Predicting Heart Disease using Machine Learning

Submitted in partial fulfillment of the requirements

for the degree of

Bachelor of Engineering

by

Sushant Kumar Mishra

Roll No.40

Shubham Ingle

Roll No.20

Sarthak Jadhav

Roll No.23

Under the Supervision of

Prof. Poonam B. Lad

DEPARTMENT OF INFORMATION TECHNOLOGY

KONKAN GYANPEETH COLLEGE OF ENGINEERING

KARJAT-410201

May 2021

Certificate

This is to certify that the project entitled **Predicting Heart Disease using Machine Learning** is a bonafide work of **Shubham Ingle (Roll No. 20),Sarthak Jadhav(Roll No.23) and Sushant Kumar Mishra(Roll No.40)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Undergraduate** in **DEPARTMENT OF INFORMATION TECHNOLOGY**.

Prof. Poonam B. Lad

Assistant Professor

Department of Information Technology

Dr. M. J. Lengare

Head of Department

Department of Information Technology

Principal

Konkan Gyanpeeth College of Engineering

Project Report Approval for B.E.

This project report entitled **Predicting Heart Disease using Machine Learning** by **Shubham Ingle (Roll No. 20),Sarthak Jadhav(Roll No.23) and Sushant Kumar Mishra(Roll No.40)** is approved for the degree of **DEPARTMENT OF INFORMATION TECHNOLOGY**.

Examiners

1.....

2.....

Date.

Place.

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature
Sushant Kumar Mishra (Roll No 40)

Signature
Sarthak Jadhav (Roll No 23)

Signature
Shubham Ingle (Roll No 20)

Date.

Abstract

Predicting and detection of heart disease has always been a critical and challenging task for healthcare practitioners. Hospitals and other clinics are offering expensive therapies and operations to treat heart diseases. So, predicting heart disease at the early stages will be useful to the people around the world so that they will take necessary actions before getting severe. Heart disease is a significant problem in recent times; the main reason for this disease is the intake of alcohol, tobacco, and lack of physical exercise. Over the years, machine learning shows effective results in making decisions and predictions from the broad set of data produced by the health care industry. Some of the supervised machine learning techniques used in this prediction of heart disease are Logistic Regression , random forest (RF), ,knearest neighbour algorithm. Furthermore,system made on the performances of these algorithms is summarized

Acknowledgements

We wish to express our profound and sincere gratitude to **Prof. P. B. Lad**, Department Information Technology, KGCE, Karjat, who guided us into the intricacies of this project with matchless magnanimity. We thank , Head of the Dept. of Information Technology, KGCE Karjat and **Dr. M. J. LENGARE**, Principal, KGCE Karjat for extending their support during the Course of this investigation. We would be failing in our duty if we don't acknowledge the co-operation Rendered during various stages of image interpretation by. We are highly grateful to who evinced keen interest and invaluable support in the progress and successful completion of our project work. We are indebted to for their constant encouragement, co-operation and help. Words of Gratitude are not enough to describe the accommodation and fortitude which they have shown throughout my endeavor.

Contents

List of Figures

List of Tables

Abbreviations

CVD	**C**ardio **V**ascular **D**iseas
EDA	**E**xploratory **D**ata **A**nalysis
CAD	**C**oronary **A**rtery **D**iseas

Chapter 1

INTRODUCTION

1.1 Introduction

The heart is one of the main parts of the human body after the brain. The primary function of the heart is to pumping blood to the whole body parts. Any disorder that can lead to disturbing the functionality of the heart is called heart disease. Several types of heart disease are there in the world; CAD, HF are the most common heart diseases that are present. The main reason behind the coronary heart disease CAD is blockage or narrowing down of the coronary arteries Coronary arteries also responsible for supplying blood to the heart. CAD is the leading cause of death over 26 million people are suffering from coronary heart disease(CAD) around the world, and it is increasing 2% annually . In the growing world, 2% of the population around the world is suffering from CAD, and 10% of the people are older than 65 years. Approximately 2% of the annual healthcare budget spent only to treat CAD disease.

Heart disease is the leading cause of death among all others diseases ,even cancers.One in 4 deaths in India are now because of CVDs with ischemic heart disease and stroke.The diagnosis is often made, based on doctor's intutions and experience ,this may lead to an unwanted result and excessive medical cost. Heart disease is a significant issue, so there is a need for diagnosis or prediction of heart disease there areseveral methods to diagnose heart disease among them. Angiography is the trending method which is used by most

of the physicians across the world. However, there are some drawbacks associated with angiography technique. It is an expensive procedure and physicians have to analyze so many factors to diagnose a patient hence this process makes physician job very difficult, so these limitations motivate to develop a non-invasive method for prediction of heart disease. These conventional methods deal with medical reports of the patients moreover these conventional methods are time-consuming, and it may give erroneous results because these conventional methods are performed by humans. To avoid these errors and to achieve better and faster results, we need an automated system. Over the past years, researchers find out that machine learning algorithms perform very well in analyzing medical data sets. These data sets will be directly given to machine learning algorithms, and machine learning algorithms will perform according to their nature, and those algorithms will give some outputs.

1.2 Objectives

Objectives of this project is :

1. to go through the data science lifecycle steps in order to build a heart disease classification web application by using a historical dataset
2. to use flask API to deploy the model and build the web application
3. to Help avoid human biasness.
4. to Reduce the cost of medical tests.

1.3 Purpose, Scope, and Applicability

Purpose, Scope and Applicability: The description of Purpose, Scope, and Applicability are given below:

1.3.1 Purpose

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms.The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data

1.3.2 Scope

Machine learning techniques have been around us and has been compared and used for analysis for many kinds of data science applications.Heart disease is the leading cause of death among all others diseases ,even cancers.One in 4 deaths in India are now because of CVDs with ischemic heart disease and stroke.Prediction of heart diseases is a difficult and risky task. Since it is directly dependent on people's health, accuracy is a major factor. If not predicted accurately it can be disastrous.This project therefore focuses on the comparison of different data mining techniques to predict it. It shows the comparative analysis of the different methods. Cross validation error is used to compare the techniques. We choose Logical Regression, Random forest, K-Nearest Neighbors,as they are the most widely used techniques in determining diseases.

1.3.3 Applicability

Machine learning is widely used now a days in many business applications like e commerce and many more. Prediction is one of area where this machine learning used, our topic is about prediction of heart disease by processing patient's data set and a data of patients to whom we need to predict the chance of occurrence of a heart disease

1.4 Achievements

Achievements: Explain what knowledge you achieved after the completion of your work. What contributions has your project made to the chosen area? Goals achieved describe the degree to which the findings support the original objectives laid out by the project. The goals may be partially or fully achieved, or exceeded.

1.5 Organisation of Report

In introduction we studied about basic over view of the project in which we learned various outcomes of the project. how the project is going to the implemented what are the prerequisites require to develop the project etc.

In literature survey we will see various presents system the which are previously developed by various experts and we had compared system which are previously design.

in survey of technologies we will see various technologies which can be use to complete the system.

In requirements and analysis we will analyse various requirements pre-quest for developing the system also see various functional requirements software requirements and hardware requirements of the system.

In system design we see basic overall design of the system.

In the implementation chapter we will see how the system is implemented.This chapter also contain implementation strategies and test strategies

In Results we will see test results which we have been obtained by testing various algorithms and methods.

In final chapters we will see a conclusion of the project it also describes limitations and future scope of the system.

Chapter 2

LITERATURE SURVEY

Literature Survey : Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis and achieved different probabilities for different methods.

2.0.1 An Optimized Stacked Support Vector Machines Based Expert System for The Effective Prediction Of Heart Failure

Authors: Liaquat Ali et al

Year: 2019

This paper recommended a model which is consists of two methods one is X^2 statistical and deep neural network(DNN). Feature refinement is done by X^2 statistical model and classification is done by a deep neural network(DNN). In their study, they have used the Cleveland dataset. There are 303 instances in that dataset, among them, 297 have no missing data, and the remaining 6 have missing data. Among 297, 207 instances are used for training data, and the remaining 90 are used as testing data. This model gives better results compared to conventional ANN models which are present earlier. As a result of this using this proposed model, they have got 93.33 % classification accuracy using DNN. It is 3.33 percent more than that of the conventional ANN model.

2.0.2 Effective Heart Disease Prediction System

Authors: Dr. Kanak Saxena et.al

Year: 2016

This paper developed a data mining model to predict heart disease efficiently. It mainly helps the medical practitioners to make efficient decisions way based on the given parameters. The author has used Cleveland dataset from UCI, and they have used age, sex, resting blood pressure, chest pain, serum cholesterol, fasting blood sugar, etc. as attributes. Furthermore, they have divided the datasets into two parts one is for testing, and the other one is for training. They have used a 10-fold method to find accuracy

2.0.3 An System Based On Support Vector Machines For Effective Dignosis Of Heart Disease

Authors : Awais Nimat et al

Year :2016

This paper proposed an expert system based on two support vector machines(SVM) to predict heart disease efficiently. These tow SVM's have their purpose; first, one is used to remove the unnecessary features, and the second one is used for prediction. Moreover, they have used the HGSA (hybrid gird search algorithm) to optimize the two methods. By using this model, they have achieved 3.3 % better accuracy than the conventional SVM models that are present earlier.

2.0.4 An Intelligent Learning System Based O Random Search Algorithm And Optimized Random Forest Model For Improved Heart Disease Detection

Authors : Ashir Javeed et al

Year :2019

This paper developed a model to improve the prediction of heart disease by overcoming the problem of overfitting; overfitting means the proposed model performs and gives better

accuracy on testing data and gives unfortunate accuracy result for training data while predicting the heart disease. To solve this problem, they have developed a model that will give the best accuracy on both training and testing data. That model consists of two algorithms one is RAS (Random search algorithm) other one is a random forest algorithm that is used to predict the model. This proposed model gave them better results in training data as well as testing data.

Paper Comparison

S.R. NO.	**Authors**	**Techniques Used**	**Accuracy**
1	Liaqat Ali et al.	X^2 statistical model, deep neural network	93.33%(holdout) 91.57%(k-fold
2	Dr.Kanak Saxena et.al	Decision tree	86.3% (testing phase) 87.3% (training phase)
3	Awais Nimat et al.	Support vector machine, Hybrid grid search algorithm (HGSA)	92.22% (L1 linear SVM+L2 linear & RBF SVM)
4	Ashir Javeed et al.	Random search algorithm (RSA), Random forest.	93.33% (RSA+RF)

TABLE 2.1: Literature survey comparison

Chapter 3

SURVEY OF TECHNOLOGIES

In this chapter we discuss about some of the methodologies that are related that are related to subject of our project

The main goal of the entire project is to predict heart disease occurrence with the highest accuracy. In order to achieve this, we have chosen three classification algorithms This section includes all brief information about these algorithms

3.0.1 Random Forest

Random Forest is a supervised learning algorithm. Random forest can be used for both classification and regression problems, by using random forest regressor we can use random forest on regression problems. But we have used random forest on classification in this project so we will only consider the classification part.

Random Forest pseudocode

1. Randomly select “k” features from total “m” features. Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until “l” number of nodes has been reached.

Random forest prediction pseudocode

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
2. Calculate the votes for each predicted target
3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

3.0.2 Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

Types of logical regression:

1. Binary (Pass/Fail)
2. Multi (Cats, Dogs, Sheep)

Sigmoid function

$S(z) = 1/1 + e^z$

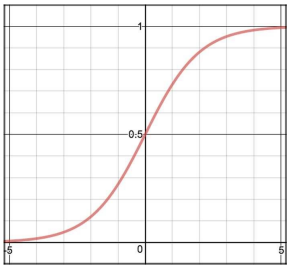

FIGURE 3.1: sigmoid function

Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line. On

the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

Procedure

1. Divide the problem into n+1 binary classification problem (+1 because the index starts at 0).
2. For each class...
3. Predict the probability the observations are in that single class.
4. prediction =< math>max (probability of the classes).

3.0.3 K-Nearest Neighbor

We can implement a KNN model by following the below steps:

1. Load the data
2. Initialize the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points

Procedure

- Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine,etc.
- Sort the calculated distances in ascending order based on distance values.
- Get top k rows from the sorted array.

- Get the most frequent class of these rows.
- Return the predicted class.

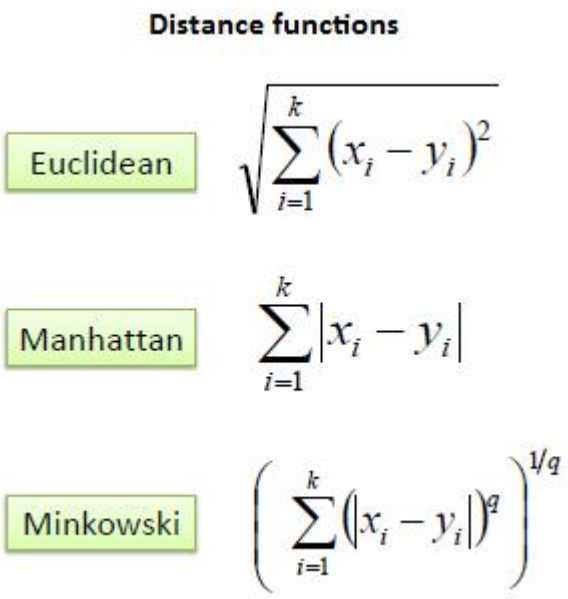

FIGURE 3.2: distance functions

Chapter 4

REQUIREMENTS AND ANALYSIS

4.1 Problem Definition

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience,time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

4.2 Requirements Specification

we will need a historic data set to study its hidden patterns and to train the models which we are going to use for prediction.and then after building the user interface of the

system ,users will need health related reports containing blood sugar level,serum cholesterol,number of major vessels effected ,maximum heart rate achieved etc.

4.3 Planning and Scheduling

Planning and scheduling is a complicated part of software development. Planning, for our purposes, can be thought of as determining all the small tasks that must be carried out in order to accomplish the goal. Planning also takes into account, rules, known as constraints, which, control when certain tasks can or cannot happen. Scheduling can be thought of as determining whether adequate resources are available to carry out the plan. You should show the Gantt chart and Program Evaluation Review Technique (PERT).

4.4 Software and Hardware Requirements

Hardware Requirement:

- 1.5 gigahertz (GHz) dual-core C.P.U
- 4 GB RAM
- 1024x768 minimum screen resolution
- 10GB Of hard disk space

Software Requirements:

- A miniconda environment
- jupyter notebook
- Python 3.8 (recommended)
- Creatly and Star ml

4.5 Conceptual Models

4.5.1 Data flow diagram

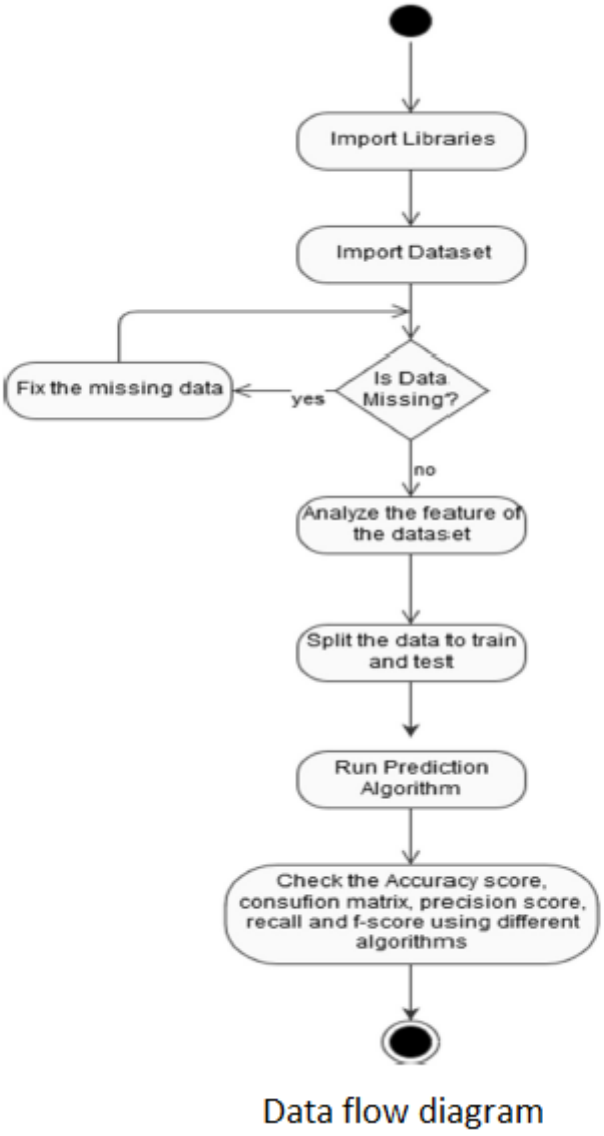

FIGURE 4.1: DFD

Chapter 5

SYSTEM DESIGN

section purpose system

- In this system we are implementing effective heart attack prediction system . We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input. After accessing data set the operation is performed and effective heart attack level is produced.
- The proposed system will add some more parameters significant to heart attack with their weight, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.
- **Data Set Information:** The original data came from the Cleveland database from UCI Machine Learning Repository. However , we've downloaded it in a formatted way from Kaggle. The original database contains 76 attributes, but here only 14 attributes will be used. Attributes (also called features) are the variables what we'll use to predict our target variable. Attributes and features are also referred to as independent variables and a target variable can be referred to as a dependent variable. We use the independent variables to predict our dependent variable.

5.1 Data Design

5.1.1 Schema Design

we are using data set from cleaveland database from UCI machine learning Repository..we are goingto use 14 important attributes from total 76 attributes from database .out of 14 attributes 1 is our target attribute. these attributes are as follows:

- age: age in years.
- sex: sex (1 = male; 0 = female).
- cp: chest pain type (Value 0: typical angina; Value 1: atypical angina; Value 2: non-anginal pain; Value 3: asymptomatic).
- trestbps: resting blood pressure in mm Hg
- chol: serum cholestoral in mg/dl.
- fbs: fasting blood sugar ¿ 120 mg/dl (1 = true; 0 = false).
- restecg: resting electrocardiographic results (Value 0: normal; Value 1: having ST-T wave abnormality; Value 2: probable or definite left ventricular hypertrophy).
- thalach: maximum heart rate achieved.
- exang: exercise induced angina (1 = yes; 0 = no).
- oldpeak: ST depression induced by exercise relative to rest.
- slope: the slope of the peak exercise ST segment (Value 0: upsloping; Value 1: flat; Value 2: downsloping).
- ca: number of major vessels (0-3) colored by flourosopy.
- thal: thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect).
- target: heart disease (1 = no, 2 = yes).

Chapter 6

CONCLUSIONS

6.1 Conclusion

After researching through various papers related to Heart Disease Prediction. We have concluded that a system can be developed that can predict if anyone has chances of heart disease based on some parameters . Various experiments had been conducted using different methodologies, the best results are seen in the methods that are based on various algorithms. Hence looking at the results, we have decided to take the same approach for developing our system. We will test our system against benchmark datasets and compare our results based on accuracy, error, and efficiency of the system.

Bibliography

[1] A. Khan M. Zhou A. Javeed L. Ali, A. Rahman and J. A. Khan. Automated diagnostic system for heart disease prediction based on 2 statistical model and optimally configured deep neural network.

[2] K. Saxena Purushottam and R. Sharma. Efficient heart disease predict ion system," procedia comput. sci. 85, pp. 962–969,2016, 2016.

[3] L. Ali et al. An optimized stacked support vector machines based expert system for the effective prediction of heart failure. 7, pp. 54007–54014, 2019, 2019.

[4] A. Javeed I. Qasim A. Noor A. Javeed, S. Zhou and R. Nour. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection.