**Heart Disease Project**

**Definition:** Heart disease is very common disease now-a-days. We see daily that there are so many people who found with some heart disease problem. In this project we have taken different attributes to measure whether that person has heart disease or not.

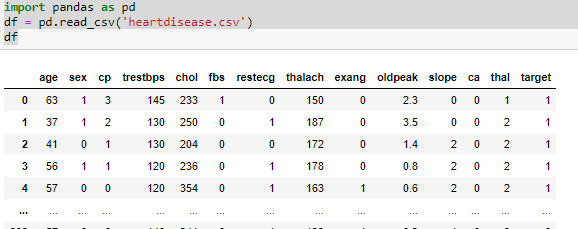
**Problem Definition:** By considering various attributes related to the heart disease I have used different machine learning algorithms to predict the problem.

For the above problem I have done some process step-by-step in the following way:

1. Importing data
2. Data wrangling
3. Data analysis
4. Data Pre-processing
5. Importing Machine Learning Algorithms
6. Model saving
7. Conclusion

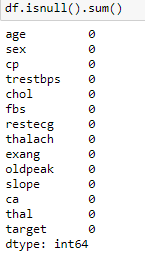
We will see all above steps in detail below-

1. **Importing Data:** First to get our data like from ‘csv file’, ‘excel file’ or any other file we have to import pandas library.



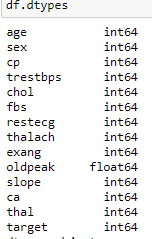
In this way we can read our data using read\_csv.

1. **Data Wrangling:** Our dataset is not always in proper format. Sometimes there maybe some missing values, sometimes there must be out of the box values. Our data is always a raw data so for that we have to do some wrangling process like.
2. Check if there are any null values are present in the dataset or not.

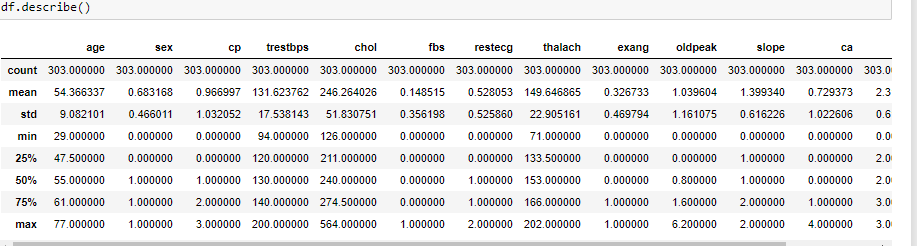


So by chance there are no null values in this dataset.

1. Check the datatype of each column.

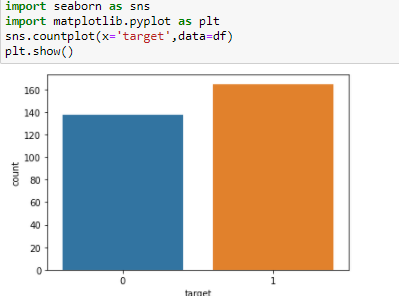


1. Check the summary of our dataset using

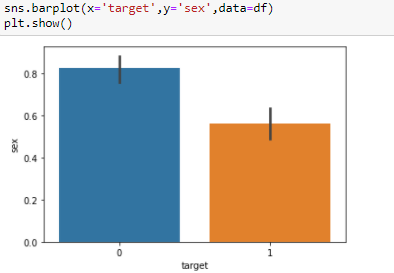


From above image we can say that all values for each column is different. For age column max value is 77 and 75% value is 61. So we can say that in such case data is ok. But for the chol column max value is 564 and 75% value is 274 so in this case we can say that some outliers are present in this column.

1. **Data Analysis:** In this step we have to do the analysis of each column using univariate method. And analysis of each column with target variable using bivariate method. So for this I have used Count method to check the number of unique values present in that particular column and also some visualization libraries to understand our data more clearly.

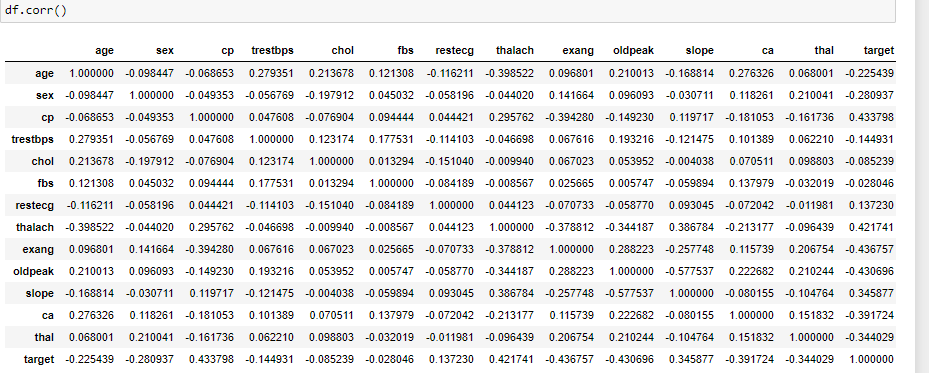


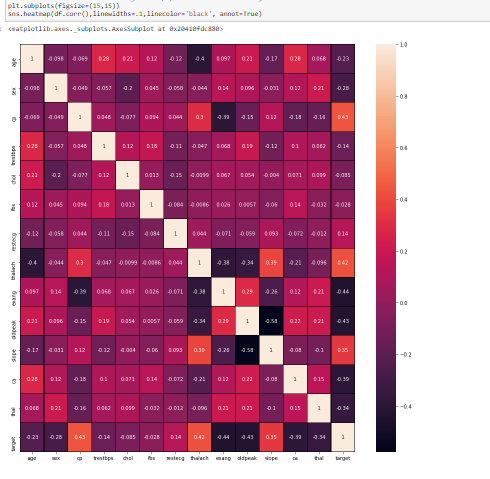
Seaborn and Matplotlib are two visualization libraries that I used to visualize graphically my data. This is the univariate method because I have visualized only one variable. The same method I have used for remaining variables.



This is the bivariate method because I have taken two variable one is target variable (target) and other variable as sex. In this way we have to visualize all variables with our target variable.

Now we have to find the correlation between all variables.



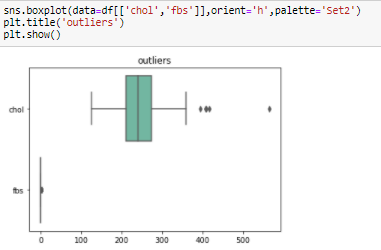


This above method is used to check the correlation between all columns. Heatmap method gives the graphical visualization from which we can easily understand the correlation. So from above diagram we can say that-

a.Thalach has direct positive relation with Target variable.

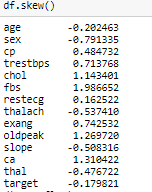
b.Cp has direct relation with Target variable

**Checking Outliers:** Our data is a raw data so we have to check whether any value is out of the box or not. (for eg. In Age column the age between 1-100 can be considered, but if there is age 200 then we can say that it is a outlier. So in that case we have remove that particular value from our datset. Otherwise it will gives us wrong predictions.)

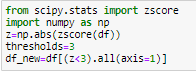


From above diagram we can say that there are some outliers in column chol and other column fbs is ok. So in the same way we have to check the outliers for ll columns.

**Skewness:** Checking the skewness in our data is important for predictions. So we have check the skewness.



From above diagram we can say that some variables have positive skewness and some has very negative skewness. So we have to remove skewness and and outliers using following method.



Zscore is used to normalize our dataset.



Yeo-Johnson method is used to remove the skewness from our data.

1. **Data Pre-Processing:** Now we have to do the standard scaling of our data which means we have to make our data in particular level, so that it will be easy for the predictions.



Splitting the data: We have to split our data for training and testing. So from this we can train some data and do the testing on remaining data.

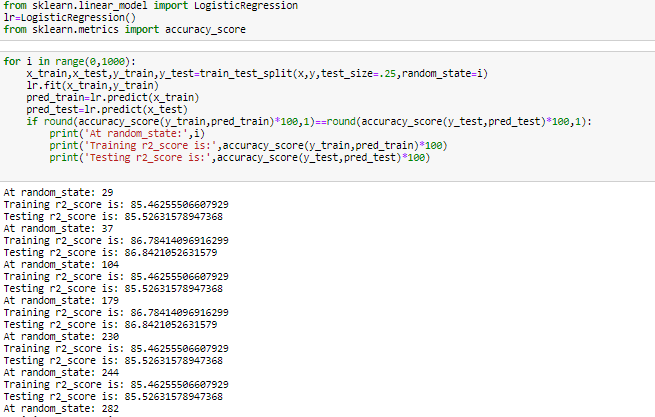


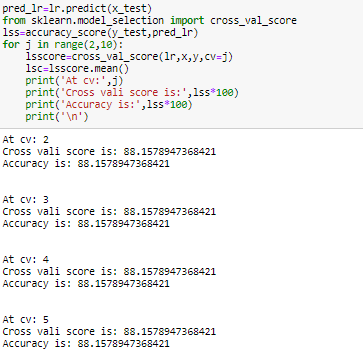


First we split our data in x and y variable. In x we have all input columns except out target variable and in y we have our target variable.

1. **Importing Machine Learning Algorithms:** In this project I have used 3 machine learning algorithms with different parameters. Then I have checked which algorithm gives the best accuracy for saving the model.

In this I have used Logistic Regression, KNeighborsclassifier, DecisionTreeclassifier. And also used Cross validation for better accuracy and AUC-ROC curve to see it graphically.





We have to do the same steps for other algorithms like KNeighborsclassifier, DecisionTreeclassifier. But in this case I have got good accuracy using Logistic regression. So I have this algorithm for saving the model. From above we can say that Cross validation score and accuracy is nearby same. So we can say that our model has performed better with accuracy 88.15%.

1. **Conclusion:** From above we concluded that Logistic Regression is better performed for this dataset with accuracy >88%.
2. **Model Saving:** Now we have to save our best model using following method.

