HEART DISEASE DETECTION USING MACHINE LEARNING

S.DHARANIDHAR SAHOO

Regd.No: 1701289273

Department of Information Technology

TRIDENT ACADEMY OF TECHNOLOGY

Bhubaneswar-751024,Odisha,India.

May2021

Project Report on

HEART DISEASE DETECTION USING MACHINE LEARNING

Submitted in Partial Fulfillment of the Requirement for the Award of the Degree of

BachelorofTechnologyin

InformationTechnology

Submittedby

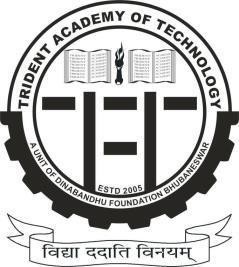
S.DHARANIDHAR SAHOO

Regd.No:1701289273

UndertheGuidanceof

Dr.BiswaranjanNayak

Asst.Professor,Dept.ofCSE

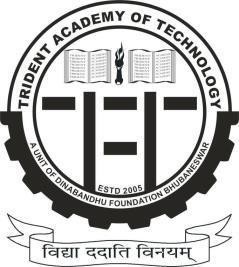


Department of Information Technology

TRIDENT ACADEMY OF TECHNOLOGY

Bhubaneswar-751024,Odisha,India.

May2021



Department of Information Technology

TRIDENT ACADEMY OF TECHNOLOGY

Bhubaneswar-751024,Odisha,India.

CERTIFICATEOFAPPROVAL

This B.Tech. Viva-Voce Examination of the Major Project work submitted by thecandidate**S.Dharanidhar Sahoo**bearingBPUTRegd.No.: **1701289273** is held during **2thMay**, **2022**and is accepted in partial fulfillment of the requirement for the award ofthe degree of **Bachelor of Technology in Information Technology** of **Biju PatnaikUniversityofTechnology,Odisha.**

Place:Bhubaneswar

Date: 07-05, 2021 (ExternalSignature)

Place:Bhubaneswar

**Dr.BiswaranjanNayak**

Date:07-05,2021 Dept.ofCSE,ProjectGuide

Place:Bhubaneswar

Date: 07-05, 2021 HOD,Dept.ofCSE

DECLARATION

I,S.Dharanidhar Sahoo declare that the Major Project Work presented through this reportwas carried out by me in accordance with the requirements and in compliance of theAcademic Regulations of the Biju Patnaik University of Technology (BAR) for theBachelor of Technology (B.Tech.) Degree Programed in Information Technology andthat it has not been submitted for any other academic award. Except where indicatedby specific reference in the text, the work is solely my own work. Work done incollaboration with, or with the assistance of, others, has been acknowledged and isindicatedas such.Any viewsexpressed inthe reportare thoseof theauthor.

Place: Bhubaneswar S.Sonu Patra

Date: 07-05, 2021 Regd.No: 1701289273

CERTIFICATE

This is to certify that the report of the Major Project Work on the topic entitled “HEART DISEASE DETECTION USING MACHINE LEARNING” which is submitted by S. Dharanidhar Sahoo inpartial fulfillment of the requirement for the award of the of Bachelor of Technology in Information Technology of Biju Patnaik University of Technology, Odisha, is a Bonafede record of the candidate's own work carried out by him under my supervision.

Supervisor Head of the Department

(Dr.BiswaranjanNayak)

Asst.Professor,Dept.ofCSE Professor & Head, Dept. of ITTridentAcademy of

Technology TridentAcademyofTechnology Bhubaneswar-751024,Odisha,

india

Bhubaneswar,India.

ABSTRACT

The detection of heart disease using machine learning algorithms and python programming. Over the post decades, heart disease is common and dangerous disease caused by fat containment. This disease occurs due to over pressure in the human body. Using different types of parameters in the dataset we can predict the cardiac-disease. We have observed a dataset consists of 12 parameters and 303 individual data values [5] to analyze the performance of patients. The main objective of the paper is to get a better accuracy to detect the heart-disease using algorithms in which the target output counts that a person having heart disease or not.

Place: Bhubaneswar S.Sonu Patra

Date: 07-05, 2021 Regd.No: 1701289273

CONTENTS

Approval i

Declaration ii

Certificate iii

Abstract iv

Acknowledgements v

Contents vi

1 TheProblemStatement 1

1.1 AbouttheProblemStatement 1

2 Introduction 2

2.1 DefinitionofRegression 2

2.1.1SimpleLinearRegression 3

2.1.2MultiplelinearRegression 4

2.1.3Non-LinearRegression 6

2.1.4OverfittinginRegression 7

2.1.5Supportvectorregression 8

2.1.6DecisionTreeregression 12

3Python & Related Libraries 16

3.1 About Python 16

3.1.1 Numpy 17

3.1.2 Pandas 19

3.1.3 SK Learn 20

3.1.4 Matplotlib 21

4EDAforPreprocessing 22

4.1.1 EDA 22

4.1.2 DataPreprocessing 24

4.1.3 Coding & Results 26

5 Conclusions&FutureScope 33

5.1 Conclusion 33

5.2 FutureScope 33

References 34

CHAPTER 1

The Problem Statement

❖ General circulation models work as tools for Heart disease Prediction. There exists lots of assumption-based analysis on some special parameters which affect have affect on heart conditions presented us with the best prediction with a machine learning models like Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbors and naïve Bayes algorithms. The machine learning model which shows the best accuracy will be taken as the machine learning model for heart disease prediction for deployment.

# About the Problem Statement

Our approach is basically developing a machine learning model using a best suitable supervised machine learning algorithm which will be able to predict the Heart Disease by taking the features “age”, “sex”, “cp”, “trestbps”, “chol”, “fbs”, “restecg”, “thalach”, “exang”, “oldpeak”, “slope”, “ca”, “thal” as input and predict the target i.e., whether the patient having heart disease or not.



CHAPTER 2

Introduction

* The heart is a kind of muscular organ which pumps blood into the body and is the central part of the body’s cardiovascular system which also contains lungs. Cardiovascular system also comprises a network of blood vessels, for example, veins, arteries, and capillaries. These blood vessels deliver blood all over the body. Abnormalities in normal blood flow from the heart cause several types of heart diseases which are commonly known as cardiovascular diseases (CVD).
* Heart diseases are the main reasons for death worldwide. According to the survey of the World Health Organization (WHO), 17.5 million total global deaths occur because of heart attacks and strokes. More than 75% of deaths from cardiovascular diseases occur mostly in middle-income and low-income countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack [1] . Therefore, detection of cardiac abnormalities at the early stage and tools for the prediction of heart diseases can save a lot of life and help doctors to design an effective treatment plan which ultimately reduces the mortality rate due to cardiovascular diseases. Due to the development of advance healthcare systems, lots of patient data are nowadays available (i.e. Big Data in Electronic Health Record System) which can be used for designing predictive models for Cardiovascular diseases.
* Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into useful information. “Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information about data” [2] . Nowadays, a huge amount of data pertaining to disease diagnosis, patients etc. are generated by healthcare industries. Data mining provides a number of techniques which discover hidden patterns or similarities from data. Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a heart disease prediction system which was validated on two open access heart disease prediction datasets.
* The Scientists are still in working process of overcoming the limitations of computer models to improvise the accuracy rate of prediction through recent technologies of adding intelligence to machine. To add intelligence the system as human we have given a study platform called Artificial Neural networks, Machine learning rule-based techniques where there exists ample impetus to study the changes in hormones or in different disease measurement units like blood pressure, body temperature, cholesterol, fatigue, etc.
* Python is most powerful programming language having numerous libraries which is used in this project with machine learning model. Machine learning is a subset model of artificial intelligence network in which uses complex algorithms and deep learning neural networks. Cardio vascular disease is a wide spread disease in all over a region. This type of disease may cause due to smoking, high blood pressure, diabetes, overweight, hyper tension, cholesterol etc. that has to be accumulated because of the fatty foods or unlimited intake of foods or non-moving to anywhere.
* This disease may occur by various heart problems such as coronary-artery disease, cardio-vascular, stroke, heart failure and much more. Chest pain (cp), resting blood pressure, cholesterol, resting electrocardiographic results, fasting blood sugar (fbs), maximum heart achieved, exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by uroscopy etc., are the major reasons for causing heart problems but we have a attributes of individual person like height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoke, alcohol, active(physically active person).
* Python libraries are the pre-requisites for making prediction in which SKLEARN is basically used in machine learning predictions. From SKLEARN, we will be able to preprocess the data by splitting the attributes and labels, test and train data, and also scale the values in the data to be values between 0 and 1 by importing the library STANDARDSCALAR. Also SEABORN is another library used in our prediction to correlate each and every attributes together. At last the confusion matrix decides accuracy perfectly by importing CONFUSION MATRIX.
* Another contribution of this paper is the presentation of a cardiac patient monitoring system using the concept of Internet of Things (IoT) with different physiological signal sensors and Arduino microcontroller. Sensor networks are currently using the Internet of Things (IoT) technology to collect, analyze and passing of information from one node to another. IoT is a recently used rapidly expanding technology, where multiple sensors/data collectors can sense, share information and communicate over a private network, Internet Protocol (IP) or public networks. The sensors collect the data after a specific time, analyze it and use it to initiate the required action, and provide an intelligent cloud-based network for analysis, planning and decision making. The products that are developed with IoT such as embedded technology, allow to exchange information among each other nodes or the Internet and it was assessed that about 8 to 50 billion devices will be connected by 2020 [3] .
* The importance and advantages of the application of Machine learning based heart disease detection and prediction system were discussed in several research findings. The application of artificial intelligence in disease detection system especially the cardiac disease system detection improves the performance of other existing widely used models like models provided by American College of Cardiology/American Heart Association (ACC/AHA) models in CVD detection and prediction [4] . The possibility and related matters of providing advanced services of a human health management system were analyzed by Zhao, Wang, and Nakahira, in 2011 and they had also given a research direction of medical technology on IoT [5] . Many types of health-related sensors and technologies were analyzed by them. They identified some issues which need to be solved. The home monitoring system and decision support system was schemed by Chiuchisan and Geman in 2014 [6] . This system contributed to home monitoring, diagnosis, medical prescriptions, medical treatment, rehabilitation and development of his patients with Parkinson’s disease. Wireless Health Monitoring System (WHMS) has attracted considerable attention from the research community and industry over the last decade. Improvement of several Machine learning algorithms and classifier performances like weighted associative classifier were reported in the detection of cardiac abnormalities [7] .
* The elderly patients monitoring from indoor or outdoor locations had been presented by a real-time mobile healthcare system in [3] . A signal sensor and a smartphone were the primary components of the system. The bio-signal sensor data was transmitted to an intelligent server via GPRS/UMTS network for data collection. The system could perform in monitoring the mobility, vital signs, location, and condition of the elderly patient from a distant location. A fully functional wireless body area network (WBAN) system had been proposed in [8] . The designed system used medical bands to obtain physiological data from sensors. The author had chosen some medical bands in order to abate the interruption between the sensors and other existing devices. To increase the operating extent, the multi-hopping technique had been implemented and a medical gateway wireless board had been used in this regard.
* Manpreet Singh et al. [9] designed a heart disease prediction system based on Structural Equation Modelling (SEM) and Fuzzy Cognitive Map (FCM). They validated the data of the Canadian Community Health Survey (CCHS) 2012 dataset. They used twenty significant attributes. To generate the weight matrix for the FCM model, SEM was used which then predicted a possibility of cardiovascular diseases. An SEM model is defined with the correlation between CCC 121 along with 20 attributes; here CCC 121 is a variable which defines whether the respondent has heart disease. Prajakta Ghadge et al. [10] researched an intelligent heart attack prediction system using big data. Heart attack needs to be diagnosed timely and effectively because of its high prevalence. The main objective of this research article was to find a prototype of an intelligent heart attack prediction system that uses big data and data mining modeling techniques. This system could gather hidden knowledge concerning heart disease from a given heart disease database.
* This work aims in developing a Decision Support System in heart disease detection that uses the data mining technique having best accuracy and performance among Naïve Bayes, Support Vector Machine, Simple Logistic Regression, Random Forest & Artificial Neural Network (ANN) etc. By using several cardiovascular system parameters such as age, blood pressure, ECG results, sex, and blood sugar, it is possible to measure the possibility of getting affected by heart disease [11] . For deriving the algorithm with the best accuracy in the detection and prediction of heart disease, a comparative analysis of chosen machine learning algorithms has been shown. This algorithm takes the medical parameters such as age, blood pressure, heartbeat, sex, ECG results, blood sugar etc. as input and shows the probability of getting affected by heart disease as output. This proposed system comprises the scheme and design of a web-based android application which uses an efficient machine learning technique to detect heart disease. It can serve as a very useful tool for doctors, patients, and medical students to diagnose heart disease. For diagnosis of fatal physiological conditions and symptoms such as heart attack requires 24 hours monitoring of patient’s health after transferring from hospital to home. Using this application, the patient can input the current parameters of heart disease from anywhere on the application interface and view the risk level of getting heart disease. All of the parameters that are not real-time like blood sugar, Serum cholesterol, ECG results will be available in the doctor’s prescribed report and there are some parameters like chest pain type and exercise-induced angina, which have to be self-measured periodically by the patient and input the values manually on the application interface. If any fatal situation is detected by the application, the patient can communicate with any doctor via video call and any registered doctor related to heart disease can be found by putting his phone number in the search bar.
* In hospitals, there are provisions for continuous monitoring of critical care heart patients whereas after the release from the hospital patients normally go out of direct supervision. These patients need continuous monitoring of their health condition to reduce the risks of unwanted complications at least for a week or so. Hence, another objective of this work is the empirical design and implementation of the prototype of a continuous real-time cardiac health monitoring system by using Arduino based sensors, which can be developed commercially so that it can be attached to the patient’s body. The sensor data can be transmitted to the server using the Wi-Fi module and saved in the server database. The sensor data can be updated every 10 seconds in the server and doctors with smartphones can view his patient’s updated health status by using this application from anywhere. While any value of the heartbeat, temperature or humidity sensors goes beyond the threshold value, doctors will receive an alert message in the application as well as his phone instantly. Besides, family members and caretakers can also visualize the patient’s real-time data through the application and if the patient remains in the ICU, the buzzer in the Arduino will alert them even though in case of network failure. When a patient will need consultancy or medication from a doctor from another distant location of the world, he/she can simply start live video streaming through the application.

### 2.1 Literature Survey

* Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. Machine Learning is concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases [1]. A major focus of Machine Learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too complex to describe generally in programming languages, so that in effect programs must automatically describe programs.
* In recent years many successful machine learning applications have been developed, ranging from data-mining programs that learn to detect fraudulent credit card transactions, to information filtering systems that learn users' reading preferences, to autonomous vehicles that learn to drive. On public highways. At the same time, there have been important advances in the theory and algorithms that form the foundations of this field [5] . The poor performance results produced by statistical estimation models have flooded the estimation area for over the last decade. Their inability to handle categorical data, cope 11 with missing data points, spread of data points and most importantly lack of reasoning capabilities has triggered an increase in the number of studies using non-traditional methods like machine learning techniques. The area of machine learning draws on concepts from diverse fields such as statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity and control theory[3] .
* There are two main types of Machine Learning algorithms. In this project, supervised learning is adopted here to build models from raw data and perform regression and classification.
* **Supervised machine learning:** Supervised Learning is a machine learning paradigm for acquiring the input output relationship information of a system based on a given set of paired input output training samples. As the output is regarded as the label of the input data or the supervision, an input-output training sample is also called labeled training data, or supervised data. Learning from Labeled Data, or Inductive Machine Learning. The goal of supervised learning is to build an artificial system that can learn the mapping between the input and the output, and can predict the output of the system given new inputs. If the output takes a finite set of discrete values that indicate the class labels of the input, the learned mapping leads to the classification of the input data[1].
* If the output takes continuous values, it leads to a regression of the input. It deduces a function from training data that maps inputs to the expected outcomes. The output of the function can be a predicted continuous value (called regression), or a predicted class label from a discrete set for the input object (called classification). The goal of the supervised learner is to predict the value of the function for any valid input object from a number of training examples[2].
* The most widely used classifiers are the Neural Network (Multilayer perceptron), Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, k-nearest neighbour, Naive Bayes algorithm, Regression Analysis, Artificial neural networks and time series analysis[5] .
* Unsupervised machine learning: Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with supervised learning or reinforcement learning, there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output. Centroid-based Clustering, Density-based Clustering, Distribution-based Clustering, Hierarchical Clustering are some example of Unsupervised Machine Learning.
* In supervised Machine Learning there are 2 types of problems one is classification where the output contains some categorical values for example (True, False), (0,1), (Red, Green, Blue), Rating form (0-5) or (0-10) and another one is Regression, in this type of problem the output generally contains some numerical values or the continuous values such as the shell’s profit, stock market price prediction etc.



CHAPTER 3

Proposed model

* Our model consists of collecting the historical weather data that includes various important factors responsible for the weather change that includes the temperature, both the maximum and minimum temperature, the moisture or humidity in the atmosphere, precipitation, UV Index of the atmosphere and the mean pressure of the atmosphere.
* In our proposed model the collected dataset is segregated into the parts which are of use and which aren’t of any use to the machine learning model. After that the dataset goes through the data preprocessing part wherein the data is passed on to a process where the missing and the error values in the dataset are replaced by the mean values or the most occurring value in that filed. Other way is not to consider those values and replacing those empty values with EAN and the carrying out the other tasks.After the data preprocessing is completed there comes the part wherein the cleaned.
* Dataset is segregated into two parts namely the training set and the test set. The training set is used to train the machine learning model to teach the model to compute the results and the testing set is the used to find the results and then comparing the actual and the calculated value and the using the error value as the benchmark to teach the machine learning model further.
* The training phase also will also contain fold cross validation wherein the dataset is divided in to k sets k times and then the dataset is divided into test and training sets such that in a set training sets are selected randomly and then in sets the model is trained. The kth set is then used as the test set for the trained machine learning model. This technique not only does helps us to reduce the condition of underfitting and the condition of overfitting as well.

**[Methodology]**

**Data Preprocessing**

**Feature Extraction**

**Training Model**

**Evaluating Model**

**Implementing Algorithm**

**Data Collection**

**[Work-Flow Diagram]**

**[Fig-3.1.]**

* + **Algorithms**
* Logistic Regression
* Support Vector Machine
* Decision Tree
* Random Forest
* K-Nearest Neighbors
* Naïve Bayes

### 3.1.1 Logistic Regression:

* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
* Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
* Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
* In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
* The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
* Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
* Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

### 3.1.2 Support Vector machine:

* Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
* The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
* SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

### 3.1.3 Decision Tree:

* Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
* In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
* There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:
* Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
* The logic behind the decision tree can be easily understood because it shows a tree-like structure.

### 3.1.4Random Forest:

* Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
* As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
* The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.
* Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

* Step-1: Select random K data points from the training set.
* Step-2: Build the decision trees associated with the selected data points (Subsets).
* Step-3: Choose the number N for decision trees that you want to build.
* Step-4: Repeat Step 1 & 2.
* Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

### 3.1.5K-Nearest Neighbors:

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
* It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

* Step-1: Select the number K of the neighbors
* Step-2: Calculate the Euclidean distance of K number of neighbors
* Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
* Step-4: Among these k neighbors, count the number of the data points in each category.
* Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
* Step-6: Our model is ready.

### 3.1.6 Naive Bayes:

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
* It is mainly used in text classification that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
* Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

**Why It is Called Naive Bayes?**

* The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

* Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

**Bayes' Theorem:**

* Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
* The formula for Bayes' theorem is given as:
* Where, P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
* P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
* P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
* P(B) is Marginal Probability: Probability of Evidence.

**Working of Naïve Bayes' Classifier:**

* Working of Naïve Bayes' Classifier can be understood with the help of the below example:

Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps:

* Convert the given dataset into frequency tables.
* Generate Likelihood table by finding the probabilities of given features.
* Now, use Bayes theorem to calculate the posterior probability.

Problem: If the heart disease is found or not. then the Player should play or not?



CHAPTER 4

Python and Related Libraries

* Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reason is its vast collection of libraries. Python libraries that used in Machine Learning are:
* Pandas
* Numpy
* Matplotlib
* Seaborn
* Scikit-learn
* Flask

### 4.1 About Python

* Python is one of the most popular and widely used programming languages and has replaced many programming languages in the industry.
* There are a lot of reasons why Python is popular among developers and one of them is that ithasan amazingly large collection of libraries that users canwork with.

Here are a few important reasons as to why Python is popular:

* Python has a huge collection of libraries.
* Python is known as the beginner’s level programming language because of its simplicity and easiness.
* FromdevelopingtodeployingandmaintainingPythonwantstheirdeveloperstobe more productive.
* Portability is another reason for huge popularity of Python.
* Python’s programming syntax is simple to learn and is of high level compared to C, Java, and C++.
* Hence, new applications can be developed by writing fewer lines of codes.
* The simplicity of Python has attracted many developers to create new libraries for machine learning.

Because of the huge collection of libraries Python is becoming hugely popular among machine learning experts.

### 4.2 NumPy

* NumPy is considered as one of the most popular machine learning library in Python.
* TensorFlow and other libraries uses NumPy internally for performing multiple operations on Tensors. Array interface is the best and the most important feature of NumPy.
* NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.
* It is the fundamental package for scientific computing with Python. It contains various features including these important ones:
* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* ToolsforintegratingC/C++andFortrancode
* Useful linear algebra, Fourier transform, and random number capabilities

#### 4.2.1 Features Of NumPy

* Interactive: NumPy is very interactive and easy to use.
* Mathematics: Makes complex mathematical implementations very simple.
* Intuitive: Makescodingrealeasyandgraspingtheconceptsiseasy.
* Lot of Interaction: Widely used, hence a lot of opensource contribution.

#### 4.2.3 Where Is Numpy Used?

* This interface can be utilized for expressing images, sound waves, and other binary raw streams as an array of real numbers in N-dimensional.
* ForimplementingthislibraryformachinelearninghavingknowledgeofNumpyisimportantfor full stack developers.
* NumPy is an open-source numerical Python library.
* NumPy contains a multi-dimensional array and matrix data structures.
* It can be utilized toper form a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines. Therefore, the library contains a large number of mathematical, algebraic, and transformation functions.
* NumPy is an extension of Numeric and Num array.
* Numpy also contains random number generators.
* NumPy is a wrapper around a library implemented in C.

*4.2.4 What Are The Most Important Numpy Data Types?*

Let’s start by understanding the most important Numpy data types. There are a large number ofNumPy objects available:

* One Dimensional Array One of the most important objects is an N dimensional array type known as ndarray.
* We can think of a one-dimensional array as a column or a row of a table with one or more elements:
* All of the items that are stored in ndarray are required to be of the same type. This implies that ndarray is a block of homogeneous data. ndarray has striding information. This numerical value is the number of bytes of the next element in a dimension.
* This helps the array to navigate through memory and does not require copying the data.
* Each ndarray contains a pointer that points to its memory location in the computer.
* It also contains its dtype, its shape, and tuples of strides.
* The strides are integers indicating the number of bytes it has to move to reach the next element in a dimension.

### 4.3 Pandas

* Pandas is a machine learning library in Python that provides data structures of high- level and a wide variety of tools for analysis.
* One of the great feature of this library is the ability to translate complex operations with data using one or two commands.
* Pandas have so many inbuilt methods for grouping, combining data, and filtering, as well as time-series functionality.

#### 4.3.1 Features Of Pandas

* Pandas make sure that the entire process of manipulating data will be easier.
* Support for operations such as Re-indexing, Iteration, Sorting, Aggregations, Concatenations and Visualizations are among the feature highlights of Pandas.

#### 4.3.2 Where Is Pandas Used?

* Currently, there are fewer releases of pandas library which includes hundred of new features, bug fixes, enhancements, and changes in API.
* The improvements in pandas regards its ability to group and sort data, select best suited output for the apply method, and provides support for performing custom types operations.
* Data Analysis among everything else takes the highlight when it comes to usage of Pandas.But, Pandas when used with other libraries and tools ensure high functionality and goodamountof flexibility.

### 4.4 Scikit-Learn

* It is a Python library is associated with NumPy and SciPy. It is considered as one of the best libraries for working with complex data.
* There are a lot of changes being made in this library. One modification is the cross-validation feature, providing the ability to use more than one metric. Lots of training methods like logistics regression and nearest neighbours havereceived some little improvements.

#### 4.3.4 Features Of Scikit-Learn

* Cross-validation: There are various methods to check the accuracy of supervised models on unseen data.
* Unsupervised learning algorithms: Again there is a large spread of algorithms in the offering–starting from clustering, factor analysis, principal component analysis to unsupervised neural networks.
* Feature extraction: Useful for extracting features from images and text (e.g. Bag ofwords)

#### 4.3.5 Where Is Scikit-Learn Used?

* It contains a numerous number of algorithms for implementing standard machine learning and datamining tasks like reducing dimensionality, classification, regression, clustering, and model selection.

### 4.5 Matplotlib

* matplotlib.pyplot is a plotting library used for 2D graphics in python programming language.
* It can be used in python scripts, shell, web application servers and other graphical user interface toolkits.
* There are several toolkits which are available that extend python matplotlib functionality.
* Some of them are separate downloads, others can be shipped with the matplotlib source code but have external dependencies.
* Basemap: It is a map plotting toolkit with various map projections, coastlines and political boundaries.
* Cartopy: It is a mapping library featuring object-oriented map projection definitions, and arbitrary point, line, polygon and image transformation capabilities.
* Excel tools: Matplotlib provides utilities for exchanging data with Microsoft Excel.
* Mplot3d: It is used for 3-D plots.
* Natgrid: It is an interface to the natgrid library for irregular gridding of the spaced data.

#### 4.5.1 Some features of Python Plot supports-

1. Font properties
2. Axes properties
3. Line styles

### 4.6 Flask

* Flask is a Python Web Frame work for developing the we b applications.
* Flask is a web framework, it’s a Python module that lets you develop web applications easily. It’s has a small and easy-to-extend core: it’s a microframework that doesn’t include an ORM (Object Relational Manager) or such features.
* It does have many cool features like url routing, template engine. It is a WSGI web app framework.

#### 4.6.1 What is a Web Framework?

* A Web Application Framework or a simply a Web Framework represents a collection of libraries and modules that enable web application developers to write applications without worrying about low-level details such as protocol, thread management, and so on.

#### 4.6.2 What is Flask?

* Flask is a web application framework written in Python. It was developed by Armin Ronacher, who led a team of international Python enthusiasts called Poocco. Flask is based on the Werkzeg WSGI toolkit and the Jinja2 template engine.Both are Pocco projects.

#### 4.6.3 WSGI

* The Web Server Gateway Interface (Web Server Gateway Interface, WSGI) has been used as a standard for Python web application development. WSGI is the specification of a common interface between web servers and web applications.

#### 4.6.4 Werkzeug

* Werkzeug is a WSGI toolkit that implements requests, response objects, and utility functions. This enables a web frame to be built on it. The Flask framework uses Werkzeg as one of its bases.

#### 4.6.5 jinja2

* jinja2 is a popular template engine for Python. A web template system combines a template with a specific data source to render a dynamic web page.
* This allows you to pass Python variables into HTML templates like this:



# Codes for Flask (App.py)

from flask import Flask, render\_template, request, redirect, url\_for, session

from flask\_mysqldb import MySQL

import MySQLdb.cursors

import re

import numpy as np

import joblib

app = Flask(\_\_name\_\_)

model = joblib.load('model.pkl')

app.secret\_key = 'your secret key'

app.config['MYSQL\_HOST'] = 'localhost'

app.config['MYSQL\_USER'] = 'root'

app.config['MYSQL\_PASSWORD'] = ''

app.config['MYSQL\_DB'] = 'login'

mysql = MySQL(app)

@app.route('/')

@app.route('/login', methods =['GET', 'POST'])

def login():

msg = ''

if request.method == 'POST' and 'username' in request.form and 'password' in request.form:

username = request.form['username']

password = request.form['password']

cursor = mysql.connection.cursor(MySQLdb.cursors.DictCursor)

cursor.execute('SELECT \* FROM account WHERE username = % s AND password = % s', (username, password, ))

account = cursor.fetchone()

if account:

session['loggedin'] = True

session['id'] = account['id']

session['username'] = account['username']

msg = 'Logged in successfully !'

return render\_template('index.html', msg = msg)

else:

msg = 'Incorrect username / password !'

return render\_template('login.html', msg = msg)

@app.route('/logout')

def logout():

session.pop('loggedin', None)

session.pop('id', None)

session.pop('username', None)

return redirect(url\_for('login'))

@app.route('/register', methods =['GET', 'POST'])

def register():

msg = ''

if request.method == 'POST' and 'username' in request.form and 'password' in request.form and 'email' in request.form :

username = request.form['username']

password = request.form['password']

email = request.form['email']

cursor = mysql.connection.cursor(MySQLdb.cursors.DictCursor)

cursor.execute('SELECT \* FROM account WHERE username = % s', (username, ))

account = cursor.fetchone()

if account:

msg = 'Account already exists !'

elif not re.match(r'[^@]+@[^@]+\.[^@]+', email):

msg = 'Invalid email address !'

elif not re.match(r'[A-Za-z0-9]+', username):

msg = 'Username must contain only characters and numbers !'

elif not username or not password or not email:

msg = 'Please fill out the form !'

else:

cursor.execute('INSERT INTO account VALUES (NULL, % s, % s, % s)', (username, password, email, ))

mysql.connection.commit()

msg = 'You have successfully registered !'

elif request.method == 'POST':

msg = 'Please fill out the form !'

return render\_template('register.html', msg = msg)

@app.route('/predict', methods=['POST'])

def predict():

if request.method=='POST':

f1 = float(request.form['f1'])

f2 = float(request.form['f2'])

f3 = float(request.form['f3'])

f4 = float(request.form['f4'])

f5 = float(request.form['f5'])

f6 = float(request.form['f6'])

f7 = float(request.form['f7'])

f8 = float(request.form['f8'])

f9 = float(request.form['f9'])

f10 = float(request.form['f10'])

f11 = float(request.form['f11'])

f12 = float(request.form['f12'])

f13 = float(request.form['f13'])

feature\_array = [f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f11,f12,f13]

feature = np.array(feature\_array).reshape(1,-1)

prediction = model.predict(feature)

if prediction==0:

prediction='No Heart Disease Found'

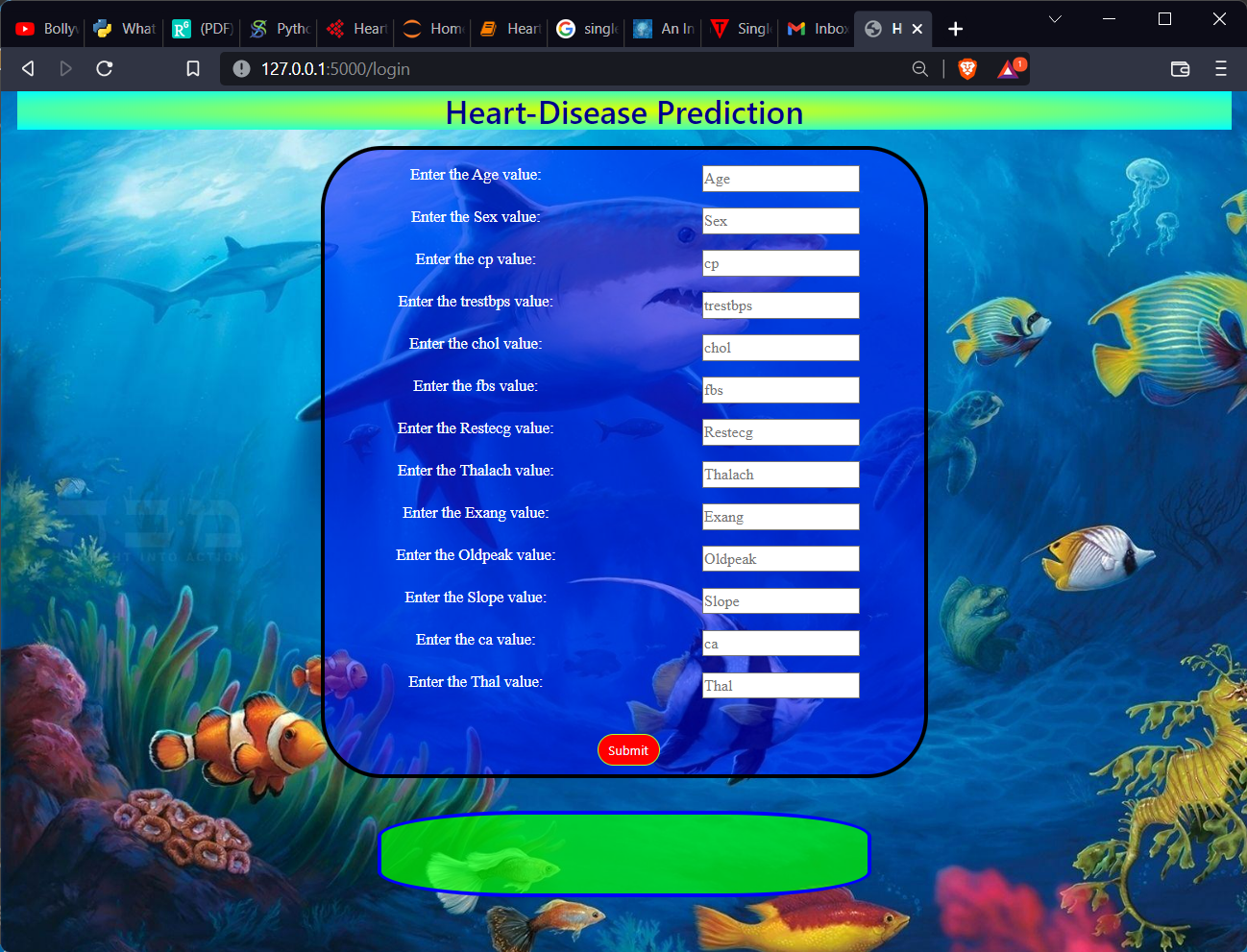
else:

prediction='Heart Disease Found'

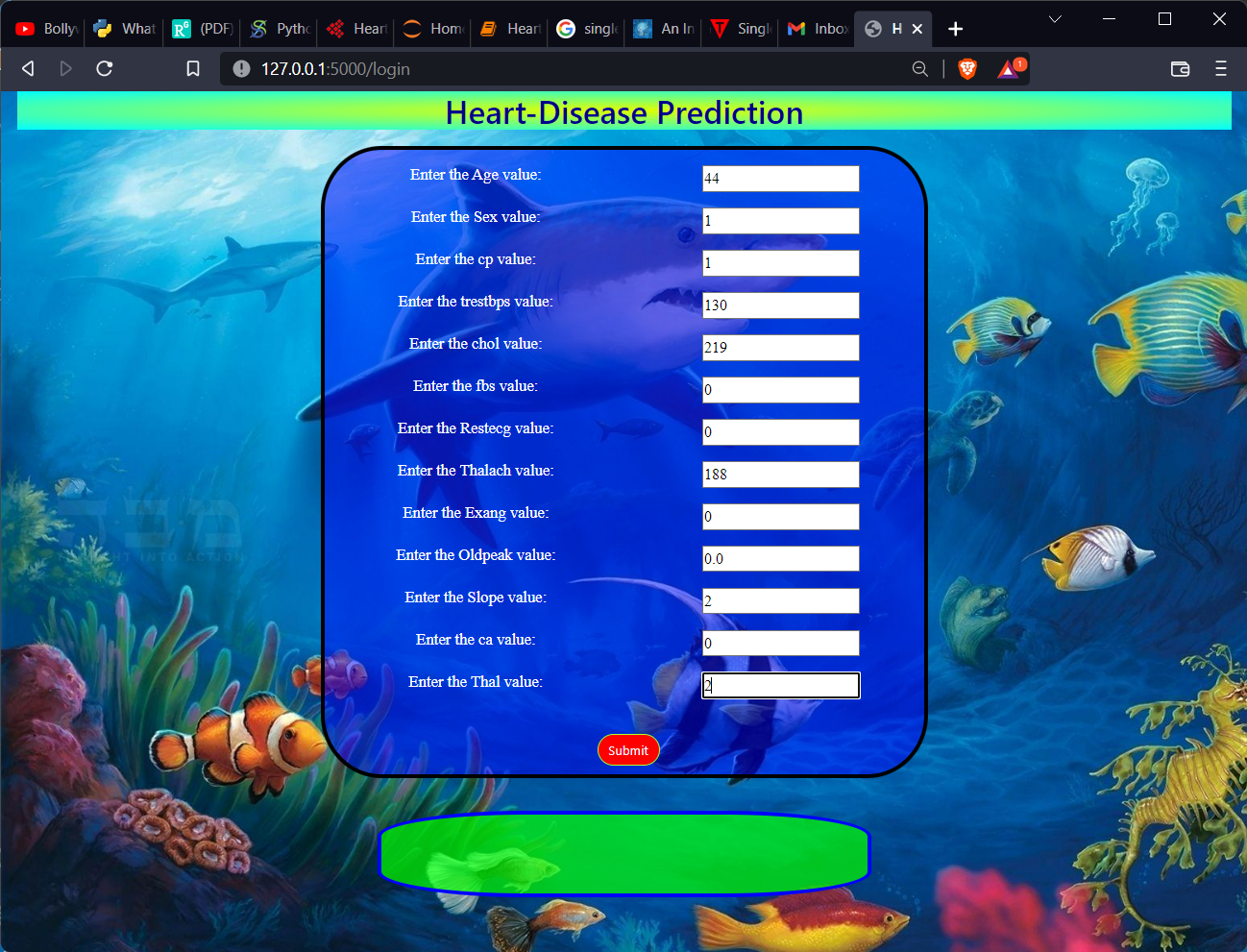
return render\_template('index.html', prediction='{}'.format(prediction))

if \_\_name\_\_ == "\_\_main\_\_":

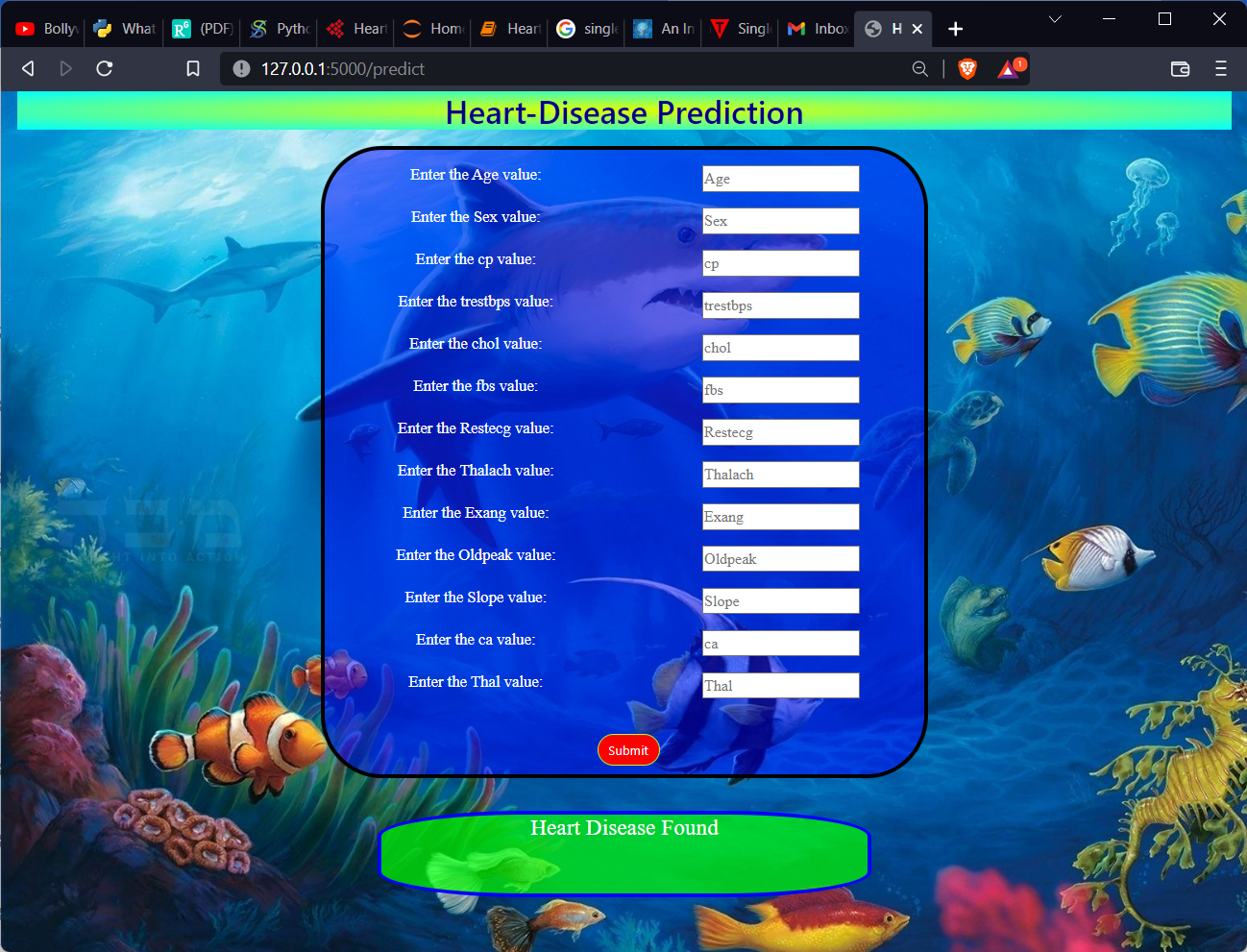
app.run(debug=True)



Taking User inputs:



Predicting the Result:





CHAPTER 5

EDA and Preprocessing

# 5.1 EDA

* It is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. Following are the different steps involved in EDA:
* Data Collection
* Data Cleaning
* Data Preprocessing
* Data Visualization
* Data Collection: Data collection is the process of gathering information in an established systematic way that enables one to test hypothesis and evaluate outcomes easily.
* Data Cleaning: Data cleaning is the process of ensuring that your data is correct and useable by identifying any errors in the data, or missing data by correcting or deleting them. Once the data is clean we can go further for data preprocessing.
* Data Preprocessing: Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. It includes normalization and standardization, transformation, feature extraction and selection, etc.
* The product of data preprocessing is the final training dataset. Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set;
* uncover underlying structure;
* extract important variables;
* detect outliers and anomalies;
* test underlying assumptions;
* develop parsimonious models;
* determine optimal factor settings.

### 5.1.1 Focus:

* The EDA approach is precisely an approach not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

### 5.1.2 Philosophy:

* EDA is not identical to statistical graphics although the two terms are used almost interchangeably. Statistical graphics is a collection of techniques—all graphically based and all focusing on one data characterization aspect. EDA encompasses a larger venue; EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques; EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call "statistical graphics", but it is not identical to statistical graphics.

### 5.1.3 History:

* The seminal work in EDA is Exploratory Data Analysis, Tukey, (1977). Over the years it has been fitted from other note worthy publications such as Data Analysis and Regression, Mosteller and Tukey (1977), Interactive Data Analysis, Hoaglin (1977), TheABC's of EDA, Velleman and Hoaglin (1981) and has gained a large following as "the" way to analyze a data set.
* Techniques: Most EDA techniques are graphical in nature with a few quantitative.
* The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain
* some new of the nun-suspected insight in to the data. In combination with the natural pattern-recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out.
* The particular graphical techniques employed in EDA are often quite simple, consisting ofvarious techniques of:
* Plotting the raw data (such as data traces, histograms, bi-histograms, probability plots, lag plots, block plots, and Youden plots.
* Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
* Positioning such plots so as to maximize our natural pattern-recognition abilities, suchas using multiple plots per page.

### 5.2 Data Preprocessing:

* Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
* Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.
* Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).
* Data goes through a series of steps during preprocessing:

#### 5.2.1 Data Cleaning:

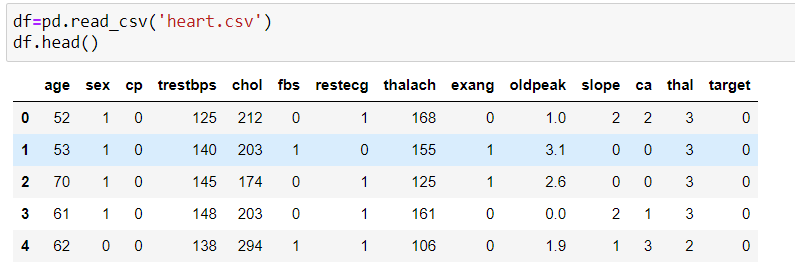
* Data is cleansed through processes such as filling in missing values and smoothing the noisy data, or resolving the inconsistencies in the data.
* Data Integration: Data with different presentations are put together and conflicts within the data are resolved.
* Data Reduction: This step aims to present a reduced representation of the data ina data warehouse.
* Data Discretization: Involves the reduction of a number of values of acontinuous attribute by dividing the range of attribute intervals.
* Data Transformation: Data is normalized, aggregated and generalized.

Some of Data Preprocessing steps are given below:

All the libraries are used for data preprocessing and building the model are given in below figure:



[Fig- 5.1.]



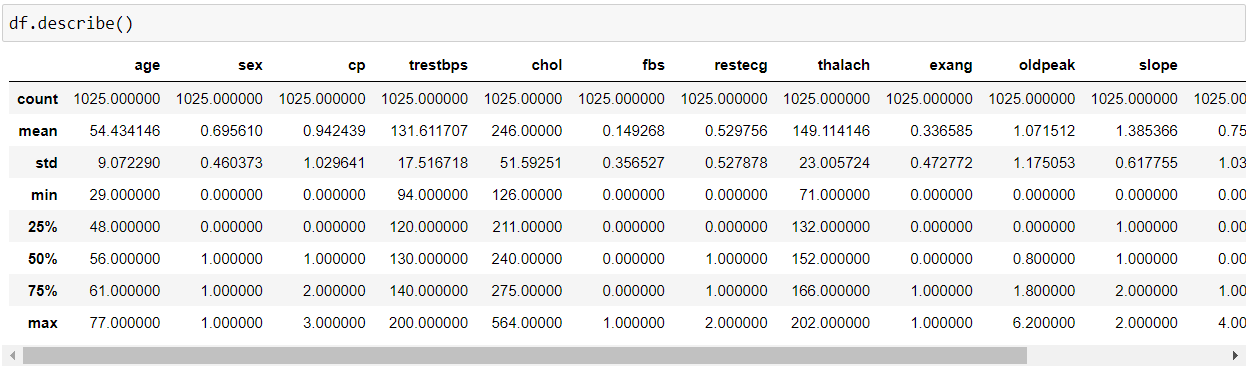
**[Fig-5.2.]**

The data frame having 1025 number of rows and 14 columns initially.



**[Fig-5.3.]**

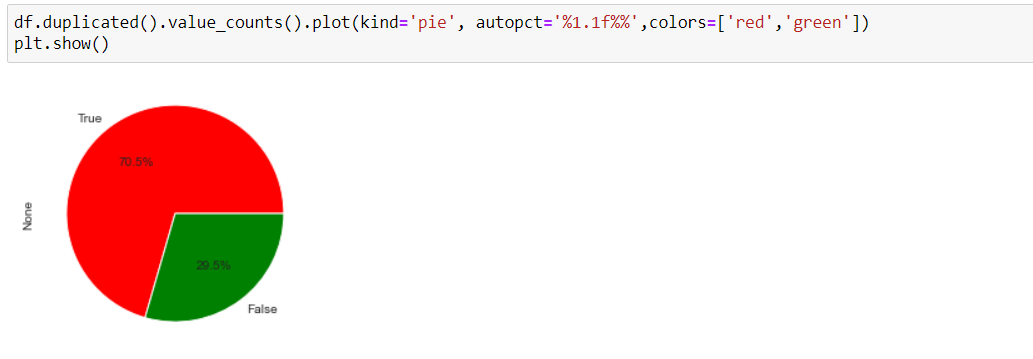
Some statistical information about the dataset is given below:



**[Fig-5.4.]**

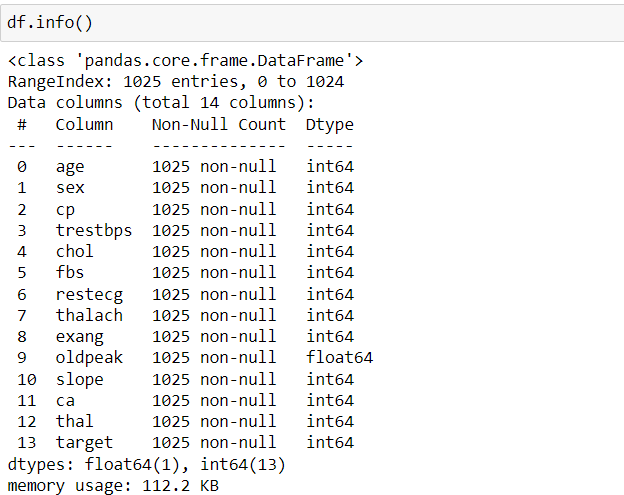
Checking for null values whether the dataset containing duplicated values and the percentage of the duplicated value is given in below pie chart figure:

70.5% of data among the whole dataset contains Duplicated values so those duplicated items must be removed from the dataset for training/ building a better machine learning model.



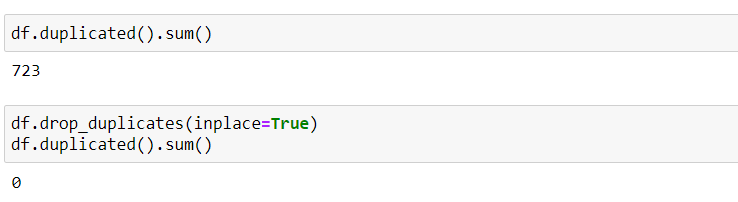
**[Fig-5.4.]**

Here is the information about the Dataset.



**[Fig-5.5.]**

Droping/ Deleting the duplicated columns from the dataset



**[Fig – 5.6.]**

Checking for null values whether the dataset containing any null values/ blank shells or not?



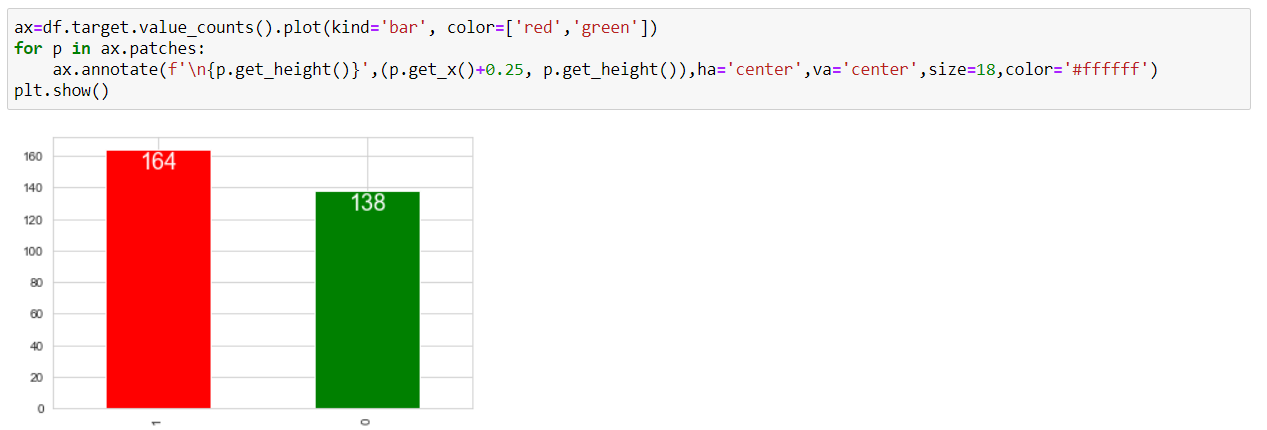
[Fig – 5.7.]

Now the shape of the dataset is:



[Fig – 5.8.]

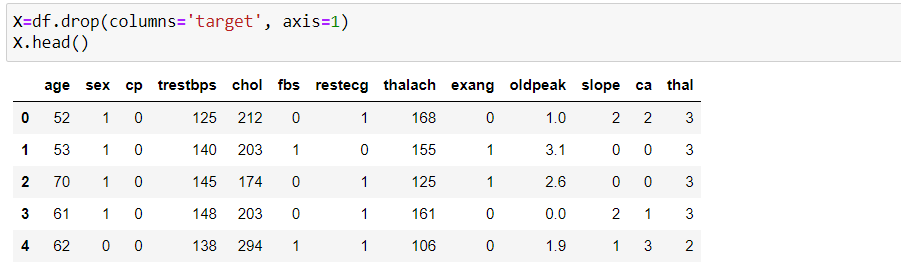
As we not this dataset is for detection of heart disease using various features of human body parameters. So the target is the patient having heart disease or not. If the patient having heart disease then the value is 1 otherwise the value is zero.



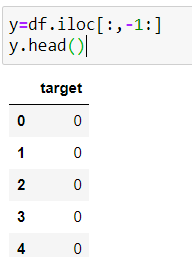
[Fig – 5.9.]

From the total data after cleaning there is 164 patients data having heart disease and 138 patients having no disease.

Separating dependent and independent features;

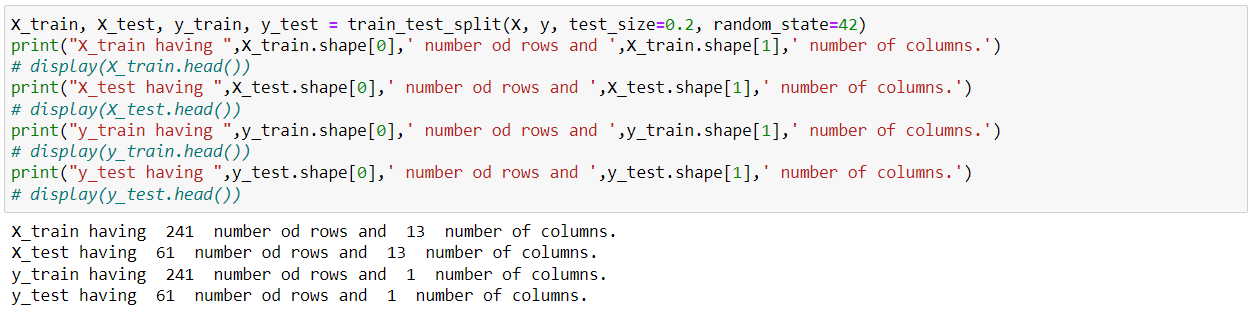


[Fig – 5.10.]



[Fig - 5.11.]

Feature extraction using the function train\_test\_split, which comes under the scikit-learn library. Here tha data is separated in to training dependent (X\_train), training dependent (y\_train), testing dependent (X\_test), testing independent (y\_test) with the traing and testing ratio 70% of data and 30% of data respectively.



**[Fig – 5.12.]**

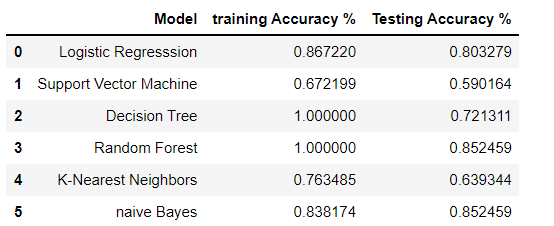
All the training independent features are contained by “X\_train” and all the testing independent features are contained by “X\_test” similarly, the one and only training dependent features are contained by “y\_train” and all the testing dependent feature is contained by “y\_test”.

The correlation-plot of given dataset is:



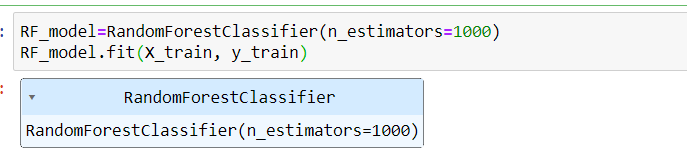
[Fig -5.13.]

We have create 6 different types of model from Sklearn library these are Logistic Regression, Support Vector Classifier from Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier from KNN and the Gaussian NB classifier from the Naïve Bayes algorithm. Among all the libraries (Algorithms) the Random Forest algorithm performs best with 100% training accuracy and 85.24% testing accuracy.



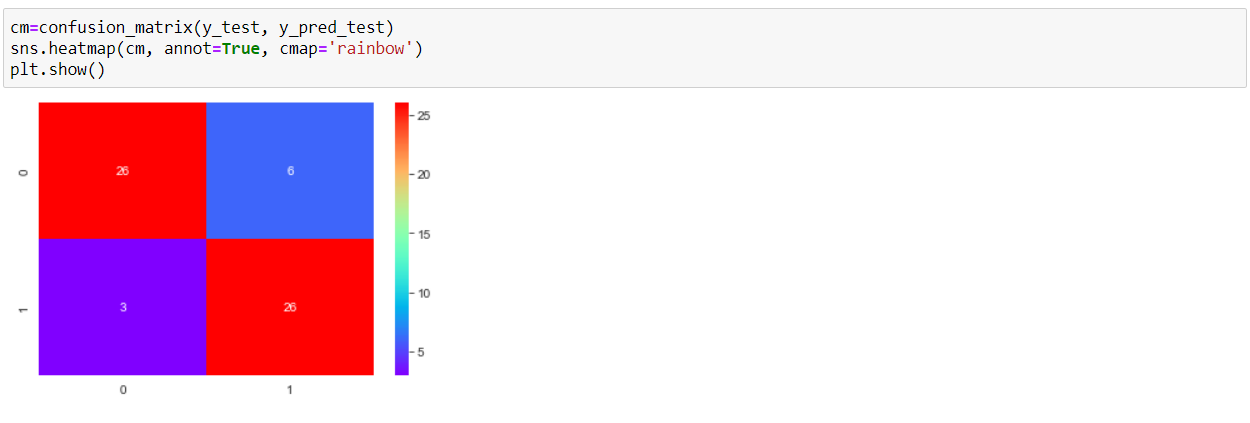
**[fig-5.14.]**

Fitting the training data in to Random Forest Classifier model with (n\_estimators = 1000) and the random\_stet is 42. figure given below:



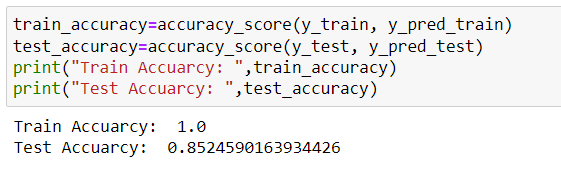
**[Fig -5.15.]**

### Confusion matrix for Random Forest Classifier model



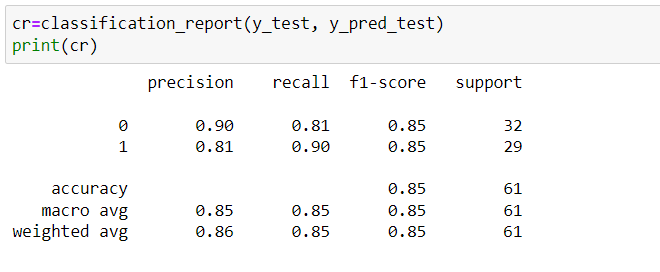
**[Fig – 5.16.]**

### Accuracy score for random Forest model



**[Fig – 5.17.]**

### Classification Report for Random Forest model



**[Fig – 5.18.]**



CHAPTER 6

Conclusions & Future Scope

# Conclusion

* We conclude that through this model we can predict the weather condition by giving four inputs to the developed model and the model can predict the weather condition with 84% of accuracy. GaussianNB classifier performs best among all the six supervised machine learning algorithms we have applied.

# FutureScope

* This has so many scope for predicting the patient having Heart Disease or not. so it can be implement on any device and useful in hospital to predict heart disease based on these parameters such as “age”, “sex”, “cp”, “trestbps”, “chol”, “fbs”, “restecg”, “thalach”, “exang”, “oldpeak”, “slope”, “ca”, “thal” and etc.

REFERENCES

* <https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda>
* https://www.researchgate.net/publication/26904238\_Crop\_Management\_Factors\_What\_is\_Important
* <https://medium.com/pursuitnotes/decision-tree-regression-in-6-steps-with-python-1a1c5aa2ee16>
* https://www.kaggle.com/datasets/ananthr1/weather-prediction