

Heart Failure Prediction using Machine Learning Algorithm

BY

Md. Ramizul Abedin

ID: 171-15-9315

Md. Golam Hafiz Shakil

ID: 171-15-8803

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

Supervised By

Md. Abbas Ali Khan

Senior Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Majidur Rahman

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2021

APPROVAL

This Project/internship titled **Heart Failure Prediction using Machine Learning algorithm** submitted by **Md. Ramizul Abedin**, ID No: 171-15-9315, **Md. Golam Hafiz Shakil**, ID No: 171-15-8803 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 02-05-2021.


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

Dr. Fizar Ahmed
Assistant Professor

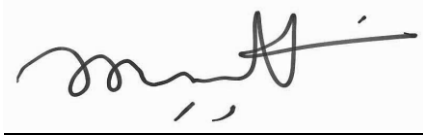
Department of Computer Science and Engineering
Faculty of Science & Information Technology



Internal Examiner

Md. Azizul Hakim
Senior Lecturer

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



External Examiner

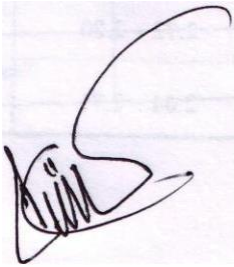
Dr. Mohammad Shorif Uddin
Professor

Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Abbas Ali Khan, Senior Lecturer, Department of CSE** Daffodil International University and Co-Supervisor **Mr. Majidur Rahman, 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:



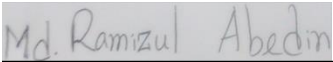
Md. Abbas Ali Khan
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Majidur Rahman
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md. Ramizul Abedin
ID: 171-15-9315
Department of CSE
Daffodil International University

Golam
Hafiz

Md.Golam Hafiz Shakil

ID: 171-15-8803

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 Abbas Ali Khan, Senior Lecturer, Department of CSE Daffodil International University and Co-Supervisor Mr. Majidur Rahman, Lecturer, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning Algorithm” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Prof. Dr. Touhid Bhuiyan, Head, Department of CSE, for his kind help to finish our project and also to the other faculty members and the staffs of CSE department of Daffodil International University.

We would like to thank our entire course mates 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

We are living in the modern age and our daily life is undergoing multiple changes that directly have positive and negative effects on our health. Different types of diseases are greatly increased for this changing nature where heart disease has become more prevalent. As a consequence, people's lives are at risk. The changes in blood pressure, cholesterol, pulse rate, etc. can lead to heart diseases that include narrowed or blocked blood vessels. It may cause Heart failure, congenital heart disease, heart disease, Myocardial infarction (Heart attack), Hypertrophic cardiomyopathy, pulmonary stenosis, and even sudden cardiac arrest. Many forms of heart disease can be detected or diagnosed with different medical tests by considering the family medical history and other factors. But it is quite hard to predict heart disease without any medical exams. But "Machine Learning" is making it a little simpler nowadays. The purpose of the current study is to predict the risk of heart diseases and to make people aware of their daily routine with high accuracy. For the prediction of heart disease risk, we use five 'Machine Learning' classification algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB) and Random Forest (RF). Our finding demonstrates that DT with greater precision outperforms the SVM, NB, KNN, and RF. Finally, we use massive algorithm features which can predict the symptoms of heart disease so that people should be take care of heart health.

TABLE OF CONTENTS

| CONTENTS | Page |
|--------------------------------|-------------|
| Board of examiners | ii-iii |
| Declaration | iv-v |
| Acknowledgements | vi |
| Abstract | vii |
| CHAPTER | 1-4 |
| CHAPTER 1: INTRODUCTION | 1-4 |
| 1.1 Introduction | 1-2 |
| 1.2 Motivation | 2-3 |
| 1.3 Objectives | 3 |
| 1.4 Rationale of the Study | 3 |
| 1.5 Research Questions | 3 |
| 1.6 Expected Output | 4 |
| 1.7 Report Layout | 4 |

| | |
|--|-------|
| CHAPTER 2: BACKGROUND & Algorithm | 5-17 |
| 2.1 Introduction | 5 |
| 2.2 Related Works | 5-8 |
| 2.3 Research Summary | 8 |
| 2.4 Scope of the Problem | 8 |
| 2.5 Challenges | 9 |
| 2.6 Algorithms | 9 |
| 2.6.1 LINEAR REGRESSION | 9-11 |
| 2.6.2 DECISION TREE ALGORITHM | 11-13 |
| 2.6.3 Support Vector Machine | 13-14 |
| 2.6.4 NAIVE BAYES ALGORITHM | 14-16 |
| 2.6.5 KNN Algorithm | 16-17 |
| CHAPTER 3: Research Methodology | 17-23 |
| 3.1 Introduction | 17 |
| 3.2 Research Subject and Instrumentation | 18 |
| 3.2.1 Research Subject | 18 |
| 3.2.2 Research Instrumentation | 18 |
| 3.3 Data Collection Procedure | 18-19 |
| 3.4 Statistical Analysis | 19-21 |
| 3.5 :Train Methodology | 21-23 |

| | |
|---|--------------|
| CHAPTER 4: DESIGN SPECIFICATIOND | 24-28 |
| 4.1 Introduction | 24 |
| 4.2 Experimental Results | 24-25 |
| 4.3 Descriptive Analysis | 25-27 |
| 4.4 Summary | 27-28 |
| | |
| CHAPTER 5: Implementation (Results and Analysis) | 28-36 |
| 5.1 Introduction | 28 |
| 5.2 Results and Analysis | 28-31 |
| 5.3 Accuracy of Models with All Features | 31 |
| 5.4 Feature Engineering | 32-35 |
| 5.6 Accuracy of Models with Selected Features | 35-36 |
| 5.7 Cross Validation | 36-37 |
| 5.8 Analysis | 37 |
| | |
| CHAPTER 6: Results and Analysis | 37-38 |
| 6.1 Conclusion | 37 |
| 6.2 Future Scope | 37-38 |
| | |
| REFERENCES | 39-40 |
| | |
| APPENDIX | 41-42 |

LIST OF FIGURES

| FIGURES | PAGE |
|---|-------------|
| Figure 3.4.1: Statistical view of collected dataset Age group (Death event) | 20 |
| Figure 3.4.2: Statistical view of collected dataset DEATH_EVENT | 21 |
| Fig3.5: Model summary | 22 |
| Figure 4.3.1: Statistical view of collected dataset normal visualization | 25 |
| Figure 4.3.2: Statistical view of collected dataset visualization | 26 |
| Fig5.2.1: Results of KNN Algorithm | 29 |
| Fig5.2.2: Percentage of heart disease patients in dataset | 29 |
| Fig5.2.3: Distribution of gender of heart disease dataset | 30 |
| Fig5.2.4: Age distribution of normal and heart patients | 30 |
| Fig5.4.1: Seraum_creatinine for the dataset of normal & heart patients | 31 |
| Fig5.4.2: DEATH_EVENT serum creatinine | 32 |
| Fig5.4.3: St slope normal and heart patients | 33 |
| Fig5.4.4: DEATH_EVENT diabetes | 33 |
| Fig5.4.5: high blood pressure | 34 |
| Fig5.4.6: DEATH_EVENT high blood pressure | 35 |
| Fig5.4.7: Heart patients | 35 |

CHAPTER 1

INTRODUCTION

1.1 Introduction

Human health is one of the world's challenges for humanity. World Health Organization (WHO) mentioned that an individual is a primary right to healthy health body. So keep it up every people need to provide proper and healthy proper healthcare services. But it is a matter of sorrow that all 31 percent Heart-related problems are the leading cause of death worldwide [1]. Heart is the main organ of our body. All People life depends on the heart itself work effectively. If the power of the heart is not great, it will affect different parts. Every human body such as the cerebrum, kidneys, etc. at that stage. If the human body is blood circulation will be wasted, organs like mind and heart will suffer. Blood binding or captured in the cerebrum is often called a stroke and it is called a heart attack. Human life is surely entirely main part to the ability of the heart and brain to function. Every people needs the power of both. Many reasons increment the danger of heart disease [2]. Lots out there main causes of coronary heart disease risk, such as coronary family history [3] like sickness, over smoking, eating less, over Vibration, cholesterol, over blood cholesterol, obesity rate, physical inertia, overweight, highest blood pressure, stress or heart disease (otherwise known as coronary), such as high blood pressure, chest pain, drug use, etc. Heart disease continues to be the leading cause of death in the world for centuries. In 2015, World Health Organization (WHO) estimates 17.7 million total casualties from congestive heart failure worldwide [4]. Running report by the World Health Organization (WHO) in 2018 shows that the results are 56.9 millions people of deaths on the planet in 2016 due to heart attack disease [5]. And 17.3 million people died of heart attack disease in 2008 [6]. In 2011, the disease alone murdered about 787,000 people and 380,000 people each year with coronary heart disease, one person has a heart attack every 30 seconds and one dies every 60 seconds heart attack disease [7]. Different types of machine learning algorithms for example, Linear Regression (LR), Logistic Regression (LR), Decision tree (DT), Multi-Level Perceptron (MLP), Nave Base (NB), K-Nearest Neighbor (K-NN), Random Forest (RF) and Support Vector Machine (SVM) [8] used to separate and use valuable data from clinical dataset with significant client input and effort [8]. Machine learning is basically the expression of learning from a huge measure of unrefined information. Machine learning is otherwise called the main extent of information the principal [9]. There are two basic models of machine learning called predictive models and descriptive models. The predictive model is identified as the created model expect a specific result or outcome using prescients display system [10]. When the narrator model is identified as a model designed to give a higher idea information without focusing on any specific variable using the investigation system factor tests and group tests and the like [11]. Human heart diseases

are the leading cause of death in public: a number of people die each year heart disease than some other illness. So if we can predict coronary disease and every person stay be careful beforehand, a lot of deaths can be reduced. In this study, we recommended a machine learning to evaluated heart disease datasets to predict heart disease danger levels based on all selected symptoms. The hypotheses of this strategy will help individuals understand their heart condition that they may be aware of it and if they have any problem then they communicate the great doctors, so that as soon as possible as a result of reducing mortality. Identification and great nursing of the heart disease is very complex, especially in increasing countries, in the absence of diagnostics lack of devices and physicians and other resources infect proper prediction and treatment of heart disease patients. With the recent advent of information technology and this concern Machine Learning (ML) techniques are being used to progress software to help doctors create early heart attack disease decision. Early detection of the disease and right predicting the likelihood of a person at danger of heart disease can reduce mortality.

Blood pressure is one of the principal factors affecting heart failure disease. Following to back studies, 13% of cardiovascular deaths are caused by high blood pressure or hypertension. A study published on Friday in Jame Cardiology reporting heart disease almost patients with coronavirus in Wuhan found that 20% of covid-19 patients had recorded of heart attack problems [12].

Caused shortness of breath and lung damage but later it was lots of patients have been found to be affected with the covid-19 expertise heart problems. Now doctors are advising that one in five deaths is due to the coronavirus because of the heart damage.

Medical data mining techniques are used in treatment data to extract semantic patterns knowledge. Redundancy in treatment information, multiple features, imperfect and so on the close relationship with time is the problem of using huge amounts of data effectively became a big problem for the health sector. Provides data mining methods and technology converts these data important into useful important or answerable information. This the heart disease prediction system will help cardiologists make quick best decisions so that more patients can receive great nursing in a small period of time as a result saving millions of lives [13].

Blood pressure is one of the principal factors affecting heart failure disease. Following to back studies, 13% of cardiovascular deaths are caused by high blood pressure or hypertension.

1.2 Motivation

The main motivation of this research is to provide a specific disease analysis system with extended list of capabilities. Actually, the number of doctors needs to be performed tests

for a particular disease analysis require a lot of time, effort and hard cash. However our proposed system will predict the danger of heart attack disease with high accuracy bring about reduction of time and labor.

1.3 Objective

The main purpose of this study is to predict a weather in which a patient has a heart attack or does not use various machine learning algorithms in qualified datasets. Look for relationships with different attribute. Gain a clear idea about our proposed data mining strategies and analyze the results and compare the results of different data mining strategies. We will analyze our strategies if there is any possibility to improve our results.

1.4 Rationale of the Study

Convenient realization of heart disease can prevent death. In each case, it is view at the last stage of heart disease or after death. So the basic thing is we can't do not differentiate or understand coronary heart disease at an early stage. For this condition, the machine learning method can detect coronary or heart damage disease. By different data mining strategies we are able to differentiate heart disease in the beginning period, that we can reduce the sudden human death.

1.5 Research Questions

There are many diseases that affect us severely, one of which is cardiovascular disease. It is a serious disease, because we can hear that most of the people die of heart attack disease and other similar diseases related to the heart [14,15,16].

Most of the medical scientists monitoring that many times heart attack patients may not be in the majority survived a long time and die of a heart failure.

Following the questions arise when we think about it implement our idea -

How to forecast the danger of heart disease with over accuracy?

How do people make people awareness of their heart health?

How to prevent accidental death by heart attack disease?

How to make every human awareness of their eating habits?

1.6 Expected Output

There are many people who are very ignorant about their heart health. Because of this they are suffering from a variety of diseases where heart disease is very common and knocking on the door of death in the long run. If people using our system always monitor the condition of their heart, they will be aware of their heart health, so that the mortality due to heart disease will be reduced. We do not provide any medical assistance through our project but we can easily make people aware of their predictable danger levels of heart attack disease by changing their lifestyle, changing their diet, quitting smoking, etc. Because they do not give us an proper rate of danger of heart failure disease, but in this study we realize a lots of characteristics such as smoking, family previous generation of heart problems etc.

Research history of heart disease, cholesterol, over blood pressure, chest pain, causes of age, gender, stress, physical exercise, drug addiction, taking unprescribed medicine etc as a result we were able to view the danger of heart failure disease with higher validity.

1.7 Report Layout

In this report, evaluating the cardiovascular dataset, we proposed a machine learning classification algorithm to forecast heart problems danger levels based on principal symptoms. Assumptions from this technology will help people understand their heart condition so that they can be aware of it and if they have any heart problems. They will go to the great doctor as soon as possible as a result of reducing the mortality rate. This report is divided into six sections. This is the first part where we talk about the motivation of our work and the expected results. In the second chapter (Chapter 2) we address related work, various topics. We address the issue of process and implementation data collection in the following chapters. The fourth chapter is for test results and evaluation. In the fifth chapter we discuss how to implement our proposed system. Towards the end, chapter six represents the conclusion and future work.

CHAPTER 2

Background & Algorithm discussion

2.1 Introduction

Heart failure disease is a term identify by a massive quantity of heart-related healthcare? Classification of algorithms for machine learning make it a short natural to forecast heart disease. Heart disease is a major problem in medical science. It is easy way to predict heart problems danger using classification algorithms in machine learning. In Machine Learning (ML), the training and test data included in the individual classes were classified using abundant systems. It seems important to predict the danger of heart problems through machine learning in diagnosing heart attack disease. The intention of this program is to evaluate the unique classification systems used in the forecast of heart failure disease. The number of experiments provided on this day to diagnoseticate the illness is provided but using the system of 'machine learning' can decrease heart disease.

2.2 Related Works

At present, the primary concern of medical science is the prognosis of heart disease. We can easily predict heart disease risk using a variety of techniques such as "machine learning" such as SVM, DT, NB, KNN (nearest neighbor), RF (random forest) and ANN (artificial neural network). Various studies have been done and coronary disease prognosis is gradually getting more and more accurate. They have applied different data mining techniques to test and predict the risk of heart disease and have obtained different results of accuracy. Table 1 shows the different types of data mining techniques used to predict heart disease. The hazard factor classes are explained in Cardiac Prediction [17]. This at work, four variables such as behavior, condition, age and gender of the patient are used to differentiate coronary disease. The work of Sudha et al. [18] proposed classification Strategies such as Naive Bayes, Decision Tree and Neural Network for forecast heart disease danger. Classification strategies were adopted, such as decision trees, Bayesian classifieds, and back ending neural systems.

| Classifier of algorithm | Accuracy (%) | Objective | Year | Reference |
|--------------------------------|---------------------|---|-------------|------------------|
| Support Vector Machine | 73.2 | forecast and diagnosis of heart disease | 2019 | [1] |

| | | | | |
|---------------------------|-------|---|------|-----|
| Artificial neural network | 48 | patients using Data Mining (DM) Technique | | |
| Support Vector Machine | 86 | risk prediction of heart disease at classification in data mining | 2015 | [7] |
| Decision Tree | 76 | | | |
| Naive Bayes | 79 | | | |
| Artificial neural network | 85 | | | |
| Decision Tree | 82.5 | exploration of heart disease and forecast of heart Attack in coal mining regions using Data Mining Techniques | 2010 | [3] |
| Support Vector Machine | 82.5 | | | |
| Decision Tree | 82.22 | heart disease forecast using feature selection approaches | 2019 | [2] |
| Naive Bayes | 84.24 | | | |
| Random Forest | 84.17 | | | |
| | | | | |

In this research study records with unnecessary knowledge were banished from the data distribution center before the mining process took place. Following to the research study [19] Intelligent Heart Disease forecasting systems using data mining techniques, Sellappan Palaniappan and others. Used three data mining techniques. Expulsion of confidential knowledge from the language and capabilities of the DMX (Data Mining Extensions) investigation and the heart disease database, buildings and models found through preparation and approval have been recorded in contrast to a test dataset. Qualifications are represented using lift charts and classification metrics. Gives the best model to expect in patients with coronary heart disease nervous networks and decisions followed by trees off an impression of being naval base. Hlaudi Daniel Masethe work [20] predicting heart disease using taxonomy the algorithm was tested to find the best system for predicting heart attack and anticipation. Who Composed by K. Sudhakar et al. [21] Study of heart disease prognosis using data mining various methods have been informed over the years to determine the rate of coronary disease prognosis. These methods include artificial neural networks, seamless bases, decision trees, and classification algorithms. Alagugowri et al. [22] created an anticipated framework for predicting heart health illness. The danger of heart attack based on the weighted fuzzy rule is explained in the prediction [23]. In this study, a hybrid algorithm involving heavy credit payment techniques as well as common mining methods was used to create vague rules. Given under vague rules, a heart attack is expected / predicted. Laura E. Burke and. Al. [24] and examined a group of researchers examining how mobile health could play an important role in preventing cardiovascular disease. They showed numerous insights and gave some thought to how multilevel cardiovascular disease can be fought.

They have written some future studies, for example, strengthening the regular physical activity, quitting smoking, multidrug application for the treatment of obesity to control hypertension, and dyslipidemia; and treatment of diabetes mellitus. Heart Disease Prediction System proposed and developed by AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin.

Using the data mining method

- a) Choice of important traits for prognosis of coronary disease.
- b) Artificial neural network for coronary disease dependent system notable highlights.

The accuracy of the expectation is about 80% and easy to use HDPS (hydraulic data processing system) and it has highlights like ROC (rate of change) band show, execution

show area [25]. Following to Abhishek Taneja [26] manages the operation of the heart disease forecast system using data mining techniques. Four attempts were made using selected order calculations in 7339 illustrated full preparation data-sets. To use Knowledge Discovery (KDD) in databases a predictive model that can predict the risk of heart disease based on estimates. Who Composed by K. Srinivas et al. [27] For example, data mining methods are proposed for regulation Data-based, Decision Tree, Multinomial Naive Base, Gaussian Naive Bayes, Bernoulli Naïve Bayes and Artificial Neural Networks, a huge amount of treatment considerations to predict the danger of heart attack. We discovered it the existing systems depicted in the research paper show fewer features than our proposed system, we used both public and medical datasets in them algorithm and we got the most noticeable accuracy.

2.3 Research Summary

There are many people who are very ignorant about their health. Because of this they are suffering from a variety of dangerous diseases where heart attack disease is very common and knocking on the door of death in the long run. If people using our system always monitor the condition of their heart, the death rate due to heart disease will be reduced so that they will be aware of their heart health. We do not provide any medical assistance through our project but we can easily make people aware of their lifestyle, changing diet, quitting smoking, etc. by showing the predictable risk level of heart disease. Which is why they don't give us any right heart disease risk rates However in this study we found that smoking, family previous history of cardio disease, cholesterol, over blood pressure, chest pain, age causes, gender as a outcome of stress, regular exercise, taking or not taking medication, we have been able to view a over danger of heart failure disease.

2.4 Scope of the Problem

Heart disease is the main principal cause of untimely mortality. The only reason many people die of heart disease is due to unhealthy everyday routine work and bad eating habits. This is why we decided to do research on heart disease danger forecasting so that we can abate mortality through our process. We proposed a system that gives people forecast values of heart disease danger so that people can be conscious of their heart disease. Because Awareness is the best way to prevent of heart disease. So the scope of the modern research study is the heart failure diseases.

2.5 Challenges

s of heart failure disease All of everything has its black end. That is why we have faced many problems in researching and implementing our system. Sometimes it was very difficult to manage but by the grace of the Almighty we overcame these problems. This makes us the following difficulty

Research is more difficult -

1. Time of data collection from different hospitals
2. Algorithm alternative or pick time
3. To apply Machine Learning (ML) classification algorithms
4. To implement our raised process
5. During the election of external cause

2.6 Algorithms

2.6.1 LINEAR REGRESSION:

Linear regression is one of the least demanding and most well known Machine Learning calculations. It is a factual strategy that is utilized for prescient investigation. Straight relapse makes expectations for ceaseless/genuine or numeric factors like deals, compensation, age, item cost, and so on

Direct relapse calculation shows a straight connection between a reliant (y) and at least one free (x) factors, thus called as straight relapse. Since direct relapse shows the straight relationship, which implies it discovers how the worth of the reliant variable is changing as indicated by the worth of the free factor.

TYPES OF LINEAR REGRESSION:

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- **Multiple Linear Regression:**

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

LINEAR REGRESSION LEARNING THE MODEL:

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available. In this section we will take a brief look at four techniques to prepare a linear regression model. This is not enough information to implement them from scratch, but enough to get a flavor of the computation and trade-offs involved.

There are many more techniques because the model is so well studied. Take note of Ordinary Least Squares because it is the most common method used in general. Also take note of Gradient Descent as it is the most common technique taught in machine learning classes.

SUMMARY:

In this post you discovered the linear regression algorithm for machine learning.

You covered a lot of ground including:

- The common names used when describing linear regression models.
- The representation used by the model.
- Learning algorithms used to estimate the coefficients in the model.
- Rules of thumb to consider when preparing data for use with linear regression.
 - **WHAT IS LOGISTIC REGRESSION:**
 -
 - Logistic Regression is a characterization calculation. It is utilized to foresee a twofold result (1/0, Yes/No, True/False) given a bunch of autonomous factors. To address parallel/straight out result, we utilize sham factors. You can likewise consider calculated relapse an exceptional instance of direct relapse when the result variable is absolute, where we are utilizing log of chances as reliant variable. In

basic words, it predicts the likelihood of event of an occasion by fitting information to a logit work.

HOW LOGISTIC REGRESSION IS USED:

It is important to understand that logistic regression should only be used when the target variables fall into discrete categories and that if there's a range of continuous values the target value might be, logistic regression should not be used. Examples of situations you might use logistic regression in include:

- Predicting if an email is spam or not spam
- Whether a tumor is malignant or benign
- Whether a mushroom is poisonous or edible.

When using logistic regression, a threshold is usually specified that indicates at what value the example will be put into one class vs. the other class. In the spam classification task, a threshold of 0.5 might be set, which would cause an email with a 50% or greater probability of being spam to be classified as "spam" and any email with probability less than 50% classified as "not spam".

LOGISTIC VS. LINEAR REGRESSION:

We may be asking yourself what the difference between logistic and linear regression is. Logistic regression gives you a discrete outcome but linear regression gives a continuous outcome. A good example of a continuous outcome would be a model that predicts the value of a house. That value will always be different based on parameters like it's size or location. A discrete outcome will always be one thing (you have cancer) or another (you have no cancer).

CONCLUSION:

Logistic regression is a powerful machine learning algorithm that utilizes a sigmoid

function and works best on binary classification problems, although it can be used on multi-class classification problems through the "one vs. all" method. Logistic regression (despite its name) is not fit for regression tasks.

In order to take your understanding of logistic regression farther, it would be a good idea to try and apply it to other datasets, to see what kinds of classification problems it performs well at. You may also wish to continue reading about [the probability theory](#) behind the algorithm.

2.6.2 DECISION TREE ALGORITHM:

The decision tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for regression problem.

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

Here are some useful terms for describing a decision tree:

- **Root Node:** A root node is at the beginning of a tree. It represents entire population being analyzed. From the root node, the population is divided according to various features, and those sub-groups are split in turn at each decision node under the root node.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Decision Node:** When a sub-node splits into further sub-nodes, it's a decision node.
- **Leaf Node or Terminal Node:** Nodes that do not split are called leaf or terminal nodes.
- **Pruning:** Removing the sub-nodes of a parent node is called pruning. A tree is grown through splitting and shrunk through pruning.
- **Branch or Sub-Tree:** A sub-section of decision tree is called branch or a sub-tree, just as a portion of a graph is called a sub-graph.
- **Parent Node and Child Node:** These are relative terms. Any node that falls under another node is a child node or sub-node, and any node which precedes those child nodes is called a parent node

Advantages of Decision Tree Classification

Enlisted below are the various merits of Decision Tree Classification:

1. Decision tree classification does not require any domain knowledge, hence, it is appropriate for the knowledge discovery process.
2. The representation of data in the form of the tree is easily understood by humans and it is intuitive.
3. It can handle multidimensional data.
4. It is a quick process with great accuracy.

Disadvantages of Decision Tree Classification

Given below are the various demerits of Decision Tree Classification:

1. Sometimes decision trees become very complex and these are called overfitted trees.
2. The decision tree algorithm may not be an optimal solution.

3. The decision trees may return a biased solution if some class label dominates it.

Decision Trees are data mining techniques for classification and regression analysis. This technique is now spanning over many areas like medical diagnosis, target marketing, etc. These trees are constructed by following an algorithm such as ID3, CART. These algorithms find different ways to split the data into partitions. It is the most widely known supervised learning technique that is used in machine learning and pattern analysis. The decision trees predict the values of the target variable by building models through learning from the training set provided to the system.

2.6.3 Support Vector Machine

Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. All of these are common tasks in machine learning. We can use them to detect cancerous cells based on millions of images or we can use them to predict future driving routes with a well-fitted regression model. There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC).

The main thing to keep in mind here is that these are just math equations tuned to give you the most accurate answer possible as quickly as possible.

SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

HOW AN SVM WORKS:

A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from.

What makes the linear SVM algorithm better than some of the other algorithms, like k-nearest neighbors, is that it chooses the best line to classify your data points. It chooses the line that separates the data and is the furthest away from the closest data points as possible.

A 2-D example helps to make sense of all the machine learning jargon. Basically you have some data points on a grid. You're trying to separate these data points by the category they should fit in, but you don't want to have any data in the wrong category. That means you're trying to find the line between the two closest points that keeps the other data points separated.

So the two closest data points give you the support vectors you'll use to find that line. That line is called the decision boundary. The decision boundary doesn't have to be a line. It's also referred to as a hyperplane because you can find the decision boundary with any number of features, not just two.

TYPES OF SVM:

There are two different types of SVMs, each used for different things:

- Simple SVM: Typically used for linear regression and classification problems.
- Kernel SVM: Has more flexibility for non-linear data because you can add more features to fit a hyperplane instead of a two-dimensional space.

WHY SVMs ARE USED IN MACHINE LEARNING:

SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. This is one of the reasons we use SVMs in machine learning. It can handle both classification and regression on linear and non-linear data.

Another reason we use SVMs is because they can find complex relationships between your data without we needing to do a lot of transformations on your own. It's a great option when we are working with smaller datasets that have tens to hundreds of thousands of features. They typically find more accurate results when compared to other algorithms because of their ability to handle small, complex datasets.

2.6.4 NAIVE BAYES ALGORITHM:

Naive bayes is probabilistic machine learning algorithm based on the bayes theorem, used in a wide variety of classification tasks.in this article,we will understand the naïve bayes algorithm and all essential concepts so that there is no room for doubts in understanding.

The simplest solutions are usually the most powerful ones, and Naïve Bayes is a good example of that. Despite the advances in Machine Learning in the last years, it has proven to not only be simple but also fast, accurate, and reliable. It has been successfully used for many purposes, but it works particularly well with natural language processing (NLP) problems.

Where the naïve bayes algorithm can be used:

Here are a portion of the regular uses of Naive Bayes for genuine errands:

Archive order. This calculation can assist you with deciding to which classification a given report has a place. It very well may be utilized to characterize messages into various dialects, classes, or subjects (through the presence of catchphrases).

Spam sifting. Credulous Bayes effectively figures out spam utilizing watchwords. For instance, in spam, you can see the word 'viagra' significantly more frequently than in ordinary mail. The calculation should be prepared to perceive such probabilities and, at that point, it can proficiently apply them for spam sifting.

Assumption investigation. In light of what feelings the words in a content express, Naive Bayes can ascertain its likelihood being good or negative. For instance, in client surveys, 'great' or 'economical' normally imply that the client is fulfilled. In any case, Naive Bayes isn't touchy to mockery.

Picture arrangement. For individual and exploration purposes, it is not difficult to fabricate a Naive Bayesian classifier. It tends to be prepared to perceive manually written digits or put pictures into classes through administered AI.

TYPES OF NAÏVE BAYES CLASSIFIERS:

There are several types of naïve bayes.

OPTIMAL NAÏVE BAYES:

This classifier chooses the class that has the greatest a posteriori probability of occurrence (so called maximum a posteriori estimation, or MAP). As follows from the name, it really is optimal but going through all possible options is rather slow and time-consuming.

GAUSSIAN NAÏVE BAYES:

Gaussian Bayes is based on Gaussian, or normal distribution. It significantly speeds up the search and, under some non-strict conditions, the error is only two times higher than in Optimal Bayes (that's good!).

MULTINOMIAL NAÏVE BAYES:

It is usually applied to document classification problems. It bases its decisions on discrete features (integers), for example, on the frequency of the words present in the document.

BERNOULLI NAÏVE BAYES:

Bernoulli is similar to the previous type but the predictors are boolean variables. Therefore, the parameters used to predict the class variable can only have yes or no values, for example, if a word occurs in the text or not.

2.6.5 KNN Algorithm

We are familiar with machine learning and the basic algorithms that are used in the field, then we've probably heard of the k-nearest neighbors algorithm, or KNN. This algorithm is one of the more simple techniques used in machine learning. It is a method preferred by many in the industry because of its ease of use and low calculation time.

Pros:

- Easy to use.
- Quick calculation time.
- Does not make assumptions about the data.

Cons:

- Accuracy depends on the quality of the data.
- Must find an optimal k value (number of nearest neighbors).
- Poor at classifying data points in a boundary where they can be classified one way or another.

WHERE TO USE KNN?

KNN is often used in simple recommendation systems, image recognition technology, and decision-making models. It is the algorithm companies like Netflix or Amazon use in order to recommend different movies to watch or books to buy. Netflix even launched the Netflix

Prize competition, awarding \$1 million to the team that created the most accurate recommendation algorithm.

Conclusion

Now you know the fundamentals of one of the most basic machine learning algorithms. It's a great place to start when first learning to build models based on different data sets. If we have a data set with a lot of different points and accurate information, this is a great place to begin exploring machine learning with KNN.

CHAPTER 3

Research Methodology

3.1 Introduction

In this study, we are going to predict heart disease risk using five machine learning algorithms such as Decision Tree, Gaussian Naive Bayes, Multinomial Naive Bayes, Artificial Neural Network, Random forest and support vector machine due to huge accuracy compared to other algorithms. To implement this algorithm, we first had to collect datasets about the external features of the symptoms of heart disease. By studying the research papers, we have already learned that these five machine learning techniques give more accuracy. Data processing is displayed at an understandable start by converting raw data into an accessible setting for some reason. Data clearing is a process where data is cleared by resolving missing data, copying and data irregularities. Subsequently, the quality of information is improved by bringing information support. Changing information or data from one organization to the next is known as data transformation. It is usually a source configuration is expected to change to the required organization for a specific reason. It is basically the conversion of numbers or alphabetically advanced data in a modified sorting and rearrangement structure temporarily or experimentally. The primary concept of data reduction is to reduce the countless Information system in helpful information. Feature selection is similarly characterized as factor resolution, feature choice or variable subset choice for model development which prevents the subset of the appropriate highlight (variable index) from being sorted. Figure 3.1.1 demonstrates the methodological framework for continuing this study.

3.2 Research Subject and Instrumentation

We have divided this section into two sub-sections so that it becomes obvious to everyone.

3.2.1 Research Subject

Heart disease is a broad term that covers many topics and conditions ranging from abnormal heartbeats and body failure related to the heart. A large number of people suffered from heart disease because of unhealthy daily routines and bad eating disorders such as smoking, eating too much oily foods and not maintaining regular exercise, we have raised a process in this current study that will make people based on their danger of heart attack disease by Machine Learning (ML) classification algorithms be conscious of. We can't surely say it will close heart failure disease but it will abate mortality.

3.2.2 Research Instrumentation

For data anthology, we planned a research paper observed consisting of three general questions and twelve yes / no questions. We have distributed these survey papers to get the prospective outcome. We have created a survey paper for the general public and cardiologists in different hospitals so that the following questionnaires - "Age?", "Gender?", "People type?", "Smoker?" "Diabetes?", "Treatment for high blood pressure?", "Did your mother or father have a heart failure disease earlier on the age of 60?", "Do you take any medications?", "Do you feel any kind of torment?", "Do you feel irritated?", "Are you sweating or feeling lightheaded?", "Do you have asphyxia?", "Do you physical exercise regularly?", "Do you feel any pain in the middle of your chest or thorax or heart?" "Do you feel any pressure, blood pressure, high blood pressure, stiffness, pain or hesitation chest that can spread to your chest, jawbone or shoulders back or shoulders arms or shoulder chest or one or both arms?" "At the same time, the idea is that data raised from the ordinary public and heart attack patients is very lawful and approved.

3.3 Data Collection Procedure

We need to pick up data to finish our experiment study. We raised all database online or virtual and offline. We have pick up our essential knowledge from ordinary public and various individual heart patients hospitals in Dhaka are called offline database collection or information collection. In addition, we created a Google form to collect data from different people by sending a link called virtual or online data collection. We asked them our following questions and in the survey paper we took answers from them.

Accordingly, we have collected all the important information. Some ordinary public were unable to give their data in that case we raise their data from their relatives and their galenical profiles. We raise data from previous research. For this cause, we had to raised

a lot of research papers, journals, websites and reports and then we have gathered this information and raised the data we need to ingredient or implement our concept.

3.4 Statistical Analysis

After all data collection, we explore and method this data in different way. We then convert the dataset to retrieve missing and irregular data. Next, we selected external factors (associated with the heart disease) to apply our selected algorithm through the collected dataset. By distributing the questionnaires, we collected our datasets and the results of all the questionnaires are shown in percentages in Figure 3.4.1.

The designed questionnaire for this study is shown in Table 3.4.

In our thesis, we can find out which factors are more influence people to heart failure and we can also find out what is the solution of heart disease problem in Bangladesh.

The research is hold on the people of Dhaka city and most of the sample of heart attack patients, we have collected from different hospitals in Dhaka city, Bangladesh. And age group of 40 years to 90 years. But we very sorry to say that in corona pandemic situation, we have not able to collect more data for our thesis purpose but we have collected total 300 samples. We have completed our thesis and implement our thesis we have to use python programing language. In this research two types of class. There are Yes or No (O or 1).

Table 3.4: Different Questionnaires of collected or raised dataset

Let,

| | |
|-----------------------------------|--|
| Question 1: Do you have anemia? | Question 11: Time? |
| Question 2: Creatinine? | Question 12: Death event? |
| Question 3: Do you have diabetes? | Question 13: Did your parents have a heart attack disease before age 60? |

| | |
|--|--|
| Question 4: Ejection fraction? | Question 14: Do you receive any kind of drug or medicine? |
| Question 5: Do you take treatment for high blood pressure? | Question 15: Do you perceive any kind of stress or torment or tension? |
| Question 6: Platelets? | Question 16: Do you perceive heartburn with lightheadedness? |
| Question 7: Serum creatinine? | Question 17: Do you have shortness of breath? |
| Question 8: Serum sodium? | Question 18: Do you perceive any pain in the center of your chest? |
| Question 9: Gender? | Question 19: Do you take regular physical exercise? |
| Question 10: Are you smoker? | Question 20: Do you perceive any pressure, torment, tightness or density, pain or squeezing in your chest that may spread to your neck, jaw or shoulder or one or both arms? |

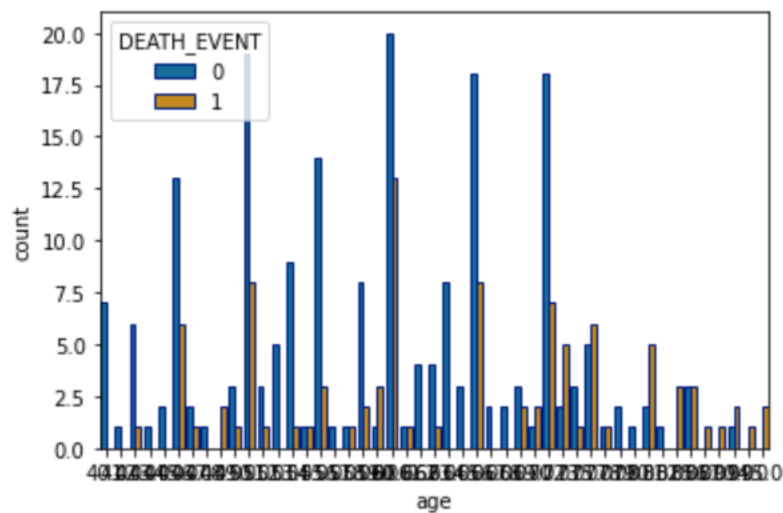


Figure 3.4.1: Statistical view of collected dataset Age group (Death event)

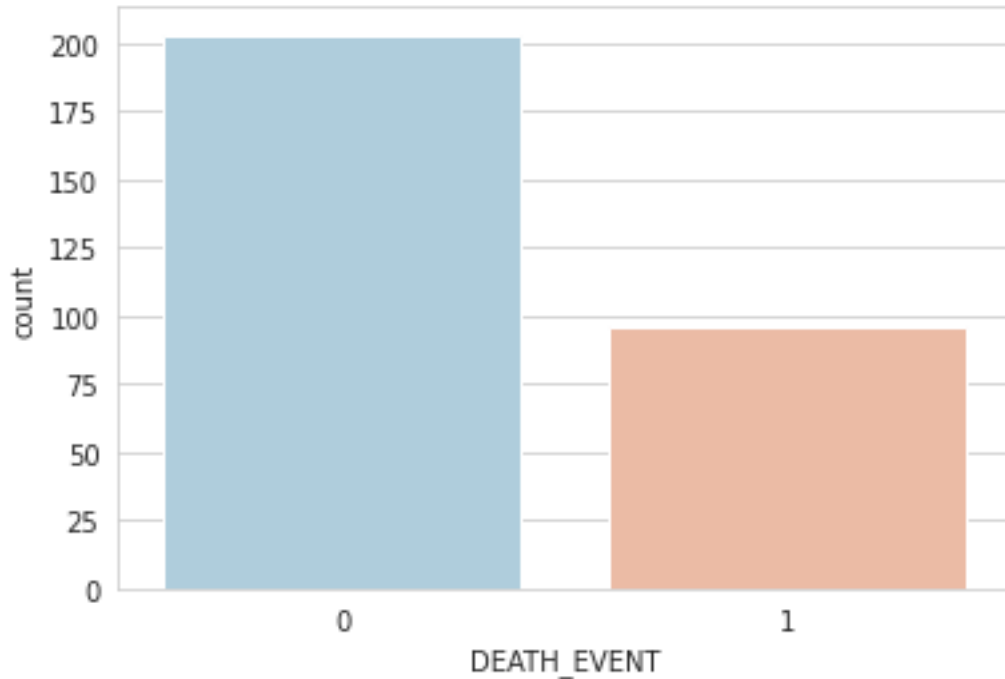


Figure 3.4.2: Statistical view of collected dataset

DEATH_EVENT

3.5 :Train Methodology

In this context, we will discuss how to train. With training, the machine will understand the data I have to enter the learning. So how to train, how to train data, and how to process it is described here. First we loaded our pre-saved dataset model and then tmake a tensorflow model for training.


```

▶ # Define a simple sequential model
def create_model():
    model = tf.keras.models.Sequential([
        keras.layers.Dense(512, activation='relu', input_shape=(784,)),
        keras.layers.Dropout(0.2),
        keras.layers.Dense(10)
    ])

    model.compile(optimizer='adam',
                  loss=tf.losses.SparseCategoricalCrossentropy(from_logits=True),
                  metrics=[tf.metrics.SparseCategoricalAccuracy()])

    return model

# Create a basic model instance
model = create_model()

# Display the model's architecture
model.summary()

```

↳ Model: "sequential_1"

| Layer (type) | Output Shape | Param # |
|-----------------------|--------------|---------|
| dense_2 (Dense) | (None, 512) | 401920 |
| dropout_1 (Dropout) | (None, 512) | 0 |
| dense_3 (Dense) | (None, 10) | 5130 |
| Total params: 407.050 | | |

Fig3.5.1: Model summary

```

Epoch 1/10
32/32 [=====] - 1s 12ms/step - loss: 1.1645 - sparse_categorical_accuracy: 0.6770 - val_loss: 0.7128 - val_sparse_categorical_

Epoch 00001: saving model to training_1/cp.ckpt
Epoch 2/10
32/32 [=====] - 0s 7ms/step - loss: 0.4517 - sparse_categorical_accuracy: 0.8720 - val_loss: 0.5511 - val_sparse_categorical_

Epoch 00002: saving model to training_1/cp.ckpt
Epoch 3/10
32/32 [=====] - 0s 7ms/step - loss: 0.2964 - sparse_categorical_accuracy: 0.9240 - val_loss: 0.4659 - val_sparse_categorical_

Epoch 00003: saving model to training_1/cp.ckpt
Epoch 4/10
32/32 [=====] - 0s 7ms/step - loss: 0.2047 - sparse_categorical_accuracy: 0.9480 - val_loss: 0.4469 - val_sparse_categorical_

Epoch 00004: saving model to training_1/cp.ckpt
Epoch 5/10
32/32 [=====] - 0s 8ms/step - loss: 0.1572 - sparse_categorical_accuracy: 0.9670 - val_loss: 0.4547 - val_sparse_categorical_

Epoch 00005: saving model to training_1/cp.ckpt
Epoch 6/10
32/32 [=====] - 0s 7ms/step - loss: 0.1259 - sparse_categorical_accuracy: 0.9750 - val_loss: 0.4355 - val_sparse_categorical_

Epoch 00006: saving model to training_1/cp.ckpt
Epoch 7/10
32/32 [=====] - 0s 8ms/step - loss: 0.0897 - sparse_categorical_accuracy: 0.9870 - val_loss: 0.4165 - val_sparse_categorical_

Epoch 00007: saving model to training_1/cp.ckpt
Epoch 8/10
32/32 [=====] - 0s 7ms/step - loss: 0.0698 - sparse_categorical_accuracy: 0.9880 - val_loss: 0.4274 - val_sparse_categorical_

Epoch 00008: saving model to training_1/cp.ckpt
Epoch 9/10

```

Fig3.5.2: Model accuracy

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

The writer created a strategy in this article to create frequent idea based on user different symptoms. It helps ordinary public to know the danger level of heart failure disease from the outer. The outcome will help doctors forecast danger levels for heart attack patients. We used three Machine Learning (ML) algorithms of which our decision tree classifier algorithm and random forest classifier algorithm was the great performers to get high accuracy.

4.2 Experimental Results

The primary goal of our forecast process is to predict heart failure disease danger. Nowadays various machine learning methods and data mining techniques practice it natural to forecast the level of danger of heart attack disease. To apply these data mining technique methods we necessity to acquire the database and then we need to system this data information very nearly. A total of 14 features have been collected. The salient features or symptoms for this study are age, gender, stress, diabetes, ejection fraction, high blood pressure, serum creatinine, platelets, serum sodium, smoking, family history of heart disease, substance use or non-use, regular exercise or not, chest pain or not, shortness of breath, etc. Then the danger of heart disease Machine Learning (ML) algorithms to estimate the level. After implementing four machine learning classification algorithms, we have achieved a unique level of accuracy.

Table 4.2: Performance research study of Machine Learning (ML) Algorithms

| Algorithms used | Accuracy |
|-----------------------|--------------------|
| Logistic Regression | 0.8 |
| Decision Tree | 0.7733333333333333 |
| Random Forest | 0.83 |
| SVM | 0.84 |
| Gaussian Naive Bayes | 0.76 |
| Bernoulli Naïve Bayes | 0.6533333333333333 |

| | |
|-------------------------|-----|
| Multinomial Naive Bayes | 0.6 |
|-------------------------|-----|

It is viewed from the above Table and Graph that Random Forest provides the greatest precision of 0.84%. We introduced our scheme with this algorithm because of its elevated precision.

4.3 Descriptive Analysis

Heart failure disease is the leading cause of death in both men and women. Analyzing our dataset, it has been viewed that men are more prone to heart disease as shown in Figure 4.3.1.

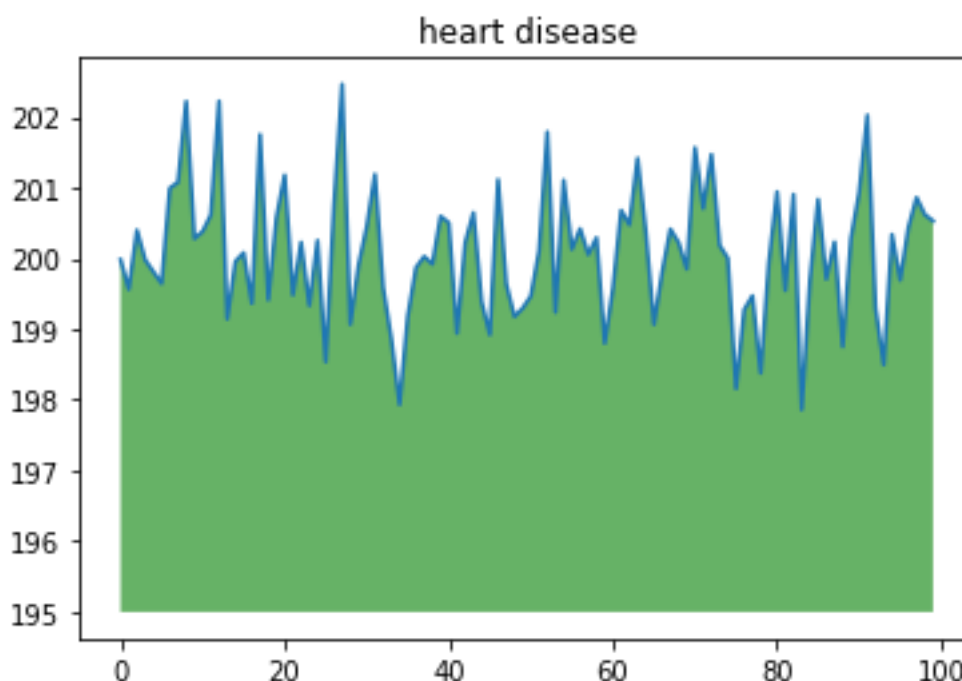


Figure 4.3.1: Statistical view of collected dataset normal visualization

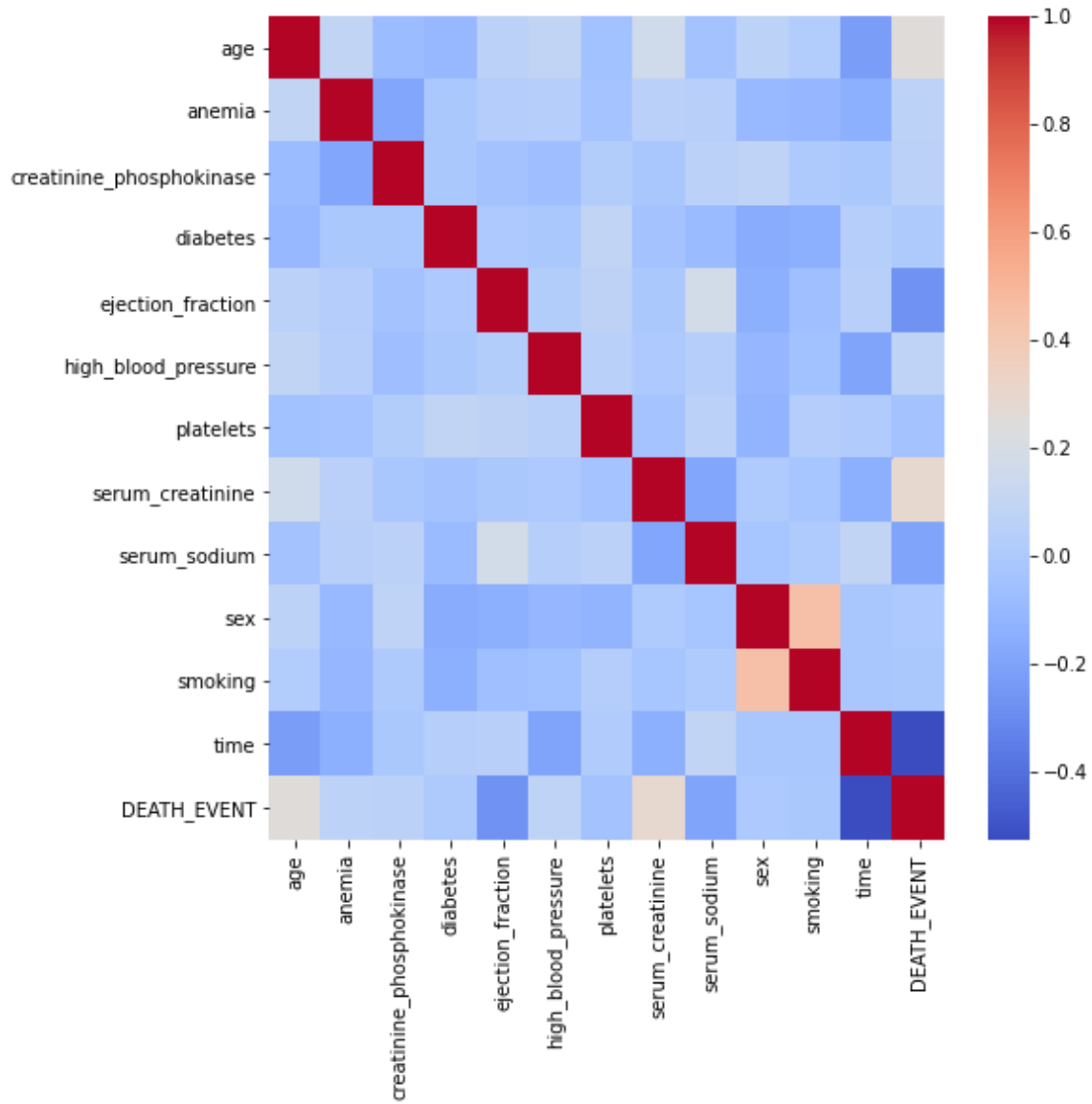
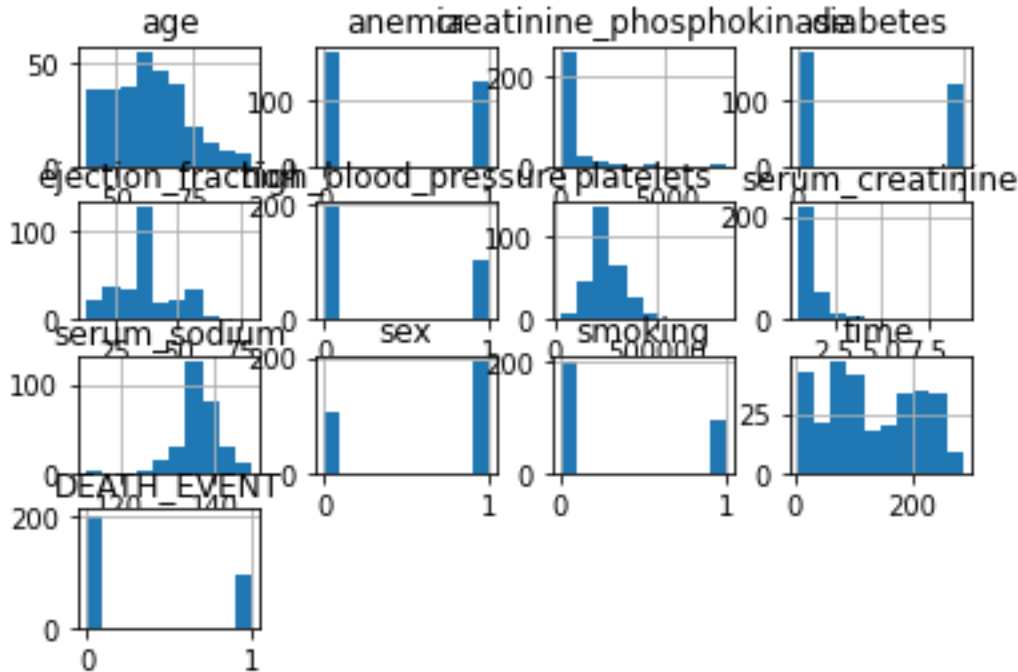


Figure 4.3.2: Statistical view of collected dataset visualization



Which further proves that the percentage of men with heart attack disease is higher than the percentage of women heart attack disease. A large numbers of ordinary public are dying every day from various diseases where heart disease has an effect. The primary cause of heart disease is living an unhealthy lifestyle like overeating, not avoiding high cholesterol oily foods, not taking regular physical exercise, everyday do smoking etc. So in this research we have proposed a system or model that can help people understand their danger of heart failure disease.

4.4 Summary

The primary objective of this research project is to make them aware of their daily life routine by understanding the danger level of heart failure disease and this will partially abate the mortality rate. Our research project unable to provide any form of treatment therapy. Suppose that someone eats a lot of oily food every day and has no practice or physical exercise of what to do. If he / she uses our project, he / she will be able to understand the danger of heart attack disease so that he becomes aware of his heart health and take regular exercise continue which will protect people from heart disease. To apply the machine learning algorithm, data were collected from different hospitals and generally written down. Another method of data collection was from summarizing the release of individual patients. Thus, an all-out 13 component of about 300 data was collected. This collected data was then sorted and arranged efficiently in an excel position. Using this information, it comes in contact with various machine learning algorithms. From the data set, fifteen features are disabled, for this example, age, gender, blood pressure and

©Daffodil International University 27

estimation of smokers and the likelihood of heart disease in such patients. These credits survive on Support Vector Machine Algorithm, Decision Tree Algorithm and Non investive Ventilation Algorithm, Random Forest Algorithm are bases group algorithms in which Random Forest and SVM gives the best outcome with the highest accuracy as viewed in Table 4.2.

Our proposed algorithms are considered better than the accuracy of other algorithms in the literature review. From past research work and Table 1, it has become clear that the achievement rate of any of the research studies for the reference algorithm of the involved heart disease dataset is not more than 86%. In light of the examination of the results, it was seen that the proposed models provided intelligent results in managing conceivable coronary disease patients. Implementing the proposed strategy to maintain the efficiency of the proposed method is the opposite of the algorithm.

CHAPTER 5

Implementation (Results and

Analysis)

5.1 Introduction

We have developed a simple strategy for predicting heart disease risk using smartphones. An Android-based prototype software was created specifically to include information on heart patients and individuals. Data from 300 heart patients and the general population have been analyzed with hypertension, diabetes, smoking, family history, hypertension, stress and danger variables. Clinical symptoms may suggest underlying heart disease. Machine learning technology has been used to gather data and predictive scores have been created.

5.2 Results and Analysis

In our before the studies we have discussed various algorithms, previous functions this field and the dataset we used for our trial research [28]. This was the basis for all this chapter. In this research studies, we consider and discuss the outcome we got after implementing the results algorithms and their analysis.

```

▶ plt.plot([k for k in range(1,21)], knn_scores, color = 'green')
  for i in range(1,21):
    plt.text(i,knn_scores[i-1], (i,knn_scores[i-1]))
    plt.xticks([i for i in range(1,21)])
    plt.xlabel('Number of Neighbors (K)')
    plt.ylabel('Scores')
    plt.title('K Neighbors Classifier scores for different k value')

```

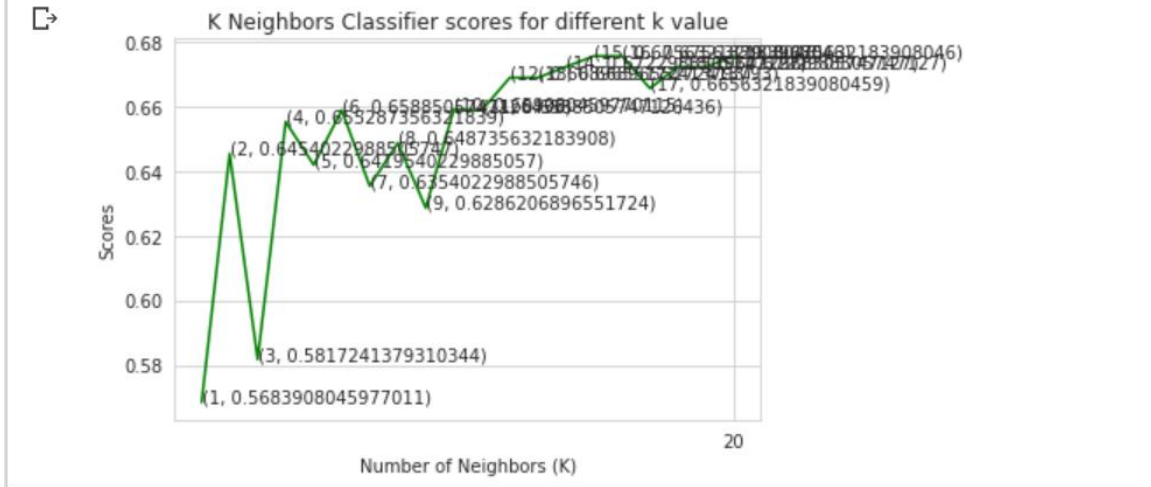


Fig5.2.1: Results of KNN Algorithm

```

▶ # Plotting attrition of employees
  fig, (ax1, ax2) = plt.subplots(nrows=1, ncols=2, sharey=False, figsize=(10,4))

  ax1 = df['DEATH_EVENT'].value_counts().plot.pie( x="Heart disease", y = 'no. of patients',
    autopct = "%1.0f%", labels=["Heart Disease","Normal"], startangle = 60,ax=ax1);
  ax1.set(title = 'Percentage of Heart disease patients in Dataset')

  ax2 = df["DEATH_EVENT"].value_counts().plot(kind="barh" ,ax =ax2)
  for i,j in enumerate(df["DEATH_EVENT"].value_counts().values):
    ax2.text(.5,i,j,fontSize=12)
  ax2.set(title = 'No. of Heart disease patients in Dataset')
  plt.show()

```

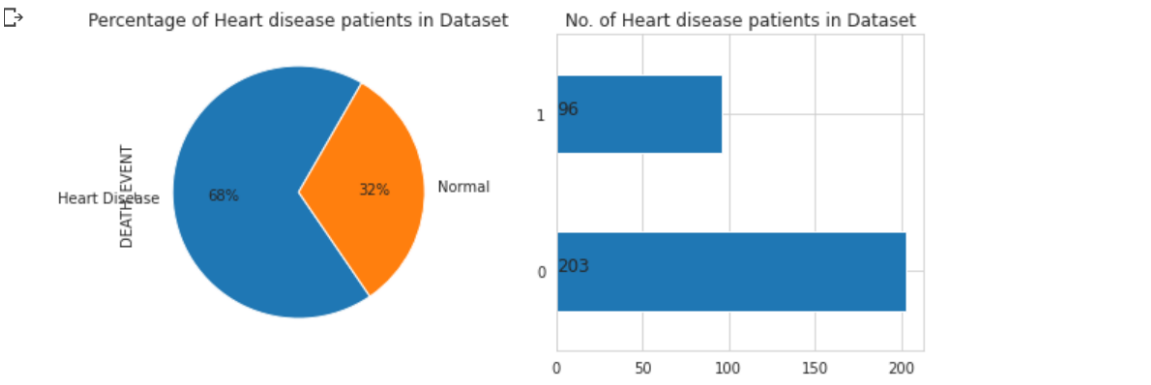


Fig5.2.2: Percentage of heart disease patients in dataset

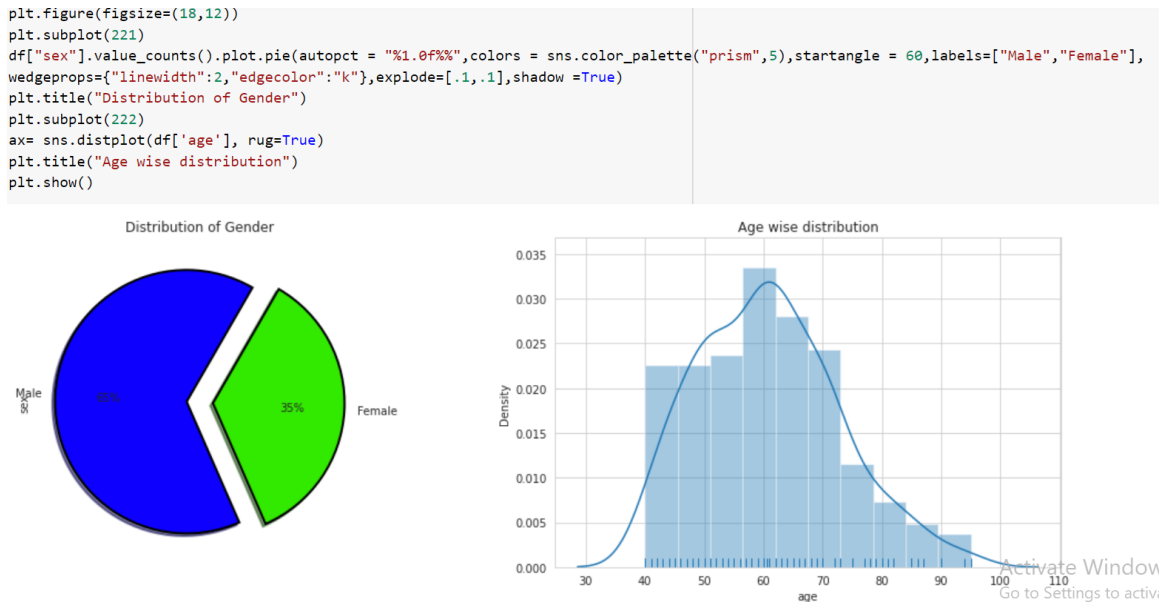
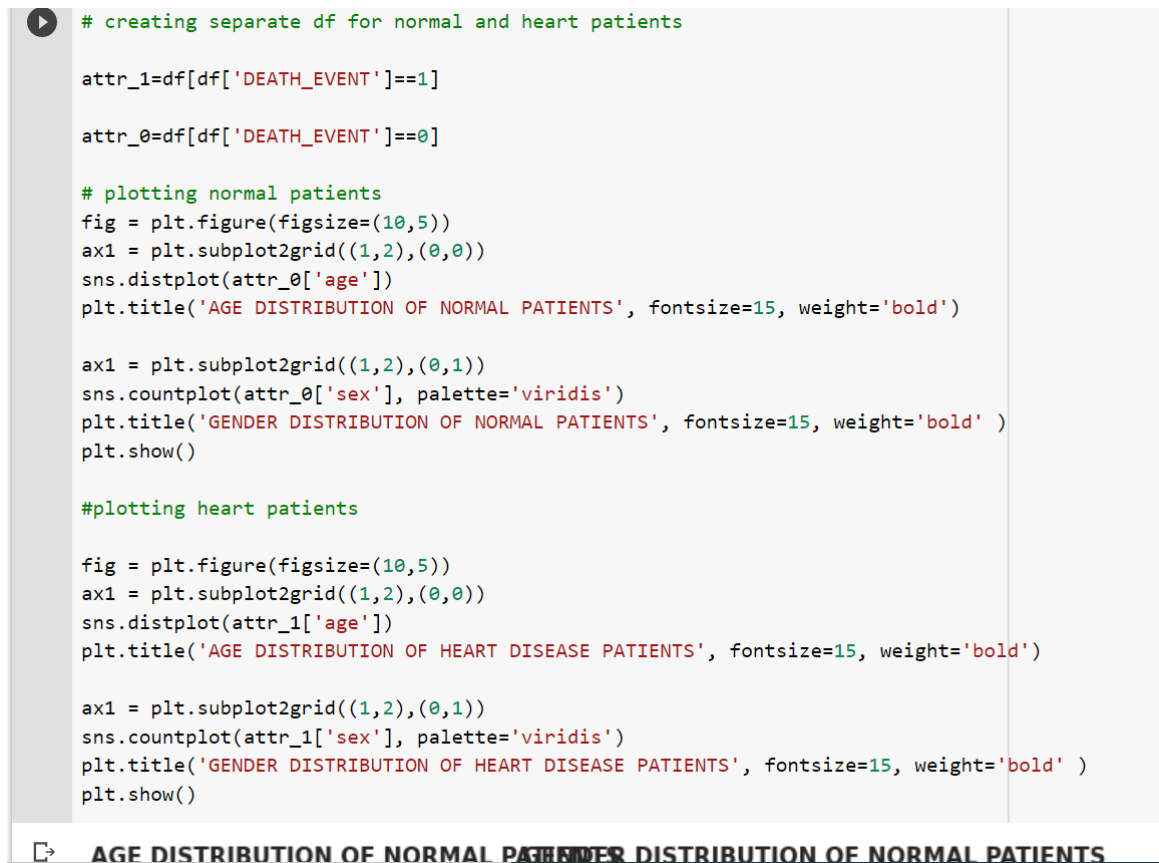


Fig5.2.3: Distribution of gender of heart disease dataset



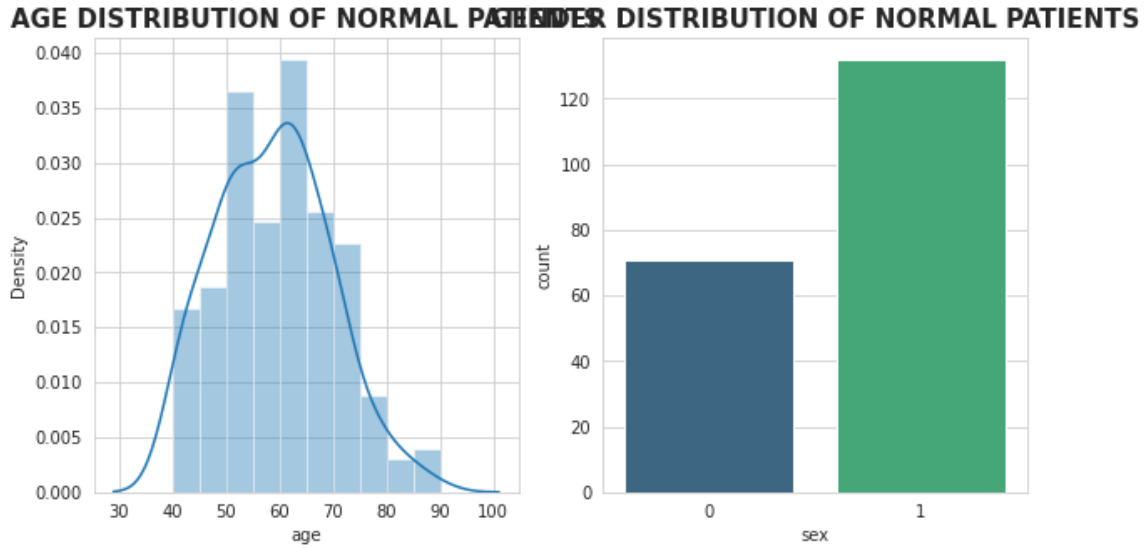


Fig5.2.4: Age distribution of normal and heart patients

5.3 Accuracy of Models with All Features

The outcome have been obtained by employing various classifications of algorithms. At first, we use the whole data set, including all features and application Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, Gaussian Naive Bayes, Bernoulli Naïve Bayes, Multinomial Naive Bayes. Table 5.1 contains the validity score or accuracy various algorithms that we have applied practical to our dataset. Figure 5.1 shows the graphical representation Table [28].

| Classifier of Algorithm | Score or Accuracy |
|-------------------------|--------------------|
| Logistic Regression | 0.8 |
| Decision Tree | 0.78 |
| Random Forest | 0.8333333333333334 |
| SVM | 0.8466666666666667 |
| Gaussian Naive Bayes | 0.7666666666666667 |
| Bernoulli Naïve Bayes | 0.6533333333333333 |
| Multinomial Naive Bayes | 0.6 |

Fig 5.1 Accuracy or outcome score of the algorithms

5.4 Feature Engineering

A good number of system can influenced the score or accuracy of the algorithm. So it is very important to work with features. There are a number of cause why some people may want to work with someone selected properties. Like fewer features support us train faster. We can use react as new features in them by sorting out the most important features. Sometimes it gives amazing development [29]. But a few features are linearly dependent concerned to others. This can put a embrace on the system. Characteristic selection means selecting only the features that are necessary to develop the score of the algorithm. It abate the training period and assuage the weight.

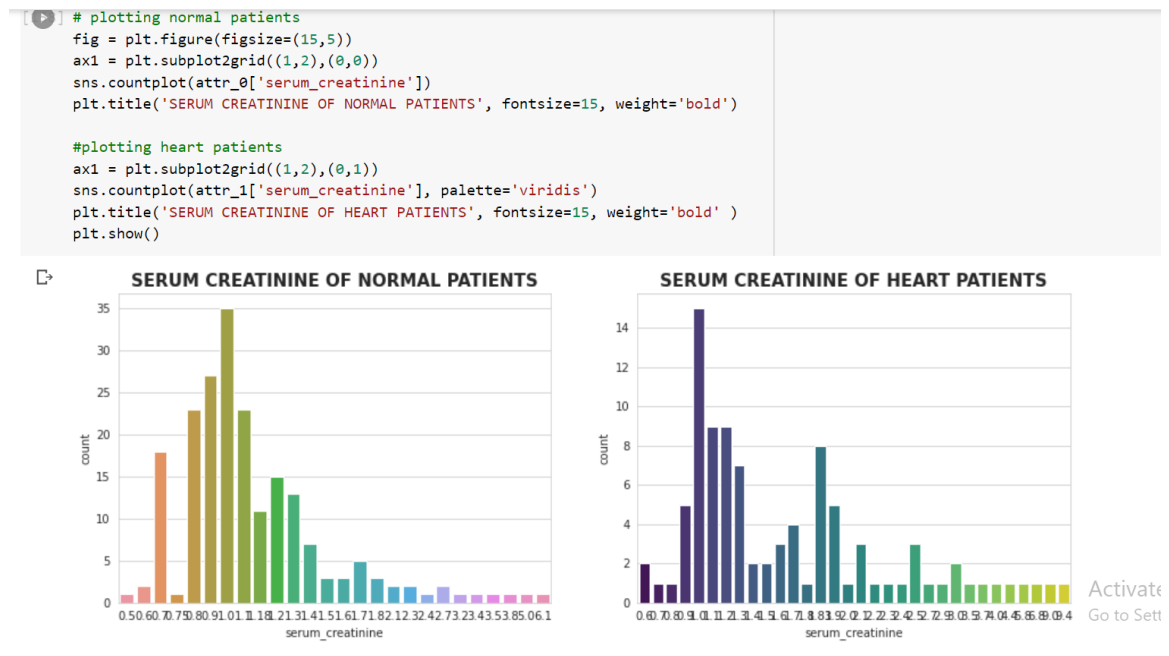


Fig5.4.1: Seraum_creatinine for the dataset of normal & heart patients.

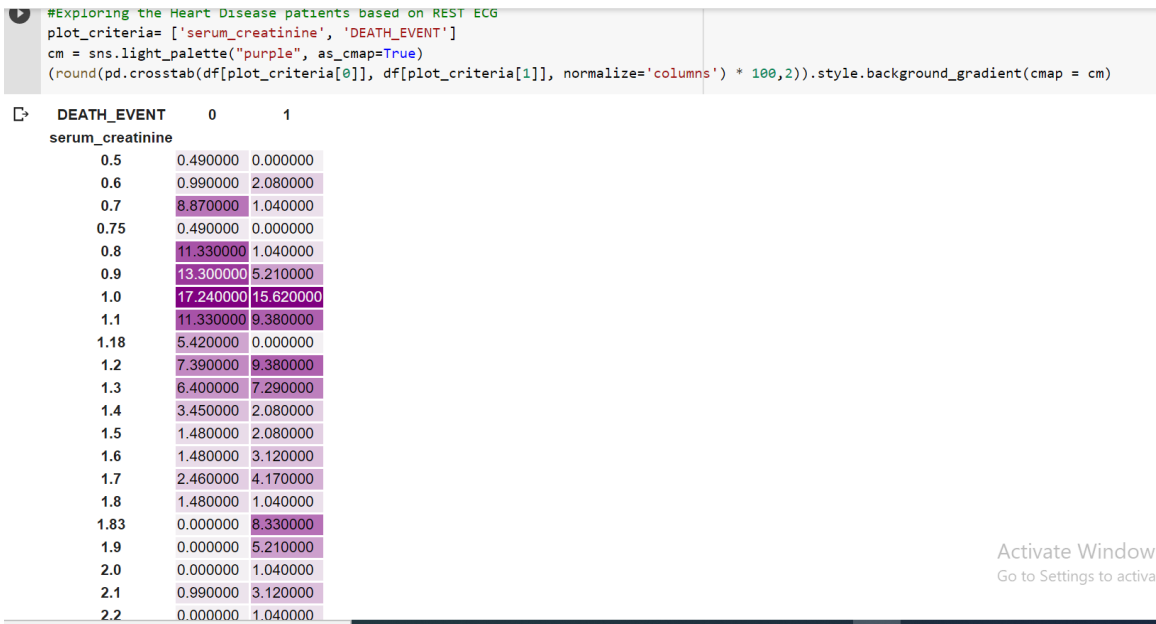


Fig5.4.2: DEATH_EVENT serum creatinine



Fig5.4.3: St slope normal and heart patients

```
[161] #Exploring the Heart Disease patients based on ST Slope
plot_criteria= ['diabetes', 'DEATH_EVENT']
cm = sns.light_palette("pink", as_cmap=True)
(round(pd.crosstab(df[plot_criteria[0]], df[plot_criteria[1]], normalize='columns') * 100,2)).style.background_gradient(cmap = cm)
```

| DEATH_EVENT | 0 | 1 |
|-------------|-----------|-----------|
| diabetes | | |
| 0 | 58.130000 | 58.330000 |
| 1 | 41.870000 | 41.670000 |

Fig5.4.4: DEATH_EVENT diabetes

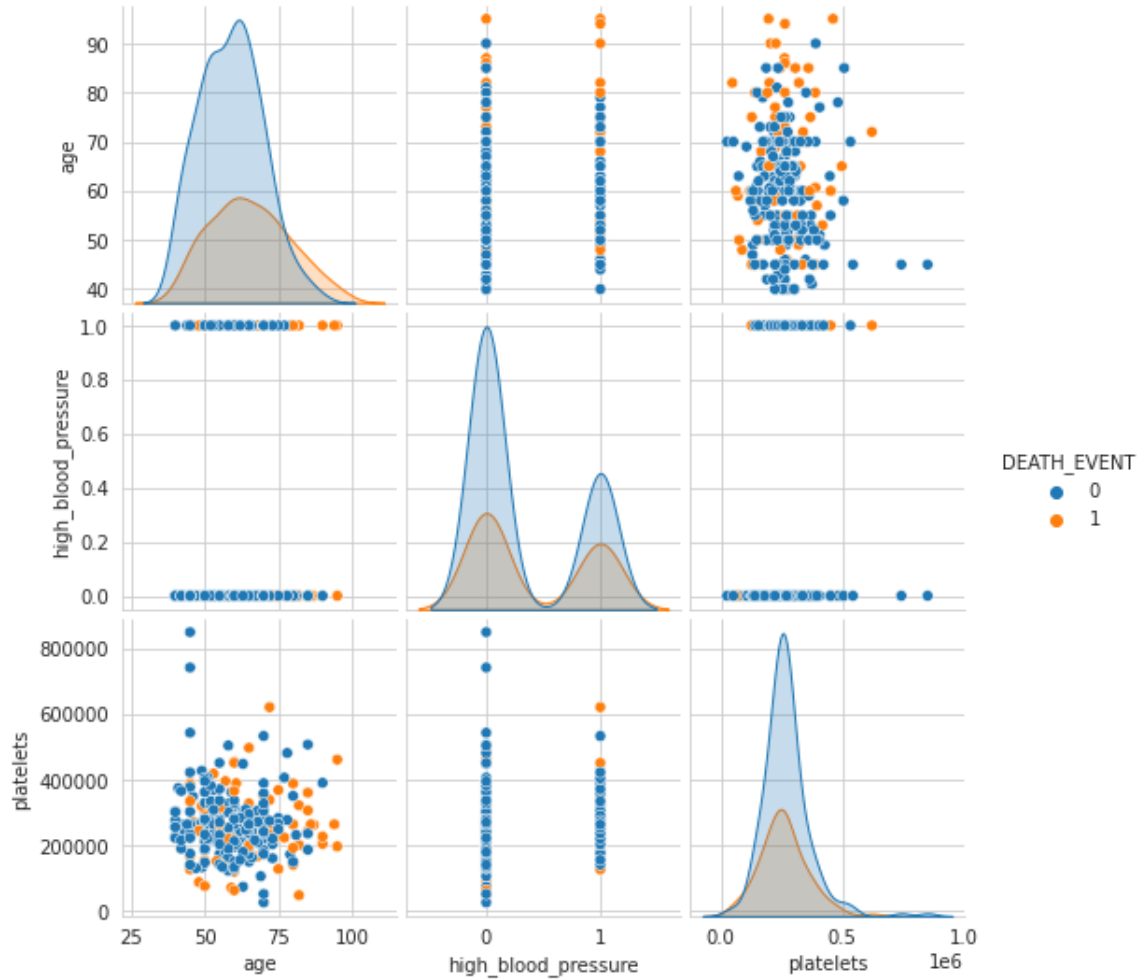


Fig5.4.5: high blood pressure

```
[163] sns.scatterplot(x = 'high_blood_pressure', y = 'platelets', hue = 'DEATH_EVENT', data = df)
```

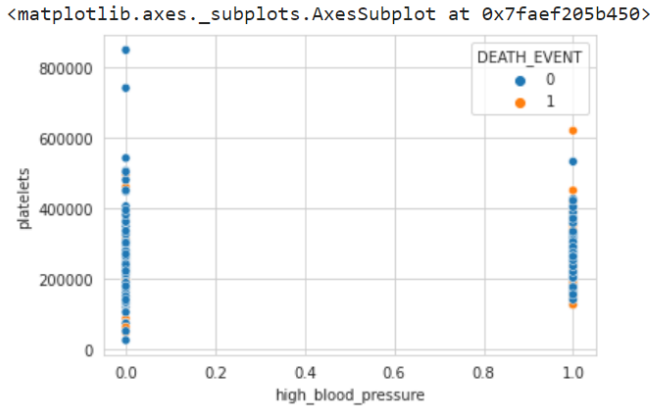


Fig5.4.6: DEATH_EVENT high blood pressure

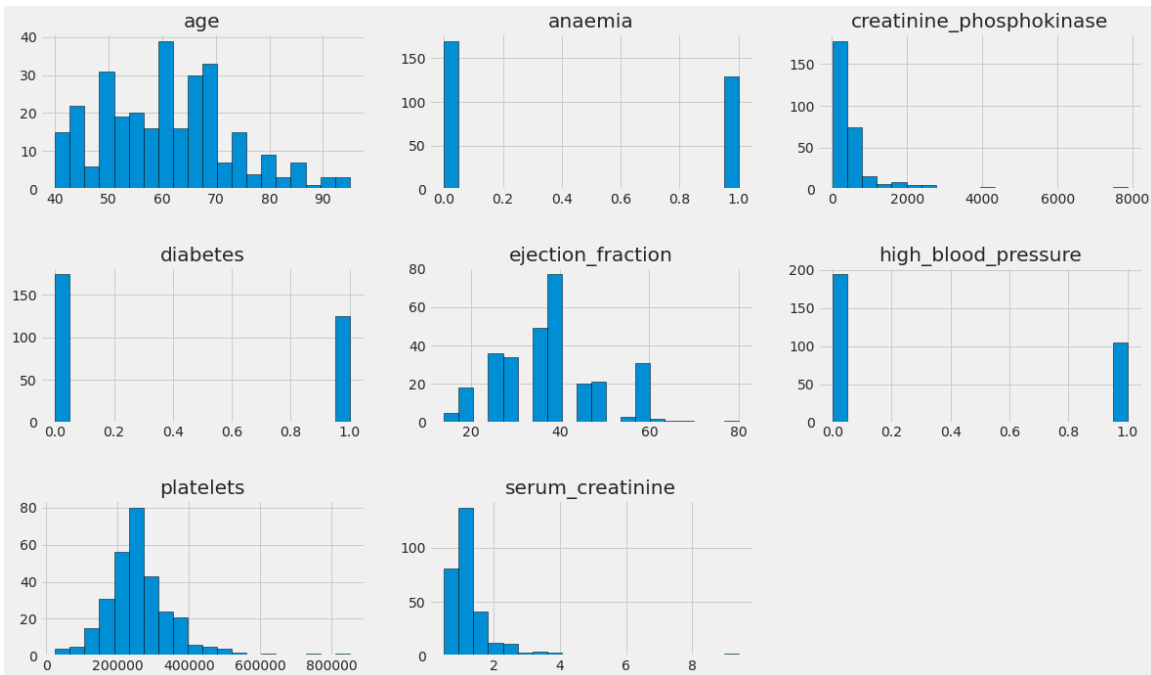


Fig5.4.7: Heart patients

5.5 Feature Importance

One of the most fundamental query we can query a model, what is the great impact on the forecast of characteristic? That the idea is named the magnitude of the characteristic or feature. There may be something in the dataset system or features that do not affect the forecast too much [30]. In many cases, a few characteristic or features can abate the score

or accuracy level of a system. So, this is significant to action with the right qualities. And until now we have main action on all the characteristic of the dataset and [5.1] charted the validity score of different system models. So, right now we want to view the suitable switch or variation of accuracy in the various classifications after that picked a subset of characteristic or multiplication.

5.6 Accuracy of Models with Selected Features

Then view the magnitude of the classification in Table 4.2 and Table 4.3, we elected the following features

And view the difference in forecasts [31]

4.2 Table 4.4 view the score variation then elected a characteristic. And the figure view visual statement of variation in score or accuracy [31].

5.7 Cross Validation

CV that means cross validation is a necessary part in system model training. This inform us if our system is at high danger of excessive advantage. Most of the struggle time, open LB accuracy not too much dependable. Once again when we develop the system and got good local CV accuracy, LB accuracy get bad. It is thoroughly trust that we should believe our CV accuracy or scores in this conditions. Overall we view the CV accuracy acquired by various methods for synchronizing with one other and with LB accuracy, but it is not ever feasible [32].

| Classifier of Algorithm | New Score or Accuracy (%) | Previous Score or Accuracy (%) | Accuracy develop or increase (%) |
|-------------------------|---------------------------|--------------------------------|----------------------------------|
| Logistic Regression | 0.8 | 0.8208333333333333 | 0.0208333333333333 |
| Decision Tree | 0.7733333333333333 | 0.7958333333333333 | 0.0225 |
| Random Forest | 0.83 | 0.7666666666666667 | 0 |
| SVM | 0.84 | 0.7541666666666667 | 0 |
| Gaussian Naive Bayes | 0.76 | 0.725 | 0 |
| Bernoulli Naive Bayes | 0.6533333333333333 | 0.6625 | 0.0091666666666666 |

| | | | |
|-------------------------|-----|--------------------|----------------|
| Multinomial Naive Bayes | 0.6 | 0.6166666666666667 | 0.016666666666 |
|-------------------------|-----|--------------------|----------------|

5.8 Analysis

Since the over table, the individual algorithms carry through the good rely on if cv and attribute election is use to helpful or not. Each algorithm has an internal ability to great work several algorithms rely on condition. For example, LR, MNB and DT data good work with a massive amount of datasets than when it is short. Although the BNB work good with a little amount of the datasets. So, RF, SVM, GNB grade default work a significant induction. In spite of that then doing the math it cannot give any outcome as it can give a best dataset. MNB are best feature on dataset [32]. DT, LR, BNB, MNB are best feature on dataset. Whether uncommitted among the feature of the dataset, this would pay the low score.

Chapter 6

Results and Analysis

6.1 Conclusion

In our reconstruction study, we have effort to balance various ML algorithms and forecast whether cardio illness will occur if a particular man, given different personal attribute and prefix. The principle purpose of our research was to balance or compare with score and explore the cause of behind the change of various algorithms. We used the Cleveland dataset for heart disease with 300 instances and used 10 part cross-verification to divide the data into two sections that are training and testing the datasets. [33] We considered 13 features and seven separate algorithms for accuracy analysis applied or implemented. At last of the implementation part, we found that GNB and RF are paying the maximum level of score accuracy in our dataset which is 91.21 percent and DT is representing of the poor level of score accuracy which is 84.62 percent. Perhaps other algorithms may be perform good for other various examples and other datasets but in our case we have got this result [34]. Overall, if we develop the features we can search many proper and perfect outcome but it will take many period and take other opportunity to procedure and the model process will be gradual than it is right now because it will be the many elaborate will represent many data. So, envisage of these potential issues we made a judgement and conclusion that is best for us to active labor on.

6.2 Future Scope

In our dataset used in our research or thesis is very short and primitive or ancient. Moreover, no new dataset related to cardio illness or heart attack disease has been launched so far. Recent modern datasets are needed and we can pick up it from different kind of hospitals in our country Bangladesh [35]. We can also appreciate or value the skills of every separate classifier and the abbreviation of such national classifications or attribute as bagging recruitment, boosting and stacking strategies.

References:

- [1] Gold, J. C. [1] Arnold, C. (1990). Heart disease. Franklin Watts. and Cutler, D. J. (2000). Heart disease. Enslow Publishers.
- [2] Yanwei X, Wang J, Zhao Z, Gao Y. Combination data mining models with new medical data to predict outcome of coronary heart disease. Proceedings International Conference on Convergence Information Technology; 2007, pp. 868–72.
- [3] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 704-706.
- [4] World Health Organization (WHO), 2017. Cardiovascular diseases (CVDs) – Key Facts. [http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)), [Accessed on 19 July 2019].
- [5] World Health Organization. The world health report 2000: health systems: improving performance. World Health Organization, 2000.
- [6] ARCHANA, BADE, AHER DIPALI, and SMITA KULKARNI PROF. "International Journal On Recent and Innovation Trends In Computing and Communication." pp. 2277-4804.
- [7] The heart foundation <http://www.theheartfoundation.org/heart-disease-facts/heart-disease-statistics/> [Accessed on 19 July 2019].
- [8] Srinivas, K., Rani, B.K., Govrdhan, A., 2010a. Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng. (IJCSE), 2010, Vol. 2, No. 2, pp. 250–255
- [9] Silwattananusarn, Tipawan, and Kulhida Tuamsuk. "Data mining and its applications for knowledge management: A literature review from 2007 to 2012." 2012.
- [10] Leventhal, Barry. "An introduction to data mining and other techniques for advanced analytics." Journal of Direct, Data and Digital Marketing Practice 2010, Vol. 12, No. 2, pp. 137-153.
- [11] Leventhal, Barry. "An introduction to data mining and other techniques for advanced analytics." Journal of Direct, Data and Digital Marketing Practice, 2010, Vol. 12, No. 2, pp. 137-153
- [12].<https://health.economicstimes.indiatimes.com/news/diagnostics/a-heartattack-no-it-was-coronavi>
- [13] Mamatha Alex P and Shaicy P Shaji, "Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique " International Conference on Communication and Signal Processing, April 4-6, 2019, India.
- [14] Babu, Sarath, "Heart disease diagnosis using data mining technique." Electronics Communication and Aerospace Technology (ICECA), 2017 Internationalconference of. Vol. 1. IEEE, 2017.

- [15] Banu, MA Nishara, and B. Gomathy. "Disease forecasting system using data mining methods." Intelligent Computing Applications (ICICA), 2014 International Conference on. IEEE, 2014.
- [16] Krishnaiah, V., "Diagnosis of heart disease patients using fuzzy classification technique." Computer and Communications Technologies (ICCCT), 2014 International Conference on. IEEE, 2014.
- [17] Shamshad Rahman Lubna, "Predicting Coronary Heart Disease through Risk Factor Categories", ASEE Conference, 2014.
- [18] Sudha A, Gayathiri P, Jaisankar N. Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications. 2012, Vol. 43, No. 14, pp. 26–31.
- [19] Ishtake, S.H., Sanap, S.A.: Intelligent heart disease prediction system using data mining techniques. Int. J. Healthc. Biomed. Res, 2013, Vol. 1, No. 3, pp. 94–101.
- [20] Goetz, T. (2010). The decision tree. Rodale.
- [21] Sudhakar, K., Manimekalai, M.: Study of heart disease prediction using data mining. Int. J. Adv. Res. Comput. Sci. Softw. Eng.
- [22] Alagugowri S, Christopher T. Enhanced Heart Disease Analysis and Prediction System [EHDAPS] Using Data Mining. International Journal of Emerging Trends in Science and Technology 2014, No. 1, pp. 1555-1560.
- [23] P.K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", Journal of King Saudi University-Computer and Information sciences, 2012, pp. 27-40.
- [24] L. Burke, J. Ma, K. Azar, G. Bennett, E. Peterson, Y. Zheng, W. Riley, J. Stephens, S. Shah, B. Suffoletto, T. Turan, B. Spring, J. Steinberger and C. Quinn, "Current Science on Consumer Use of Mobile Health for Cardiovascular Disease Prevention", Circulation, vol. 132, No. 12, pp. 1157-1213, 2015.
- [25] AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin, "HDPS: Heart Disease Prediction System" Department of Medical Informatics, Tzu Chi University, Hualien City, Taiwan.
- [26] Taneja, A.: Heart disease prediction system using data mining techniques. Orient. J. Comput. Sci. Technol.
- [27] Srinivas, K., Kavihta Rani, B., Govrdhan, A.: Applications of data mining techniques in healthcare and prediction of heart attacks. (IJCSE) Int. J. Comput. Sci. Eng, 2010, Vol. 2, No. 2, pp. 250–255.
- [28] Tiger, S. and Reingold, M. (1986). Heart disease. J. Messner.
- [29] Healey, J. (2005). Heart disease. Spinney Press.
- [30] Arnold, C. (1990). Heart disease. Franklin Watts. se. Spinney Press [28.37] Tiger, S. and Reingold, M. (1986). Heart disease. J. Messner.
- [31] Canfield, J., Hansen, M. V., and Rackner, V. (2005). Heart disease. Health Communications.
- [32] Dittmer, L. (2012). Heart disease. Creative Education.
- [34] Gold, J. C. and Cutler, D. J. (2000). Heart disease. Enslow Publishers.

APPENDIX

Heart Failure Prediction

| | | | | | | | |
|--------------|----------------------|--------|----------------------|---------------|----------------------|-----|----------------------|
| Age: | <input type="text"/> | Gender | <input type="text"/> | Mar | <input type="text"/> | Fem | <input type="text"/> |
| People type: | <input type="text"/> | Normal | <input type="text"/> | Heart patient | | | |

| | |
|-----------------------------------|--|
| Question 1: Do you have anemia? | <input type="text"/> Yes <input type="text"/> No |
| Question 2: Creatinine? | <input type="text"/> Yes <input type="text"/> No |
| Question 3: Do you have diabetes? | <input type="text"/> Yes <input type="text"/> No |
| Question 4: Ejection fraction? | <input type="text"/> Yes <input type="text"/> No |
| Question 6: Platelets? | <input type="text"/> Yes <input type="text"/> No |
| Question 7: Serum creatinine? | <input type="text"/> Yes <input type="text"/> No |
| Question 8: Serum sodium? | <input type="text"/> Yes <input type="text"/> No |
| Question 9: Gender? | <input type="text"/> Yes <input type="text"/> No |
| Question 11: Time? | <input type="text"/> Yes <input type="text"/> No |
| Question 12: Death event? | <input type="text"/> Yes <input type="text"/> No |

| | |
|--|--|
| Question 13: Did your parents have a heart attack disease before age 60? | <input type="checkbox"/> Yes <input type="checkbox"/> No |
| Question 14: Do you receive any kind of drug or medicine? | <input type="checkbox"/> Yes <input type="checkbox"/> No |
| Question 16: Do you perceive heartburn with lightheadedness? | <input type="checkbox"/> Yes <input type="checkbox"/> No |
| Question 17: Do you have shortness of breath? | <input type="checkbox"/> Yes <input type="checkbox"/> No |
| Question 18: Do you perceive any pain in the center of your chest? | <input type="checkbox"/> Yes <input type="checkbox"/> No |
| Question 19: Do you take regular physical exercise? | <input type="checkbox"/> Yes <input type="checkbox"/> No |
| Question 20: Do you perceive any pressure, torment, tightness or density, pain or squeezing in your chest that may spread to your neck, jaw or shoulder or one or both arms? | <input type="checkbox"/> Ye <input type="checkbox"/> No |

Heart Failure Prediction using Machine Learning Algorithm

ORIGINALITY REPORT

23%

SIMILARITY INDEX

%

INTERNET SOURCES

%

PUBLICATIONS

23%

STUDENT PAPERS

PRIMARY SOURCES

| | | |
|---|---|----|
| 1 | Submitted to University of West Florida Student Paper | 6% |
| 2 | Submitted to Ain Shams University Student Paper | 2% |
| 3 | Submitted to CSU, San Jose State University Student Paper | 2% |
| 4 | Submitted to Daffodil International University Student Paper | 2% |
| 5 | Submitted to Nottingham Trent University Student Paper | 2% |
| 6 | Submitted to Delhi Technological University Student Paper | 1% |
| 7 | Submitted to Middlesex University Student Paper | 1% |
| 8 | Submitted to Gitam University Student Paper | 1% |
| 9 | Submitted to Chester College of Higher Education Student Paper | 1% |