HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

A PROJECT REPORT

Submitted by

Abhay Kumar [Reg No: 42310001]

Under the guidance of

DR. Anto Praveena M.D

(Assistant Professor, Department of Computer Science and Engineering)

In partial fulfillment of the Requirements for the Degree

Of

MASTER OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

A logo of a university

Description automatically generated

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING AND TECHNOLOGY

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

JEPPIAAR NAGAR, SH 49A, CHENNAI, TAMIL NADU 600119

APRIL- 2024

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

JEPPIAAR NAGAR, SH 49A, CHENNAI, TAMIL NADU 600119

BONAFIDE CERTIFICATE

Certified that this project report titled “HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS” is the Bonafide work of “ABHAY KKUMAR [Reg No: 42310001], who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

DR. ANTO PRAVEENA M. D DR. L. LAKSHMANAN

GUIDE HEAD OF THE DEPARTMENT

Assistant Professor HOD

Dept. of Computer Science and Engineering Dept. OF CSE

Signature of Internal Examiner Signature of External Examiner

ABSTRACT

* The application of Machine Learning is highly successful in fields like business, retail and marketing and has led to a vast spread in its application in various other industries. Among these sectors is the healthcare industry. This industry is known to be “information rich” and has a great scope for effective decision making & discovering hidden patterns. Advanced machine learning techniques can help us with this situation. Talking about the healthcare industry, one of the most talked about is heart diseases. The heart is one of the most important parts of the human body. It helps to purify and circulate blood to all parts of the human body. Most millions of number deaths in the world are due to heart diseases. Some symptoms like chest pain, faster heartbeat, discomfort in breathing, are recorded in medical history. This all data will be analyzed on a regular basis for heart disease. In this Research review an overview of heart disease and its current modern procedures is firstly introduced. An in-depth analysis of the most relevant OF machine learning techniques Algorithms available in the literature for heart disease prediction System is briefly elaborated. The discussed machine learning algorithms are Decision Tree, SVM, ANN, Naive Bayes, Random Forest, KNN.DNN XG-BOOST, ADA-BOOST. The algorithms are compared based on features. We are working on the algorithm with best accuracy Results. This will be helping the doctors to assist the heart patient's problem easily.

ACKNOWLEDGEMENTS

The success and the final outcome of this project required guidance and assistance from different sources, and we feel extremely fortunate to have got this all through the completion of our project. Whatever we have done is largely due to such guidance and assistance and we should not forget to thank them. We would like to thank the Director (Engineering & Technology), Dr.Mariazeena\_Johnson for her constant support and facilities in the department of Computer science And Engineering for completing the project towards M.E degree. Secondly we express our sincere thanks to Dr.L.LAKSHMAN Professor and the Head of the Department CSE for all the help and infrastructure provided to us to complete this project successfully. I am owe our profound gratitude to our project guide DR. ANTO PRAVEENA M.D, Assistant Professor, who took a keen interest in our project work and guided us all along, till the completion of our project. We are thankful to all the teaching and non-teaching staff of our department for helping us to bring this project to success.

Abhay Kumar [Reg-42310001]

TABLE OF CONTENTS:

CHAPTER NO TITLE PAGE NO.

ABSTRACT 3

LIST OF TABLES 8

LIST OF FIGURES 9

LIST OF ABREVIATION 11

LIST OF SYSMBOLS 12

1 INTRODUCTION

1.1 OVERVIEW 13

1.2 DATA ANALYTICS IN HEART DISEASE PREDICTION 14

1.3 Motivation 15

1.4 PROPOSED RESEARCH 16

1.5 RESEARCH OBJECTIVES 16

1.6 REPORT ORGANISATION 16

2 LITERATURE STUDY 17

3 Methodology 18 3.1 Existing System 18

3.2 Proposed System 19

3.2.1 Collection of Dataset 19

3.2.2 Selection of Attributes 20

3.2.3 Pre-Processing of Data 22

3.2.4 Balancing of Data 23

3.2.5 Predication of Disease 24

3.3 Application of heart disease Predication model 24

4 System Analysis

4.1 Overview of the system 25

4.2 Advantage of Proposed System 26

4.3 Machine Learning 26

4.4 Machine Learning Algorithms 28

4.4.1 SVM 28

4.4.2 Naïve Bayes 30

4.4.3 KNN 31

4.4.4 Decision Tree 32

4.4.5 Random Forest 33

4.4.6 Logistic Regression 34

4.4.7 Linear Regression 35

4.4.8 ANN 37

4.4.9 Deep Neural Network 38

4.4.10 Ada-boost 39

4.4.11 Xg-boost 40

5 System Design

5.1 System Architecture 42

5.2 Architecture of Predication System 43

6 Coding and Testing

6.1 Software Requirements 44

6.2 Objective and type of Testing 45

6.2.1 Unit Testing 45

6.2.2 Integration Testing 45

6.2.3 Functional Testing 45

6.2.4 System Testing 45

6.2.5 White Testing 45

6.2.6 Black box Testing 45

6.3 Testing in Machine Learning 46

6.4 Test Cases 47

7 Results and Analysis

7.1 Data Set 48

7.2 Metrics for Performance Analysis 49

7.3 Performance of Machine Learning Algorithms 51

7.3.1 SVM 51

7.3.2 Naïve Bayes 53

7.3.3 KNN 54

7.3.4 Decision Tree 55

7.3.5 Random Forest 56

7.3.6 Logistic Regression 5

7.3.7 ANN 58

7.3.8 DNN 59

7.3.9 Ada-boost 63

7.3.10 Xg-boost

7.4 Results from the test Data set 63

7.5 Result and Discussion 70

8 Conclusion and Future Work

8.1 Conclusion Remarks 72

8.2 Future Scope 72

References 73

Appendix 74

LIST OF TABLES

TABLE NO. TABLE NAME PAGE NO

1.1 Different types of heart diseases 13

6.4 Test cases. 47

7.3 Confusion Matrix 51

7.3.1 SVM Confusion Matrix 51

7.3.2 Naïve Bayes Confusion Matrix 53

7.3.3 KNN Confusion Matrix 54

7.3.4 Decision Tree Confusion Matrix 55

7.4.5 Random Forest Confusion Matrix 56

7.4.6 Logistic Regression Confusion Matrix 57

7.4.7 ANN Confusion Matrix 58

7.4.8 DNN Confusion Matrix 59

7.4.9 Ada-boost Confusion Matrix 60

7.4.10 Xg-boost Confusion Matrix 63

LIST OF FIGURES

FIGURE NO. FIGURE NAME PAGE NO.

4.1 Possible hyperplanes 11

4.2 Hyperplane with the maximum margin 12

4.4.1 SVM 28

4.4.2 Naïve Bayes 30

4.4.3 KNN Clustering 31

4.4.4 Decision tree 32

4.4.5 Random Forest with two trees 33

4.4.6 Sigmoid function graph 34

4.4.6 Logistic Regression 34

4.4.7 Linear Regression 35

4.4.8 ANN 37

4.4.9 Deep Neural Network 38

5.1 System Architecture 42

5.2 Architecture of Predication System 43

7.1 Dataset used. 44

7.3.1 SVM Results 51

7.3.2 Naïve Bayes results 53

7.3.3 KNN results 54

7.3.4 Decision tree results 55

7.3.5 Random Forest results 56

7.3.6 Logistic regression results 57

7.3.7 ANN Results 58

7.3.8 DNN 59

7.3.9 Ada-boost Results 60

7.3.10 Xg-boost Results 62

7.4.1 Percentage of men and women in dataset 64

7.4.1 Risk of men and women having a heart attack 64

7.4.2 Heart attack in men out of total men 65

7.4.2 Heart attack in women out of total women 65

7.4.3 Distribution of people vs age 66

7.4.3 Age 66

7.4.4 Distribution of serum cholesterol 67

7.4.4 Cholesterol 67

7.4.5 Distribution of max heartrate 68

7.4.5 Thalach 68

7.4.6 Distribution of rest blood pressure 69

7.4.6 Trestbps 69

7.5 Results and Discussion 70

7.5.1 Input 70

7.5.2 Output 71

ABBREVIATIONS

DNN Deep Neural Network

ANN Artificial Neural Network

SVM Support Vector Machine

KNN K Nearest Neighbors

CVD Cardiovascular Disease

HDFS Hadoop Distributed File System

ECG Electrocardiogram

BP Blood Pressure

WEKA Waikato Environment for Knowledge Analysis

CNN Convolutional Neural Network

SMO Sequential Minimal Optimization

TP True Positive

TN True Negative

FP False Positive

FN False Negative

EKG Electrocardiogram

ReLU Rectified Linear Unit

LIST OF SYMBOLS

\* Multiplied by

% Percentage

/ Divided by

𝜎 Scale

√ Under root

CHAPTER 1

INTRODUCTION

* 1. OVERVIEW

Heart disease is the kind of disease which can be cause the death. Every year too many peoples are dying due to heart Attack diseases. Heart disease can be occurred due to the weakening of heart muscles cells. Also, the heart failure can be described as the failure of heart to pump the blood in human artery. Heart diseases is also known as coronary artery disease (CAD). means CAD can be occurred due to insufficient blood supply to body arteries parts. Heart disease can be detected using some symptoms like high blood pressure, chest pain, hypertension, cardiac arrest, discomfort in breathing etc. There are many types of heart diseases with different types of symptoms are recorded. Like: 1) heart disease in blood vessels: chest pain, shortness of breath, 2) heart disease caused by abnormally heartbeats: slow heartbeat, discomfort in breathing, chest pain etc. Most common symptoms are recorded chest pain, shortness of breathed. Most common symptoms are chest pain, shortness of breath, fainting are reordered due to heart disease patients. Causes of heart disease are defects you’re born with, high blood pressure, diabetes, smoking, drugs, alcohol. Sometimes in heart disease the infection also which affects the inner membrane artery which is identified by symptoms like high fever, fatigue, dry cough, skin rashes. Causes of heart infection are bacteria, viruses, parasites and parasitism. Types of heart disease: Cardiac arrest, Hypertension, Coronary artery disease, Heart failure, Heart infection, Congenital heart disease, Slow heartbeat, stroke this type of heart disease are recorded in medical history, ang. Now at a present days there are so many automated techniques are detect the heart disease like data mining, And machine learning, deep learning, etc. So, in this paper review we will be brief introduction about machine learning techniques. In this we train and test the datasets using the machine learning repositories. There are so many risk factors on the basis of that the heart disease are predicted. Risk factors are like Age, Sex, Blood pressure, Cholesterol level, Family history of coronary illness, Diabetes, Smoking, Alcohol, Being overweight, Heart rate, Chest Pain etc.

|  |  |
| --- | --- |
| Arrhythmia | The heartbeat is improper whether it may irregular, too slow, or too fast. |
| Cardiac arrest | An unexpected loss of heart failure function, consciousness and Shortness of breath occur suddenly |
| Congestive heart failure | The heart does not pump blood as well as human artery it should be, it is the condition of chronic. |
| Congenital heart disease | The heart’s abnormally which develops before birth. |
| Coronary artery disease | The heart’s major blood vessels can be Damage, or any disease can be occurring in the blood vessels. |
| High Blood Pressure | It has been a condition that the force of the blood against the artery walls is too high. |
| Peripheral artery disease | The narrowed blood vessels which can be reduce flow of blood in the limbs, is the circulatory condition. |
| Stroke | Interruption of blood supply or occur damage to the brain. |

Fig Different types of heart disease

1.2 DATA ANALYTICS IN HEART DISEASE PREDICTION:

Data analytics and machine learning algorithms have been used across various datasets to extract patterns and predict the occurrence of heart diseases by numerous papers. The most used algorithms have been SVM, Logistic Regression, Naive Bayes, KNN and decision tree, out of which SVM and Logistic regression and Ada-boost and xg-boost have been known to give more accurate Accuracy results. Some Based papers has also used to Hadoop based MapReduce and HDFS algorithms to store dataset across different nodes and run the system parallelly. The most used attributes against the datasets have been Blood pressure and heart rate and the age group used in the datasets is generally over the age of 25. Although all known based papers have to use the above-mentioned algorithms, and none have employed machine learning in their studies, which tends to give a higher accuracy rate result. Our system, along with all these algorithms uses a Logistic Regression and more attributes dataset like cholesterol, angina, age, sex etc. for more

1.3 MOTIVATION

It is well known fact that patients when has affected with heart disease are subjected to lot of tests like ECG, EKG, and so many things on. But these tests are carried out only when the person gets chest pain or the kind of symptom leading to the heart disease. In today’s modern world there are lot of body wearable devices which could be give the pulse rate, Blood pressure and so. There is nothing like a person get heart disease only after the age of 30. The current generation is under lot of stress and pressure due to work and other factors. So, there is an urgent need to perform dataset and analysis of the physiological parameter towards possibility of heart disease before heart attack happens disease. So, towards this some amount of research carried out by employing Machine learning algorithm for prediction of heart disease like Regression, KNN, SVM, Bayes, Decision Tree and Ada-boost and xg-boost, Logistic Regression and so many. So, with the upcoming machine learning, we have here proposed to develop a heart disease prediction system with highest accuracy Results as compared to other machine learning algorithm based on thirteen physiological parameters as compared to most important ones which are heart rate, age, and sex alone.

1.4 PROPOSED RESEARCH:

Based on above mentioned motivation, we here have proposed to develop a Machine learning-based prediction system of heart diseases which would be based on historical physiological data set. It is clear a good amount of work done in the past in predicting system heart diseases with the help of machine learning algorithms like SVM, Regression, Bayes, KNN and even Artificial Neural Network Ada-boost, Xg-boost and Logistic regression. Most of the system has taken two parameters’ variables data set like heart rate, age and so. But we are in our work are taking thirteen parameters including cholesterol, Blood pressure, heart rate which can be unique in prediction of heart disease. So, our system encompasses the Deep Learning framework called Deep Artificial Neural Network and logistic regression for prediction of heart disease system based on physiological data set of people based on age group and sex which includes thirteen parameters like cholesterol, resting BP, resting ECG, maximum heart rate, age, sex, chest pain, fasting blood sugar, exercise included angina ,slope also, the most number of major vessel, slope of peak exercise, thalassemia, depression. This is compared with other machine learning algorithms like SVM, Bayes, KNN, Decision Tree, Random Forest, Logistic Regression and ANN ,Xg-boost ,Ada-boost for prediction of heart disease with accuracy Results and error.

1.5 RESEARCH OBJECTIVES

The major objective of our project work is focused on the following Steps.

• Collection of heart disease related all data set and preprocessed

• Validation of SVM, Linear Regression, Bayes, KNN algorithm for heart disease prediction on the attribute’s dataset.

• Validation of random forest and decision tree on the given set of data for heart disease prediction • Validation of ANN and Deep Neural Network on data set for prediction of heart disease.

• Accuracy as well as error computation on above all mentioned algorithm.

• Comparative analysis of all algorithms in terms of accuracy results.

1.6 REPORT ORAGANISATION:

This based report deals with various aspects of the proposed system. We can begin with chapter 1, which gives the basic overview of the project reports. It is clearly explaining how the data analytics is used in heart disease prediction systems and gives our research objectives. Chapter 2 starts with explaining all the different research-based papers that gave us ideas to carry forward their work. These papers gave us a strong background. The next chapter will be analyses the present system as well as the proposed system. It gives a basic idea as to how the project will work and what objectives it will work with. It is also specifying all the algorithms used in the project. In chapter 4 we have described the data set used, its attributes, workflow of the system and system architecture. Chapter 5 gives the system configuration including hardware and software requirements along with testing. Chapter 6 gives the results and accurate results of our project. Conclusion gives us a review about this project and describes how this system is useful to society. Next chapter deals with the enhancements of this project and how it can be modified for better results with accuracy and the last chapter lists the references.

CHAPTER 2

LITERATURE REVIEW

K. Polara Ju, [1] did the work of Predicting Heart Disease dataset using the model of Multiple Regression, which implied that Multiple Linear Regression was more suitable for heart disease prediction system. The dataset used to consist of 32000 entries with 13 attributes and was partitioned into training and testing, where training data was 70% and testing data was 30%. Upon considering the accurate results, it was concluded that the regression algorithm was more accurate than others. Marjia,

[2] did the heart disease prediction using SMO, j47, K-Star, Multilayer perception and Bayes Net using WEKA. From the results of the prediction results, it was found that the performance of Bayes Net and SMO was optimum, by using the k-fold cross validation technique. The other algorithms did not give a satisfactory performance. And hence, accuracy performance was improved for better diagnostic decisions. S. Seema,

[3] focused on predicting general chronic diseases rather than just heart disease can be prediction. She utilized the hospital health records by machine learning using Decision tree, Support Vector Machine (SVM), Artificial Neural Network and Naive Bayes. From this prediction, on applying the mentioned algorithms, it was concluded that highest accuracy results were given by SVM, while for diabetes, the highest accuracy rate was achieved by Naive Bayes. Ashok Kumar Dwivedi,

[4] used various algorithms are of SVM, Classification Tree, Logistic Regression, Naive Bayes, SVM, Naive Bayes, KNN and ANN. The highest accuracy results rate among these was attained by Logistic regression. And Megha Shahi,

[5] used machine learning for heart disease prediction. WEKA was used for the diagnosis of heart

Purushottam,

[6] the data set is used to machine learning techniques for the prediction of heart diseases. This is a study to aid the medical practitioner to take better decisions based on a particular parameter attribute. By training and testing a particular parameter, it is retrieved an accuracy rate of 86.3% in testing, and an accuracy rate of 87.3% during training. Gomathi

, [7] it is also proposed a prediction of multiple disease using the techniques of machine learning. These days, machine learning is being extensively used for multi disease prediction, as the number of tests is subsequently reduced. The study of majorly and focused on predicting breast cancer, heart diseases and diabetes etc.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM:

Heart disease has been highlighted as a silent killer which has leads to the deaths of a person without any obvious symptoms. And The nature of the diseases is the cause of growing anxiety and depression about the heart disease & it is consequences conditions. Hence The continued efforts have being done to predict the possibility of this deadly diseases in prior time. So that the various tools & techniques are regularly being experimented with used to suit the present-day time to health needs. Machine Learning Algorithms techniques that can be a good method in this regard function. Even though heart disease can be occurred in different types of forms there is a one common set of core risk factors that influence whether the someone will be ultimately at risk for heart disease or not. By collecting the data from various sources and classifying them under suitable headings & finally will be analyzing to extract the desired dataset details and we can be concluded. Machine learning Algorithms technique could be used very well adapted to the do the prediction of heart diseases problem. As the well-known quote says, “Prevention is better than cure”, early prediction & its control can be helpful to prevent & decrease the death ratio rates due to heart disease problems.

3.2 PROPOSED SYSTEM:

* The working of the system starts with the collection of data and selecting the all-important attributes. Then the required data set is preprocessed into the required format. The data set are then divided into two parts training and testing data set. This algorithms are applied, and the model is trained using the training data set. The accuracy of the system is obtained by testing the system using the testing data set. This system is implemented using the following modules method.
* 1.) Collection of Dataset
* 2.) Selection of attributes
* 3.) Data Pre-Processing
* 4.) Balancing of Data
* 5.) Disease Prediction

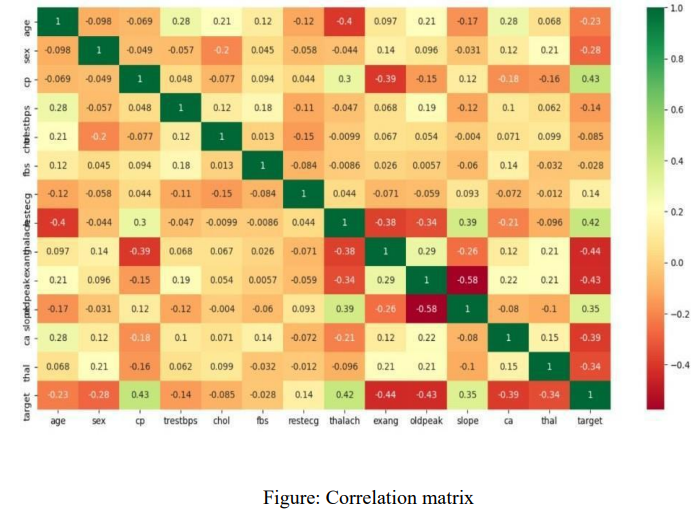
3.2.1 Collection of datasets:

* Initially, we will be collecting a dataset for our heart disease prediction system model. After that the collection of the dataset, we are split the dataset into training data set and testing data set. The training data set is used for prediction model learning and testing data set is used for evaluating the prediction model. For this project, 70% of training data set is used and 30% of data set is used for testing. This The data set used for this project of Heart Disease UCI. This data set are consisting of 76 attributes; out of which, 14 attributes are used for the system.
* A diagram of a model

  Description automatically generated

Fig no 3.1.1 Collection of Data

3.2.2 Selection of attributes:

* Attribute or Feature selections includes the selection of most appropriate attributes for the prediction systems. This is used to increase the efficiency of the system. The Various attributes of the patients. Like age, sex(gender), chest pain, fasting blood pressures, serum cholesterol ,Fbs, restecg,slope etc. are selected for the prediction system. The Correlation matrix are used for attribute selection for this model system.
* 
* Fig number 3.2.2 Correlation matrix
* SELECTION OF THE ATTRIBUTES DATASET:

|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO.** | **Attribute** | **Description** | **Type** |
| 1 | Age | Patient’s age (29 to 77) | Numerical |
| 2 | Sex | Gender of patient (male-0 female-1) | Nominal |
| 3 | Cp | Chest pain type | Nominal |
| 4 | Trestbps | Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200) | Numerical |
| 5 | Chol | Serum cholesterol in mg/dl, values from 126 to 564) | Numerical |
| 6 | Fbs | Fasting blood sugar>120 mg/dl, true- 1 false-0) | Nominal |
| 7 | Resting | Resting electrocardio graphics result (0 to 1) | Nominal |
| 8 | Thali | Maximum heart rate achieved(71 to 202) | Numerical |
| 9 | Exang | Exercise included agina(1-yes 0- no) | Nominal |
| 10 | Oldpeak | ST depression introduced by exercise relative to rest (0 to .2) | Numerical |
| 11 | Slope | The slop of the peak exercise ST segment (0 to 1) | Nominal |
| 12 | ca | Number of major vessels (0-3) | Numerical |
| 13 | Thal | 3-normal | Nominal |
|  |  |  |  |

3.2.3 Pre-processing of Data:

* Data pre-processing is all most important step for the creation of a machine learning algorithms model. Initially, the data may not be clean or in the required format for the model which can cause misleading outcomes data. In pre-processing of data, we transform the data into our required format model. And it is used to deal with noises, duplicates, and missing values of the dataset models. Data pre-processing has been the activities like importing datasets attributes parameters, and splitting datasets to attribute scaling, etc. Preprocessing of data set is required for improving the accuracy of the model System.
* A diagram of data processing

  Description automatically generated
* Fig number 3.2.3 Data Pre-processing
* 3.2.4 Balancing of Data:
* Imbalanced datasets can be balanced in two ways methods. They are Under Sampling and Over Sampling
* (a) Under Sampling: In Under Sampling, dataset balance is done by the reduction of the size of the ample class of data set and This process is considered when the amount of data is adequate.
* (b) Over Sampling: In Over Sampling, dataset balance is done by increasing the size of the scarce samples class of data set .and This process is considered when the amount of data is inadequate.
* A graph of a person's target

  Description automatically generated with medium confidence
* 3.2.5 Prediction of Disease:
* The too many Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification. And Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy results is used for heart disease prediction systems.
* A diagram of a diagram of a person's reaction

  Description automatically generated with medium confidence

Fig number 3.2.5 Predication of Disease

3.3 APPLICATIONAPPLICATION OF HEART DISEASPREDICATION MODEL:

* It is utilizing the Data such as blood pressure, cholesterol, diabetes and then tries to predict the possible heart disease in patients' details (future Predication by diagnosis).
* This may be help in talking to Preventive measure and hence can try to avoid the possibility of heart disease in patient.
* So, When a patient is predicated as positive for heart disease, then the medical history data for the patient can be analyzed by the doctors

CHAPTER 4

SYSTEM ANALYSIS

4.1 OVERVIEW OF THE SYSTEM:

Heart Disease prediction system is done at hospital based on patient visiting hospital based on current heart condition like high pulse rate, chest pain or so many reason. Accordingly, patients are tested for heart blockage using ECG, and Angiogram and so. Based on the test, patients are treated by performing pacemaker, Angioplasty and so. Now with ever increasing pressure of work and stress, there is no such thing as a heart problem that comes only after age of 45 or so. Now even adults at the age of 20, 30s, 40s get heart disease due to work pressure and stress so many things. Some even get in their childhood too. So, there is an urgent need to come up with an automated medical diagnosis system for heart disease based on physiological dataset based on Blood pressure, heart rate, blood sugar, cholesterol and so for predicting the heart disease. Some amount of work done by applying machine learning algorithm based on data and collected for predicting the heart disease using algorithms like Regression, SVM, Bayes, KNN, Decision Tree and ANN Logistic regression for heart disease prediction. But in all these systems, only a few parameters are taken for prediction which are heart rate, age and so. Also, no one have tried to be using logistic regression in thus field of heart disease prediction. So, we in this work have applied Logistic Regression for the heart disease prediction based on 13 physiological parameters and compared with other algorithms like Regression, SVM, Bayes, KNN, Decision Tree, Random Forest and ANN in terms of error and accuracy. Now for our work, we have taken a dataset that was publicly available on Kaggle for predicting heart diseases. The parameters attributes dataset used as input for data analysis using Machine learning algorithm are as follows:

1. age: which is taken in years

2. sex: 1 for male, 0 for female

3. cp: short for chest pain.

4. trestbps: blood pressure taken when the body is resting

5. chol: level of cholesterol

6. Fasting blood sugar: (1 for true; 0 for false)

7. resting electrocardiographic results Value 0: normal Value 1: having the ST-T wave which is not normal Value 2: shows the probability of having left ventricular hypertrophy

8. thalach: the max level of heart beats achieved by the heart

9. exercise induced angina: 1 for yes; 0 for no

10. oldpeak = the depression by exercise compared to the one at rest

11. slope: peak of the ST segment during exercise Value 1: no slope Value 2: straight line Value 3: down sloping

12. ca: major vessels (0-3) coloured by fluoroscopy

13. thal: 3 = perfect; 6 = permanently defected; 7 = defect that can be altered

14. num: heart disease diagnosis

4.2 ADVANTAGES OF THE SYSTEM:

Our project-based paper study to focuses not only on Blood pressure and heart rate of the patient, but also on other factors like age, sex, angina, chest pain location, cholesterol etc., that has not usually taken in consideration.

• With a greater number of attributes parameters against which the system is being tested and trained, the data can be divided into various clusters according to different factors systems.

• Our system based tends to predict when the chances of occurrence of heart disease are more, prior to the actual occurrence, unlike the other systems, which generate an alert only when the extreme values had been attained.

• Our system divides the dataset according to factors like age and sex.

• Logistic Regression has been used in our system, which hasn't been employed in any other paper dealing with heart disease prediction system, yet. Logistic regression has the highest accuracy among all the machine learning Algorithms.

• Our project-based system uses all the viable machine learning and data mining algorithms.

4.3 MACHINE LEARNING:

• Machine learning (ML) are the subset of artificial intelligences (AI). And It uses statistical techniques to enable computers to learn and make decisions without being explicitly programmed.

• ML are allowing machines to automatically to learn from data and past experiences to identify patterns and make predictions with minimal human intervention. For example, personal voice assistants like Siri, Alexa, and Cortana use to ML-based speech recognition systems.

• ML are also used in social media platform

• Types of machine Learning

• 1) Supervised Machine Learning.

• 2) Unsupervised Machine Learning.

• 3) Reinforcement machine Learning

1. Supervised MACHINE Learning:

Supervised machine learning is a type of machine learning that are uses to labeled data to train algorithms and classify data or predict outcomes results.in this supervised machine learning algorithms that is the training dataset are provided to the machines work as the supervisor that the teaches in machines learning algorithms to predict the output results correctly accuracy. And it applies to the same concept as a student learns under the supervision of the teacher. Supervised machine learning is a process of providing the input dataset as well as correct output dataset to the machine learning model function. The aim of a supervised machine learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

1. UNSUPERVISED MACHINE LEARNING:

This Unsupervised machine learning is the type of machine learning technique method way that is used to algorithms and find patterns in the unlabeled data set. Because Unsupervised machine learning could not be directly applied in a regression and classification problem because unlike supervised machine learning, we have the input data but not have the corresponding output data. The goal of unsupervised machine learning is to find the underlying structure of dataset, and group that data according to similarities, and represent that dataset in a compressed format method.

• Unsupervised machine learning is helpful for finding to useful insights from the data.

• Unsupervised machine learning is much too like how a human learns and thinks by their own experiences, which makes it closer to the real AI.

• In The real-world basis, we don’t always have input data with the corresponding output results so to solve such as cases, so we need unsupervised machine learning.

3)REINFORCEMENT MACHINE LEARNING:

• Reinforcement machine learning is an area of Machine Learning. And it is regarding taking suitable action and maximizing reward in a particular condition. It are employed by multiple type of software in machines to find the best possible behavior and path it would be taken in in a specific condition.

4.3 MACHINE LEARNING ALGORITHMS:

* In this work, I have choice of Eleven machine leaning algorithms which are include
* 1) Random Forest
* 2) SVM
* 3) Decision Tree
* 4) Naïve Bayes
* 5) ANN
* 6) DNN
* 7) KNN
* 8) Logistic Regression
* 9) Linear Regression
* 10) Ada-boost
* 11) Xg-boost
* These algorithms are used to unfold a prediction system which will be analyze to predict whether the patient is pertaining to any heart disease or not with best accuracy Results

4.4.1 Support Vector Machine:

• Support Vector Machine Learning Algorithms or SVM is one of the most important popular Supervised machine Learning algorithms, which can be used for Classification as well as And Regression problems. However, it is primarily, to use for Classification problems in Machine Learning.

• The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can be a separate data points into different classes. The hyperplane tries to maximize the margin between the closest points of different classes point. The optimal divider is the one that is equidistant from the boundaries of each group class.

• SVMs were first introduced in the 1960s and refined in 1990. They are broadly classified into two types: simple or linear SVM and kernel or non-linear SVM .

• To train and test a SVM model data set , the data set are divided into train and test data. The data is also scaled to lie between 0 and 1

The followings are important concepts in SVM:

* 1) Support Vectors - Data Points that are very closest to the hyperplane are called support vectors. And Separating line will be defined with the help of these data points.
* 2) Hyperplane - As we can see in the above diagram, it is a decision plane or space which is divided between a set data of objects having different classes groups.
* 3) Margin - It must be defined as the gap between two lines on the closest data points of different classes and different groups. It must be calculated as the perpendicular distance from the line to the support vectors machine. And Large margin is considered as a good margin and small margin is considered as a bad margin.

A diagram of a support vector

Description automatically generated with medium confidence

Fig Number 4.4.1 SVM

Advantages And Disadvantages of support vector machines:

ADVANTAGES:

● Effective in high dimensional spaces.

● Still effective in cases where the number of a dimensions is greater than the number of samples.

DISADVANTAGES:

● If they must number of features is much greater than the number of samples classes and avoid the over-fitting in choosing Kernel functions and regularization term is crucial. SVMs had not been directly applied to provide the probability of estimates, and these will be calculated using an expensive five-fold cross-validation methods model.

4.4.2 Naïve Bayes:

* It is a machine learning technique that are works on this strategy of the Bayes’ Theorem. It is basically assuming that there would be no attributes And dependent on each other. And It is a group of algorithms that have a common principle that every feature is independent of the other group. Bayes’ Theorem tells us the probability of an event that will be occur when another event has already occurred. The mathematical equation is:

• Probability(a|z) = (Probability(z|a) \* Probability(a)) / Probability(z)

• Where :

• Probability(a|z): Gives us the probability of a (the hypothesis) gives the data is a.

• Probability (z|a): Gives the probability of the data when the hypothesis is true.

• Probability (a): Regardless of the data, the hypothesis is said to be true.

• Probability (d): Regardless if the data, the probability of the hypothesis is given.

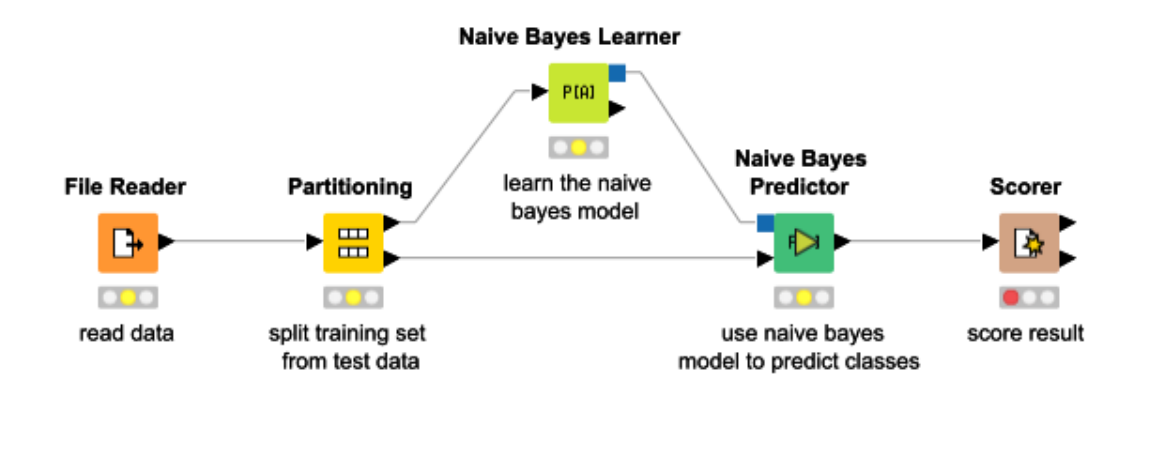


Fig Number 4.4.2 Naïve Bayes learner

4.4.3 KNN:

• K nearest neighbors abbreviated as KNN is an algorithm that clusters data into classes and then classifies it as per their similarity measures. Classification is based on most of the votes to its neighbors. Data is assigned to the classes that have the nearest neighbors. As we increase the number of nearest neighbors, I.e., the value of k, the accuracy might be increase. KNN is broadly used for pattern recognition and statistical prediction methods.

• It divides the data into clusters And based upon the distance from the nearest neighbors.

A diagram of a training course

Description automatically generated

Fig No. 4.4.3 KNN

4.4.4 Decision Tree:

• Decision Tree is one of the most supervised learning algorithms. In this the data is continuously split based on a certain parameter after which we end up getting the decision nodes and the leaves. What makes it different from the other supervised algorithms is that it can be also solve the regression and classification problems easily way. The main aim is to create a system that can be predict the results that we desire just by learning the decision rules from the prior data, i.e., the training set.

Working:

* In this algorithm a Decision Tree, for predicting the class of the given dataset Attributes parameters, this can be the algorithm starts from the root node of the tree. And This algorithm method compares the values of the root attribute set with the record (real dataset) attribute and based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute dataset value with the other sub-nodes and moves further. It can continue the processing until it reaches the leaf node of the tree. The complete process it can be better understood using the below algorithm steps:

● Step-1: Begin the decision tree with the root node, says S, which can be contains the complete dataset attributes.

● Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM) dataset attributes.

● Step-3: Divide the S into the subsets that it can be contains possible values for the best attributes.

● Step-4: Generate the Decision Tree node, which can be contains the best attribute.

● Step-5: Recursively make new decision trees are using the subsets of the dataset attributes and created in step -3. method is Continue this process until a stage and reached where you cannot be further classifying the nodes and call the final node as a leaf node.

A diagram of decision making

Description automatically generated

Fig No. 4.4.4 decision tree

4.4.5 Random Forest:

This Random Forest algorithms, just as the name suggest creates several number of a random forest due to which it is also called the random decision forest. It is one of the most supervised learning algorithms. It is builds random forests that are basically just a group of decision trees. It is mostly trained with the bagging method as it is the most efficient. What are the bagging method does is, it combines all the learning models which can be turn in helps us with the overall result. Just like decision trees, it can be used for both regression and classification problems.

A diagram of a tree

Description automatically generated

Fig No. 4.4.5 random Forest

Algorithm Steps:

* It works in four steps method way:
* Select the random samples from a given dataset.
* Construct a Decision Tree for each sample and get a prediction result from each other Decision Tree.
* Perform a vote for each group to predict the result.
* Select the prediction result with the most votes as the final prediction system.
* Advantages And Disadvantages:
* ADVANTAGES:
* ● Random Forest can perform both Classification and Regression tasks.
* ● It is capable of handling larger datasets with high dimensionality.
* ● It enhances the accuracy results of the model and prevents the overfitting issue.
* DISADVANTAGES:
* Although Random Forest it can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

4.4.6 Logistic Regression:

Logistic Regression algorithm is rather a predictive analysis. It is used to analyze the relation between the data among which one is a dependent variable, and the other one is independent variable. What doses logistic regression , it is rounds off all the values of the results to the closest binary value for the ease of prediction. This can be sometimes causing a few errors, and not predict very precisely. To remove this error, multiple regression is used which helps in getting the closest to the true result. Logistic Regression uses complex cost functions such as the ‘Sigmoid function’. In machine learning sigmoid is used to predict the probabilities.

A graph of a function

Description automatically generated

Fig No. 4.4.6 Logistic Regression

4.4.7 Linear Regression:

• Linear regression Algorithms is a fundamental machine learning algorithm it is used for predicting numerical values based on input features. It is assuming a linear relationship between the features and the target variable. The model learns the coefficients that best fit the data and can be make predictions for new inputs.

• in this Linear regression algorithm has shown a linear relationship between a two variables like dependent (y) and one or more independent (x) variables, hence called as linear regression. Since linear regression has shown the linear relationship, which means it is to finds how the value of the dependent variable is changing according to the value of the independent variables.

• The linear regression algorithms model has provided a sloped straight line and representing the relationship between the variables. Consider the below image:

y= Dependent Variable (Target Variable) y= Independent Variable (predictor Variable)

A diagram of a line of regression

Description automatically generated

Fig No.4.4.7 Linear Regression

4.4.8 Artificial Neural Network (ANN):

* An artificial neural network abbreviated as ANN is a model that works as the human brain or neural network I.e., the neurons. It is called a computational model that processes all the complex data. It isn’t given any task specific goal, but it learns from the examples or data that is given to it just like the brain. It is based on a collection of nodes called the artificial neuron. More the no. Of neurons, better the system. The neurons transmit signals from one to another making a proper connection which resembles the human neural network.
* A neural network has the following 3 layers:
* • Input layer – It is consisting of the complex raw info that we feed to the neurons.
* • Hidden layer – These are the computational layers which has been taken the input and the weight of a node from the previous layers, processes it with the activation function will be and sends the output to the next layer.
* • Output layer – This depends on the output of the hidden layers, and the functions taking place in there.

# The basic computational unit is the neuron. It is receiving inputs from the sources provided, and each input carries a weight, which has been given according to the relative importance of all the other inputs. Then the function is applied to

A diagram of a brain function

Description automatically generated with medium confidence

Fig No.4.4.8 ANN

4.4.9 DEEP NEURAL NETWORK:

* Deep Neural Network Algorithms abbreviated as a DNN, is a complex neural network consisting of more than 2 layers. This algorithm uses to high level of mathematical models for data processing in complex manner. Deep learning is derived from a vaster family of networks of neural methods, such as CNNs. Deep learning can either be unsupervised, supervised or semi- supervised.
* Deep Convolutional and recurrent neural networks are utilized in natural language processing computer vision audio recognition video recognition, social network filtering bioinformatics machine translation image analysis, drug design, and material inspection where the outcomes were comparable/superior to some human professionals.
* Use the different layers of nonlinear units of processing for featured extraction as well as transformation. Each of the successive layers uses the previous layer’s output as the input.
* • Can be either too supervised and unsupervised or semi-supervised also.
* • learn the different hierarchies of representations which relate to different hierarchies of abstractions.
* • It is consisting of an input layer, an output layer and between many variable hidden layers.
* • Output of one layer goes as the input to other. The activation function used in our module is the Re LU between hidden layers and SoftMax at the output layer which takes 0 or 1 .

nA diagram of a network

Description automatically generated

Fig No. 4.4.9 Deep Neural Network

4.4.10 ADABOOST ALGORITHM:

• Ada-boost was the first really successful boosting algorithm developed for the purpose of binary classification. Ada-boost are short for Adaptive Boosting techniques and it is a very popular boosting technique methods which can be combines multiple “weak classifiers” into a single “strong classifier”

• ALGORITHM:

• 1. Initially, Ada-boost is selecting and training subset randomly.

• 2. It is iteratively training and testing the Ada-boost techniques machine learning algorithms model can be selecting the training data set and based on the accurate prediction results of the last training.

• 3. It has been assigning the higher weight to wrong the classified observations models methods so that is the next iteration these observations will be get the high probability for classification and regression accurate results.

• 4. Also, it is assigning to the weight and the trained classifier dataset in each iteration according to the accuracy results of the classifier. The more accurate to classifier will get high weight.

• 5. This process will be iterated until the complete training data fits without any error or until reached to the specified maximum number of estimators.

• 6. To classify, and perform a "vote" to across all the machine learning algorithms built you.

Advantages And Disadvantages:

Advantages:

• It's easy to implement

• It can be used with any dataset

• It's less prone to overfitting than other algorithms

• It can improve the accuracy of weak classifiers

• It can be used with many base classifiers

• It's easier to use than other algorithms

• It can be used to classify text and images

• Disadvantages:

• AdaBoost uses a progressively learning boosting technique. Hence the high-quality of data is needed in examples of Ada-Boost vs Random Forest. And it is also very sensitive to outliers and noise in data requiring the elimination of these factors before using the data set model.

4.4.11 XGBOOST ALGORITHM:

XG-Boost is a robust machine-learning algorithm that can be help you to understand your data and make better decisions. XG-Boost is an implementation of gradient-boosting decision trees.

• The Algorithms of XG-boost is an implementation of Gradient Boosted decision trees. It is an also type of Software library that was designed basically to improve speed and model performance. In this machine learning algorithm and decision trees are created in sequential form. Weight plays an important role in XG-boost. Weights are assigned to all the independent variables which are then fed into the decision tree which can be predicts results. The Weight of variables parameters are predicted wrong by the tree is increased and these the variables parameters are then fed to the second decision tree. These is individual classifiers/predictors then assembled to give a strong and more precise model. It can be work on a regression, classification, ranking, and user-defined predictions.

1. Regularization:

• Regularization: XG-Boost can incorporate regularization techniques to prevent overfitting. It can include both L1 and L2 regularization terms in its objective function, which will help is controlling the complexity of the model and reducing variance.

• gamma: The minimum loss reduction allowed for a split to occur. The optimal value of a gamma depends on the data set and other parameter values.

• alpha: L1 regularization on leaf weights. Larger values mean more regularization.

• lambda: L2 regularization. This term is a constant that is added to the second derivative function (Hessian) of the loss function during gain and weight (prediction) calculations.

• The regularization term (Ω) is a combination of L1 and L2 regularization terms. It is defined as:

• Ω(F) = λ \* Ω1(F) + 0.5 \* γ \* Ω2(F)

• WHERE:

• Ω1(F) is the represent of L1 norm of the leaf weights

• Ω2(F) is the represent of L2 norm of the leaf weights

• Regularization parameters are usually to impact tree structure as well as weight values

1. Parallel processing:

• Parallelization – The process of sequential tree building is done using the parallelized implementation in the XG-Boost algorithm. This had made the possible due to the outer and inner loops that are interchangeable. The outer loop have listed the leaf nodes of a tree, while the inner loop will be calculating the features.

• : XG-boost will be utilizing the power of parallel processing and that is why it is much faster than GBM. It is uses multiple CPU cores to execute the model. While using Scikit Learn library, N thread hyper-parameter is used for parallel processing. N thread represents number of CPU cores to be used. If you want to use all the available cores, don't mention any value for N thread and the algorithm will detect automatically.

1. Handling missing values:

• XG-boost has been an in-built capability to handle missing variable values. When XG-boost can be encounters a missing value at a node, it tries both the left and right hand split and learns the way leading to higher loss for each node. It then does the same when working on the testing dataset.

1. Cross Validation:

The XG-boost can be allowing user to run a cross-validation at each iteration of the boosting process and thus it is way too easy to get the exact optimum number of boosting iterations in a single run. This is unlike GBM where we must do grid-search and only a limited values can be tested.

1. Effective Tree Pruning:

Tree Pruning: The XG-Boost is used to max-depth parameter as specified the stopping criteria for the splitting of the branch and starts pruning trees backward. This Is depth-first approach to improves computational performance an significantly.

• A GBM will be stop the splitting of a node when it encounters a negative loss in the split function. Thus, it is a more of a greedy algorithm. XG-boost on the other hand makes splits up to the max-depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain.

A diagram of a classifier

Description automatically generated

Fig No. 4.4.11 Effective Tree Pruning

CHAPTER 5

SYSTEM DESIGN

This chapter gives the system architecture of our project towards heart disease analysis using Machine Learning. Followed by the architecture, and we give predication of architecture systems Diagram and Data set used.

5.1 SYSTEM ARCHITECTURE:

A diagram of a system

Description automatically generated

* This system architecture gives an overview of the working of the system.
* This is the working of this system is described as follows:
* Dataset collection is collecting the most important data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease systems. After that identifying the available of data resources, And they are further selected, And cleaned, made into the desired form also. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease systems. Accuracy measure compares the accuracy of different classifiers results.

5.2 Architecture Of Predication systems:

A diagram of a software company

Description automatically generated with medium confidence

1)Our system is first divided into training and testing data.

2)The training data is 70% and is given supervised inputs and outputs.

3)The testing data is 30% and shows us how well the system is trained.

4) The dataset we have chosen consists of 13 attributes according to which various algorithms perform their calculations and approximations.

5) The system starts with first preprocessing of the dataset we have fed to it.

6) It studies and analysis it, and then applies the required machine learning algorithm.

7) If it finds that the dataset is supervised, it will separate it into training data and testing data.

8) Otherwise, it will stop.

9) The algorithms we are using are all supervised.

10) After the application of algorithm, internal validation is done.

11) Accuracy is printed in the code itself.

12) Different accuracy is given for different algorithms.

13) We compare the accuracies of all the algorithms and the algorithm that gives the highest accuracy is the one, which is chosen for prediction, eventually.

14) In our system, the algorithm that gets the highest accuracy rate is Logistic Regression.

CHAPTER 6

CODING & TESTING:

6.1) Software Requirement:

1. Anaconda : It is an open-source software available to us which has been enables us to easily code in using python or R on different operating systems such as the windows, Linux, and Mac OS. It has been millions of users worldwide and is well known as the industry which helps us in developing the systems, testing them, and training the machines. This is further enables us to:

• Manage all the imported libraries, And their dependencies, and the environments of developing with Anaconda.

• In developing techniques method to train our machine with TensorFlow, And scikit-learn, etc.

• it is Analyze the datasets and manipulate the with Dask, NumPy, pandas, and Numb

• The Visualize or plot the results with Matplotlib, Holoviews, Bokeh, and Data shader. It is also providing us with jupyter notebooks which has all the in-built libraries embedded in the already. This eases our coding stress, and also helps us code with more efficiency results.

2. Python: The most abundantly used general high level programming language. It is used for both a small scale and big scale systems. It can be easily be interpreted. And It is said to support multiple programming paradigms. It is including features of procedural, object-oriented, and functional programming together. It is already garbage-collected which makes it more efficient.

3. Numpy: It is a python programming library. This is basically helps us to deal with large datasets, matrices, and multi-dimensional arrays. It is also providing us with several mathematical functions which help us and ease the calculations. It is an open-source software available to all.

4. Pandas- It is a library that is written in python Programming language. It is helps us with the analyzing of data. It is also providing us with tools and functions to manipulate a large amount of dataset.

5. Sklearn : It is a library used in machine learning Algorithms in python programming language. It is mainly helps us with the classification of data,And regression of models, and in clustering algorithms. These algorithms include SVM, random forest etc.

6. Tensorflow- It is a highly know as open-source software, which is free and available to all. It is used for differentiable programming with a large number of tasks dataset. It is used against large data sets to help us with the dataflow, and its manipulation. It is a basically a math library with various features.

6.2 OBJECTIVES AND TYPES OF TESTING:

* Testing is the performed so that we can be locate problems. It is the method of identifying all the possible faults of any system. Also, it is acts as method to test the functionality of all the individual components and the system. It is the process of verifying that the hardware system And performs its functionalities and meets user needs and expectations. Testing ensures that even if the product fails, it is fails in a controlled condition that can be managed later. There are various types of test. Each of the test is used to address a particular testing needs.

6.2.1 UNIT TESTING:

* TESTING This testing means it is including the creation of test scenarios that verify that program is working in the way it can be supposed to. In this type of testing, the objective is to validate that each module or unit is performing the tasks that it is supposed to do. A unit is the smallest part that must be tested.

6.2.2 INTEGRATION TESTING:

* These tests are made just to check the different parts after that they have been integrated to determine whether they are functioning properly as a unit or not. This type of testing is driven by events and is more curious about the outcomes of the component. Integration tests show that although the components are working properly as individual but also the combination of components is also working properly.
* 6.2.3 FUNCTIONAL TESTING:
* These kind of tests are providing systematic demonstration of the functions tested and the proof of their functioning being successful. Functional testing is targeted at a valid input, invalid input, functions, output, procedures. And The preparation on functional tests is focused on the requirements, key functionalities, different and extreme test cases
* 6.2.4 SYSTEM TESTING:
* This is type of tests tells us whether the whole system after all its parts has been joined together meets the need of shareholders. It is checks and confirms the known and expected results. These kinds of testing are done to check whether the system delivers what it was supposed to. It can be checks whether the all the parts work well with each other or not.
* 6.2.5 WHITE BOX TESTING:
* It is done in a way such that the tester has been information about all the components, workings, structure and architecture of the hardware/software. It is used for deep level testing to ensure that those places that are not accessible by a black box test are covered in white box testing. It basically means that there shall be no blind spots in the system.
* 6.2.6 BLACK BOX TESTING:
* TESTING Here the testing that is done on the software/hardware by someone who has been no information on how to operate it. Before a product is developed some documentation is done, they include requirement analysis documents that has all the details about what a product must be. Now these tests are based on those sources only. Imagine a black colored box, you can be not seeing into it. You know nothing about what it contains. This is exactly how the testing is done.

6.3 TESTING IN MACHINE LEARNING:

* Very easily put, in machine learning Algorithms is a type of system/application which is based upon models that are use prediction and analysis. Building these systems is a hard task, but testing and validating the entire system is even more difficult. And The Traditional methods of testing are always based on stationary inputs. And The output will also be fixed as the input remains fixed, but in the case of machine learning, the system has been built along with the input, which will be change or modify as we come to know more and more about it, therefore unlike the traditional system, the output here will be never remain fixed. Hence, the strategies for testing a machine learning system must be a bit different. Below are some of these strategies mentioned that can be used in testing the machine learning systems.
* 1. Development of training datasets: Here, a subset of dataset that is used for training purpose, which means it can be used to train the system to obtain the given prediction. It is supervised, i.e., the output is given along with the input.
* 2. Development of testing data sets: The testing data set is a training data subset that has been built in an intelligent way to check the how robustly our system has been trained and to check all the combinations that are possible. The resultant model will be finely tuned according to the outputs of the testing dataset.
* 3. Development of validation and correction test suite: This is based on algorithms and test datasets. For e.g., in our Project system scenarios consist of clustering results based upon the different factors/attributes and creating profiles of risk depending upon behaviors and demography.
* 4. Understanding the algorithm: All of this is dependent upon calculations and how the different algorithms use different calculations. Some algorithms like regression-based algorithms give some numeric results regressively, or continuously. On the other hand, And there are some algorithms that are calculate the results by dividing the outcomes into different parts or behaviors, and then some algorithms are create to multiple layers between the input and output layer. Continuous mathematically numeric variables such as return on investment.
* 5. Using Statistics for conveying the results : The Outcomes are generally predicted in the form of working model, or a numeric value or a working interface or something like that. But in machine learning algorithms, the use of bars, histograms and other statistical diagrams it is used to depict the prediction or analysis in a more comparative or mathematical method way. (The machine learning models will usually have approximate, and not actual results upon validation. In conclusion, software testing is as crucial and important a task in machine learning, as is in any other traditional system, but unlike those systems , our testing is based on more dynamic factors and generally will produce a relative or approximate result which can be best shown statistically).

6.4 TEST CASES:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SR NO | TEST CASE NAME | TEST CASE DESCRIPTION | EXAMPLE TEST CASE | EXPECTED OUTPUT |
| 1 | Equivalence Partitioning | This divides the input data of a software unit into partitions of equivalent data from which test cases can be derived. | Our dataset can be divided into two classes on the basis of the number mapped with state. The valid partitions are : | If valid: Successful If invalid: Unsuccessful |
| 2 | Boundary Value Analysis | In this we focus on the boundary values of the various valid and invalid inputs. | We want to test the value of the age to be a positive integer. | If valid: Successful If invalid: Unsuccessful |
| 3 | Data input | To check what happens when data is inputted in the wrong format | 1)Instead of giving the input in 0’s and 1’s. The data is fed into the system as alphabets. For example Sex of the person.  2)For thalach If it is 140 (1) = high thalach Else (2) = normal thalach  3)For trestbp If 120 = Normal Else if < 120 = High Blood pressure | 1)The code must throw an error pertaining to change of format of the data inputted.  2)Depending on the value of thalach the testing of data set takes place. If negative value is entered then throw error.  3) If negative value is entered then throw error |

CHAPTER 7

RESULTS AND ANALYSIS

This chapter discusses the comparative analysis of different machine learning towards the prediction of heart disease. In addition we also have give the graphical representation of data set on basis of age, Serum cholesterol, BP and so on .

7.1 DATA SET:

* For e.g.: Our system here, uses a dataset to predict the occurrence of heart diseases by dividing the dataset into test data and training data. 70% data set is used for training and the rest 30%data set are useing for testing. Now, the input used in our case is the attributes we have used that correspond to factors that result in heart diseases. The output would be a binary digit indicating whether a person is susceptible to heart diseases or not. The algorithms used in the system for prediction, are learning algorithms which will always change over periods of time based on various input factors . Therefore, the results might change when we learn even more about the data fed as input.
* https://github.com/Abhayku18113211/Heart-Disease-Prediction-System-Using-ML

Data set used:

* For our predicated system , we have taken dataset that was publicly available on online Kaggle for predicting the heart diseases. The parameters used as input for data analysis using of Machine learning algorithm are as follows:
* 1. age: which has taken in years
* 2. sex: 1 for male, 0 for female
* 3. cp: short for chest pain
* 4. Trest bps : blood pressure taken when the body is resting
* 5. Chol : level of cholesterol
* 6. fasting blood sugar: (1 for true; 0 for false)
* 7. resting electrocardiographic results Value 0: normal Value 1: having the ST-T wave which is not normal Value 2: shows the probability of having left ventricular hypertrophy
* 8. Thalach: the max level of heartbeats are achieved by heart
* 9. Exercise induced angina: 1 for yes and 0 for no
* 10. Oldpeak: the depression by exercise compared to the one at rest
* 11. slope: peak of the ST segment during exercise
* Value 1: no slope
* Value 2: straight line
* Value 3: down sloping
* 12. ca: major vessels (0-3) colored by fluoroscopy
* 13. Thal: 3= perfect; 6=permanently defected; 7=defect can be altered 14. num: heart disease diagnoses.

7.2 METRICS FOR PERFORMANCE ANALYSIS:

• In this project review various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Ada-boost Xg-boost are used to predict heart disease. Heart Disease Predication UCI dataset, has been a total of 76 attributes, out of those only 14 attributes dataset are considered for the prediction of heart disease. Various So many attributes dataset of the patient like sex , chest pain type, fasting blood pressure, serum cholesterol,Slope,fbs Exang etc. are considered for this project. The accuracy for individual algorithms must be measure and which ever algorithm is giving the best accuracy results that is considered for the heart disease prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered.

Accuracy- Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset.

• 1) The accuracy rate of each algorithm has been measured and selects the algorithm with the highest accuracy Results. The accuracy rate is a correct prediction ratio to the total number of given datasets. It is expressed as:

• Accuracy = (TP + TN) /(TP+FP+FN+TN)

• Where:

• TP: True Positive

• TN: True Negative

• FP: False Positive

• FN: False Negative

1)Confusion Matrix::

• It is gives us a matrix as output and gives the total performance of the S system.

• It is a binary classifier. A confusion matrix that can be of any size depending upon the different number of parameters inputted (labels in our case).

• Where:

• TP: True positive

• FP: False Positive

• FN: False Negative

• TN: True Negative

• The Mathematically numerical value of TP, FP, TN, FN defines as:

• TP= The occur when the model accurately predicts a positive data point.

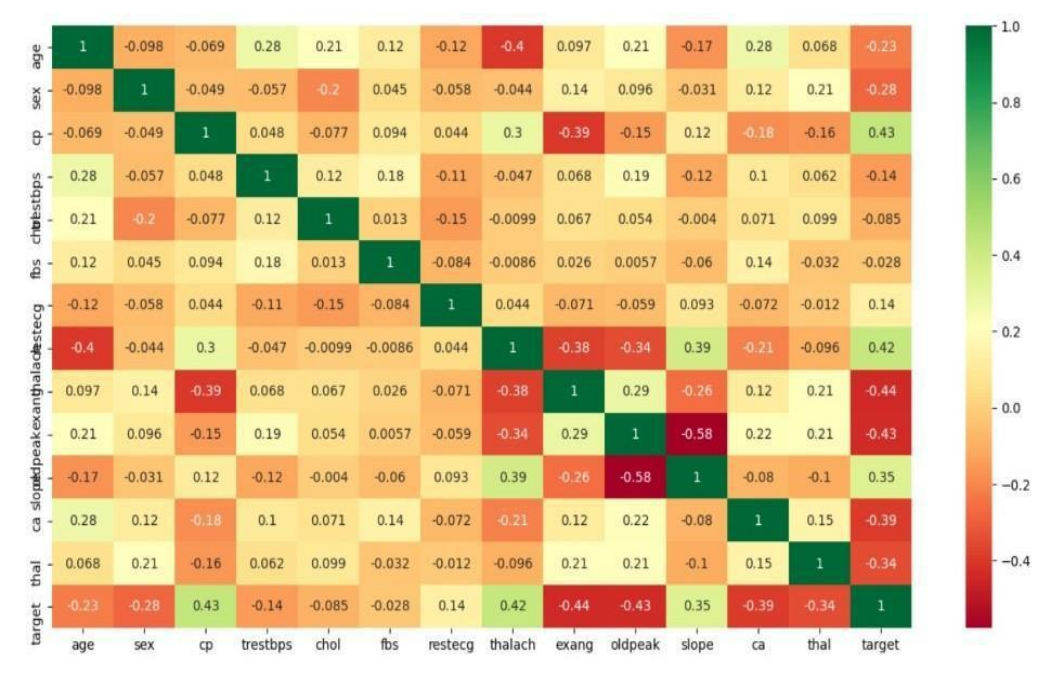
• TN=  The occur when the model accurately predicts a negative data point.

FP= The occur when the model predicts a positive dataset point incorrectly.

FN=  The occur when the model mis predicts a negative data point.

1. Correlation Matrix:

The correlation matrix in machine learning Algorithm that is used for feature selection. And It is representing dependency between various attributes dataset. Selection of attributes dataset: The Dataset of Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. The Various attributes of the patients like sex,age, chest pain type, fasting blood pressure, serum cholesterol, fbs,slope,exang etc. are selected for the prediction. The Correlation matrix are used for attribute selection for this model System.



1. Precision:

• It can be used to determine the ratio of the correct positive results to the total number of positive results predicted by the Predication system.

• 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 = 𝑇𝑃/(𝑇𝑃 + 𝐹𝑃)

• It is expressed as:

• 4) Recall-

It is using the ratio of correct positive results to the total number of positive results are predicted by the Predication system.

• 𝑅𝑒𝑐𝑎𝑙𝑙 = 𝑇𝑃/(𝑇𝑃 + 𝐹𝑁)

• 5) F1 score-:

This metric is taken into a recall and precision model and is calculated as below

It has been in the Precision of a and Recall. It is measuring the test accuracy of Results. And The range of this metrics between 0 to 1.

𝐹1 𝑠𝑐𝑜𝑟𝑒 = 2 ∗ (𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 ∗ 𝑅𝑒𝑐𝑎𝑙𝑙)/(𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛- + 𝑅𝑒𝑐𝑎𝑙𝑙)

7.3 PERFORMANCE OF MACHINE LEARNING ALGORITHMS:

7.3.1 SVM RESULTS:

* Accuracy : 0.901
* Precision : 0.904
* Recall : 0.922
* Fi Score : 0.909

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicted Value |  | TP:47 | TN:35 |
| Predicted Value |  | FP:5 | FN:4 |

Fig No.7.3.1 SVM CONFUSION MATRIX

A screenshot of a computer

Description automatically generated

Fig No. 7.3.1 SVM RESULTS

7.3.2 NAÏVE BAYES RESULTS:

* Accuracy : 0.869
* Precision: 0.900
* Recall : 0.844
* F 1 Score : 0.871

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated values | true | TP:27 | TN:26 |
| Predicated values | false | FP:3 | FN:5 |
|  |  |  |  |

Table 7.3.2 Naïve Bayes Confusion Matrix

A screenshot of a computer

Description automatically generated

Fig 7.3.3 Naïve Bayes Results

7.3.3 KNN RESULT:

ACCURACY: 0.639

RECALL:0.4666

PRECISION:0.76

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value A | Actual Value A |
|  |  | Positive | Negative |
| Predicated Values | TRUE | TP:14 | TN:16 |
| Predicated Values | FALSE | FP:6 | FN:25 |

Table 7.3.3. KNN Confusion Matrix

A screenshot of a computer program

Description automatically generated

Fig 7.3.3. KNN Results:

7.3.4 DECISION TREE RESULTS

ACCURACY: 0.8033

RECALL: 0.7666

PRECISION: 0.821

F1 score :0.271

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated Values: | TRUE | TP:23 | TN:7 |
| Predicated Values: | FALSE | FP:5 | FN:26 |
|  |  |  |  |

Table 7.3.4 Decision tree Confusion Matrix

A screenshot of a computer

Description automatically generated

Fig 7.3.4 Decision tree results:

7.3.5 RANDOM FOREST RESULTS:

* ACCURACY: 0.803
* RECALL:0.8 04
* F1Score 0.832
* PRECISION:0.860

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated Values | TRUE | TP:24 | TN:6 |
| Predicated Values | FALSE | FP:9 | FN:37 |

Table 7.3.5 Random forest Confusion Matrix:

A screenshot of a computer

Description automatically generated

Fig 7.3.5Random Forest Results:

7.3.6 LOGISTIC REGRESSION RESULTS:

* ACCURACY: 0.829
* RECALL:0.907
* PRECISION:0.813
* F1 Score: 0.856

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated Values: | TRUE | TP:24 | TN:9 |
| Predicated Values: | FALSE | FP:4 | FN:39 |
|  |  |  |  |

Table 7.3.6 Logistic regression Confusion Matrix:

A screenshot of a computer

Description automatically generated

Fig 7.3.6 Logistic Regression Results:

7.3.7 ANN RESULTS:

ACCURACY: 0.8852

RECALL: 0.8666

PRECISION:0.8965

F1 score :0.84699

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated Values | TRUE | TP:26 | TN:4 |
| Predicated Values | FALSE | FP:3 | FN:28 |

Table 7.3.7 ANN Confusion Matrix:

A screenshot of a computer program

Description automatically generated Fig 7.3.7 ANN Results:

7.3.8 DNN RESULTS:

ACCURACY: 0.918

RECALL: 0.9033

PRECISION: 0.9655

F1 score :0.5320

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated values: | TRUE | TP:28 | TN:3 |
| Predicated Values: | FALSE | FP:1 | FN:28 |

Table 7.3.8 DNN Confusion Matrix:

A screenshot of a computer code

Description automatically generated

Fig 7.3.8 DNN:

7.3.9 ADA-BOOST RESULTS:

* Accuracy:0.829
* Precision:0.860
* Recall:0.841
* F1 Score:0.849

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated Values: | TRUE | TP:26 | TN:6 |
| Predicated Values: | FALSE | FP:7 | FN:37 |

Fig 7.3.9 Ada-boost confusion matrix:

A screenshot of a computer

Description automatically generated

Fig 7.3.9 Ada-boost Results:

7.3.10 XG-BOOST RESULTS:

* Accuracy: 0.918
* Precision:0.960
* Recall:0.857
* F1 Score:0.905

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual Value | Actual Value |
|  |  | Positive | Negative |
| Predicated Values | TRUE | TP:32 | TN:1 |
| Predicated Values | FALSE | FP:4 | FN:24 |

Fig 7.3.10 Xg- boot confusion matrix

A screenshot of a computer

Description automatically generated

Fig 7.3.10 Xg-boost Results:

7.4 RESULTS FROM THE TEST DATA SET:

* We have done extrapolation of data set in regards to heart disease on basis of various factors which are Sex, Age, Blood pressure, Serum cholesterol and so.

A close-up of a pie chart

Description automatically generated

Fig no. 7.4.1 Percentage of Men and Women in our Dataset

Fig no. 7.4.1 Risk of men and women having a heart attack

ANALYSIS OF POSITIVE HEART ATTACK IN MEN OUT OF TOTAL MEN AND WOMEN OUT OF TOTAL WOMEN:

A diagram of a heart attack

Description automatically generated

Fig no.7.4.2 Analysis of Positive Heart Attack in Men Out of total men

Fig no. 7.4.2 Analysis of Positive Heart Attack in Women Out of Total women

Distribution of people vs age:

A comparison of a graph

Description automatically generated with medium confidence

Fig No. 7.4.3 Distribution of People Vs Age

Fig No. 7.4.3 Age

Distribution serum cholesterol in mg/dl:

A comparison of a normal distribution graph

Description automatically generated

Fig no. 7.4.4 Distribution serum cholesterol in mg/dl

Fig no .7.4.4 Cholesterol

Distribution of maximum heart rate achieved:

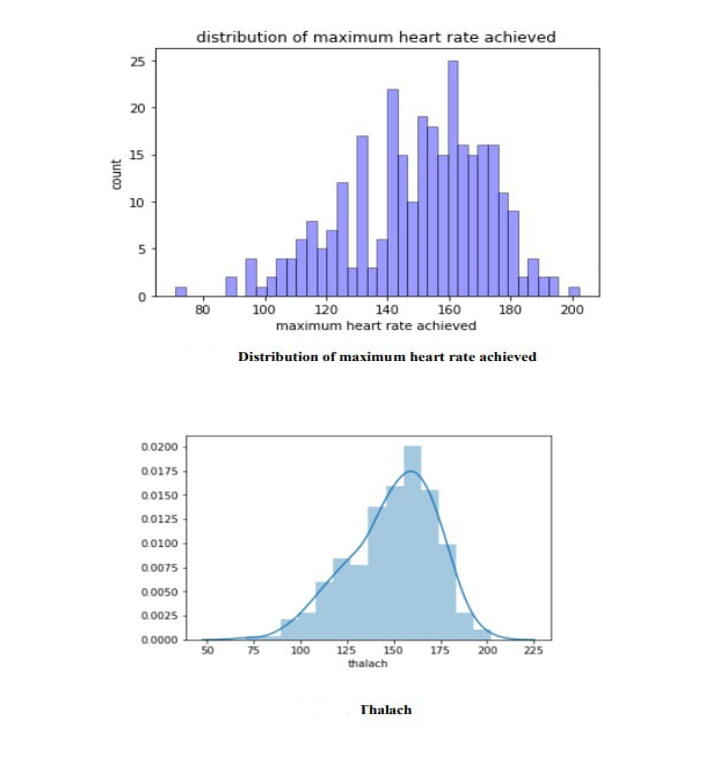


Fig no. 7.4.5 Distribution of maximum heart rate achieved.

Fig no. 7.4.5 thalach

Distribution of resting blood pressure:

A comparison of blood pressure and distribution of blood pressure

Description automatically generated

Fig no.7.4.6 Distribution of resting blood pressure:

Fig no.7.4.6 Trestbps

7.5 Result and Discussion: INPUT AND OUTPUT GUI:

* Here we are going to create and predict the patients details through GUI for our project. So, anyone can perform predictions using this GUI.

INPUT:

A screenshot of a computer

Description automatically generated

Fig No.7.5.1

OUTPUT:

A screenshot of a computer

Description automatically generated

Fig no.7.5.2

A screenshot of a computer

Description automatically generated

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 CONCLUDING REMARKS:

• Heart diseases is a major silent killer in the India and throughout the world, and it’s the application of promising technology like machine learning algorithms is to be the initial prediction system of heart diseases will be to a profound impact on society. The early prognosis of heart disease can be aid in making decisions of lifestyle changes in a high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise every year. This prompts for its early diagnosis and treatment also. The utilization of suitable technology support in this regard can be prove to be highly beneficial to the medical fraternity and patients. In this paper, the eleven different machine learning algorithms used to measure the performance are SVM, Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Adaptive Boosting, and XG Boost ANN , KNN, Linear Regression, DNN Extreme gradient boosting are applied on the dataset. The expected attributes leading to heart disease in patients are available in the dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features Attributes Parameters has been taken into the consideration, then the efficiency of the system the author gets is less. To increase are efficiency, attribute selection is done. In this n features have to be selected for evaluating the model which are gives more accuracy results. In this The correlation method of some features in the dataset is almost equal and so they are removed. If all the attributes dataset present in the dataset are taken into account, then the efficiency decreases considerably. All the Eleven machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to be use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all Eleven the extreme gradient boosting classifier gives the highest accuracy of 95.08%

8.2 FUTURE WORK:

The work done here trains the system with a limited number of datasets. The machine learning algorithms become more accurate once they are fed with a huge number of data sets. So, this system can be trained with a huge number of data sets that would increase the accuracy in predicting the heart diseases. The analysis part of the system is done, and in order to be more useful can be integrated with electronic systems which give the system real-time inputs and would help the patient get the results then and there. There are multiple combinations of algorithms that can be tested against these data sets in order to yield better results.

For the Future Scope so many more machine learning algorithms approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness regarding the diseases. The analysis part of the system is done, and to be more useful can be integrated with electronic systems which give the system real-time inputs and would help the patient get the results then and there. There are multiple combinations of algorithms that can be tested against these data sets in order to yield better results.

REFERENCES:

* [1] Amines Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, “Heart Disease Prediction Using Machine Learning and Data Mining Techniques: A Review”, Research Gate Publications, July 2017, pp.2137-2149
* [2] V. Krishnaiah, Narsimha, Heart Disease Prediction System using Machine learning Techniques and it is utilizing the Intelligent of The Fuzzy Approach methods way in a Review paper International Journal based paper of Computer are Applied in Applications, 14 February 2017
* [3] Guizhou Hu, Martin M. Root, "Building heart disease Prediction Models for Coronary artery Heart Disease by Synthesizing Multiple Longitudinal Research Findings”, European Science of Cardiology, 11 May 2005.
* [4] Mythili, Dev Mukherji, Nikita Padilla and Abhi ram Naidu, “A Heart Disease Prediction Model using SVM- Decision Trees- Logistic Regression (SDL)”,International Journal of Computer Applications, vol.68,15 April 2013
* [5] Nimai Chand Das Adhikari, Arpana Alka, and Rajat Garg, “HPPS: Heart disease Problem Prediction System can be used to Machine Learning Algotithms”.
* [6] K. Polaraju, D. Durga Prasad, “Heart Disease Predication of using Multiple Linear Regression Model", International Journal of Engineering Development and Research Development, ISSN:2322-9939, 2017.
* [7] Marjia Sultana, Afrin Haider, “Heart Disease Prediction model using WEKA tool and 10-Fold cross-validation”, The Institute of Electrical and Electronics Engineers, 18 March 2017.
* 8] Dr.S.Seema Shedole, Kumari Deepika, “Predictive analytics to prevent and control chronic of heart disease”, https://www.researchgate.net/punlication/316530781,January 2016.
* [9] Ashok kumar Dwivedi, “To Evaluate the performance of different types of machine learning algorithms techniques can be used for prediction of heart disease using ten-fold And cross-validation 16 September 2016.
* [10] Megha Shahi, R. Kaur Gurm, “Heart Disease Prediction System using Data Mining Techniques and Machine Learning Algorithms Techniques ”, Orient J. Computer Science Technology, vol.5 2018, pp.467-466
* 11] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, “ The Efficient Problem of Heart Disease Prediction System”, 2017, pp.962-969

APPENDIX

SIMPLE CODE:

* https://github.com/Abhayku18113211/Heart-Disease-Prediction-System-Using-ML

## import numpy as np

## import pandas as pd

## import matplotlib.pyplot as plt

## import re

## train = pd.read\_csv("..heart\_disease\_data.csv")

## # loading data from heart.csv

## train.dropna()

## # deletes rows with empty cells

## # Percentage of men and women in our dataset

## label = 'Men','Women' sizes = [0.6831, 0.3168] colors = ['gold','lightskyblue'] explode = (0.1, 0) # explode 1st slice

# Plot

plt.pie(sizes, explode=explode, labels=label, colors=colors, autopct='%1.1f%%', shadow=True, startangle=140) plt.title('% of men and women in our dataset\n\n\n') plt.axis('equal') plt.show()

# Plot of men at risk vs women at risk

label = 'risk of Men of having heartattack','risk of women having heartattack' sizes = [(0.449), (0. 75)] colors = ['gold','lightskyblue'] explode = (0.08, 0) # explode 1st slice

# Plot

plt.pie(sizes, explode=explode, labels=label, colors=colors, autopct='%1.1f%%', shadow=True, startangle=0) plt.axis('equal') plt.show()

# percentage of men at risk

slices\_hours = [207-93, 93] activities = ['% Men not at risk', '% Men at risk'] colors = ['r', 'g','b',' y'] plt.pie(slices\_hours, labels=activities, colors=colors, startangle=90, autopct='%.1f%%') plt.ti tle('ANALYSIS OF POSITIVE HEART ATTACK IN MEN OUT OF TOTAL MEN') plt.show( )

# percentage of women at risk

slices\_hours = [96-72, 72] activities = ['% Women not at risk', '% Women at risk'] colors = ['y','b' ] plt.pie(slices\_hours, labels=activities, colors=colors, startangle=0, autopct='%.1f%%') plt.title ('ANALYSIS OF POSITIVE HEART ATTACK IN WOMEN OUT OF TOTAL WOMEN') plt.s how()

# age distribution of people affected

import seaborn as sns

# seaborn histogram

sns.distplot(train['age'], hist=True, kde=False, bins=int(180/5), color = 'blue', hist\_kws={'edgeco lor':'black'})

# Add labels

plt.title('distribution of people count vs age') plt.xlabel('age(in years)') plt.ylabel('number of pers on') plt.show() sns.distplot(train['age'])

# distribution of cholesterol level in affected people

sns.distplot(train['chol'], hist=True, kde=False, bins=int(200/5), color = 'blue', hist\_kws={'edgec olor':'black'}) plt.title('distribution serum cholesterol in mg/dl') plt.xlabel('serum cholesterol in m g/dl') plt.ylabel('count') plt.show() sns.distplot(train['chol'])

# max heart rate distribution

sns.distplot(train['thalach'], hist=True, kde=False, bins=int(200/5), color = 'blue', hist\_kws={'ed gecolor':'black'}) plt.title('distribution of maximum heart rate achieved') plt.xlabel('maximum hea rt rate achieved') plt.ylabel('count') plt.show() sns.distplot(train['thalach'])

# resting heart rate distribution

sns.distplot(train['trestbps'], hist=True, kde=False, bins=int(100/5), color = 'blue', hist\_kws={'ed gecolor':'red'}) plt.title('distribution of resting blood pressure (in mm Hg on admission to the hos pital)') plt.xlabel('resting blood pressure (in mm Hg on admission to the hospital)') plt.ylabel('cou nt') plt.show() sns.distplot(train['trestbps'])

# IMPORTING REQUIRED LIBRARIES

from sklearn.linear\_model import LogisticRegression from sklearn.svm import SVC from sklear n.tree import DecisionTreeClassifier from sklearn.ensemble import RandomForestClassifier fro m sklearn.neighbors import KNeighborsClassifier from sklearn.naive\_bayes import GaussianNB

# SPLITTING DATA

from sklearn.model\_selection import train\_test\_split train1 = train.copy() feature\_df = train1[["a ge","sex","cp","trestbps","thalach","chol","restecg","exang","oldpeak","slope","ca","thal"]] x = n p.asarray(feature\_df) y = np.asarray(train["target"].astype('int')) x\_train,x\_test,y\_train,y\_test = tr ain\_test\_split(x,y,test\_size = 0.2,random\_state = 5)

def train\_model(x\_train, y\_train, x\_test, y\_test, classifier, \*\*kwargs)

# Instantiate model

model = classifier(\*\*kwargs) # train model model. Fit(x\_train,y\_train) y\_pred = model. Predict (x\_ test) from sklearn.metrics import confusion matrix print("Confusion matrix") print(confusion matrix “)print (confusion\_m atrix(y\_test,y\_pred))

# check accuracy and print out the results

Fit\_ accuracy = model. Score (x\_train, y\_train) test\_accuracy = model.score(x\_test, y\_test)

print(f" Train accuracy: {fit\_accuracy:0.2%}") print(f"Test accuracy: {test\_accuracy:0.2%}") return m model

# KNN

model= train\_model(x\_train, y\_train, x\_test, y\_test, KNeighborsClassifier)

# Seek optimal 'n\_neighbours' parameter

for i in range(1,10): print("n\_neigbors = "+str(i)) train\_model(x\_train, y\_train, x\_test, y\_test, K NeighborsClassifier, n\_neighbors=i)

# Decision Tree

model = train\_model(x\_train, y\_train, x\_test, y\_test, DecisionTreeClassifier, random\_state=2606 )

# Check optimal 'max\_depth' parameter

for i in range (1,8): print("max\_depth = "+str(i)) train\_model(x\_train, y\_train, x\_test, y\_test, Deci sionTreeClassifier, max\_depth=i, random\_state=2606)

# LOGISTIC REGRESSION

model = train\_model(x\_train, y\_train, x\_test, y\_test, LogisticRegression)

#Gaussian Naive Bayes

model = train\_model(x\_train, y\_train, x\_test, y\_test, GaussianNB)

# Support Vector Machines

model = train\_model(x\_train, y\_train, x\_test, y\_test, SVC)

# Support Vector Machine Linear

model = train\_model(x\_train, y\_train, x\_test, y\_test, SVC, C=0.05, kernel='linear')

# Random Forests

model = train\_model(x\_train, y\_train, x\_test, y\_test, RandomForestClassifier, random\_state=26 06)

# USING ANN Multilayer Perceptron Classifier

from sklearn.metrics import make\_scorer, accuracy\_score from sklearn.model\_selection import GridSearchCV from sklearn.neural\_network import MLPClassifier ann\_clf = MLPClassifier() pa rameters = {'solver': ['lbfgs'], 'alpha':[1e-4], 'hidden\_layer\_sizes':(9,14,14,2),'random\_state':[1]} a cc\_scorer = make\_scorer(accuracy\_score) grid\_obj = GridSearchCV(ann\_clf, parameters, scorin g=acc\_scorer) grid\_obj = grid\_obj.fit(x\_train, y\_train) ann\_clf= grid\_obj.best\_estimator\_ ann\_clf.fit(x\_train,y\_train)

y\_pred\_ann = ann\_clf.predict(x\_test) from sklearn.metrics import confusion\_matrix cm\_ann = c onfusion\_matrix(y\_test, y\_pred\_ann) cm\_ann

ann\_result = accuracy\_score(y\_test, y\_pred\_ann) ann\_result

FOR TEST CASES:

import pandas.

import os

import numpy as np

from sklearn.linear\_model import Linear Regression

from sklearn.cross\_validation import KFold

from sklearn import cross\_validation

from sklearn.linear\_model import Logistic Regression

heart = pandas.read\_csv("heart\_disease\_data.csv")

heart.loc[heart["heartpred"]==2,"heartpred"]=1

heart.loc[heart["heartpred"]==3,"heartpred"]=1

heart.loc[heart["heart pred"]==4,"heartpred"]=1

heart["slope"] = heart["slope"].fillna(heart["slope"].median())

heart["thal"] = heart["thal"].fillna(heart["thal"].median()) heart["ca"] = heart["ca"].fillna(heart["ca"].median())

print (heart. Describe()) predictors=["age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slop e","ca","thal"]

alg=Logistic Regression(random\_state=1)

predictions = []

train\_predictors = (heart[predictors].iloc[:,:])

train\_target = heart["heartpred"].iloc[:]

alg.fit(train\_predictors, train\_target)

l=len(heart.index)

while True:

x1=raw\_input("Input age (67.0):")

x2=raw\_input("Input sex (1.0):")

x3=raw\_input("Input cp (4.0):")

x4=raw\_input("Input trestbps (160.0):")

x5=raw\_input("Input chol (286.0):")

x6=raw\_input("Input fbs (0.0):")

x7=raw\_input("Input restecg (2.0):")

x8=raw\_input("Input thalach (108.0):")

x9=raw\_input("Input exang (1.0):")

x10=raw\_input("Input oldpeak (1.5):")

x11=raw\_input("Input slope (2.0):")

x12=raw\_input("Input ca (3.0):")

x13=raw\_input("Input thal (3.0):")

f = open('heart\_disease\_data.csv')

f. write (x1+','+x2+','+x3+','+x4+','+x5+','+x6+','+x7+','+x8+','+x9+','+x10+','+x11+','+x1 2+','+x13+',')

f.close()

heart = pandas.read\_csv("heart\_disease\_data.csv")

#print(l+1)

print(heart[predictors].iloc[l,:])

test\_predictions = alg.predict(heart[predictors].iloc[l,:])

test\_predictions[test\_predictions > .5] = 1

test\_predictions[test\_predictions <=.5] = 0

print(test\_predictions.item(0))

x14=test\_predictions.item(0)

print(type(x14))

f = open('heart\_disease\_data.csv')

f.write("%s"%(x14))

f.write(os.linesep)

f.close()

l+=1