**KSI: *Machine Learning & AI Dual certification***

Project build under the guidance of Professor **Nitesh Karmakar**

Project 1: **Heart Disease diagnosis**

*Machine Learning: Logistic Regression*

**Name**: Mohammed Faraz Hussain

**College**: Terna Engineering College, Nerul

**Roll number**: TE-B-24

**ID number**: TU3F1920077

**Email**: [farazhussain503@gmail.com](mailto:farazhussain503@gmail.com)

**Terna Mail**: [mohammedhussain@ternaengg.ac.in](mailto:mohammedhussain@ternaengg.ac.in)

**Phone number**: +91 9082905016 / +91 7303272780

**Table of Contents**

**Sr.No Title**

| 1 | | Introduction | |
| --- | --- | --- | --- |
|  | |  | |
| 2 | | Aim and Objective | |
|  | |  | |
| 3 | | About Dataset | |
|  | |  | |
| 4 | | Design | |
|  | |  | |
| 5 | | Implementation | |
|  | |  | |
| 6 | | Conclusion | |
|  | |  | |
| 7 | | References | |
|  | |  | |
|  | |  | |
|  | | | |
|  | |  | |
|  | |  | |
|  | | | |
|  | | | |

**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.

The World Health Organization has estimated that four out of five cardiovascular diseases(CVD) deaths are due to heart attacks.

**Aim and Objective**

The prediction of heart disease is considered one of the most important topics in the health domain.

Also, The amount of data in the healthcare industry is huge.

Heart disease is a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and many other factors. Due to such constraints, scientists have turned towards modern approaches like Data Mining and Machine Learning for predicting the disease.

Hence,

With machine learning algorithms and having large amounts of data, it is possible to extrapolate information that can help doctors make more accurate predictions.

**About Dataset**

Heart disease describes a range of conditions that affect your heart.

* **Age:** The person's age in years
* **Sex:** The person's sex (1 = male, 0 = female)
* **Chest Pain/Angina:** The chest pain experienced
  + Value 1: typical
  + Value 2: nontypical
  + Value 3: asymptomatic
  + Value 4: nonanginal
* **RestBP:** 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 measurement
  + 0 = normal
  + 1 = having ST-T wave abnormality
  + 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
* **MaxHR:** 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.
* **Slope**: the slope of the peak exercise ST segment
  + Value 1: upsloping
  + Value 2: flat
  + Value 3: downsloping
* **Ca**: The number of major vessels (0-3)
* **Thal**: A blood disorder called thalassemia
  + Value 1: normal
  + Value 2: fixed defect
  + Value 3: reversible defect

**Design**

Since the AHD column exists, which depends on the independent 13 features. Hence, it is a dependent variable.

As a dependent variable exists, it falls into the category of supervised learning.

Also, the output of the model is either **Yes** or **No.** It is a binary classification.

Hence, we will use Logistic Regression for this problem.

Logistic Regression is a statistical and machine-learning technique classifying records of a dataset based on the values of the input fields.

It predicts a dependent variable based on one or more sets of independent variables to predict outcomes .

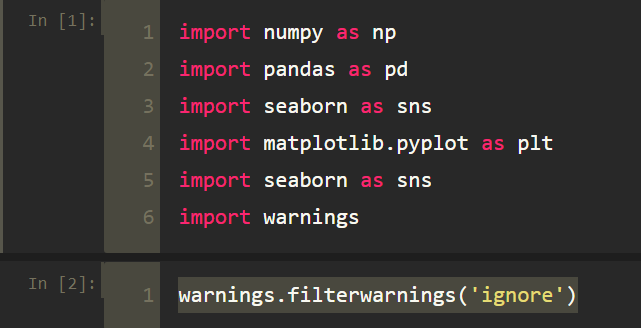
It can be used both for binary classification and multi-class classification

In this project, we have to predict whether the person has heart disease or not, which means, it falls into the category of binary classification(Since, we have to predict- Yes/No).

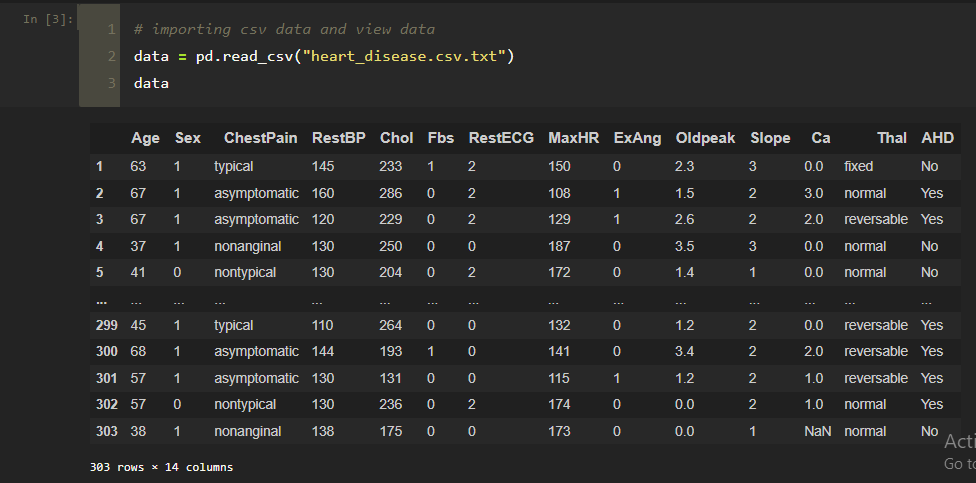
**Implementation**

To solve a machine learning model, we need to follow few steps. They are as follows:

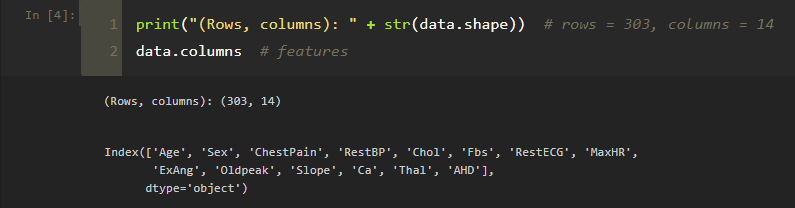
1. Importing libraries
2. Importing Dataset and Read the data (from csv)
3. Identify the dependent and independent variables.
4. Check if the data has missing values or the data is categorical or not.
5. Visualize the data.
6. Now split the data into the groups of training and testing for the respective purpose.
7. After splitting data, fit it to a most suitable model.
8. Prediction
9. Model Evaluation
10. **Importing Libraries**



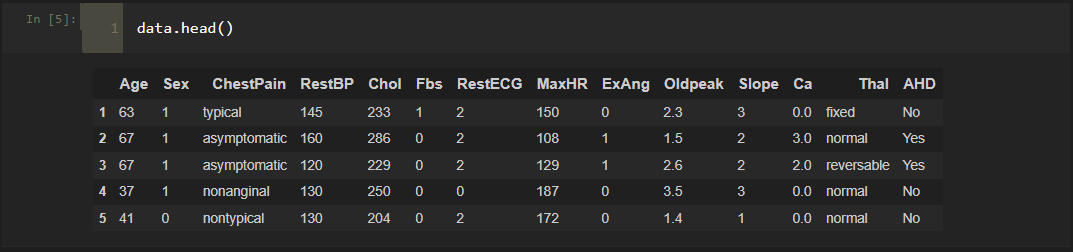
1. **Importing Dataset and Read the data (from csv)**

****

**2.1 Checking rows and columns**

****

**2.2 Checking first 5 values from the imported dataset**

****

**3. Identify the dependent and independent variables.**

**Independent variables:**

****

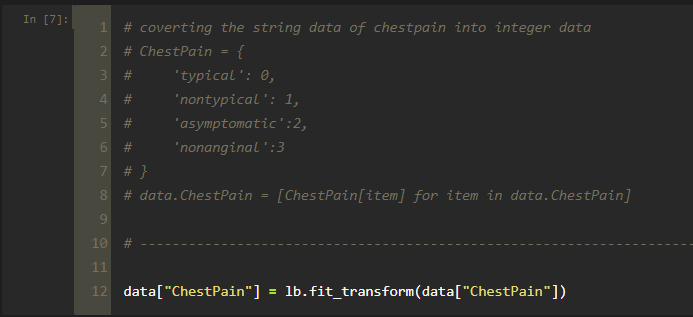
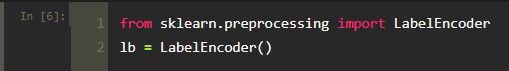
**Dependent variables:**

****

**4. Check if the data has missing values or the data is categorical or not.**

**Changing string data type to numeric data type:**

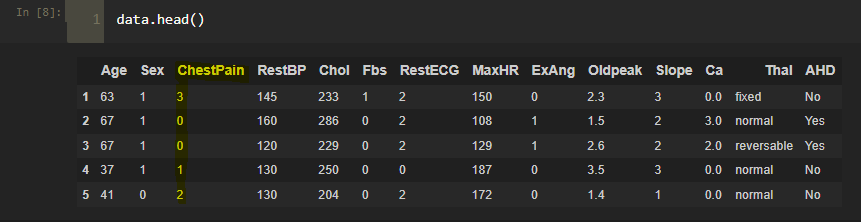
* Data generally needs to be put into numeric form for machine learning algorithms to use the data to make predictions.
* Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.

****

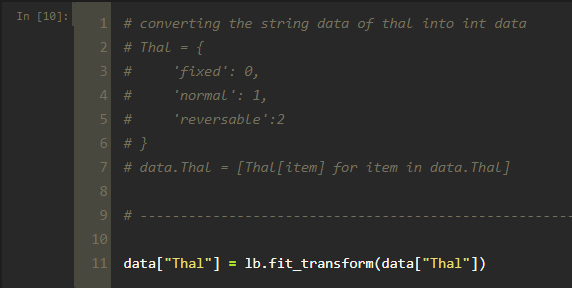
**Before LabelEncoding:**

****

**After LabelEncoding:**

****

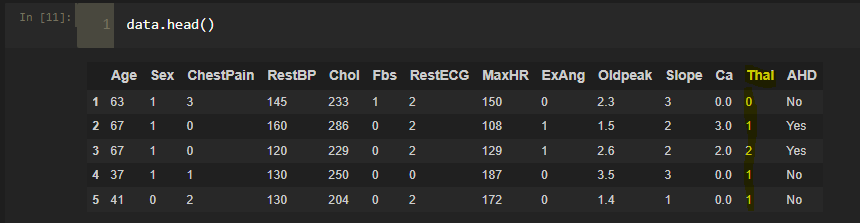
**-------------------------------------------------------------**

****

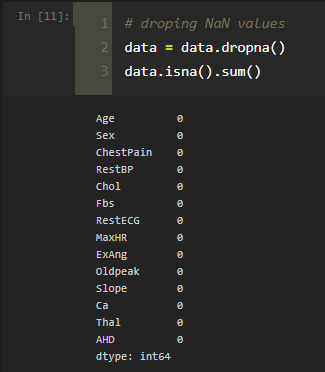
**Before LabelEncoding:**

****

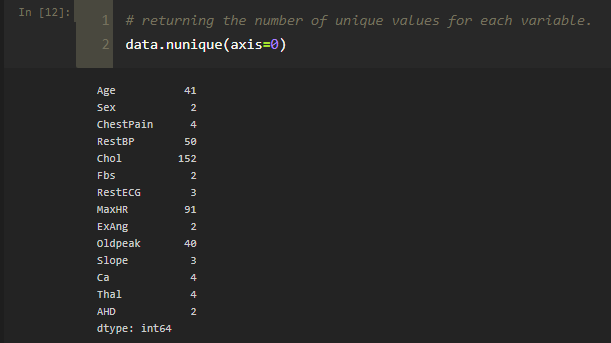
**After LabelEncoding:**

****

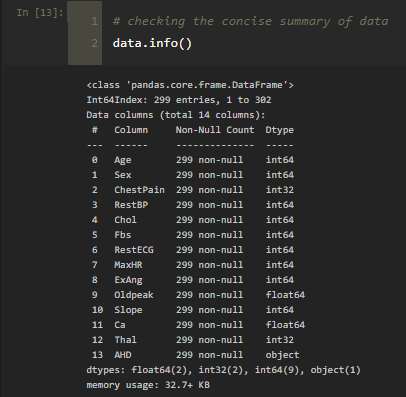
**4.1: Dropping NaN/Null Values**

****

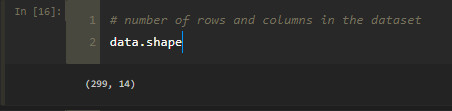
**4.2: Unique values for each variable:**

****

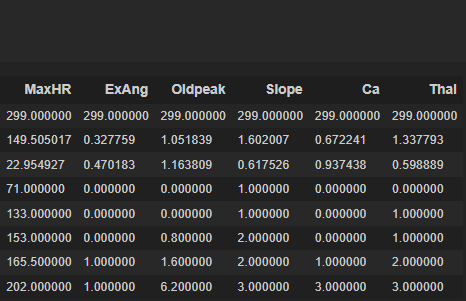
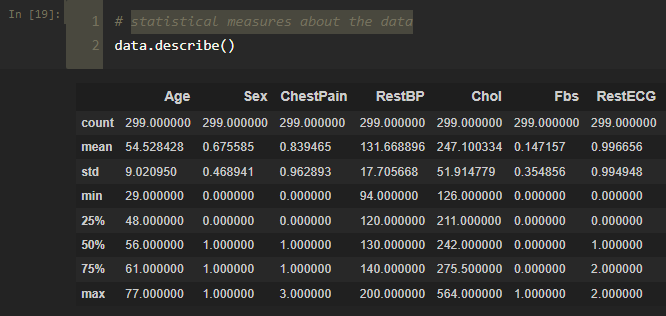
**4.3: Concise summary of data**

****

**4.4: Rows and columns in the dataset after dropping NaN values**

****

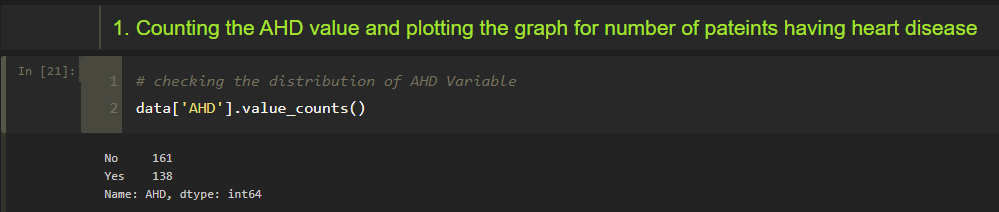
**4.5: Statistical measures about the data**

****

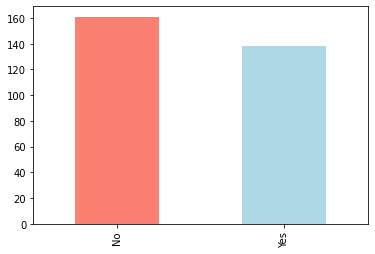
**4.6: Correlation matrix: Correlation indicates how the features are related to each other or to the target variable.**

****

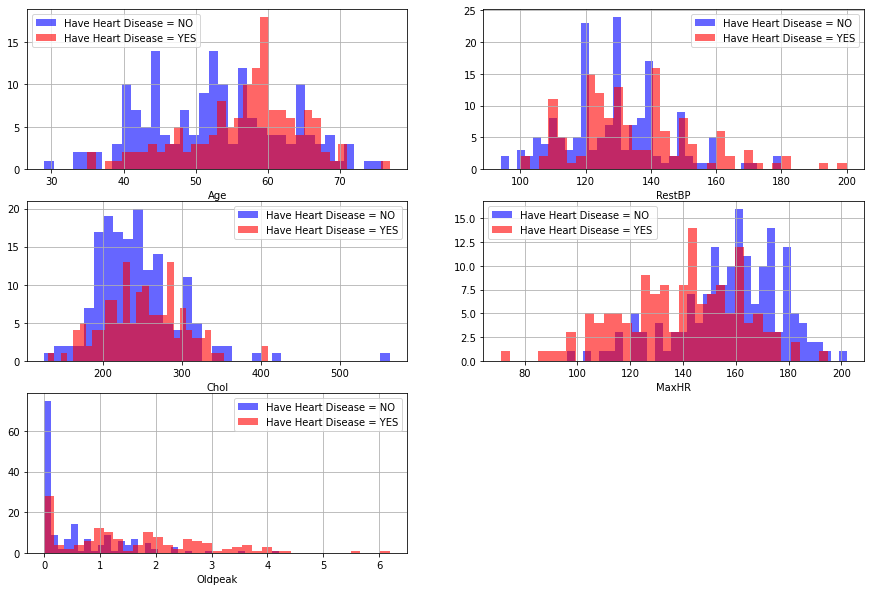
**5. Visualizing the data**

****

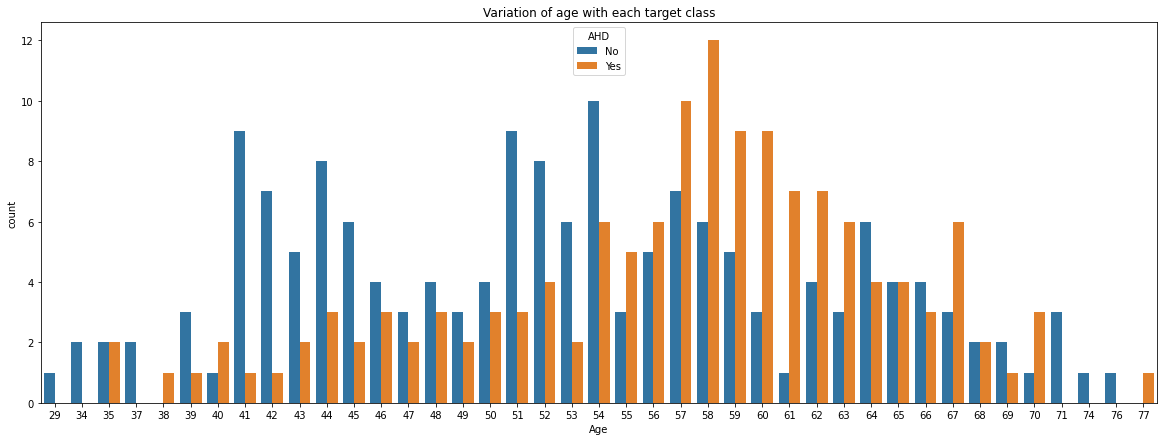
**Graph for number of patients having heart disease**

****

**Relation of various attributes with the target**

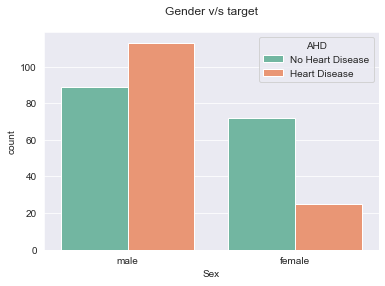
****

**Variation of age with each target class**

****

**According to this dataset, Males are more susceptible to get Heart Disease than Females.**

**Proof:**

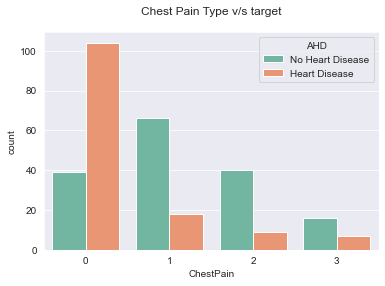
****

**There are four types of chest pain, typical, non typical, asymptomatic and nonanginal.**

**Most of the Heart Disease patients are found to have asymptomatic chest pain**

* + Value 1: typical
  + Value 2: non typical
  + Value 3: asymptomatic
  + Value 4: nonanginal

**Proof:**

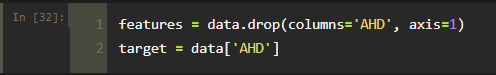
****

**Heart Disease is very common in the seniors which is composed of age group 60 and above and common among adults which belong to the age group of 41 to 60. But it’s rare among the age group of 19 to 40 and very rare among the age group of 0 to 18.**

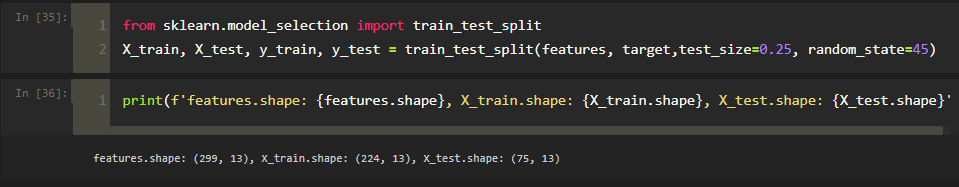
**Proof:**

****

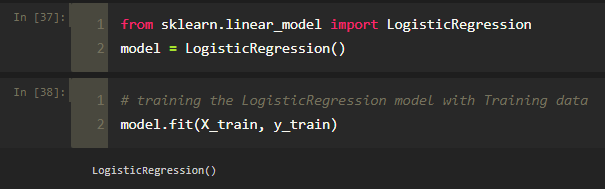
**5. Prepare Data for Modeling**

****

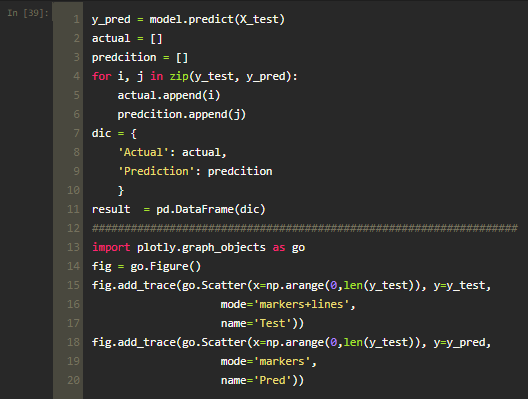
**6. Splitting the data**

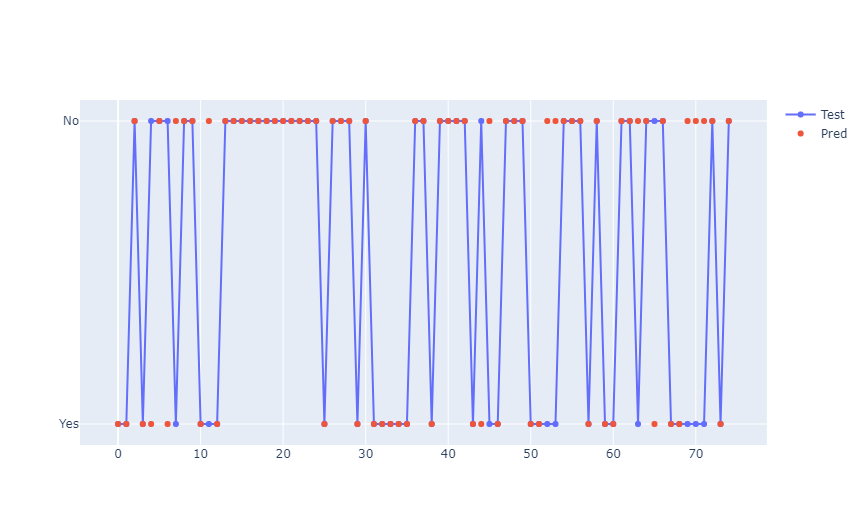
****

**7. Fitting into Model**

****

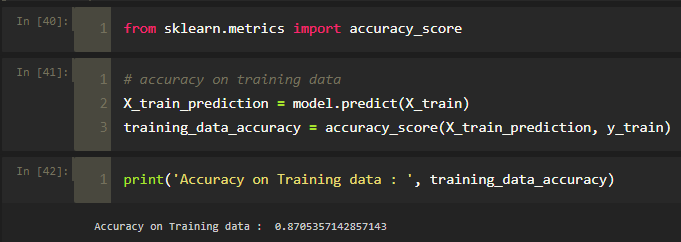
**8. Prediction and plotting ‘predicted and actual’ graph**

****

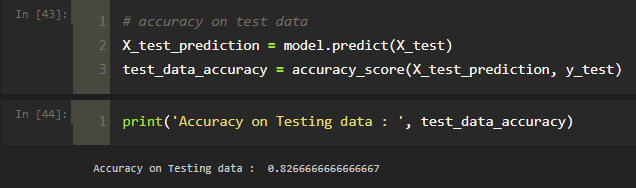
****

* The red dots represent the predicted values that are either 0 or 1 and the blue line & and dot represents the actual value of that particular patient.
* In the places where the red dot and blue dot do not overlap are the wrong predictions and where the both dots overlap those are the right predicted values.

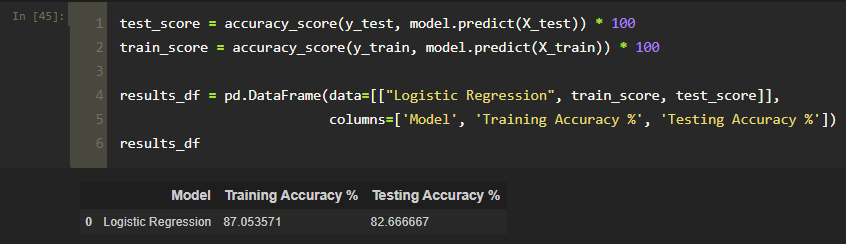
**9. Model Evaluation**

****

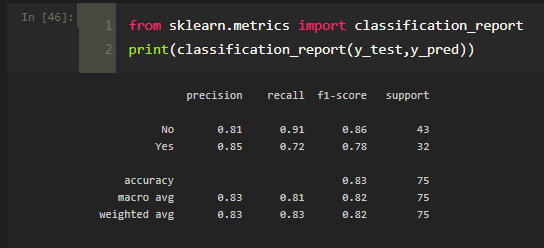
**Accuracy on Training data : 0.8705357142857143**

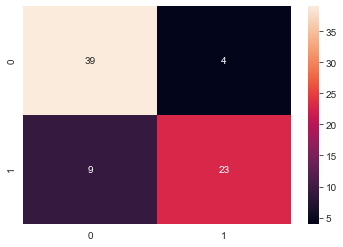
****

**Accuracy on Testing data : 0.8266666666666667**

****

**Classification report:**

****

**Confusion Matrix:**

**Conclusion**

Heart Disease is one of the major concerns for society today.

It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data.

Thus, we have successfully completed our project by choosing Logistic Regression.

**References**

* [**https://towardsdatascience.com/heart-disease-prediction-73468d630cfc**](https://towardsdatascience.com/heart-disease-prediction-73468d630cfc)
* [**https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7**](https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7)