattern classification algorithmfor diagnosis of heart disease." In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1-4. IEEE, 2017.

[20] Karayılan, Tülay, and Özkan Kılıç. "Prediction of heart disease using neural network." In *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 719-723. IEEE, 2017.

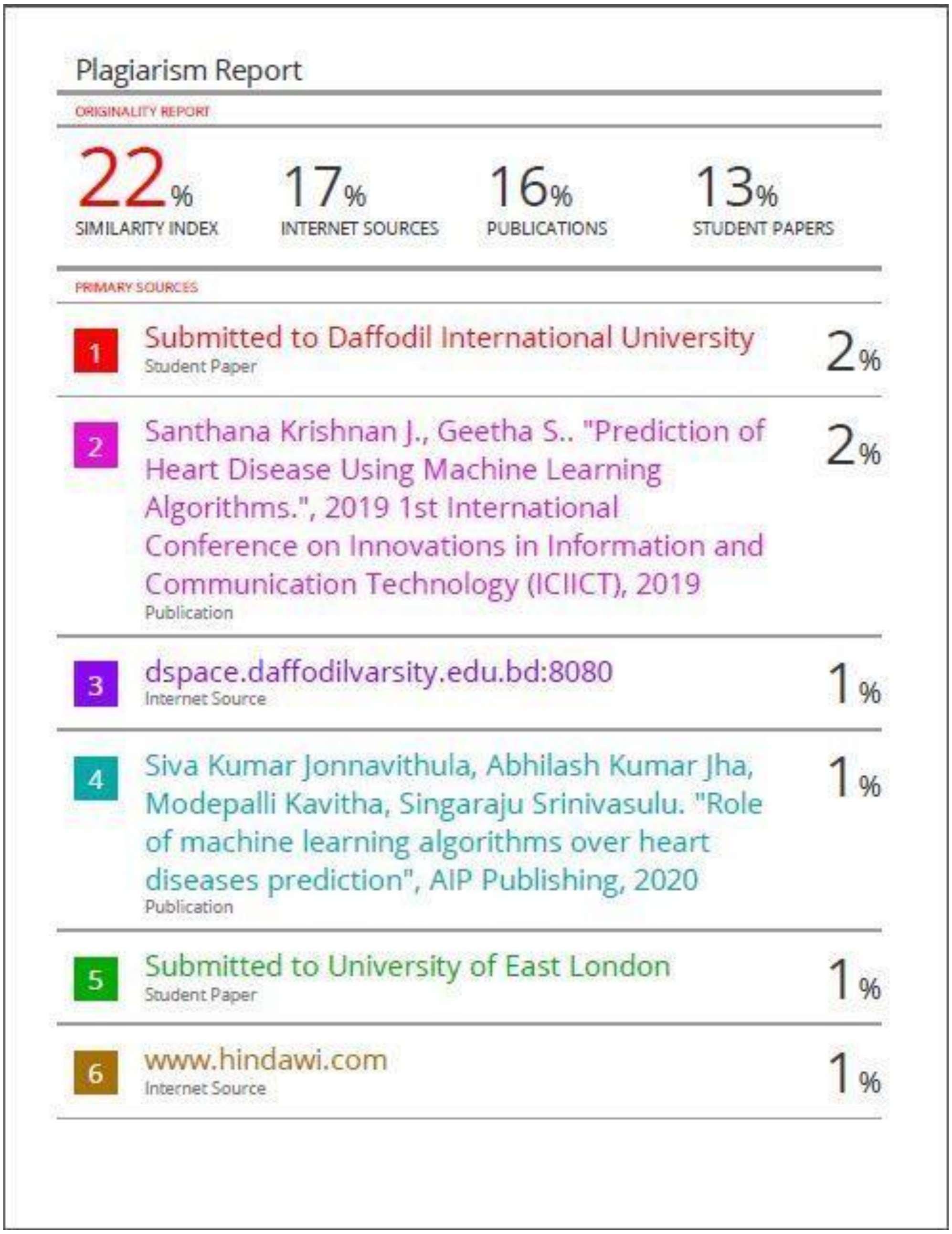
[21] Esfahani, Hamidreza Ashrafi, and Morteza Ghazanfari. "Cardiovascular disease detection using a new ensemble classifier." In *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, pp. 1011-1014. IEEE, 2017.

[22] Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining." In *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)*, pp. 1-5. IEEE, 2017.

[23] Sultana, Marjia, Afrin Haider, and Mohammad Shorif Uddin. "Analysis of data mining techniques for heart disease prediction." In *2016 3rd international conference on electrical engineering and information communication technology (ICEEICT)*, pp. 1-5. IEEE, 2016.

[24] Aldallal, Ammar, and Amina Abdul Aziz Al-Moosa. "Using Data Mining Techniques to Predict Diabetes and Heart Diseases." In *2018 4th International Conference on Frontiers of Signal Processing (ICFSP)*, pp. 150-154. IEEE, 2018.

[25] 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)*, pp. 520-525. IEEE, 2015.



Plagiarism Report

ORIGINALITY REPORT

22%	17%	16%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Santhana Krishnan J., Geetha S.. "Prediction of Heart Disease Using Machine Learning Algorithms.", 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019 Publication	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	Siva Kumar Jonnavithula, Abhilash Kumar Jha, Modepalli Kavitha, Singaraju Srinivasulu. "Role of machine learning algorithms over heart diseases prediction", AIP Publishing, 2020 Publication	1%
5	Submitted to University of East London Student Paper	1%
6	www.hindawi.com Internet Source	1%