**HEART DISEASE PREDICTION**

Submitted in the partial fulfillment of the requirement for the award of degree of

**Bachelors of Technology in Computer Science & Engineering**

****

**Under the Guidance of**

Mr. Kaustav Bandyopadhyay

**Submitted By**

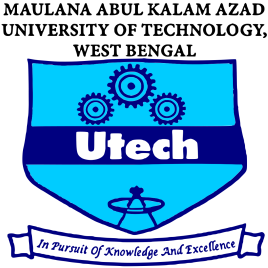
Dhriti Ojha 20600118038

Prarona Das 20600118033

Riya Majee 20600118028

Department of Computer Science & Engineering

Seacom Engineering College Howrah, West Bengal(India)



MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY

Kolkata, West Bengal India

**DECLARATION**

We hereby declare that this project report is based on our original work except for citations and quotations which have been duly acknowledged. We also declare that is has not been previously and concurrently submitted for any other degree or award at any university or any other institution.

Signature Name University Roll No.

* 20600118038
* 20600118033
* 20600118028

**CERTIFICATE**

Certified that this project entitled “ **Heart Disease Prediction”** submittedby **Dhriti Ojha** (20600118038), **Prarona** **Das** (20600118033) & **Riya Majee** (20600118028) students of Computer Science & Engineering Department, Seacom Engineering College, Howrah, West Bengal in the partial fulfillment of the requirement for the award of Bachelors of Technology (Computer Science & Engineering) Degree of West Bengal University of Technology, West Bengal ,is a record of students own study carried under my supervision & guidance. This report has not been submitted to any other university or institution for the award of any degree.

Kaustav Bandyopadhyay Sagar Chakraborty

(HOD)

(Seacom Engineering College) (Seacom Engineering College)

**ABSTRACT**

Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart disease and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. We work on predicting possible Heart Disease in people using Machine Learning algorithms.

**INTRODUCTION**

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researchers 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.. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease .

**PROBLEM STATEMENT**

Heart attack risk prediction using SVM, Decision Tree, Naive Bayes, Logistic Regression , Supervised Machine learning algorithm.

**CHALLENGES**

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either it are expensive or are not efficient to calculate chance of heart disease in human. 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.

**OBJECTIVES**

* To Detect heart attacks using supervised machine learning.
* To accurately give results.
* To identify the correct algorithm in detecting the risk of heart diseases.

**Hardware requirements:**

Processer Any Updated Processer

Ram Min 4GB

Hard Disk Min 100GB

**Software requirements:**

Machine learning

Windows

Python 3.9.13

Jupiter notebook

Visual Studio Code

Flask Application

Html & Css

**PROPOSED SYSTEM**

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1. ***Collection of Dataset***

***2. Selection of attributes***

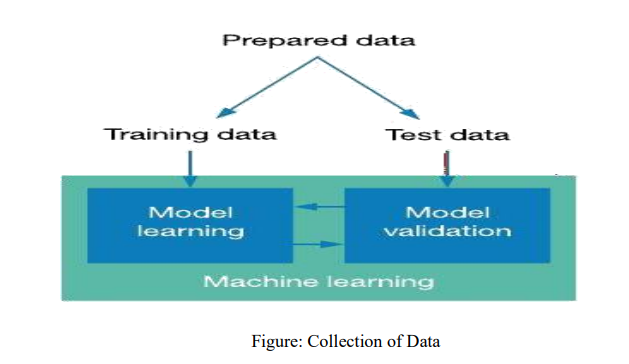
***3. Data Pre-Processing***

***4. Balancing of Data***

***5. Disease Prediction***

**Collection of Dataset :**

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

  
**Selection of Attributes:**

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

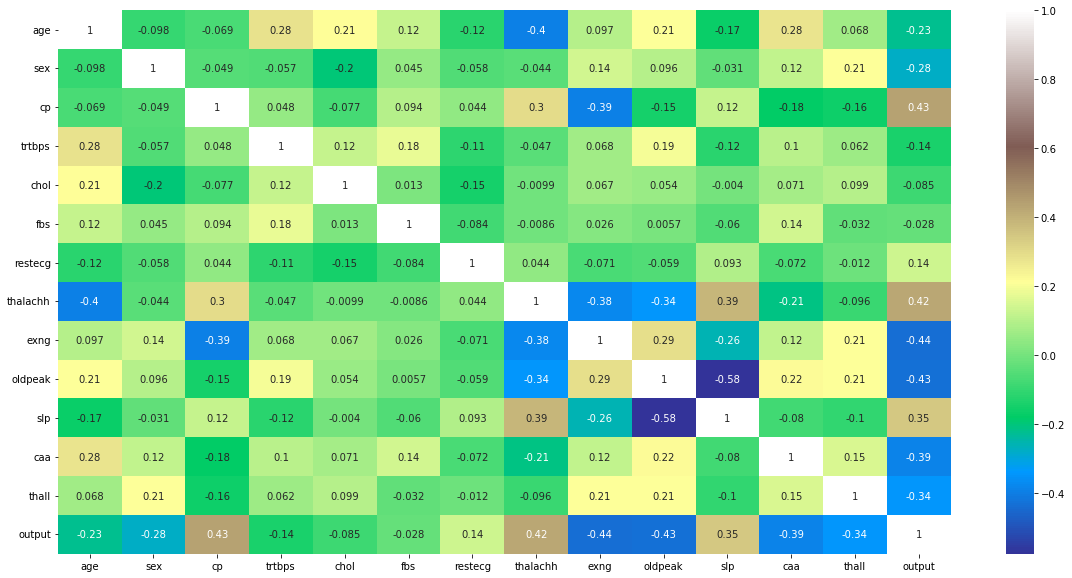


Figure: Correlation Matrix

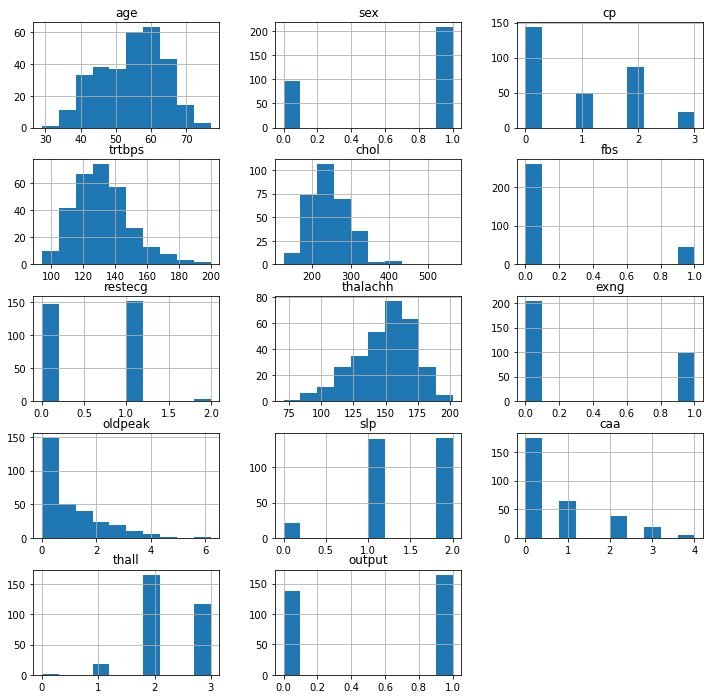
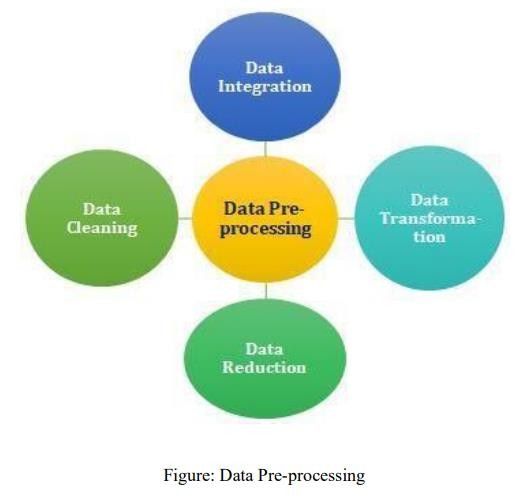
****

Figure: All Features

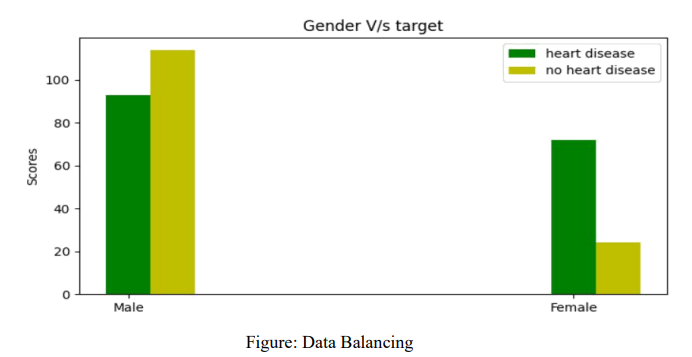
**Pre-processing of Data**

Data pre-processing is an important step for the creation of a machine learning model. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.



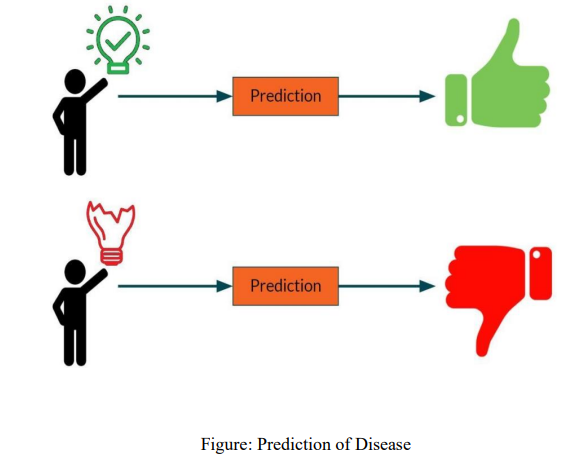
**Balancing of Data**

Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling In Under Sampling, dataset balance is done by the reduction of the size of the ample class. 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. This process is considered when the amount of data is inadequate.

****

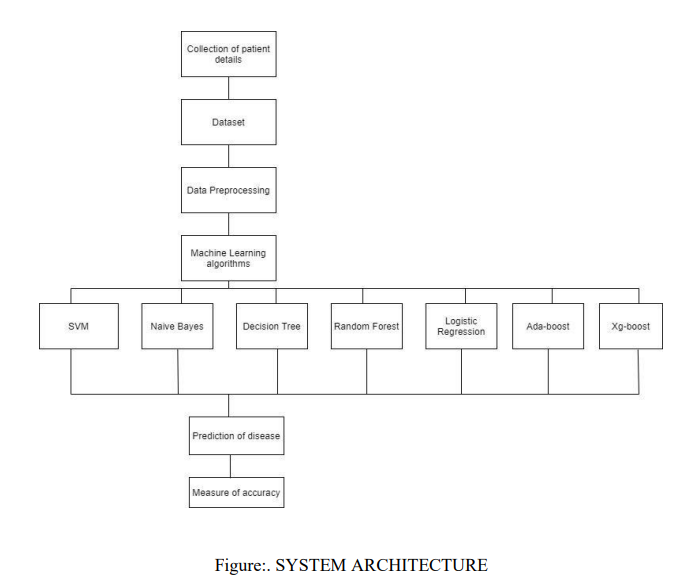
**Prediction of Disease**

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.



**WORKING OF SYSTEM & SYSTEM ARCHITECTURE :**

Dataset collection is collecting data which contains patient details. Attributes selection process selects the useful attributes for the prediction of heart disease. After identifying the available data resources, they are further selected, cleaned, made into the desired form. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease. Accuracy measure compares the accuracy of different classifiers.



**MACHINE LEARNING**

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

Supervised Learning: Supervised learning is the type of machine learning in which machines are trained using well "labeled" training data, and on the basis of that data, machines predict the output. The labeled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly.. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

***TESTING METHODS:***

**DECISION TREE ALGORITHM**

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..

**Workflow:**

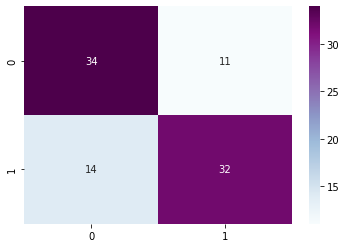
● Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

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

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

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

● Step-5: Recursively make new decision trees using the subsets of the dataset created in

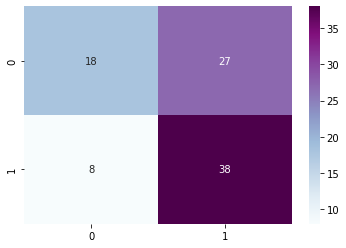
****

***SUPPORT VECTOR MACHINE (SVM):***

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.

**Workflow:**

* **Step** 1: Load the important libraries. ...
* **Step** 2: Import dataset and extract the X variables and Y separately. ...
* **Step** 3: Divide the dataset into train and test. ...
* **Step** 4: Initializing the **SVM** classifier model. ...
* **Step** 5: Fitting the **SVM** classifier model. ...
* **Step** 6: Coming up with predictions.

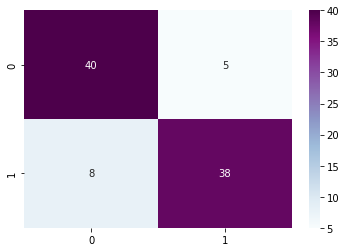
****

**NAIVE BAYES ALGORITHM:**

Naive 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. Naive 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.

**Workflow:**

1. **Step** 1: Separate By Class.
2. **Step** 2: Summarize Dataset.
3. **Step** 3: Summarize Data By Class.
4. **Step** 4: Gaussian Probability Density Function.
5. **Step** 5: Class Probabilities

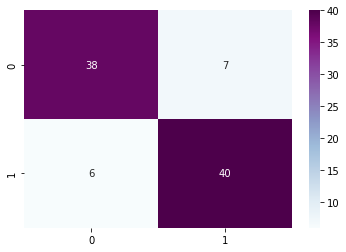


**LOGISTIC REGRESSION ALGORITHM**

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.

**Workflow:**

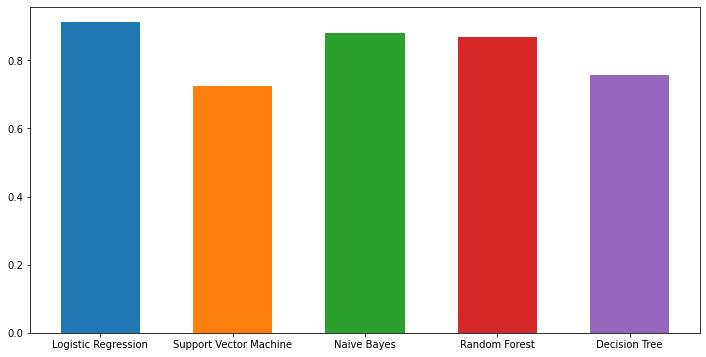
1. Data Pre-processing **step**.
2. Fitting **Logistic Regression** to the Training set.
3. Predicting the test result.
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

****

**Dataset Attributes**

* 1. Age
  2. Sex
  3. Chest pain type (4 values)
  4. Resting blood pressure
  5. Serum cholestoral in mg/dl
  6. Fasting blood sugar > 120 mg/dl
  7. Resting electrocadiographic results (values 0,1,2)
  8. Maximum heart rate achieved
  9. Exercise induced angina
  10. Oldpeak = ST depression induced by exercise relative to rest
  11. The slope of the peak exercise ST segment
  12. Number of major vessels (0-3) colored by flourosopy
  13. Thal : 1 =normal ; 2=fixed defect ; 3 = reversable defect
  14. Output(target): 0 = no heart disease ; 1=heart disease

**Comparision Analysis**



**Result:**

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy Score** |
| Decision Tree | 74% |
| Logistic Regression | 85% |
| SVM | 61% |
| Naïve Bayes | 85% |

**PERFORMANCE ANALYSIS**

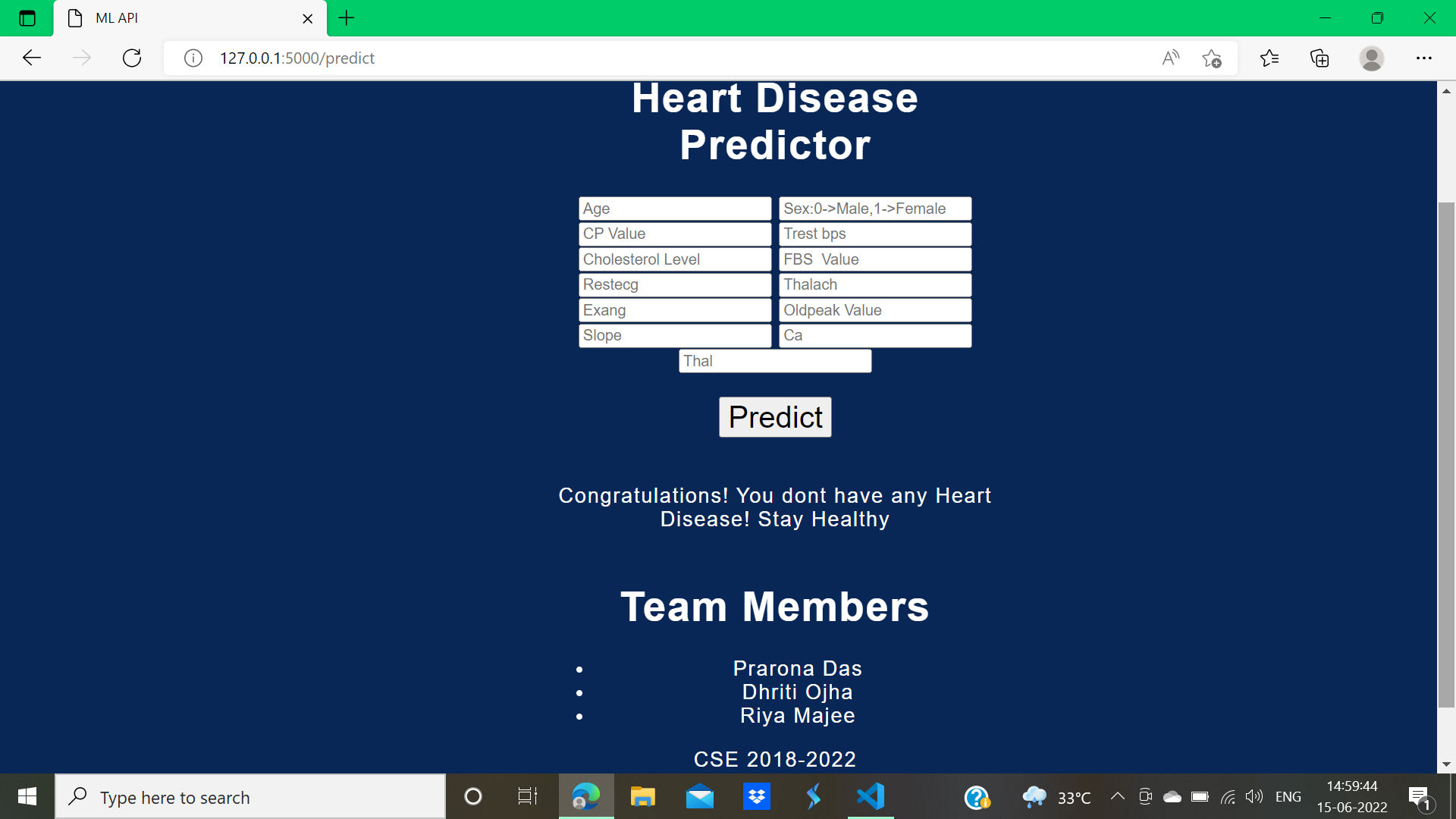
In this project we have used DECISION TREE, SVM, LOGISTIC REGRESSION , NAÏVE BAYES Algorithm

But Logistic Regression and Naïve bayes have been come out with more accuracy percentage compare to the rest two:

* We have taken different attributes like gender, chest pain type, fasting blood sugar, serum cholesterol and put all these in decision tree algorithm, as decision tree uses CART algorithm, and asks a question and based on the answer (Yes/No), it furthur splits into subtrees and since decision tree works like human brain it is very easy to understand and hence we used it in this project.
* Logistic regression is **a process of modeling the probability of a discrete outcome given an input variable**. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.
* 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**.

**WEBSITE OUTPUT**

The website has been created using visual studio code by inputting the flask application in python code advanced adding Html & Css. With the help of website and adding the accurate data anyone can check their heart health.



**CONCLUSION AND FUTURE WORK**

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, The correlation of some features in the dataset is almost equal and so they are removed. If all the attributes present in the dataset are taken into account then the efficiency decreases considerably. All the four machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision which predicts the disease efficiently. Comparing all four Logistic Regression and Naïve Bayes Algorithm give the highest accuracy of 85%.

**MOTIVATION FOR THE WORK**

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. Hence, the four algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better.