Heart conditions that include diseased vessels, structural problems and blood clots are called as Heart Disease or termed as Cardiovascular Disease.

Machine learning a subfield of Artificial Intelligence plays a vital role in medical diagnosis, it detects patterns of certain diseases with in patient electronic health care records and informs clinicians of any irregularities. In this article we will be analyzing and predicting potential heart disease in people with the help of machine learning algorithms.

Heart Disease dataset is available in Kaggle and UCI.

PROBLEM DEFINITION

While looking into the introduction information and description of this dataset we could understand, the aim of this dataset is to predict whether the patient has heart disease or not, Heart Disease = yes or no, so our target variable should have either one of this two as answer hence target outcome is categorical type we will be analyzing with classifiers and if necessary draw a AUC\_ROC curve in order to predict with optimal accuracy.

* The target column has been depicted as:-

0 representing no heart disease

1 representing heart disease is present

* Classifiers adopted

Logistic Regression: It is an appropriate analysis to conduct when when the dependent variable is binary type, keeping target column positive heart disease or negative heart disease which is converted to binary type we choose logistic regression for predictive analysis.

K-Nearest Neighbors: The classifier which derives value of k from root of total number of samples and classification is made by giving majority votes to its neighbors which is more apt for our dataset as we can observe for similar age group, sex, type of chest pain etc., how people have been effected.

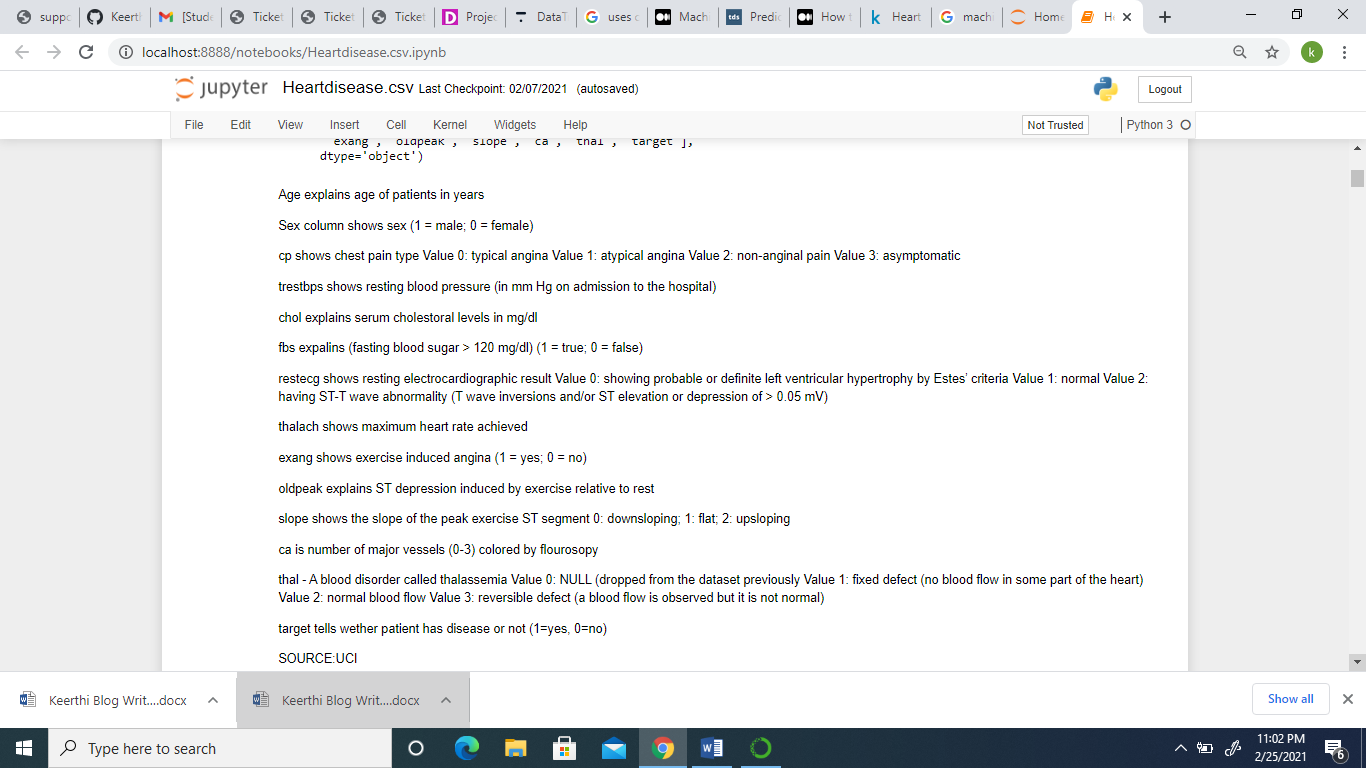
Decision Tree Classifier: It is basically analyzing data in a tree like structure, which is widely used in medical Dataset and best on categorical data.

GaussianNB Classifier: It is a classification algorithm which adopts Gaussian’s probability distribution which means it is a latent variable probabilistically related to observed variables. I prefer using naïve bayes whenever knn is used so that we can observe supervised learning and probabilistic estimation hand in hand.

Support Vector Classifier: The main objective of support vector classifier is to the data in hyperplane it leads in categorizing the data which helps to make predictions accurately.

UNDERSTANDING THE COLUMNS

Instead of importing all the libraries, we shall call when required hence started with importing pandas and data set. Understanding the columns descriptions is very important to analyze the data



From the above data we can understand there are 14 columns (13 attributes, 1 class) explained elaborately and further we passed df.shape, df.info, df.isnull, df.dtypes and found there is no null data, all the columns are integer type except old peak column so we don’t have to do the transformation procedure of converting data formats.

Exploratory Dataset Analysis (EOA)

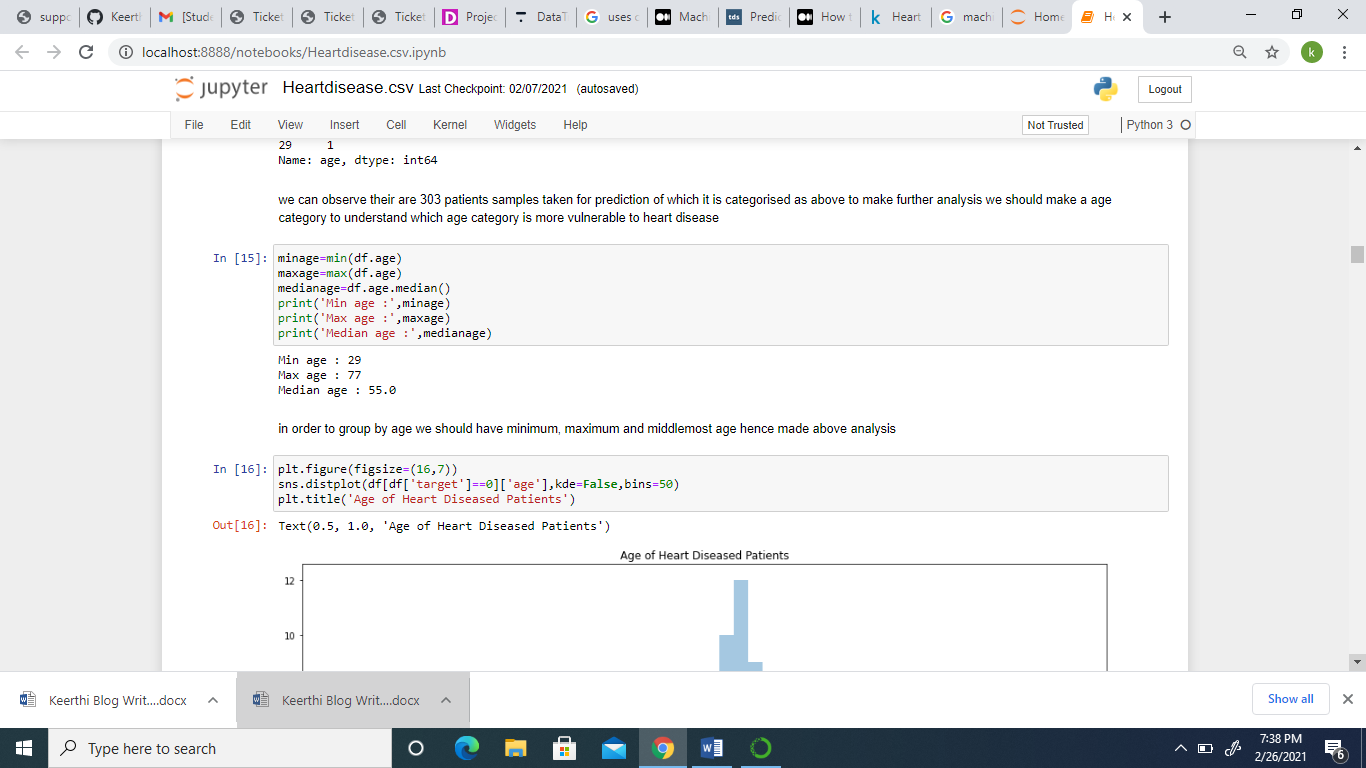
Let us now analyze the data set and summarize important columns in the form of visualization.

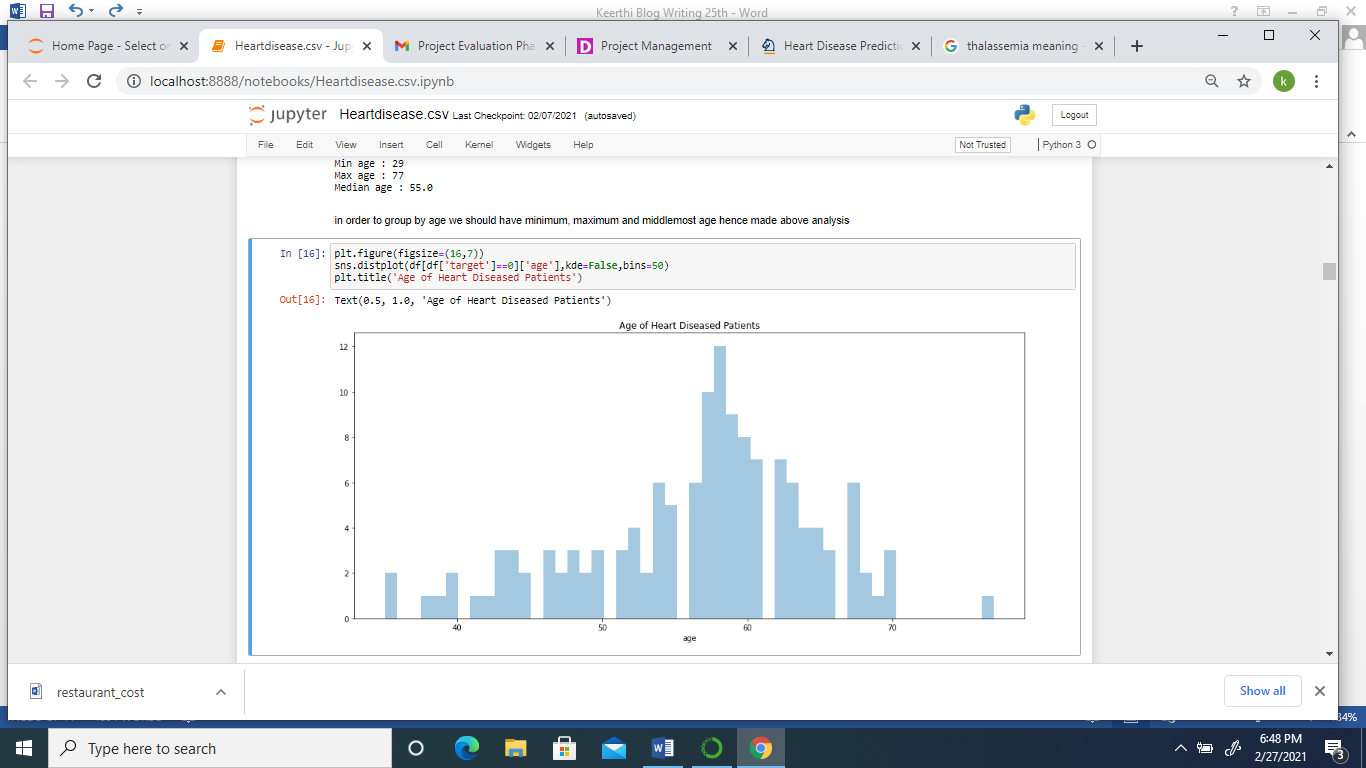
Firstly summary statistics:-

* df.corr() ,showing correlation of the target column with other attributes and while looking into columns which are positively correlated we can see chest pain is 40% related to heart disease, thalach (depicts maximum heart rate achieved) is 40%,Slope =30% and we can conclude no attributes has very high or very low correlation in particular.
* df.describe() help us understand how data is distributed, whether outliers or the data is normally distributed, the observation shows cholesterol , thalach has been disbursed widely so chances of outliers between maximum and Q3 data is possible.

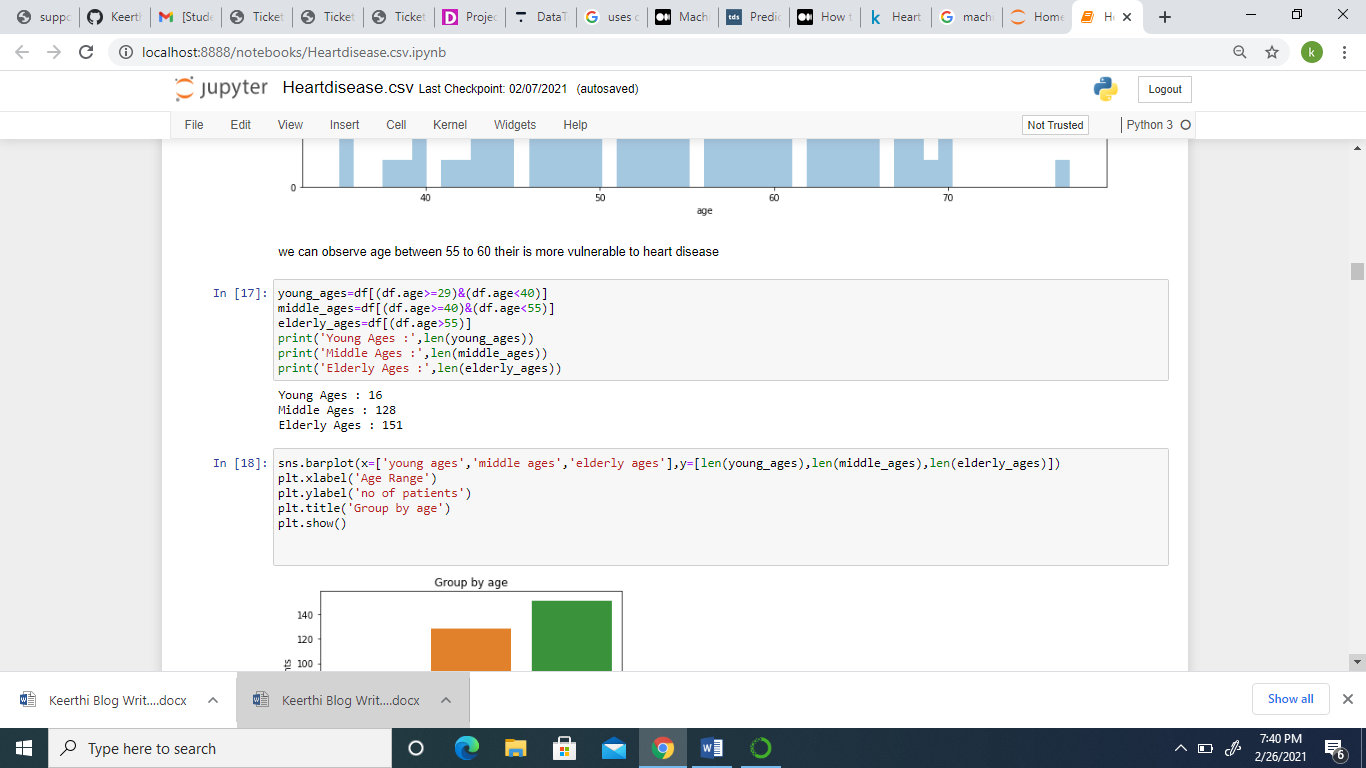
Further we are analyzing the data and understood there is no null or NaN values in the dataset and all the columns are integer type accept oldpeak which is float type, with this input about the columns we will move further for data visualization.

In order to observe the age range, we have considered Q1 as minimum age = 29, Q3 as maximum age = 77 and Median as middle age = 55 and categorized as young, middle and elderly age shown in the picture below



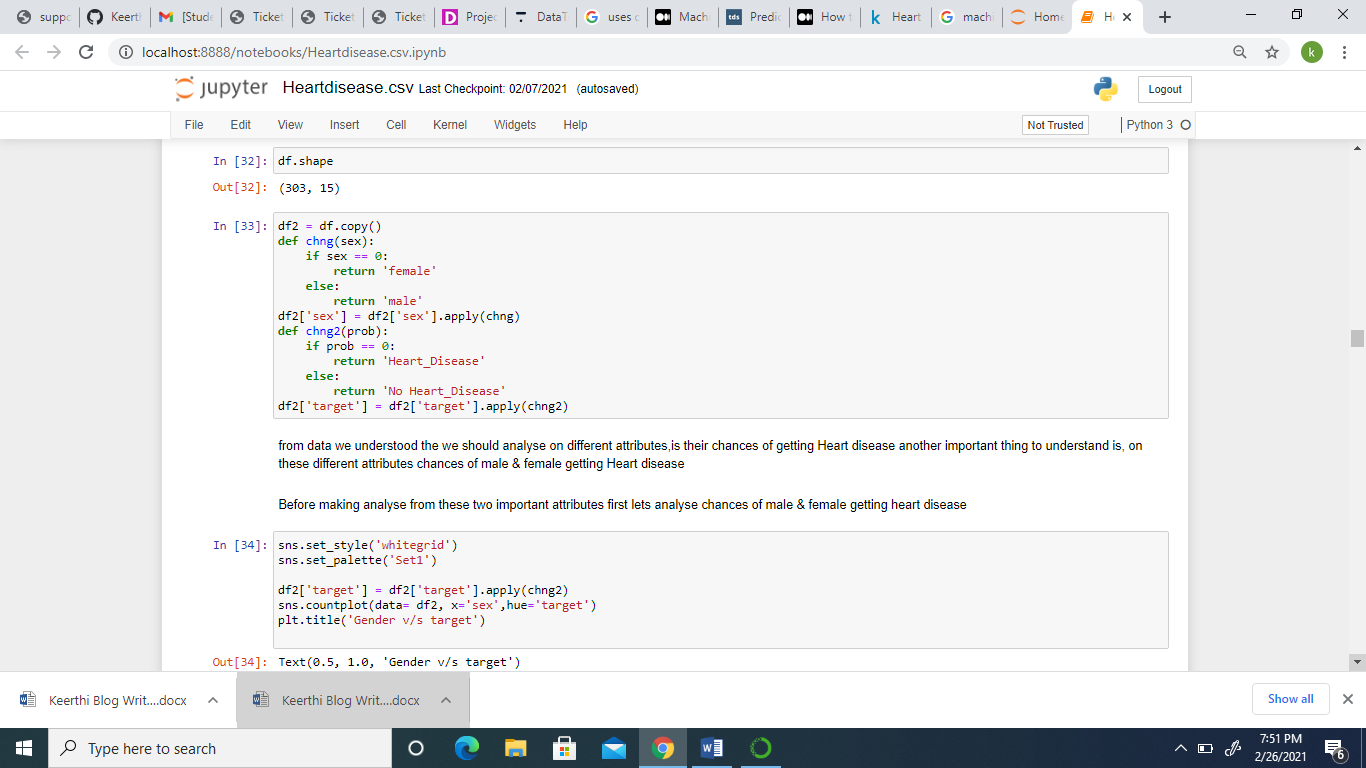


Here to divide as young age, middle age and elderly age category I have made distplot to understand how data has been distributed among different age ranges, we can the pattern has increased from 55 age range and so am considering 55 and above as elderly age and 29 to 55 mid-point approximately 40 has been considered for another partition of data so lowest 29 to 40 as young age and remaining 40 to 55 is considered as middle age, this way I segregated young, middle and elderly age category.



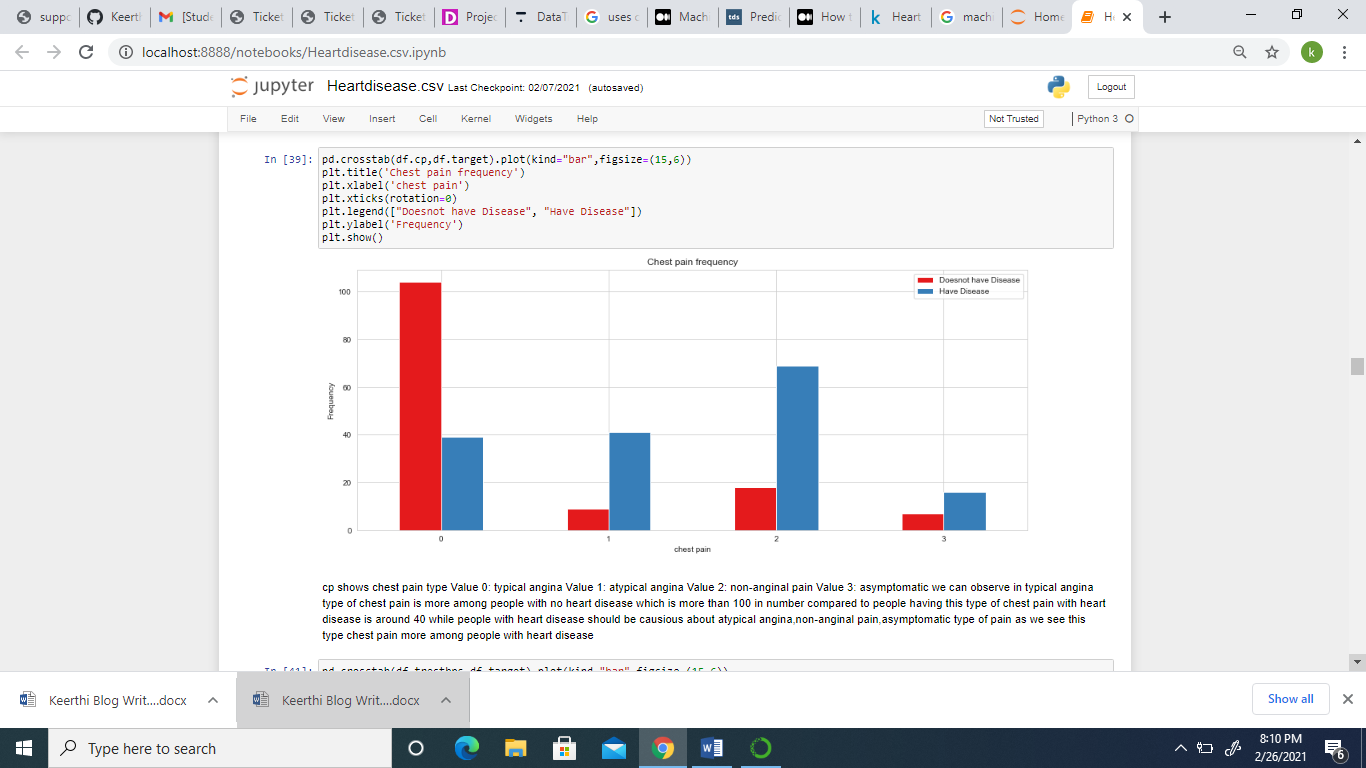
After segregating data age wise we have analyzed number of people under each age category, with age range is more vulnerable to Heart Disease and found that people in elderly age has high chances of Heart Disease followed by middle age.

Analysis on basis of sex also has to be categorized from binary to male and female in order visualize and drive a meaningful conclusion based on sex and target column, I have used def function to convert 0 to female and 1 to male and named 0 in target column as heart disease and 1 as no heart disease, the code used for the same is given below.

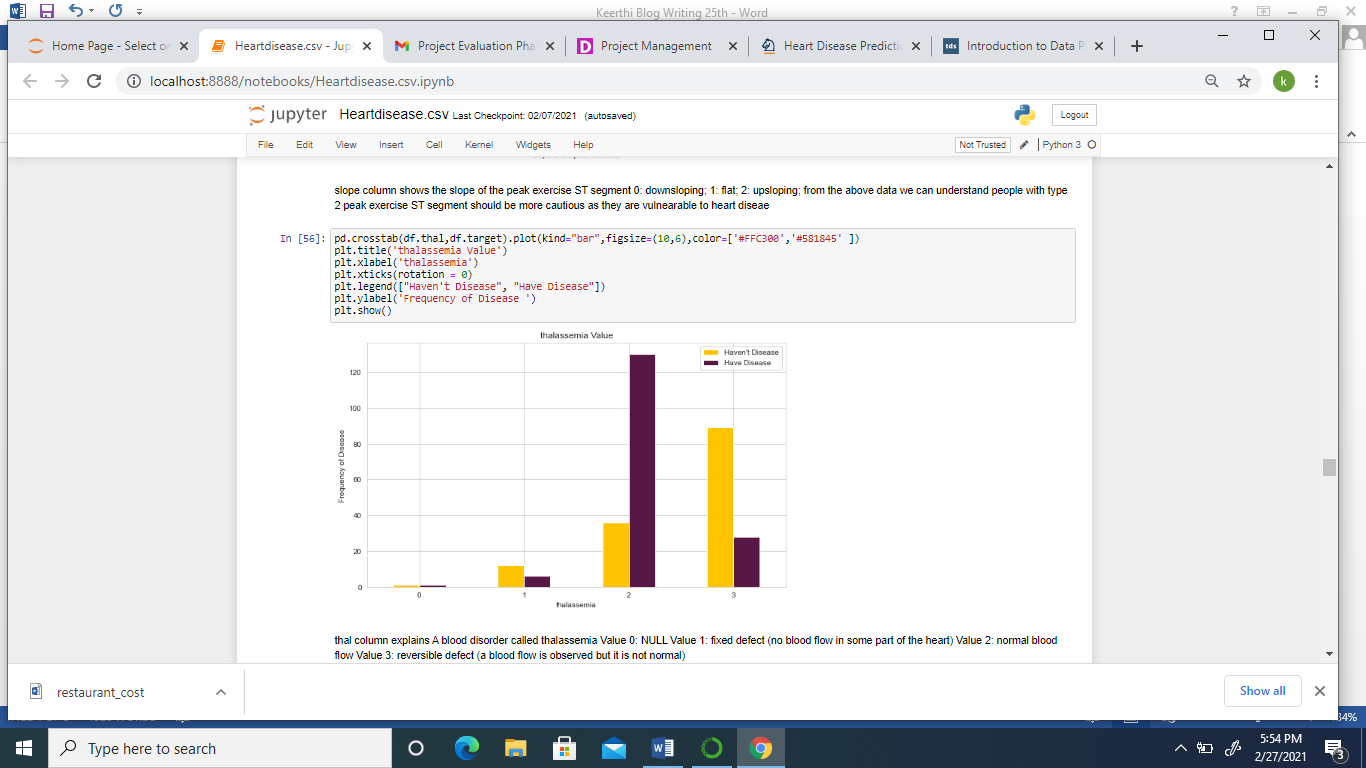


After segregating data on basis of male and female, heart disease and no heart disease we have passed count plot and understood with its result Male have more chance of getting Heart Disease compared to Female.

We will be observing all other attributes and their chances of getting heart disease with the help of crosstab as below:



In the above figure we can understand chest pain type is categorized as value 0,1,2 & 3 and found that among people who had got value0 type of chest pain over 100 people had negative result of heart disease and 38 people approximately had positive result of heart disease, while in value 1 type of chest pain 40 people had positive result and less than 10 had no heart disease as result, in case of value 2 type of chest pain we should be cautious as 70% facing this type of pain have got heart disease 18% .



Thal column explains a blood disorder called thalassemia which is divided as

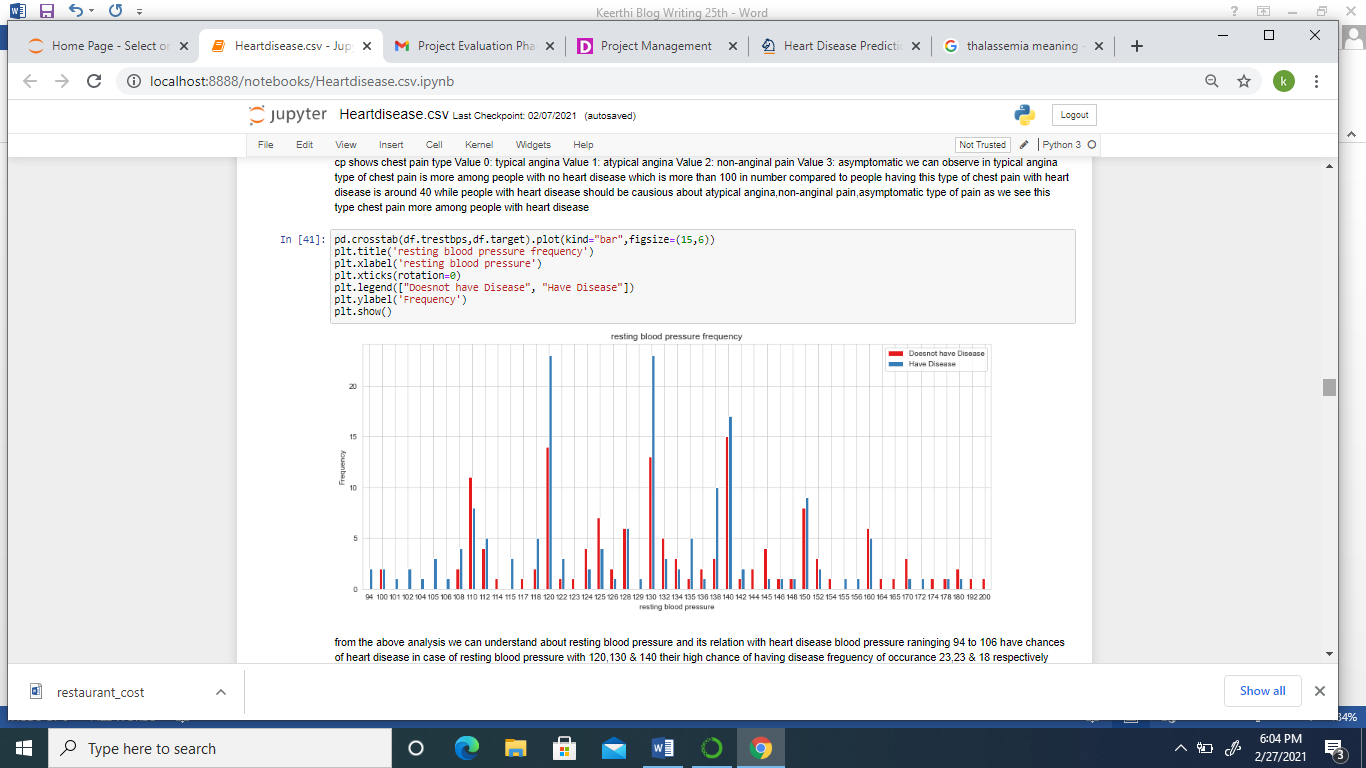
Value 0: NULL

Value 1: fixed defect (no blood flow in some part of the heart)

Value 2: normal blood flow

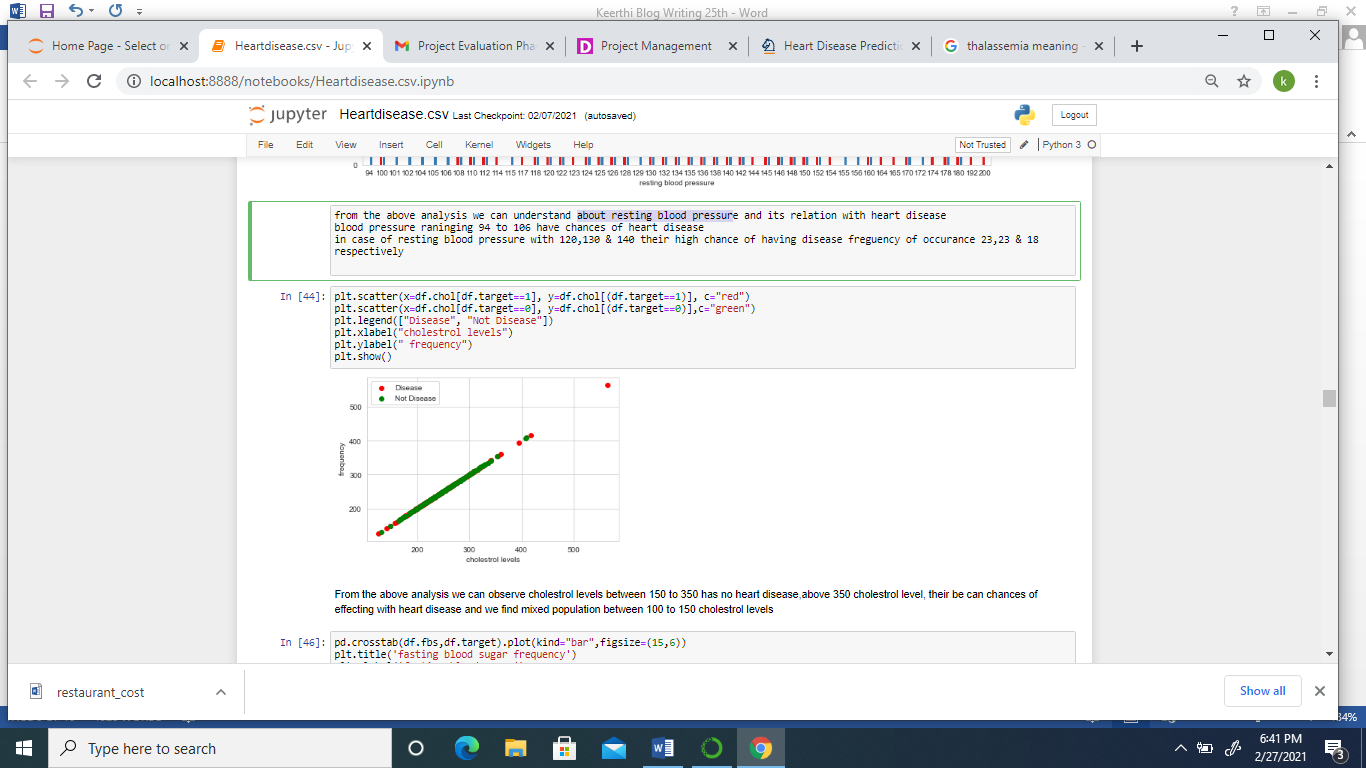
Value 3: reversible defect (a blood flow is observed but it is not normal)

Here thalassemia diseased person with value 2 type has to be more cautious as approximately 150 people have positive result for heart disease compared to 35 people have negative result to heart disease and value 3 type of thalassemia diseased people around 30 people have positive result and 90 people have had negative result.

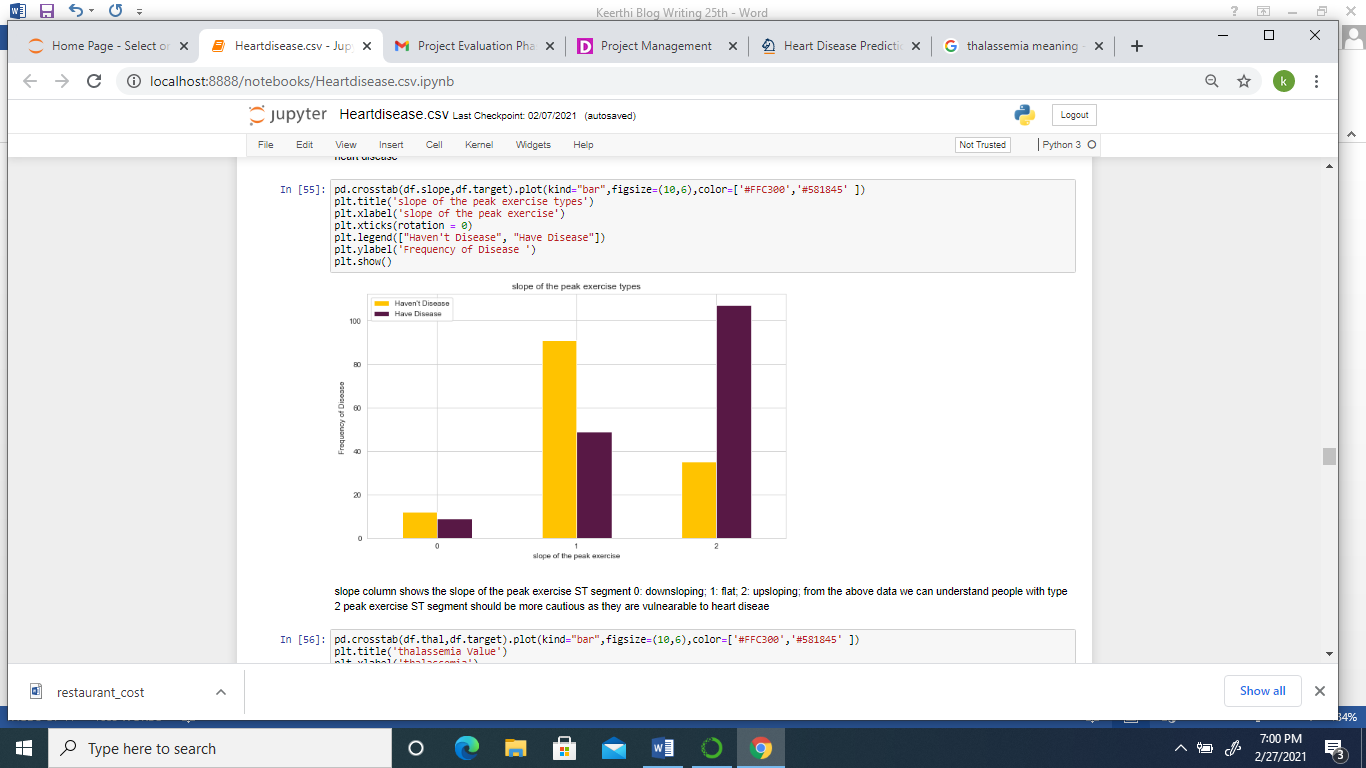


In this chart trestbps column explains the resting blood pressure, here we will analyze at what rate of blood pressure rate there is high chance of heart disease.

120 and 130 resting blood pressure we have people with high chance of heart disease and very important thing we can derive from this chart is people with low blood pressure have high chances of heart disease than people with high blood pressure, if we observe between 94 to 110 blood pressured people have chances of heart disease compared to the next extreme 160 to 200 blood pressure people.



In the chol column which tells about cholesterol levels.in this scatter plot red indicates heart disease and green indicates no heart disease further we can observe cholesterol levels between 150 to 350 has no heart disease, above 350 cholesterol level, there is high chances of people being effected with heart disease and we can find mixed population between 100 to 150 cholesterol levels, this tells we should be cautious if cholesterol levels are high.



Slope column shows the slope of the peak exercise ST segment which is being categorized as 0, 1 & 2 in following manner:

0: down sloping

1: flat

2: up sloping

Here we can understand up sloping reports are to be cautious about because 110 people with up sloping category is having heart disease while 35 – 38 range people have got negative result in up sloping, in case flat category are around 50 people with positive heart disease and 90 people with negative results for heart disease while in case of down sloping neutral 10 to 15% have negative and below 10% have positive results for heart disease.

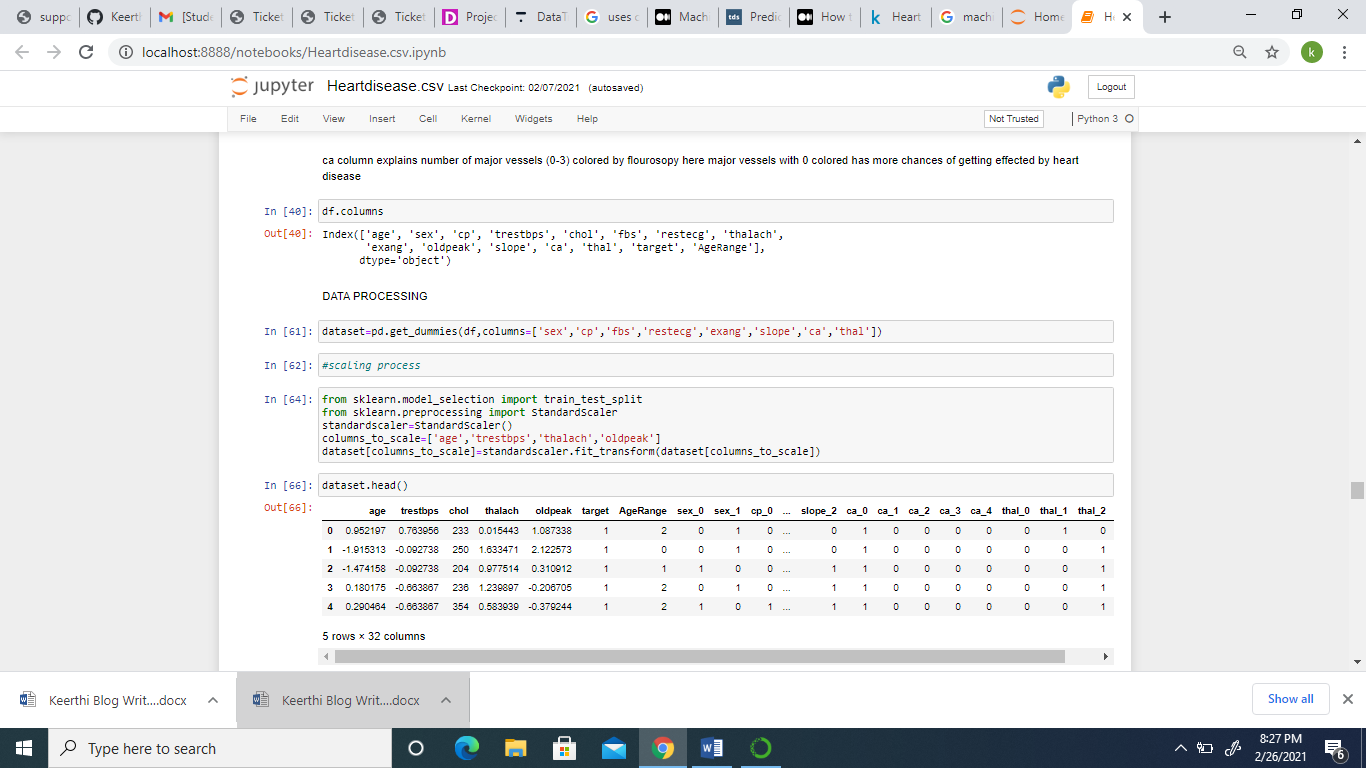
This way we have analyzed chances of heart disease (target class) on the all the remaining attributes with count plot, crosstab, bar plot, scatterplot and crosstab.

DATA PROCESSING

Data processing being an integral part of machine learning since the quality of data out of which we will be deriving a meaningful result is structured by data processing. In this process we will have to handle null values, standardization of data, handle categorical variables convert columns to One hot encoder, label encoding where ever necessary.

While calculating outliers we found large data being lost due to removal of outliers in z score method and quantile method hence we have not removed outliers from our dataset, so next step is assign data to x and y variable in order to train and test the data.

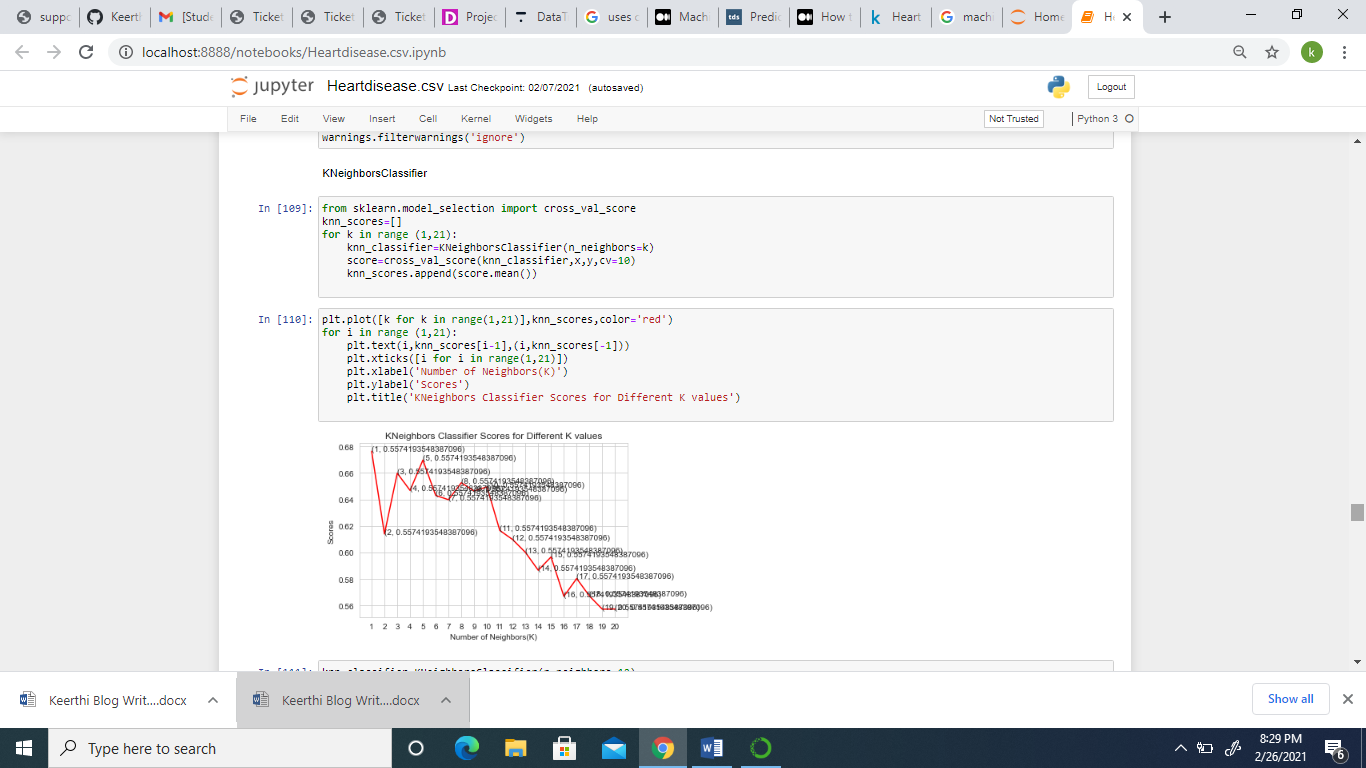
In order to build a model we have first broke the categorical columns for which we are using Onehot encoder for those columns with more than two category with the help of get\_dummies () from pandas, in case columns with two category wit is already in binary type so we are not making any changes.

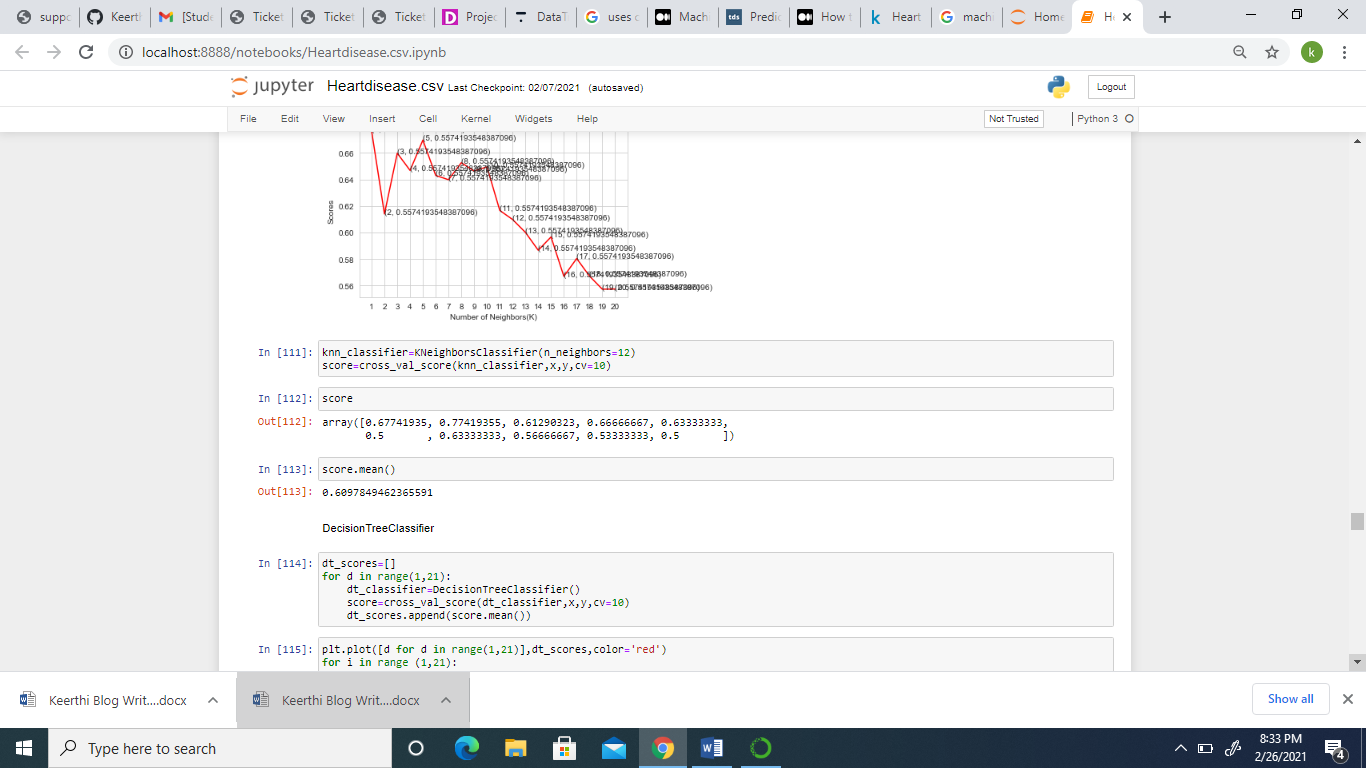


Next we are scaling the data with standard scaler and feed X with the complete data set except target column, while y with target column from data set and import libraries for model selection, classification, metrics and auc\_roc analysis

Training and testing the classifiers:

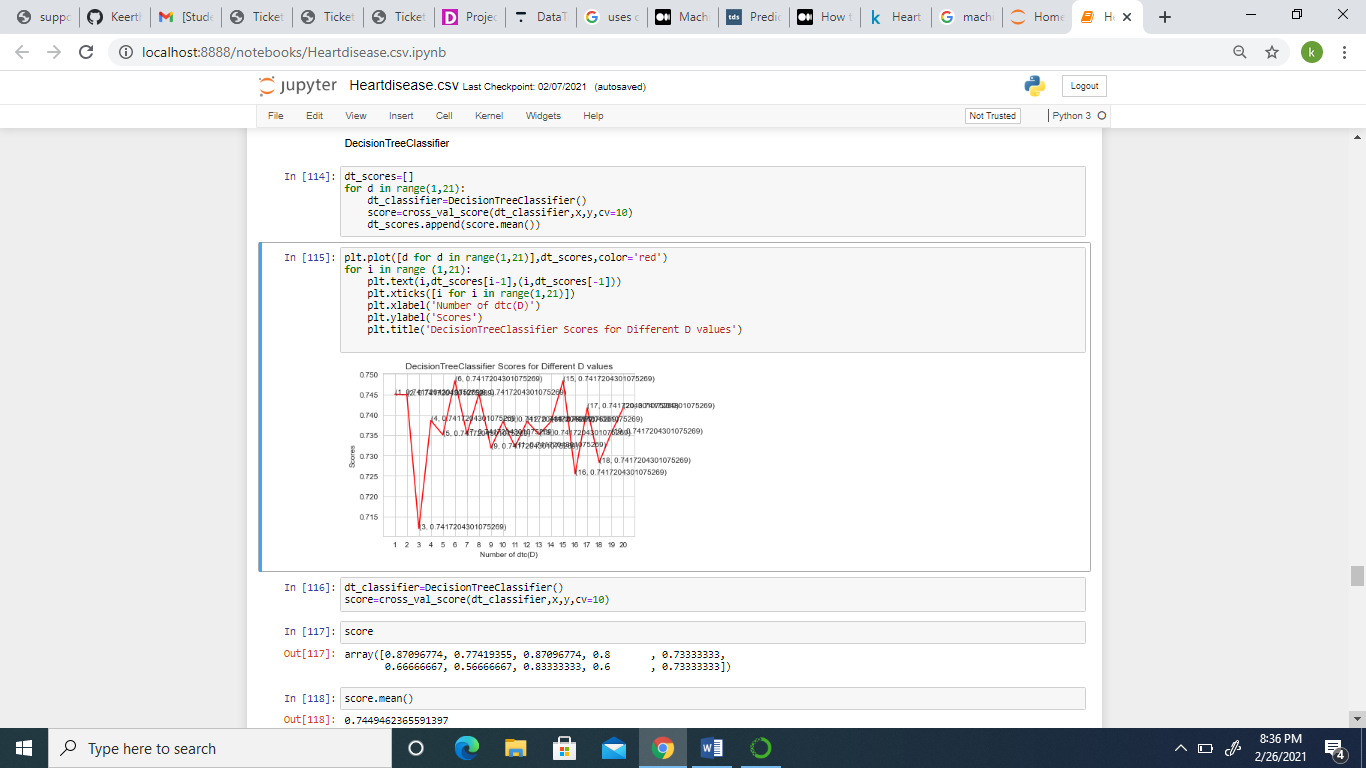
In K neighbor’s classifier we are plotting the data with K nearest neighbors and based on the majority class the data is plotted, here the numbers are varied from 1 to 21 and then the test score is calculated and line graph is plotted for the same





The prediction score with knn classifier is 60%.

In case Decision Tree algorithm, the classifier assigns values to each data point and observe maximum number of feature between 1-21 and plot the result in line graph



In the above plot we can observe the maximum score is 74%.

This way we have plotted for Logistic Regression, GaussianNB, SVC, ADAboost with test score range (1-21) and found logistic regression gives highest score

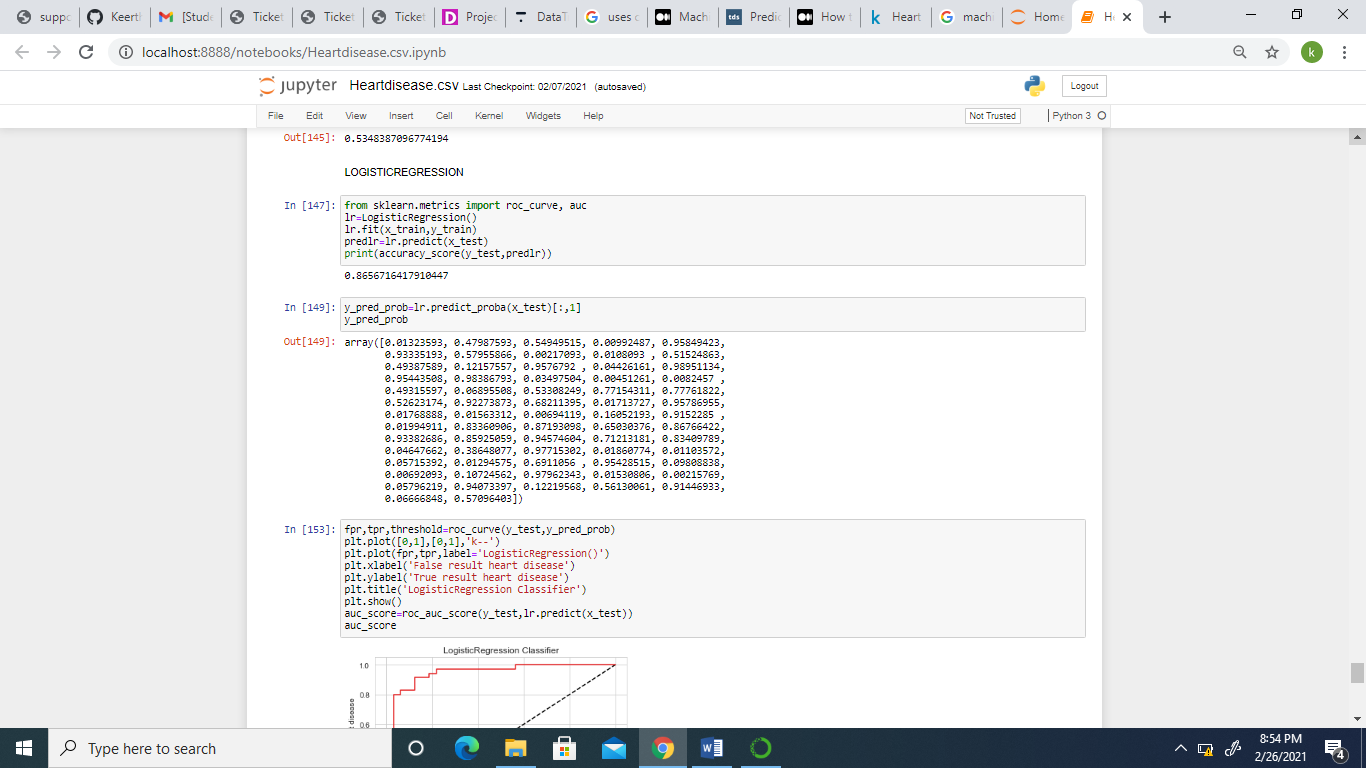
The list of classifiers and mean scores are as follows:

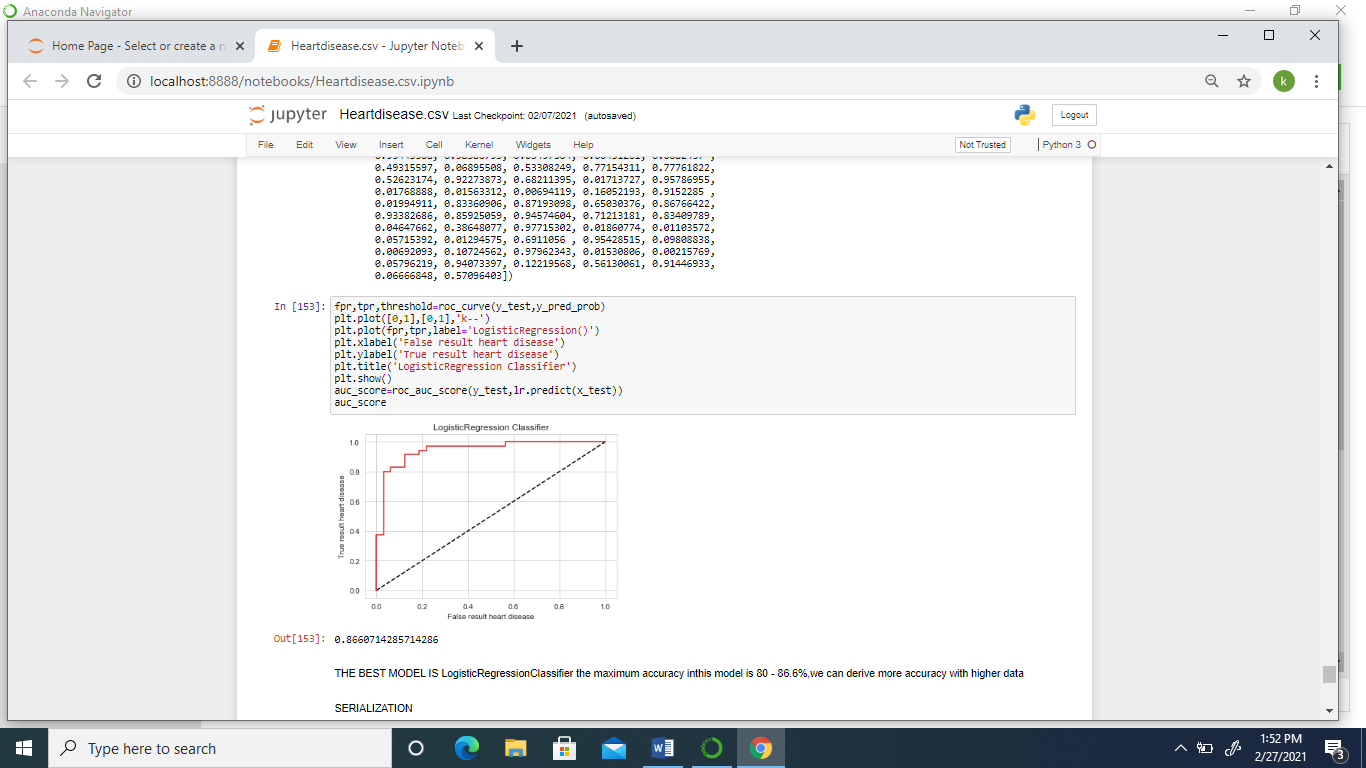
|  |  |
| --- | --- |
| K Nearest Neighbors | 60% |
| Decision Tree | 84% |
| GaussionNB | 80% |
| Support Vector Classifier | 53% |
| ADAboost | 53% |
| Logistic Regression | 86% |

From above table we can conclude logistic regression gives highest score of prediction.

Once we built the classifiers and select the best model among them next step is to evaluate it and validate it so that we can come to an conclusion that this is the best model for prediction, the abbreviations for AUC is area under the curve and ROC is receiver characteristic operator. The higher the AUC, better is the performance of the model.

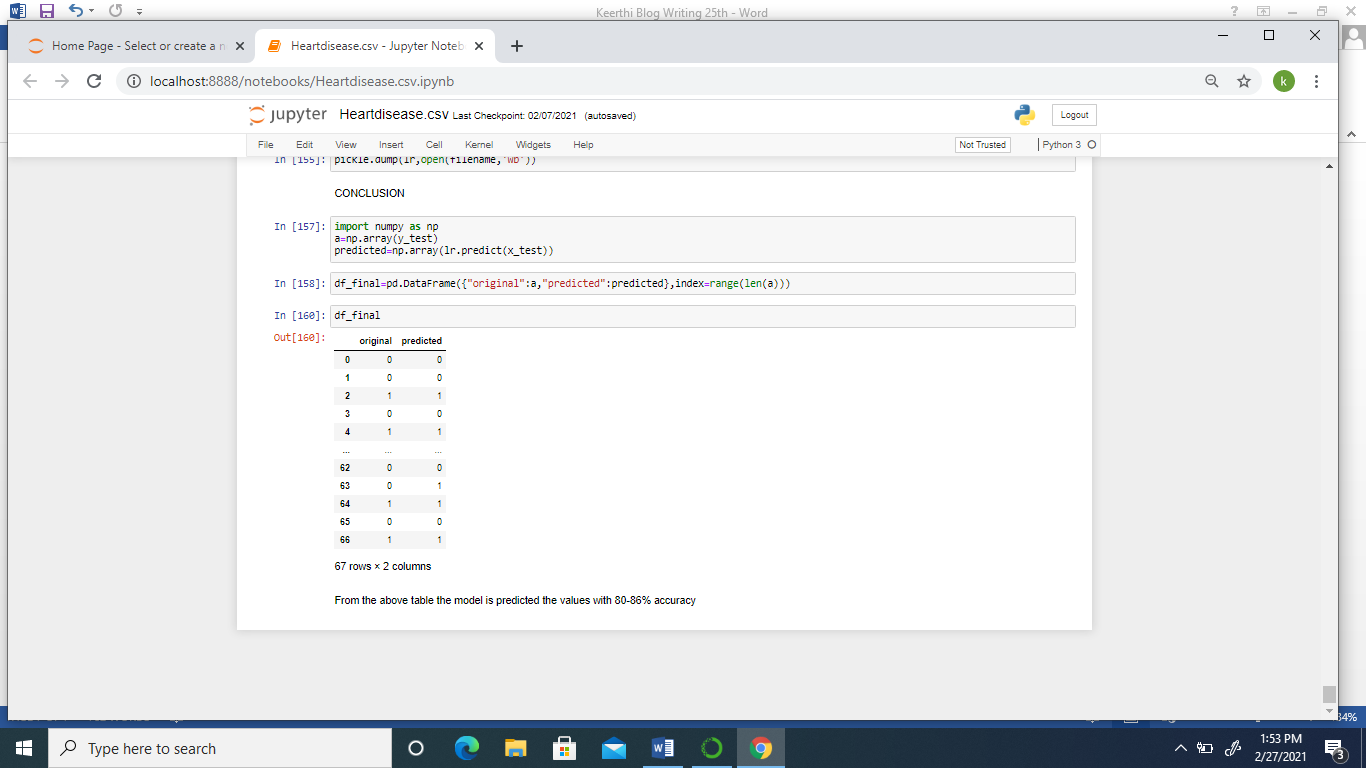
Now we are fitting logistic regression into auc\_roc\_curve.





It is evident from the plot that the AUC for logistic regression is higher, if we had fitted in other classifiers into the same chart ROC curve for other classifiers would be lower that logistic regression, and higher the AUC the better the model is for distinguishing between positive and negative classes, therefore we are choosing logistic regression which has done a better job in classifying the positive class in our Heart Disease dataset

Lastly through serialization process we are importing pickle file and dumping logistic regression fitted model in pickle file and finally conclusion chart of predicted results, y\_test which has been assigned for testing our target column of the model is being named “a”, x\_test has been assigned with all the attributes other than target column for testing our model and “predicted” is named to call the x\_test predicted with the logistic regression and same has been called together i.e., a and predicted under df\_final and derived chart for trained and tested sample from our dataset.



The accuracy of the above predicted chart is 86% we can get above 90% accuracy with more population added to the data set.

References :

analyticsvidhya.com

springer.com

scikit-learn.org