**Name : Kajal Rajendra Gomle**

**College : Terna Engineering College**

**Course : Data Science & Analytics**

Heart Disease Prediction using

Logical Regression

Introduction

Every day, the average human heart beats around 100,000 times, pumping 2,000 gallons of blood through the body. Inside your body there are 60,000 miles of blood vessels.

The signs of a woman having a heart attack are much less noticeable than the signs of a male. In women, heart attacks may feel uncomfortable squeezing, pressure, fullness, or pain in the center of the chest. It may also cause pain in one or both arms, the back, neck, jaw or stomach, shortness of breath, nausea and other symptoms. Men experience typical symptoms of heart attack, such as chest pain , discomfort, and stress. They may also experience pain in other areas, such as arms, neck , back, and jaw, and shortness of breath, sweating, and discomfort that mimics heartburn.

It’s a lot of work for an organ which is just like a large fist and weighs between 8 and 12 ounces.

Theory

**Dataset:**

heart\_disease.csv

**Dataset Columns:**

* age: The person’s age in years
* sex: The person’s sex (1 = male, 0 = female)
* cp: chest pain type  
  — Value 0: asymptomatic  
  — Value 1: atypical angina  
  — Value 2: non-anginal pain  
  — Value 3: typical angina
* trestbps: The person’s resting blood pressure (mm Hg on admission to the hospital)
* chol: The person’s cholesterol measurement in mg/dl
* fbs: The person’s fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
* restecg: resting electrocardiographic results  
  — Value 0: showing probable or definite left ventricular hypertrophy by Estes’ criteria  
  — Value 1: normal  
  — Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
* thalach: The person’s maximum heart rate achieved
* exang: Exercise induced angina (1 = yes; 0 = no)
* oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot. See more here)
* slope: the slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2: upsloping  
  0: downsloping; 1: flat; 2: upsloping
* ca: The number of major vessels (0–3)
* thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously  
  Value 1: fixed defect (no blood flow in some part of the heart)  
  Value 2: normal blood flow  
  Value 3: reversible defect (a blood flow is observed but it is not normal)
* target: Heart disease (1 = no, 0= yes)

**Context:**

This is multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has a heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

Algorithm

**Machine Learning (ML):**

Machine Learning (ML) Machine learning is widely used in almost many fields in the world including healthcare sector. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Further, machine learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world. There are two major categories of problems often solved by machine learning i.e. regression and classification. Mainly, the regression algorithms are used for numeric data and classification problems include binary and multicategory problems.Machine learning algorithms are further divided into two categories such as supervised learning and unsupervised learning. Basically, supervised learning is performed by using prior knowledge in output values whereas unsupervised learning does not predefined labels hence the goal of this is to infer the natural structures within the dataset. Therefore, selection of machine learning algorithm need to carefully evaluated.

**Logistic Regression Model:**

Logistic regression is a one of the machine learning classification algorithm for analyzing a dataset in which there are one or more independent variables (IVs) that determine an outcome and also categorical dependent variable (DV). Linear regression uses output in continuous numeric whereas 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. The logistics regression forms three types as below.

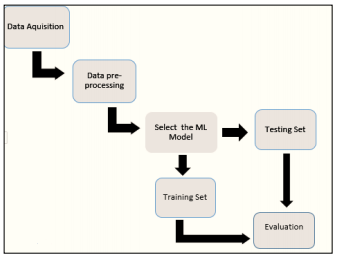
a) Binary logistics regression (two possible outcomes in a DV)

b) Multinomial logistics regression (three or more categories in DV without ordering)

c) Ordinal logistics regression (three or more categories in DV with ordering)

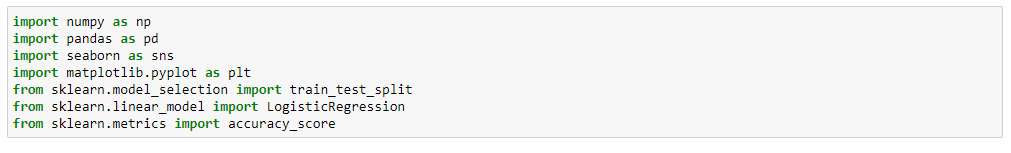
**Methodology:**

Workflow of Machine Learning ModelBuilding Figure indicates the steps followed in order to build the logistic regression model in machine learning.



**System Designing**

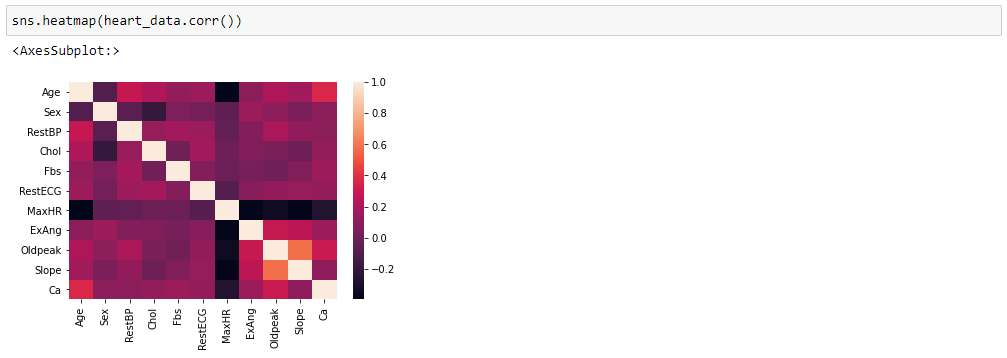
**Importing the Libraries:**

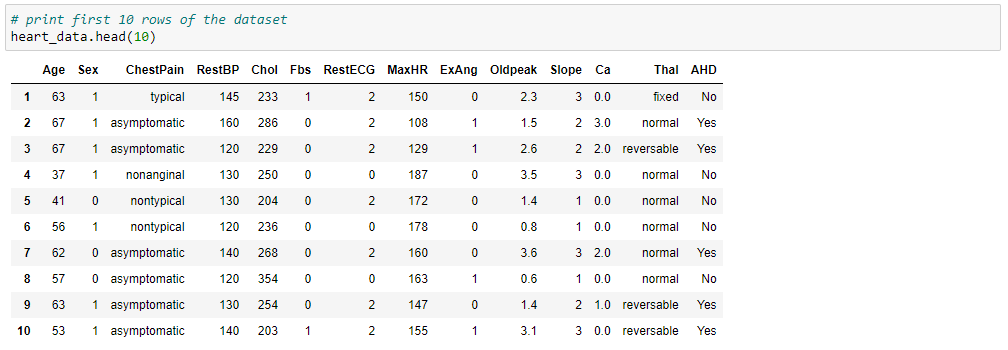


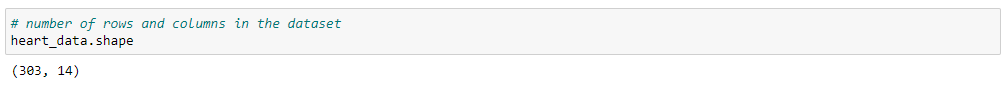
**Data Collection:**

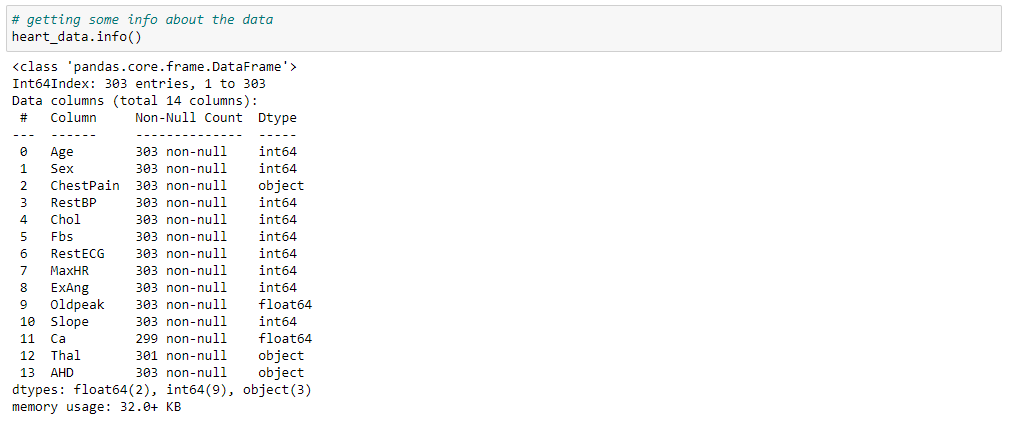


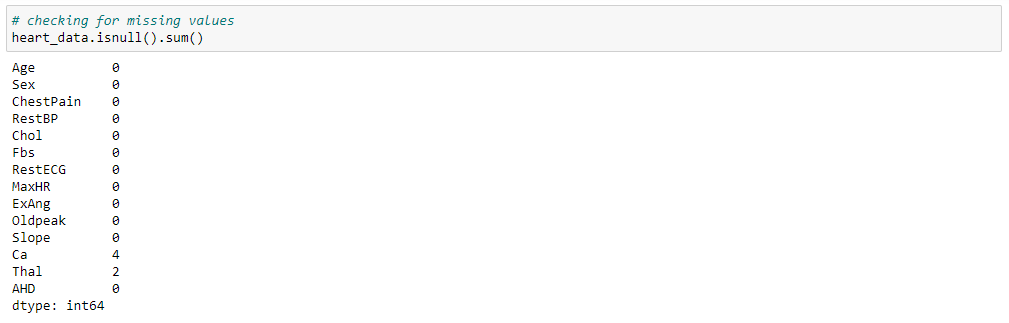
**Data Processing:**

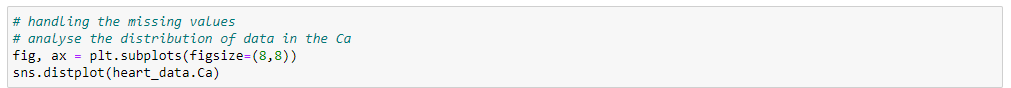


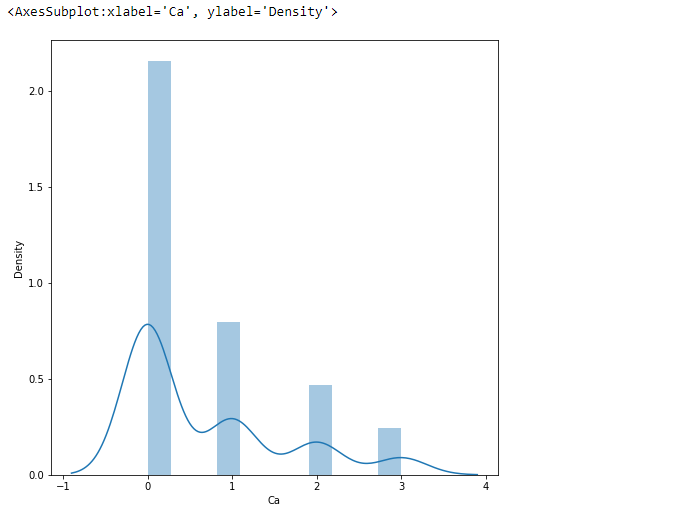


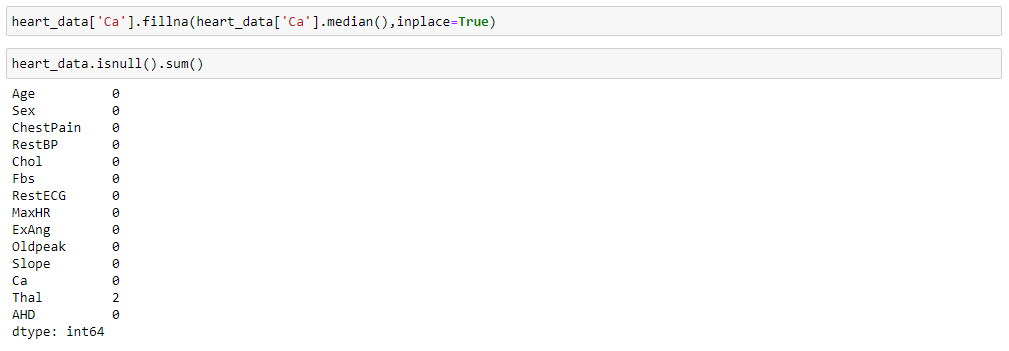




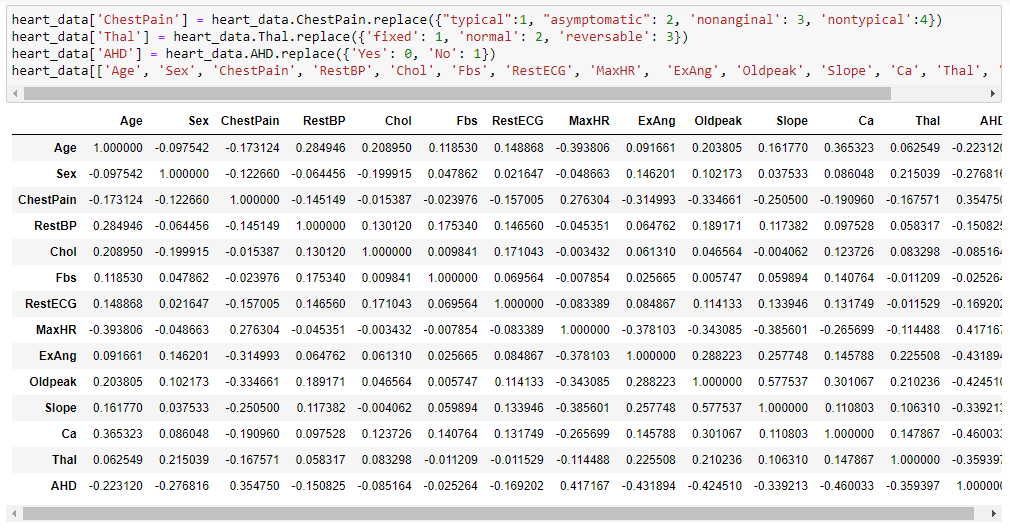


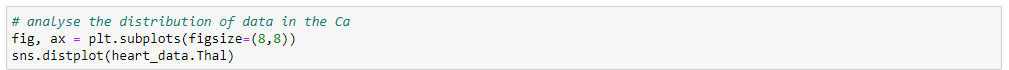


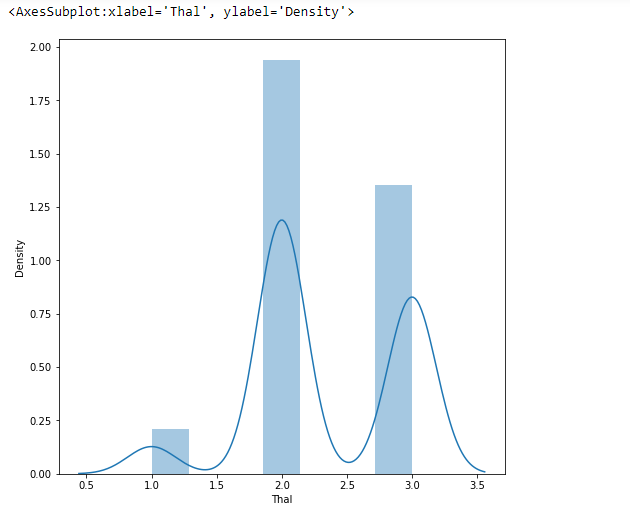


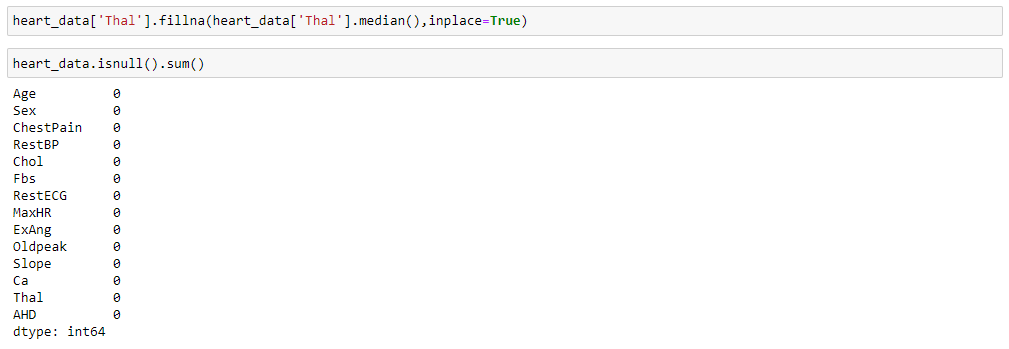


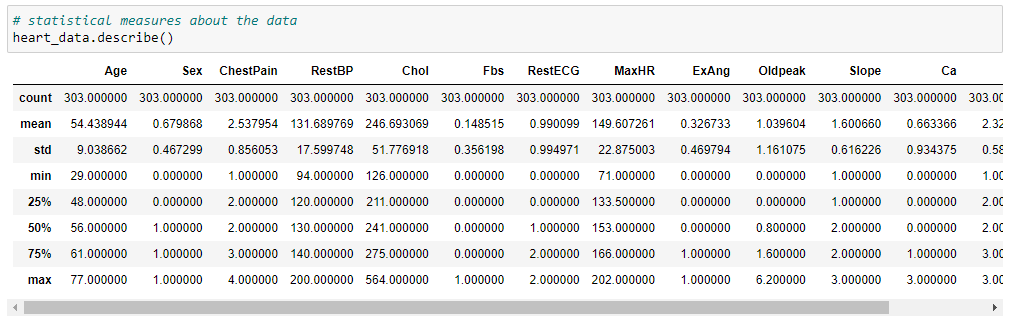
**Variable Correlation:**

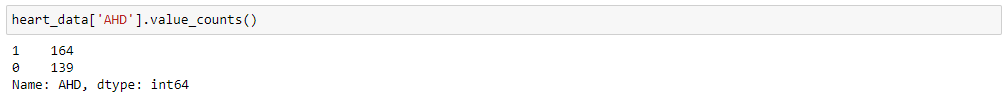






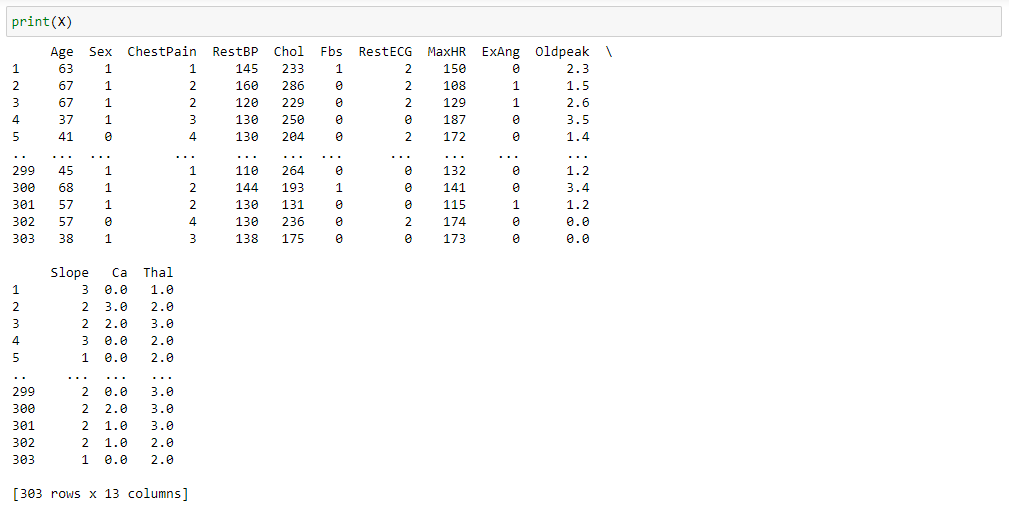


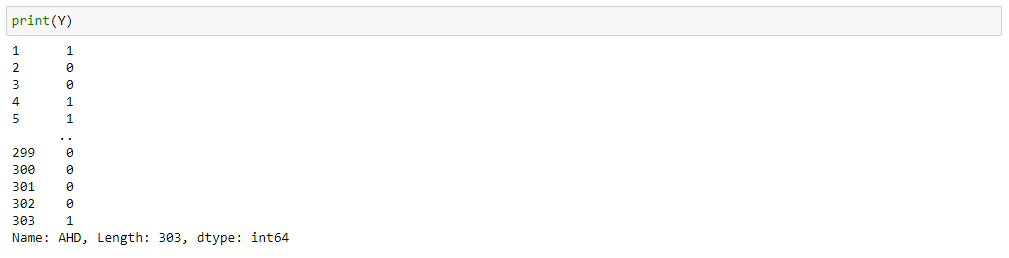




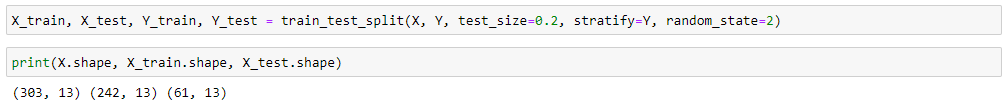
**Splitting the Features and Target:**



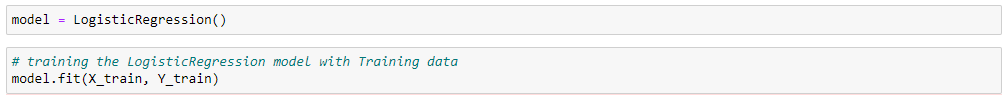




**Splitting the Data into Training data and Test data:**



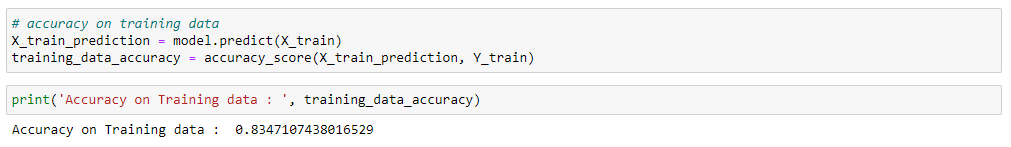
**Model Training:**



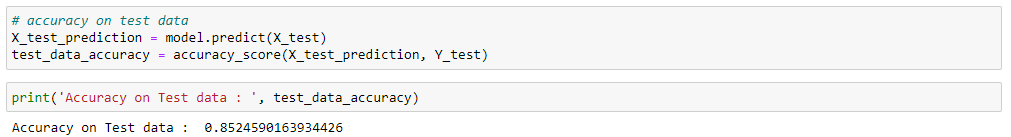
Result

**Accuracy Score:**

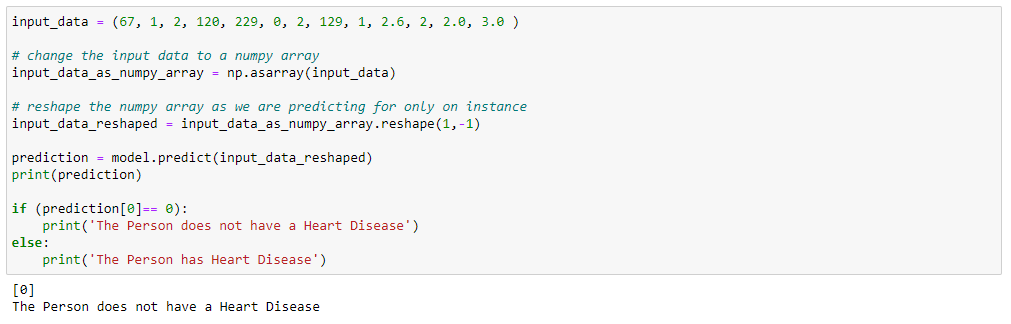
Accuracy on Training data is 0.83%



Accuracy on Test data is 0.85%



**Result of Predictive System:**



Conclusion

The amount of Heart diseases can exceed the control line and reach to maximum point. Heart disease are complicated andeachandeveryyearlotsof people are dying with this disease By using this all systems one of the major drawbacks of these works is mainly focus only to the application of classify techniques and algorithms for heart disease prediction, by all these studying various data cleaning and mining techniques that prepare and build a dataset appropriate for data mining. So that I can use this Machine Learning in that logistic regression algorithms by predicting if patient has heart disease or not. Any nonmedical employee can use this software and predict the heart disease and reduce the time complexity of the doctors.

Summary

To make this Heart Disease Prediction System I have firstly analyse the dataset completely by importing all the libraries which will be needed, collect the data and process the data. Then I splited the data into two Train data and Test data. Afterwards, I defined the machine learning model that I was using that is “Logical Regression Model”. Then I calculated the accuracy score which was 0.83 of train data and 0.85 of test data, probably the good accuracy score. Lastly, I printed the result by taking some input values from the dataset and cross check it with the dataset and I got the accurate result.