Heart Disease Prediction



1. **Problem Definition**

The term “heart disease” is often used interchangeably with the term “cardiovascular disease”. Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina) or stroke.

**Heart disease** describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you’re born with (congenital heart defects), among others.

In this article, I will be applying Machine Learning approaches(and eventually comparing them) for classifying whether a person is suffering from heart disease or not, using one of the most used dataset — [Cleveland Heart Disease dataset](https://archive.ics.uci.edu/ml/datasets/Heart+Disease) from the [UCI Repository](https://archive.ics.uci.edu/ml/index.php).

**The Data**

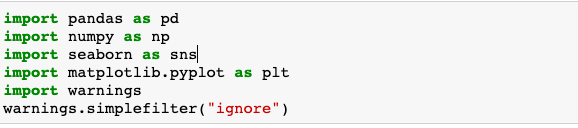
The dataset used in this article is the Cleveland Heart Disease dataset taken from the UCI repository.

The dataset consists of 303 individuals data. There are 14 columns in the dataset, which are described below.

|  |  |
| --- | --- |
| **Columns /Features** | **Description** |
| **age** | **The person’s age in years** |
| **sex** | **The person’s sex**  **Male=1, Female=0** |
| **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)** |
| **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— 0: Downsloping**  **Value-- 1: Flat**  **Value- 2: Upsloping** |
| **ca** | **The Target bar 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 = Yes, 0= No)** |

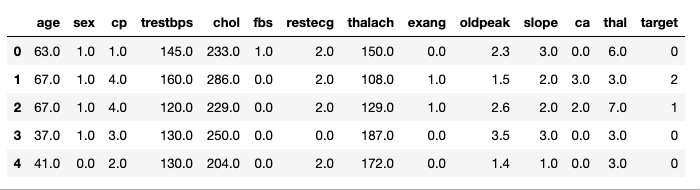
1. **Explanatory Data Analysis**

To start with, importing all dependencies and the dataset to get the information about the subject.

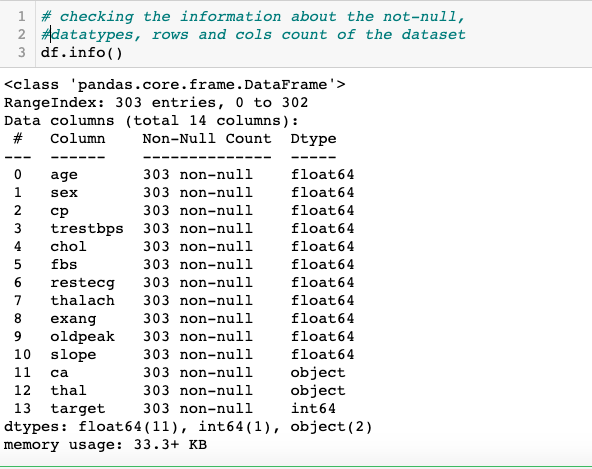




By using the above code, following data will appear on screen to have a sense of the information as per records held.



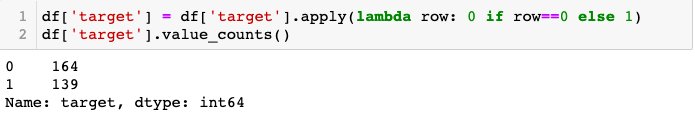
Now, to have an idea about the columns and their datatypes I will check dataset information



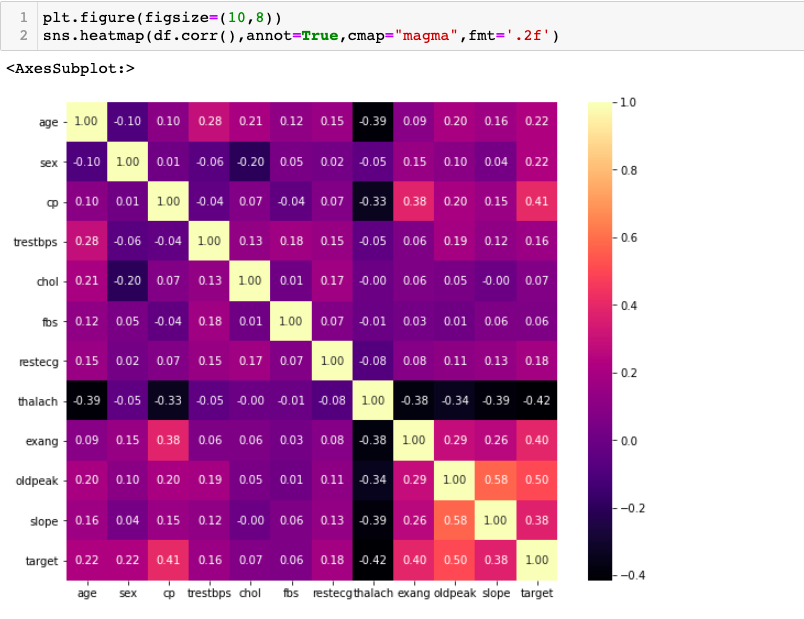
The dataset is looking pretty clean. No data is missing in any of the features as the count for all of them is same and same as number of rows too. There are no missing or null values in the dataset.

Target variable displays whether the individual is suffering from heart disease or not. Here, I will assign 0 to absence of disease and 1 to presence of disease.

|  |  |
| --- | --- |
| Attribute | Value |
| Absence | 0 |
| Presence (1,2,3,4) | 1 |



Now we have absence and presence as 0 and 1 , we will try to see the correlation between the features .



Some key Observations from the correlation :-

1. cp(chest pain ) and oldpeak(ST depression induced by exercise) are highly correlated with the target variable which means chest pain and ST depression are the main common factors behind the absence and presence of heart attack.

2. all variables are positively correlated with target variable (target) other than thalach.

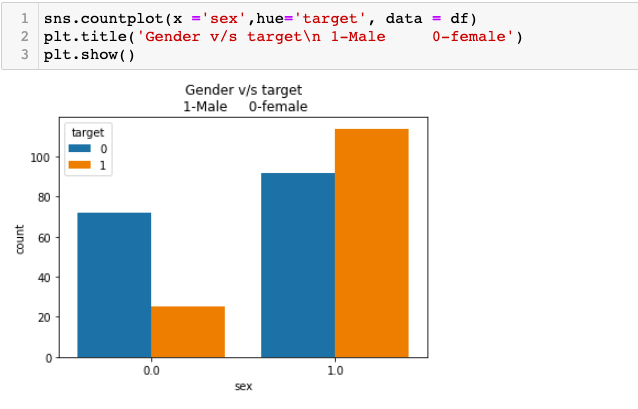
3. thalach is negatively correlated with target variable (target) which means there are very less chances of heart attack due to person’s maximum heart rate achieved .

I will try to discover some more dependencies by visualizing different plots .

**Data Visualization**

**Univariate Analysis & Bivariate Analysis**

Let us look at the people’s gender who are suffering from the disease or not.  
Here, target = 1 implies that the person is suffering from heart disease and target = 0 implies the person is not suffering.



Here, we can see that count of male is more who are suffering from heart disease as compared to female .

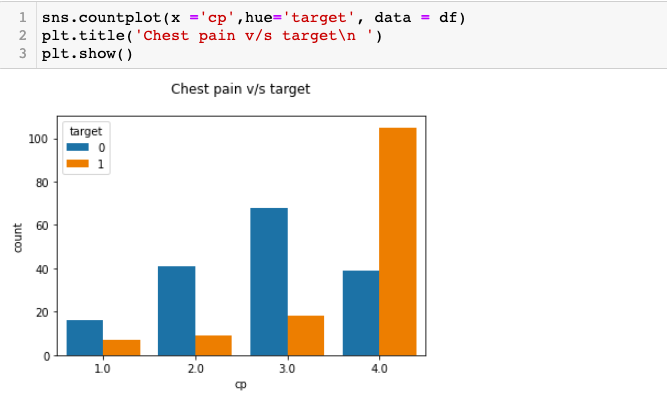
Let us look at the people with chest pain who are suffering from the disease or not.

Here, Chest Pain Type =1 implies that pain is Asymptomatic,

Chest Pain Type =2 implies that pain is Atypical Angina

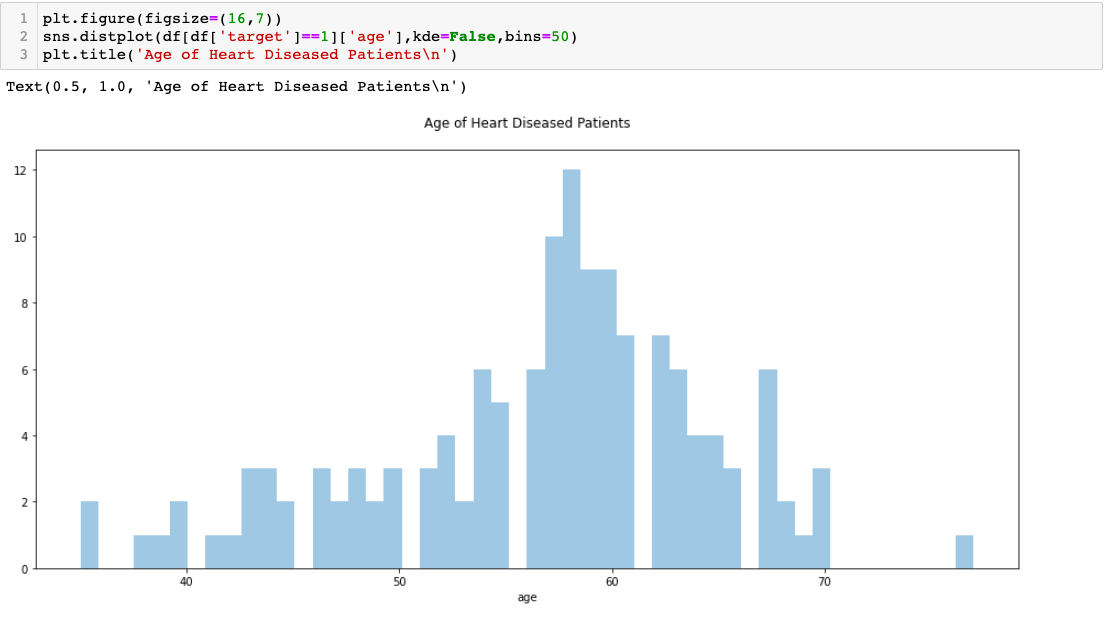
Chest Pain Type =3 implies that pain is Non-Anginal Pain

Chest Pain Type =4 implies that pain is Typical Angina



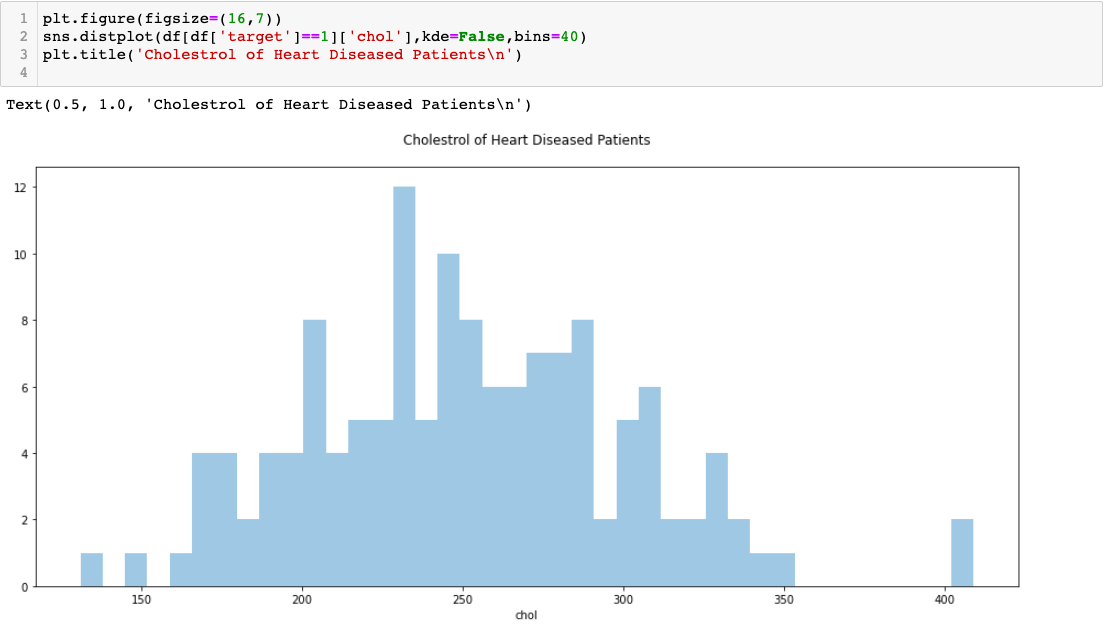
Here, we can observe that max males have the chest pain of type Typical Angina and max females have chest pain of type Non-Anginal.

Next , I will try to find out the age at which a person gets heart disease occurs the most.



Here, we can see that most people who are suffering are of the age of 58, followed by 57.Majorly, people belonging to the age group 50+ are suffering from the disease.

Let us have a look at the people having cholesterol who are suffering from the heart disease.

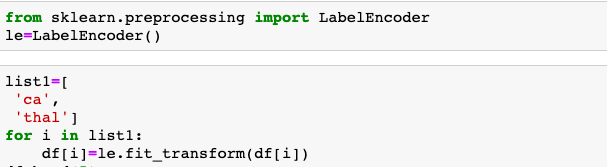


1. **EDA Concluding Remarks**

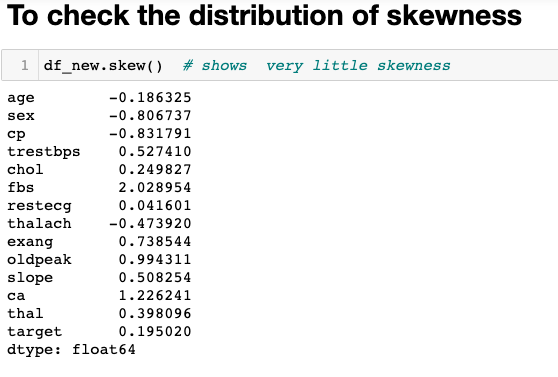
At this stage we can conclude that :-

1. Males are more who are suffering heart disease as compared to females.
2. Maximum males have the chest pain of type Typical Angina and maximum females have chest pain of type Non-Anginal.
3. People who are suffering from heart disease are of the age of 58, followed by 57.Majorly, people belonging to the age group 50+ are suffering from the disease.
4. Cholesterol level gives a big impact on heart disease.
5. **Pre-Processing Pipeline**

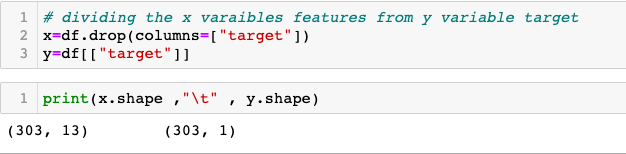
Maximum columns have the numerical values other than ca and thal. For moving further for model training, we need to transform the nominal values into numerical values by encoding the data.



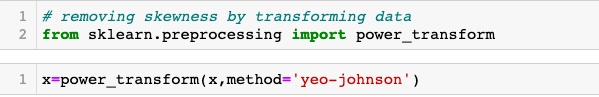
Now, the dataset is all numerical. I will check the distribution of skewness in the data. If it is there, I will remove it.



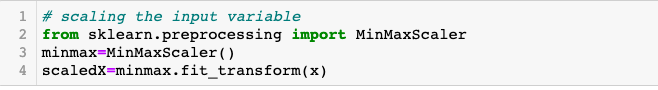
As we remove the skewness from x variables only, first I will split the x and y variable from dataset.



After splitting, we will remove the skewness from the data.



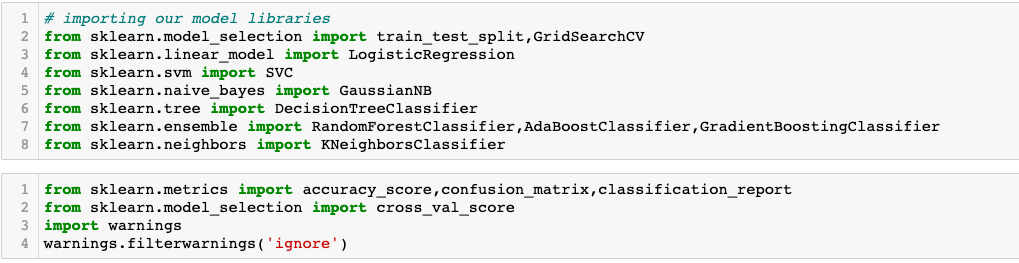
Once skewness removal has been done, I will do scaling of the dataset. **Scaling**is a technique to standardize the independent features present in the data in a fixed range. If feature **scaling**is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.



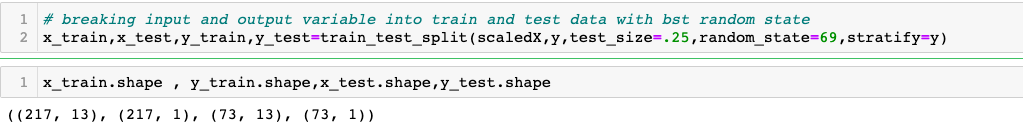
Now, all the data cleaning and preprocessing has been completed . Dataset is now ready for model training .

1. **Building Machine Learning Models**

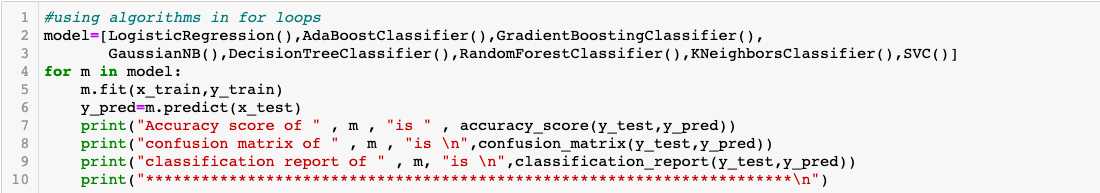
Let’s start with importing the dependencies.

****

Now let us divide the data in the test and train set. In this project, I have divided the data into 75: 25 ratio. That is, the training size is 75% and testing size is 25% of the whole data.

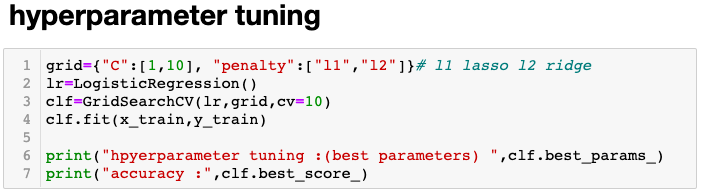


Using different algorithms, I will try to find the best model.

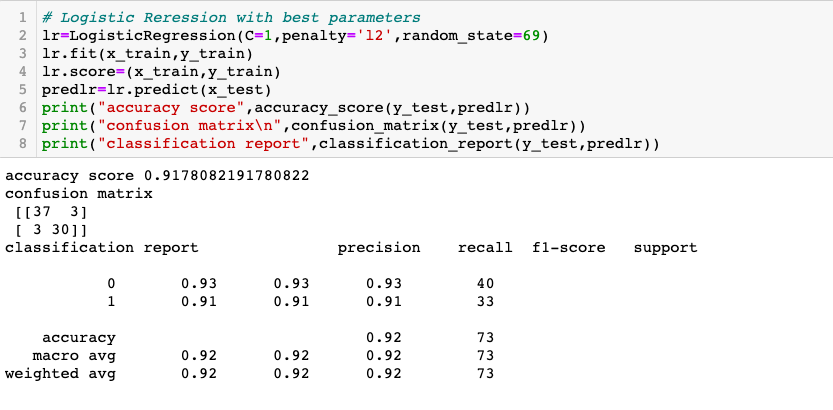


Here, LogisticRegression performed the best with accuracy score round(2) is 0.90

Doing a GridSearchCV is a great way to do hyperparameters tunning.



Finally using the best parameters.

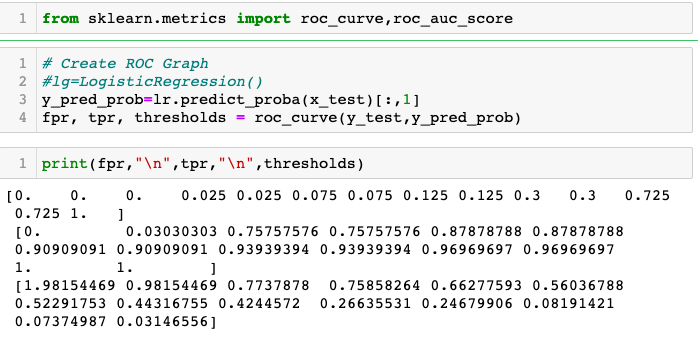


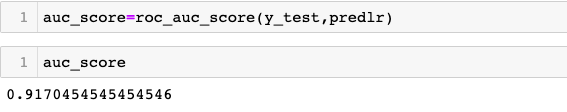
The accuracy score obtained here is  0.9178082191780822

A slight improvement but an improvement nonetheless.

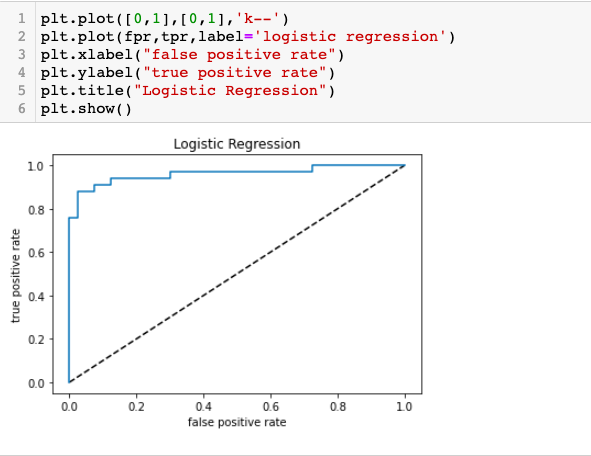
Next step will be find the roc\_auc score and plotting roc\_auc curve.

Let’s start with importing the dependency followed by finding roc\_auc score

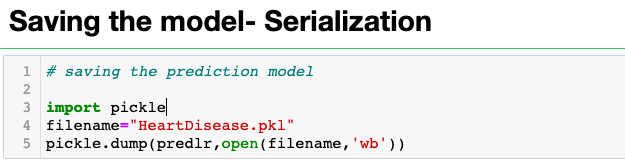




auc\_score is 0.9170454545454546. Let’s have a look at the graph.



Above graph is covering the maximum area of true positive rate. Next step will be to save the best fit model. For that we will do serialization using pickle.dump method and will load the fitted model .



1. **Concluding Remarks**

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.

I hope you find this article useful. Thanks

All the code can be found in my GitHub [here](https://github.com/piplani-bhavna/ProjectsDSR/blob/ProjectDSR/Project5_HeartDisease_Modelling.ipynb)