Sloan digital Sky survey (SDSS) offers to public data about space observations made by them and the same has been used in our project to predict astronomical objects.

artificial intelligence is the intelligence demonstrated by machines after programming it for learning and problem solving this has a wide range of applicability and space industry is no exception for that. since we have used data from SDSS it is been considered as primary data in raw format which contains 10,000 observations in 17 columns and one target column which explains about the class of prediction.

the details on this dataset is available on the Kaggle, UCI & SDSS.

PROBLEM DEFINITION

we're supposed to predict astronomical object which could be one among

Star, Galaxy, Quasars

The target column has been depicted as: -

0 representing Star

1 representing Galaxy

2 representing Quasars

Since it belongs to categorical column, we will be using classification models to predict the data

Classifiers adopted

Logistic Regression: It is an appropriate analysis to conduct when the dependent variable is binary type, keeping target column star, galaxy and quasars 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.

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.

Random Forest Classifier: It is an algorithm consisting of many decision trees which uses bagging and feature randomness while building each individual tree to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

DATA ANALYSIS

As a first step to start the project we have imported basic libraries like pandas, numpy, matplotlib, and seaborn and imported data and named as df and got into the task of understanding columns

In this data it is very important to explain each column as it contains terminologies which is not self-explanatory the details of the columns have been explained as below

objid is object identifier

ra, dec is right ascension and declination respectively

plate is plate number

u, g, r, i, z are filter bands (a.k.a. photometric system astronomical magnitudes)

run, rerun, camcol, field are descriptors of fields (i.e. 2048 x 1489 pixels) within image

redshift is increase in wavelength due to motion of astronomical object

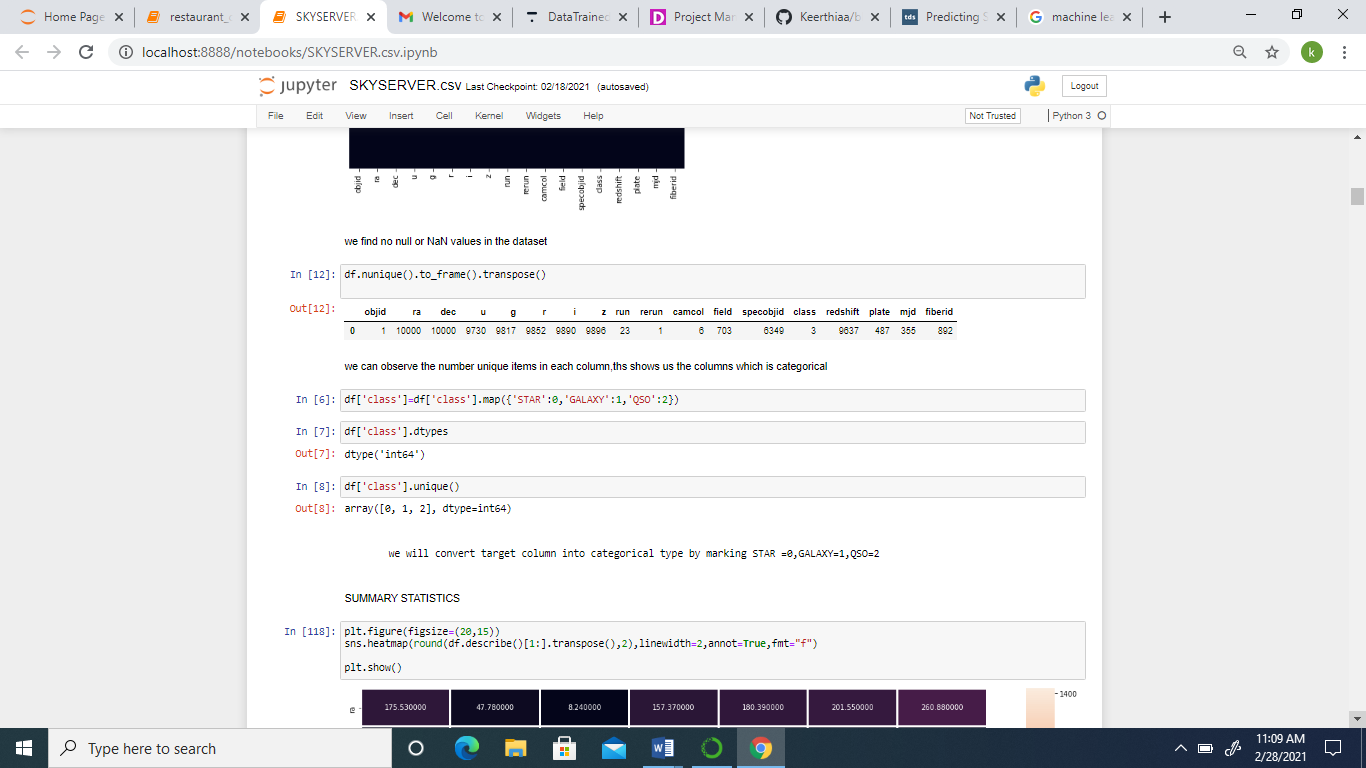
mjd is modified Julian date observation

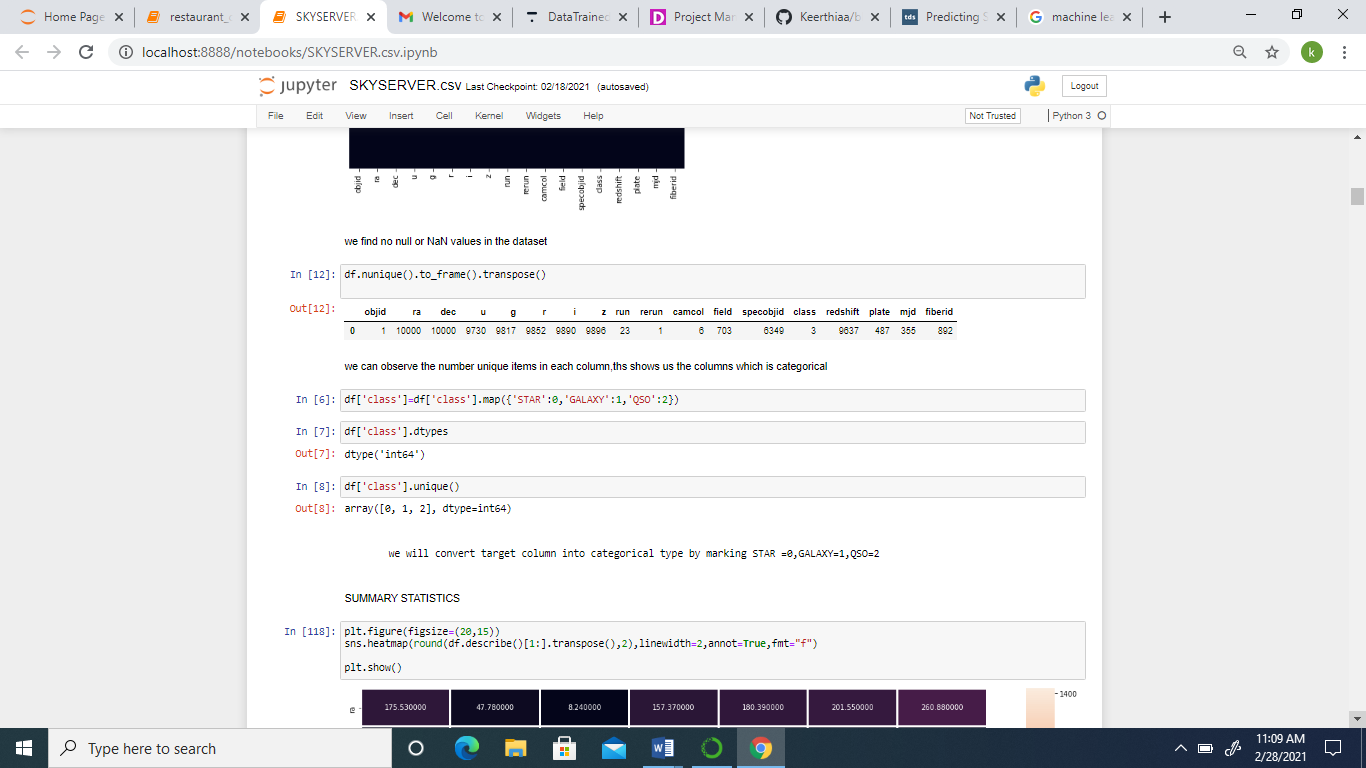
fiberid means optic fiber ID

specobjid = Object Identifier

class = object class (galaxy, star or quasars)

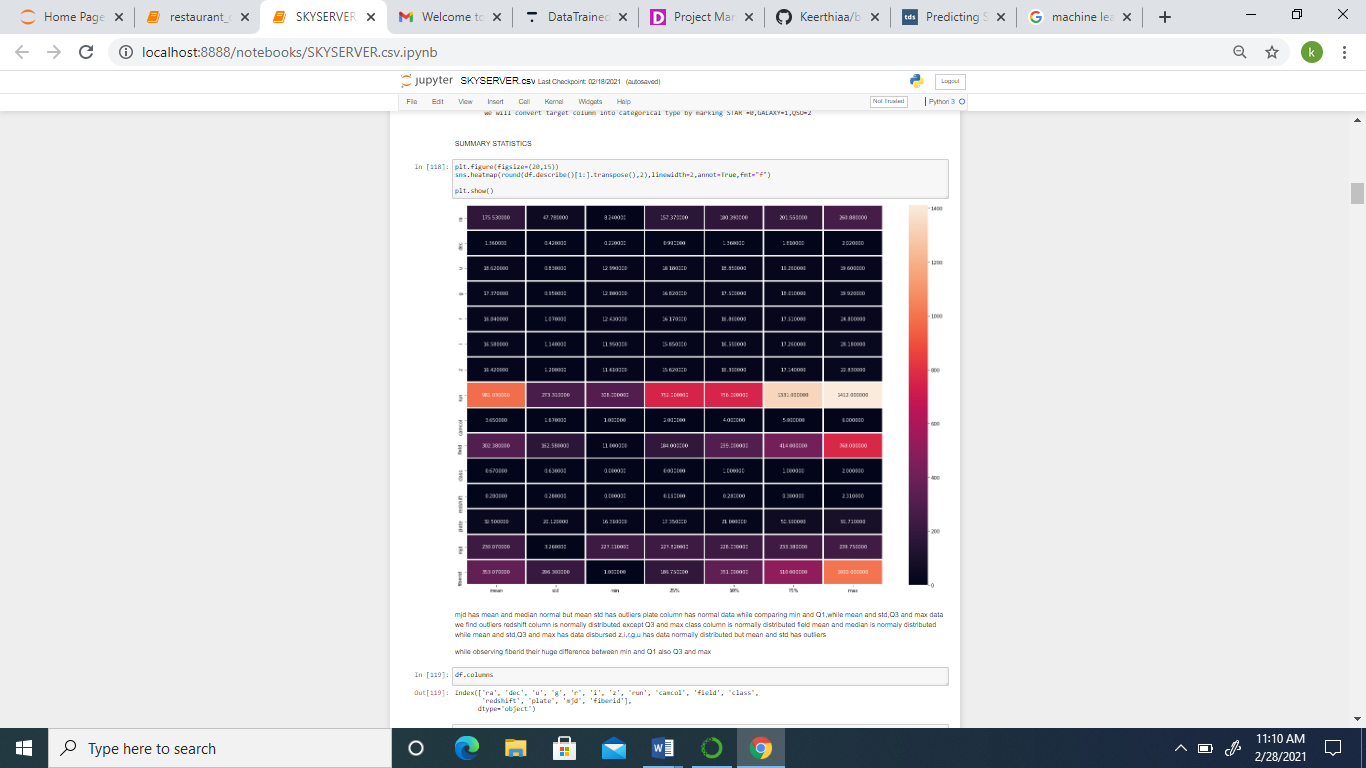
After column labels elaboration we can start with data exploration by looking into data shape and data information and data type we found that there are 18 columns with 10,000 observations and all columns are integer or float type except class column(object type) which is our Target column prediction. With df. isnull().any() we found there is no any null values and with below code we found unique characters of data in each column.

Here objid and rerun columns have only 1 unique data which is repeated for all 10,000 observations so has no impact whether the class belongs to start , galaxy or quasars hence we can ignore this 2 columns for analysis and prediction while class column have three unique items so we can convert class column to integer by labeling them as :- Star = 0, Galaxy = 1, QSO = 2,shown below.



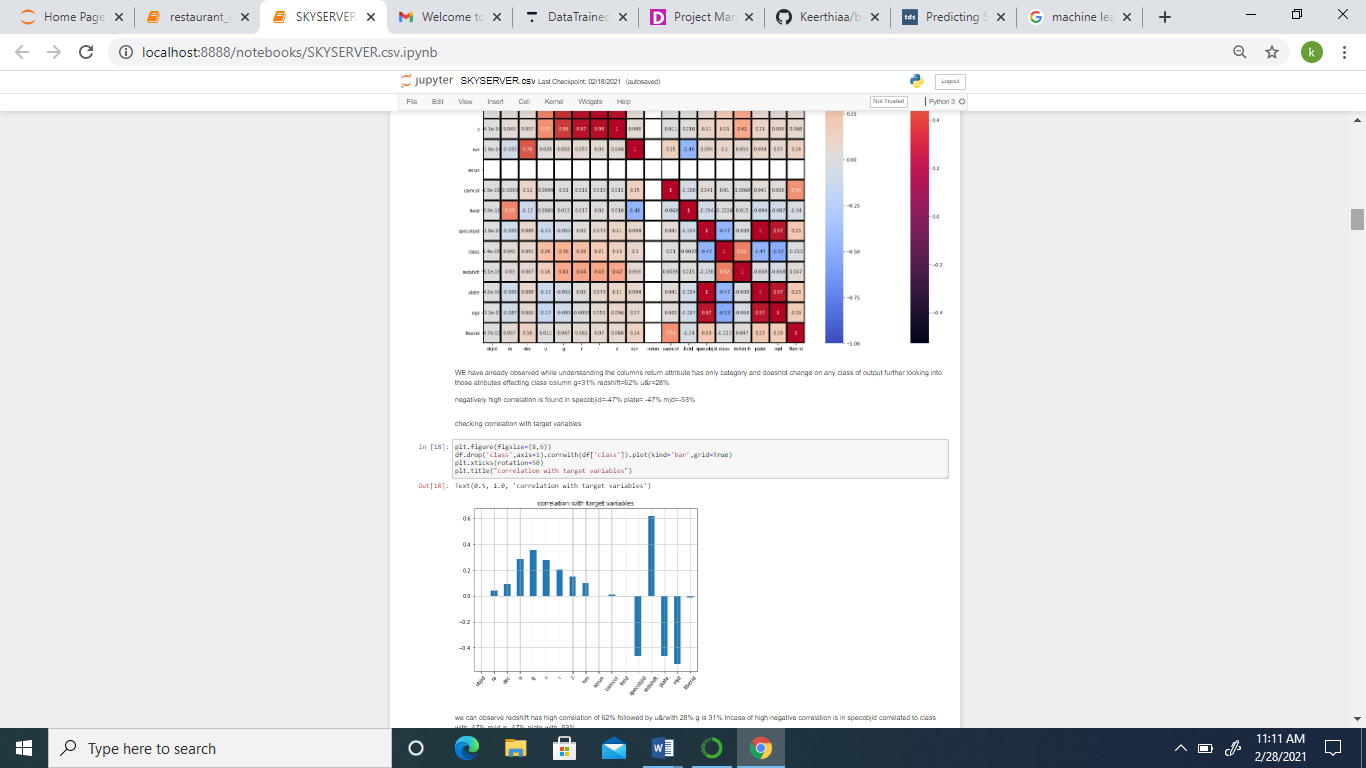
EXPLORATORY DATA ANALYSIS

Firstly let us start with Summary statistics, I have used below code to describe how data has been distributed and with the help of below figure let us analyze if the data is normally distributed and is their any outliers present and if so where is the outliers detected.



* mjd has mean and median normal but mean and standard deviation has outliers.
* We can see plate column is normally distributed as minimum value and Q1 has not much difference, but while looking into the difference between mean and std, Q3 and max data we find outliers.
* redshift column is normally distributed except Q3 and max where data is disbursed.
* class column is normally distributed
* In field column mean and median is normally distributed while mean and std, Q3 and max has data disbursed,
* While looking into z, i, r, g, u has data normally distributed but mean and std has outliers.
* In case of fiberid column there is huge difference between min and Q1 also Q3 and max.

Next step in summary statistics is to find correlation of class column with other attributes which is depicted in the chart below:



We can observe objid and rerun column has no correlation since it has only variable or data repeated for all 10,000 observations.

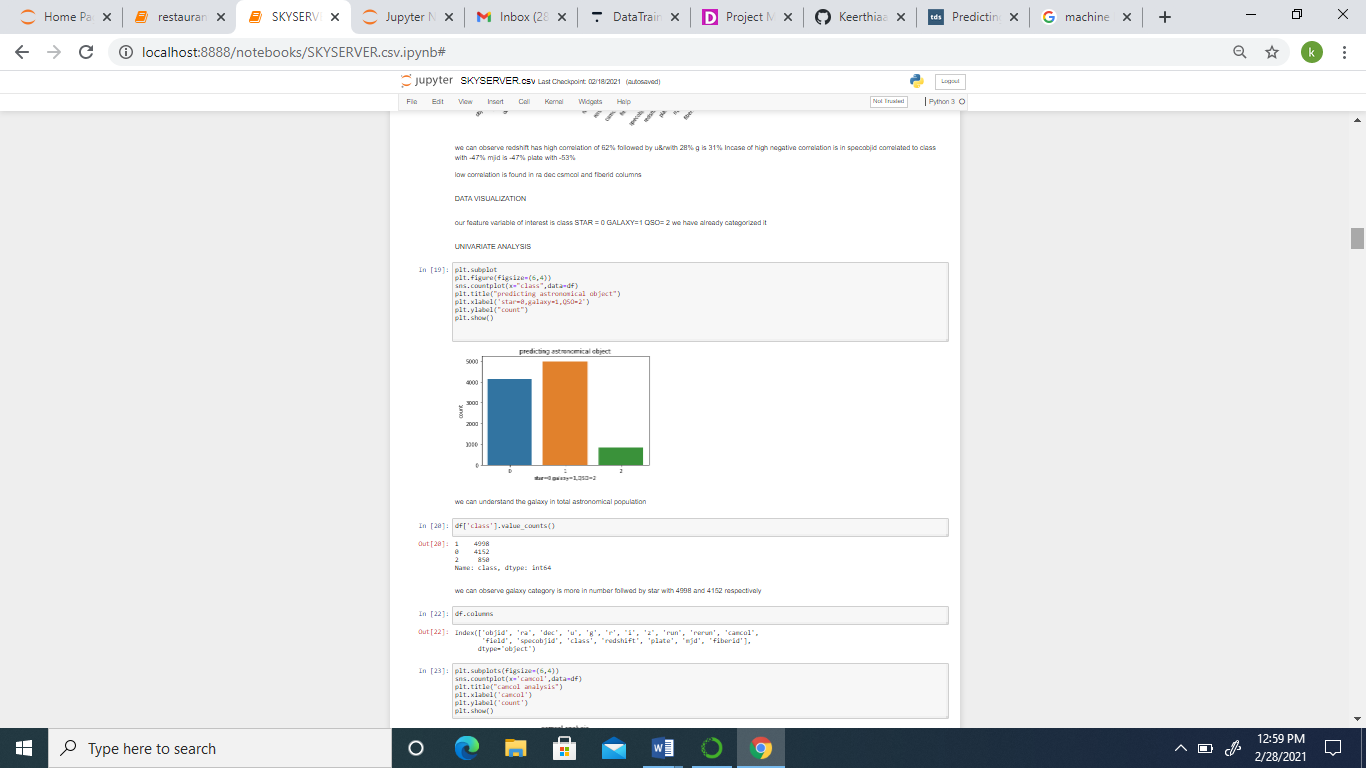
we can observe redshift has high correlation off about 62% followed by u and r with 28% and g with 31%.

In case of negative correlation, we can find specobjid that 47% mjid 47% and plate with 53% correlated to class column.

Low correlation is found in added ra, dec, camcol and fiberid.

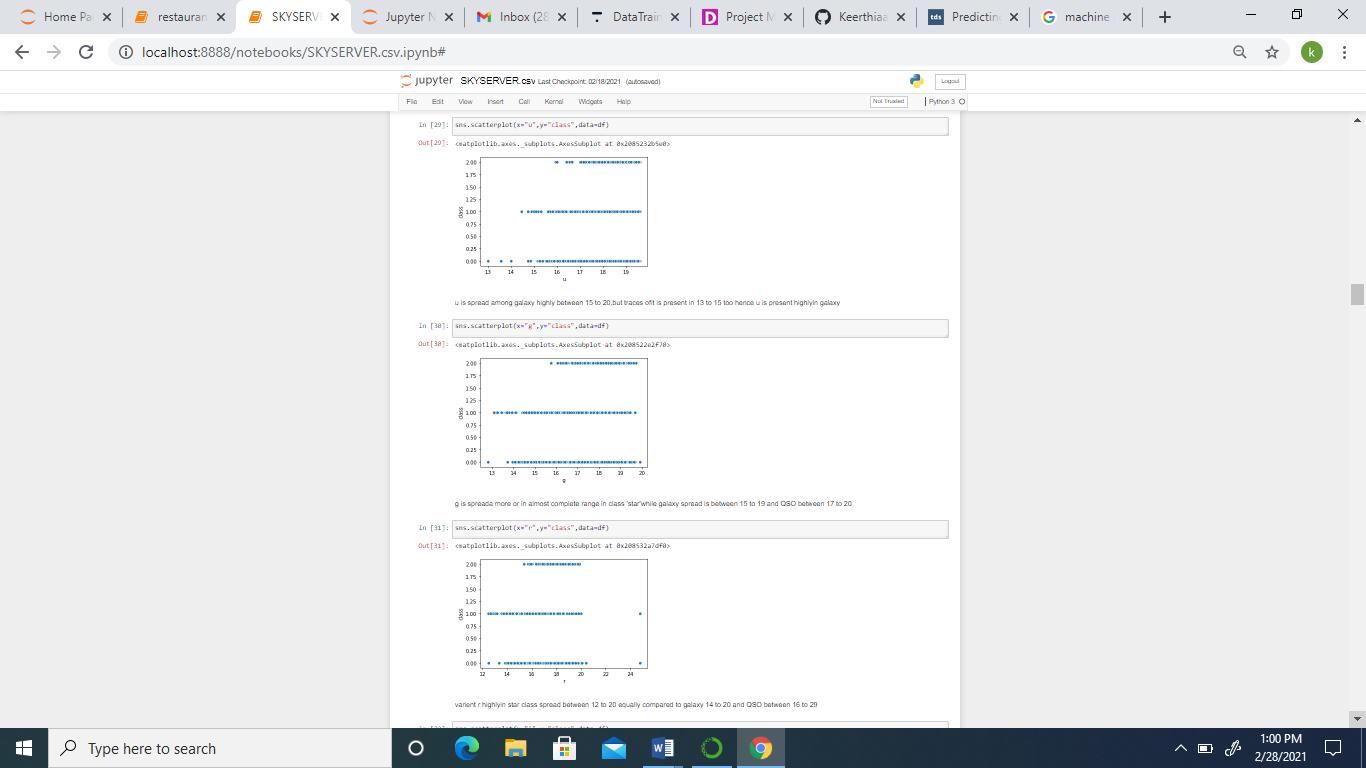
Now let us get into data visualization process which is being divided as univariate analysis, bivariate analysis and multivariate analysis in order to understand charts better we should remember our feature variable of interest is class column which is being coded as star = 0, Galaxy = 1 and QSO = 2 with this information let us look into the charts.

Firstly we have extracted information about our class column and observed how data has been distributed in each category and found galaxy has highest data of 4998 followed by star with 4152 and lastly QSO with 850 number of data the code extract this information is given below



Same way we will analyze for those columns which has impact on class column, now whether it has impact on class column or not to understand one has to look into correlation matrix, keeping that data as base we will do univariate analysis, bivariate analysis and multi variate analysis.

In univariate analysis we have used count plot, bar plot and sub plot. In case of bi variate analysis, we will be using scatter plot for each column and its relationship with the class column will be observed the diagram for the same is given below



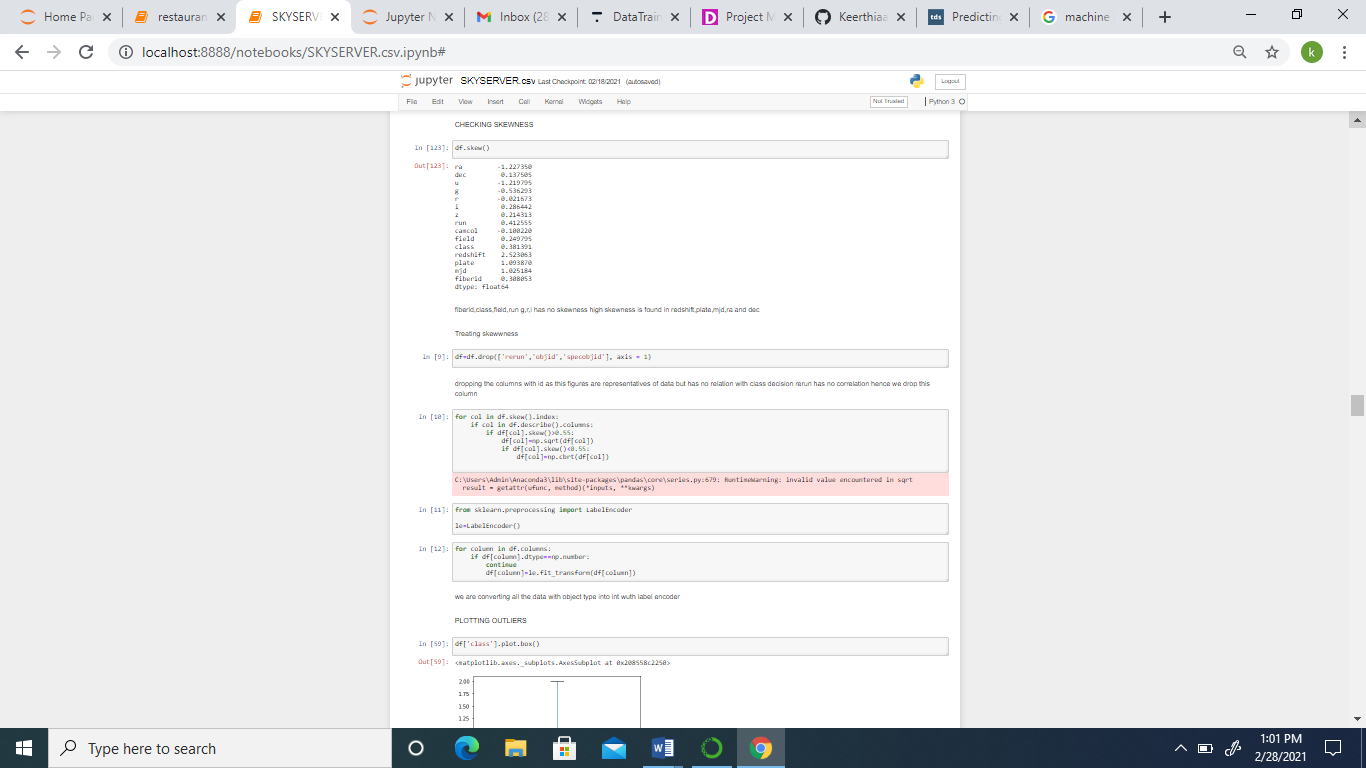
In the scatterplot above we can observe column G, G is one type of photometric system of astronomical magnitudes this is being distributed to the class as follows

star at the range 14 two 20

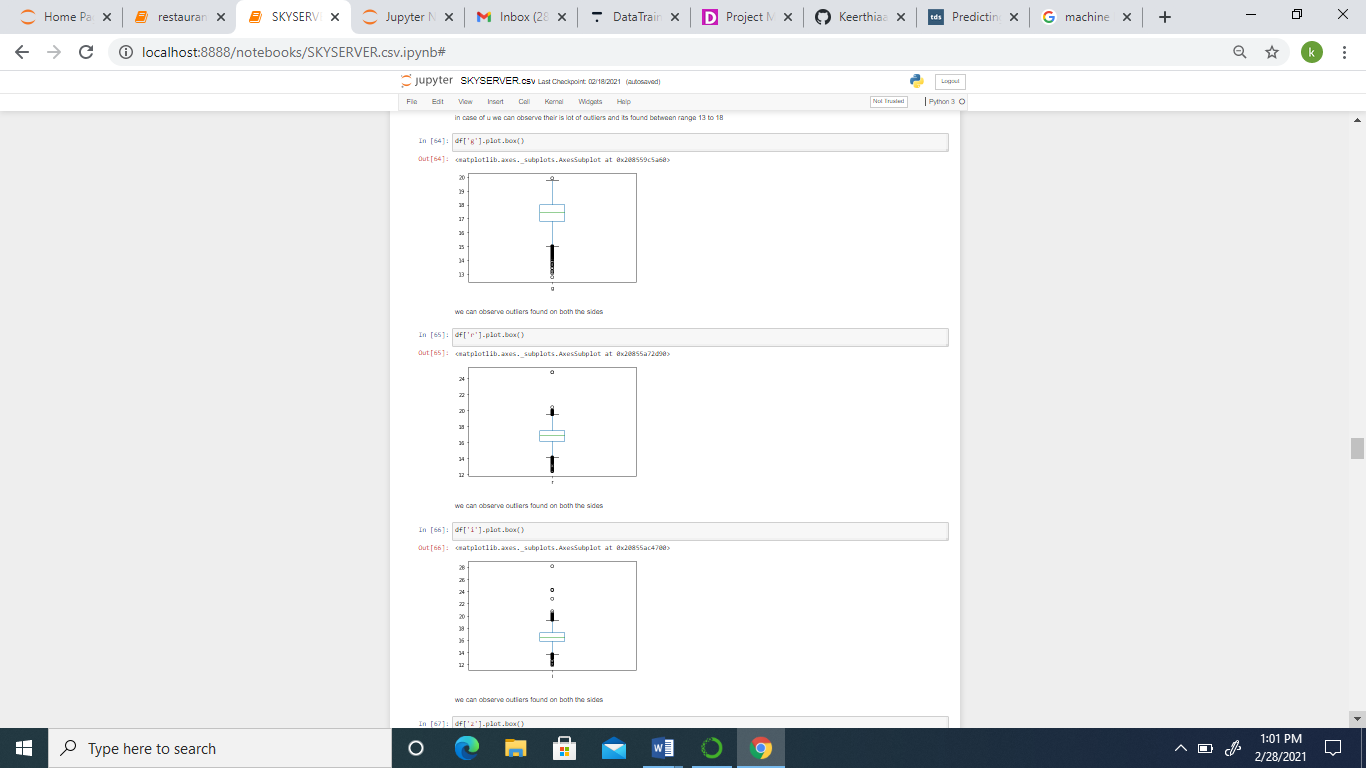
Galaxy spread between 15 to 19

QSO is spread between 17 to 20 range.

After visualizing the data we have to find if there is any skewness in each column and if found outliers are to be treated we will now observe whether skewness is present or not this can be found by seeing if any column skewness is ranging above or below .5 i.e. +/- 0.5 as skewed data therefore from the below code we found redshift, plate and mjd columns have high positive skewness while ra & u has high negative skewness.



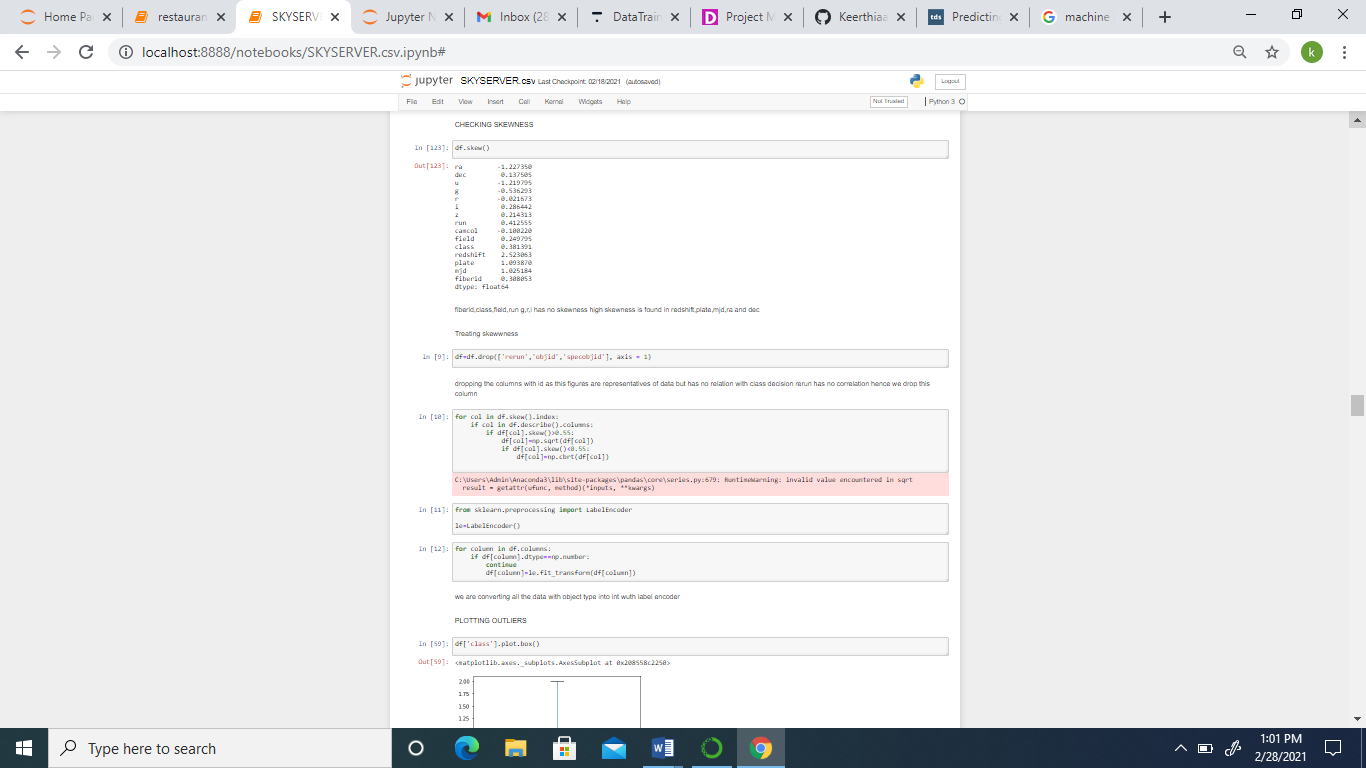
Now we are plotting the outliers for each column to visualize which column has outliers



As shown in the above chart we will have to pass for all columns to find if there is outliers , from the above figure column ‘g’ has got many outliers between minimum value and Q1 and in case of column ‘r’ outliers are present in between maximum value and Q3 also minimum value and Q1.

Before I remove skewness from our dataset let us first clean the data for this, I have dropped the ‘rerun’, ‘objid’, ‘specobjid’ columns since it has no impact on class column.

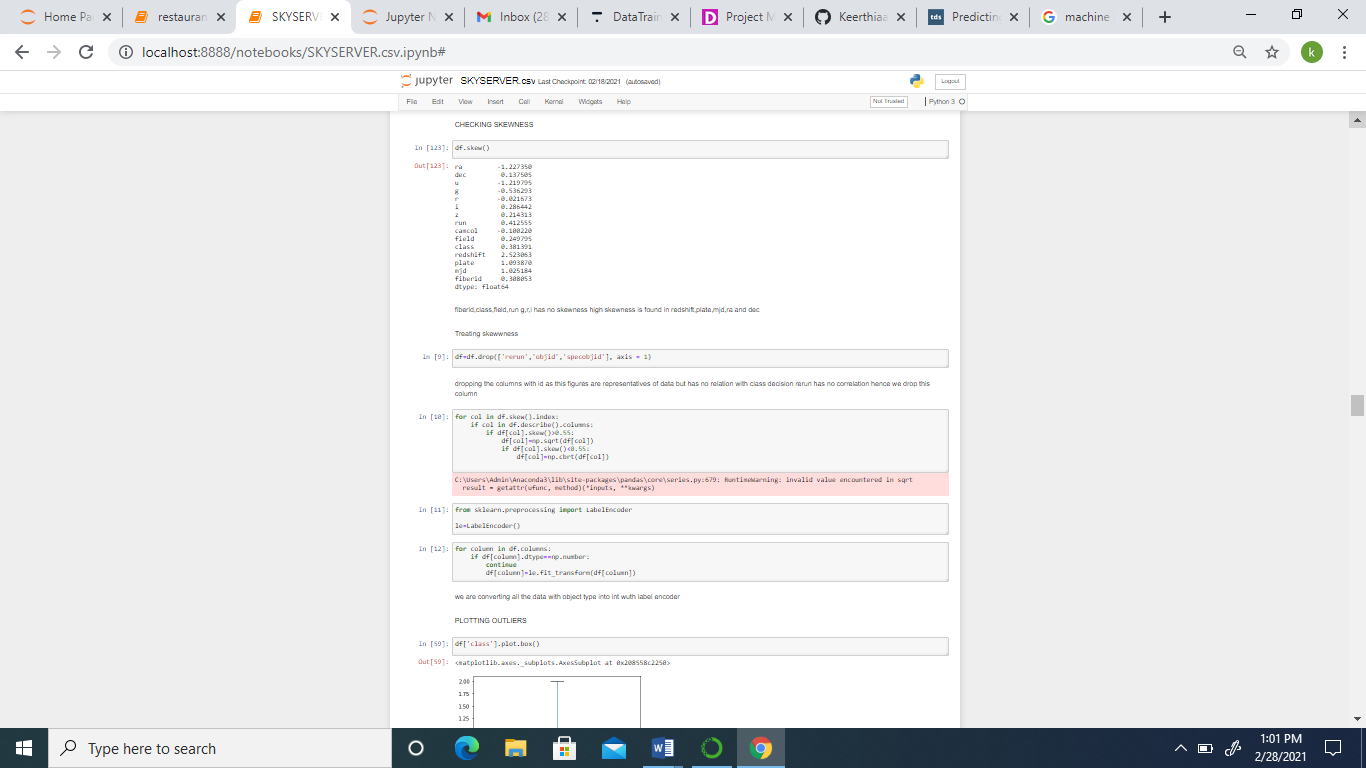
Since there is skewness present in most of the data lets remove the skewness, I have used square root, cube root method to remove skewness for this I have choose those data which has results above 0.55 and passed square root for those data and the columns which has results less than 0.55 is treated with cube root to remove outliers.



DATA PROCESSING

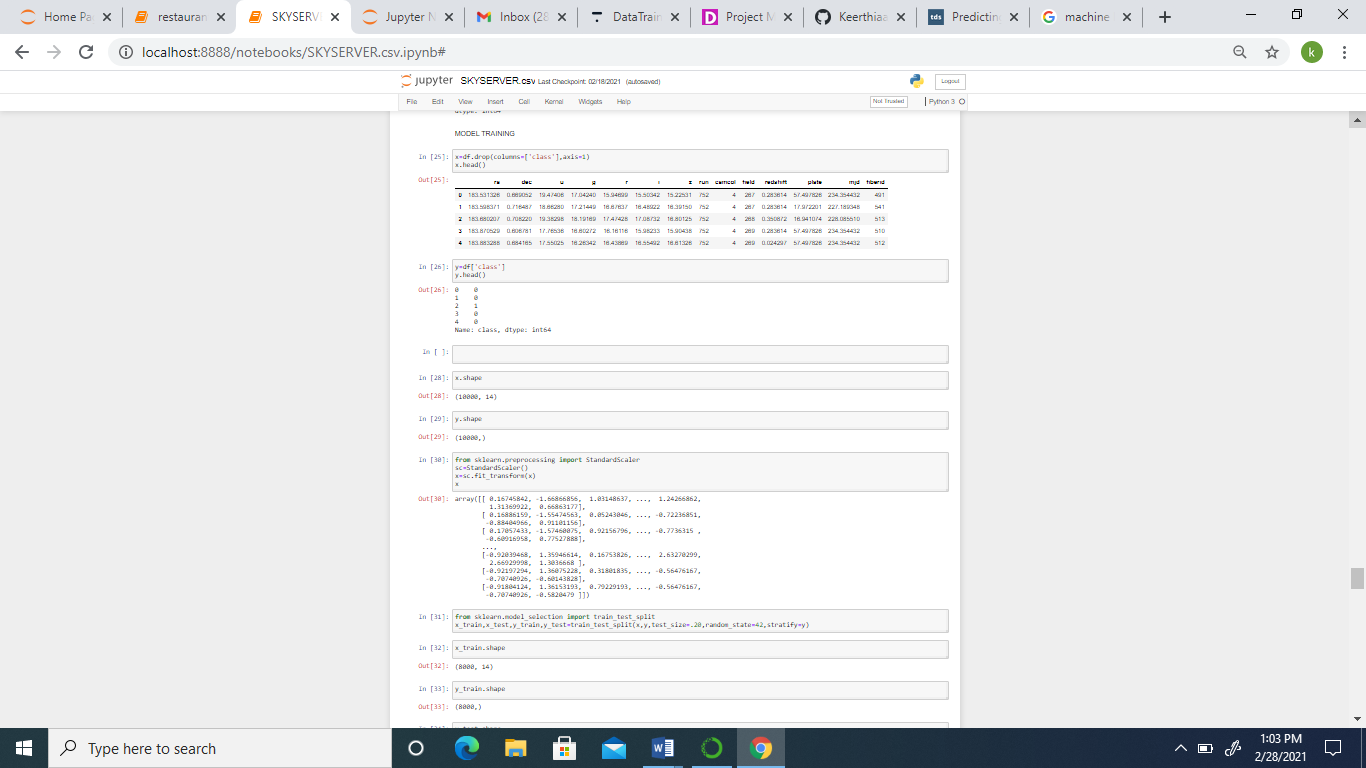
Data processing being an integral part of machine learning since the quality of data out of which we will be deriving a meaningful result is structured by data processing. In this process we will have to handle null values, standardization of data, handle categorical variables convert columns to One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

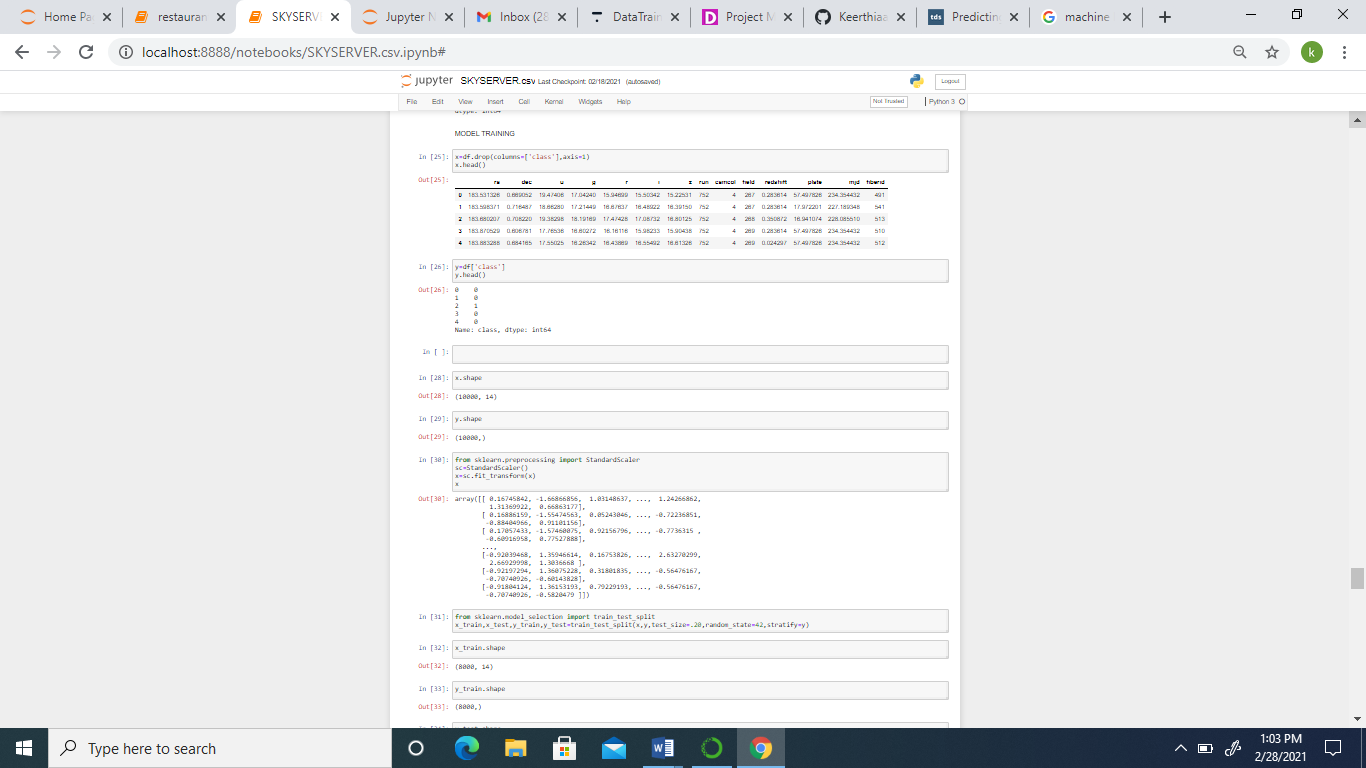
As a part of data processing we are now doing data cleansing procedure where we will convert all the data to integer type so that we can train and test the data. For this purpose, we have passed label encoder code for all the columns whose dtype is float and after fitting it into label encoder the columns will be converted to integer type.



We can also test how much data will be lost if we remove skewness through z-score method and quantile method in this dataset we found 9.5% data will be lost if we remove skewness through z- score method and in case quantile method 34% data will be lost and for 10,000 observations losing 950 observations is actual not a good decision at the best of my knowledge. Apart from it we also checked if any null data exists and removed the same

Next, we are scaling the dataset with standard scaler and feed X with the complete data set except class column, while y with class column from space server dataset as shown below

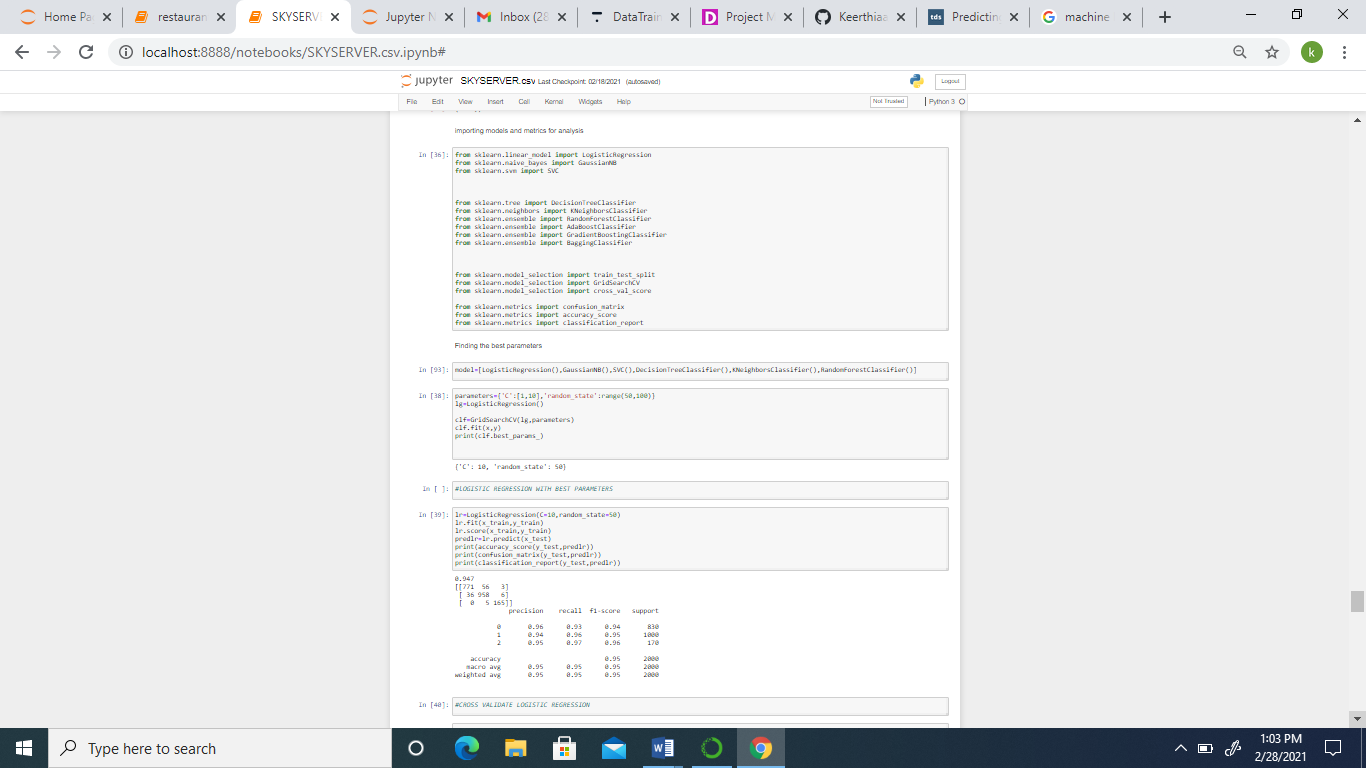




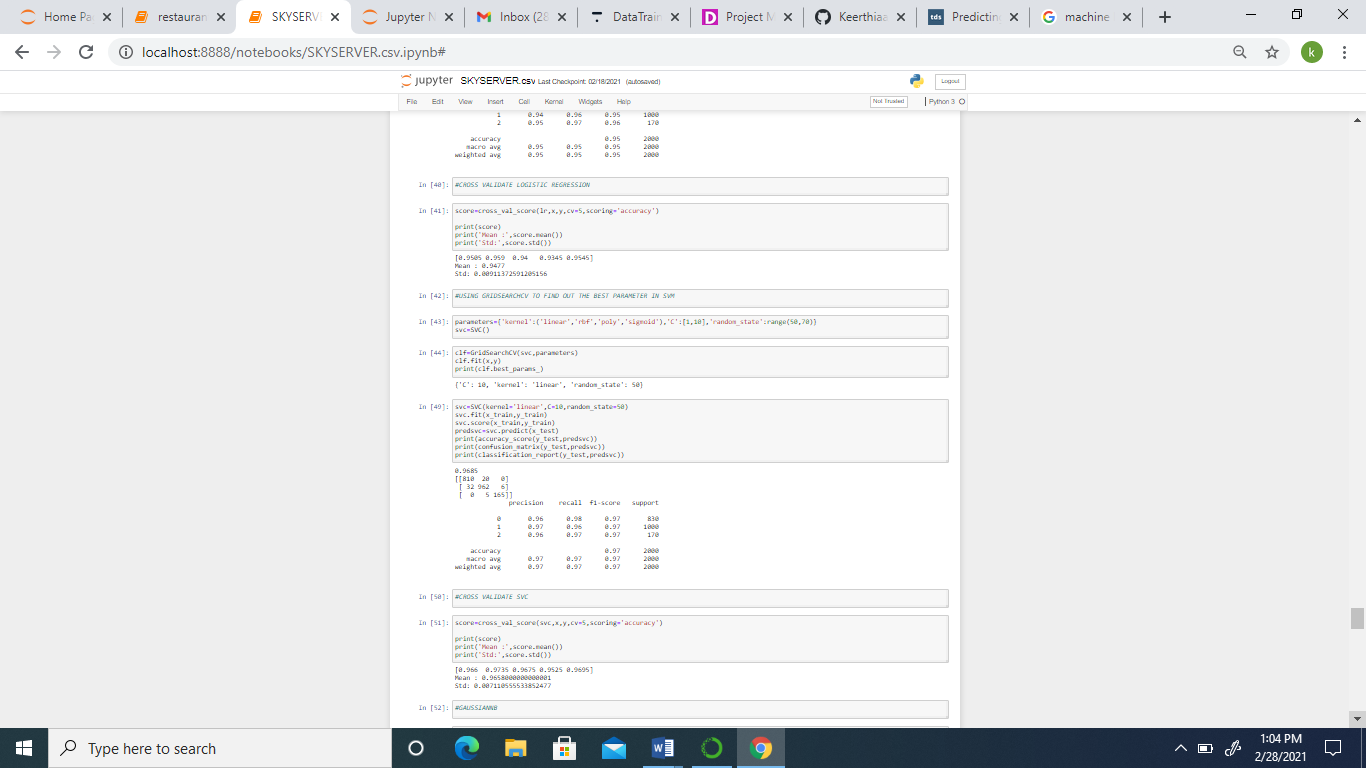
After data processing and model building part we will train and test the data I will be testing 20% of the data as we have to find predict among star, galaxy and QSO it is a categorical type hence we will predicting using classifiers for this purpose we will be importing classifiers, metrics and Grid Search CV for hyper tunning the models in our project I will be using Logistic regression, Gaussian NB, decision tree classifier, SVC, K neighbors & random forest classifier.

Logistic Regression: It is an appropriate analysis to conduct 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.

Through Grid Search CV, using best parameters we could predict through logistic regression at 94% accuracy and cross validate the score simultaneously, here is a glimpse of the trained model below



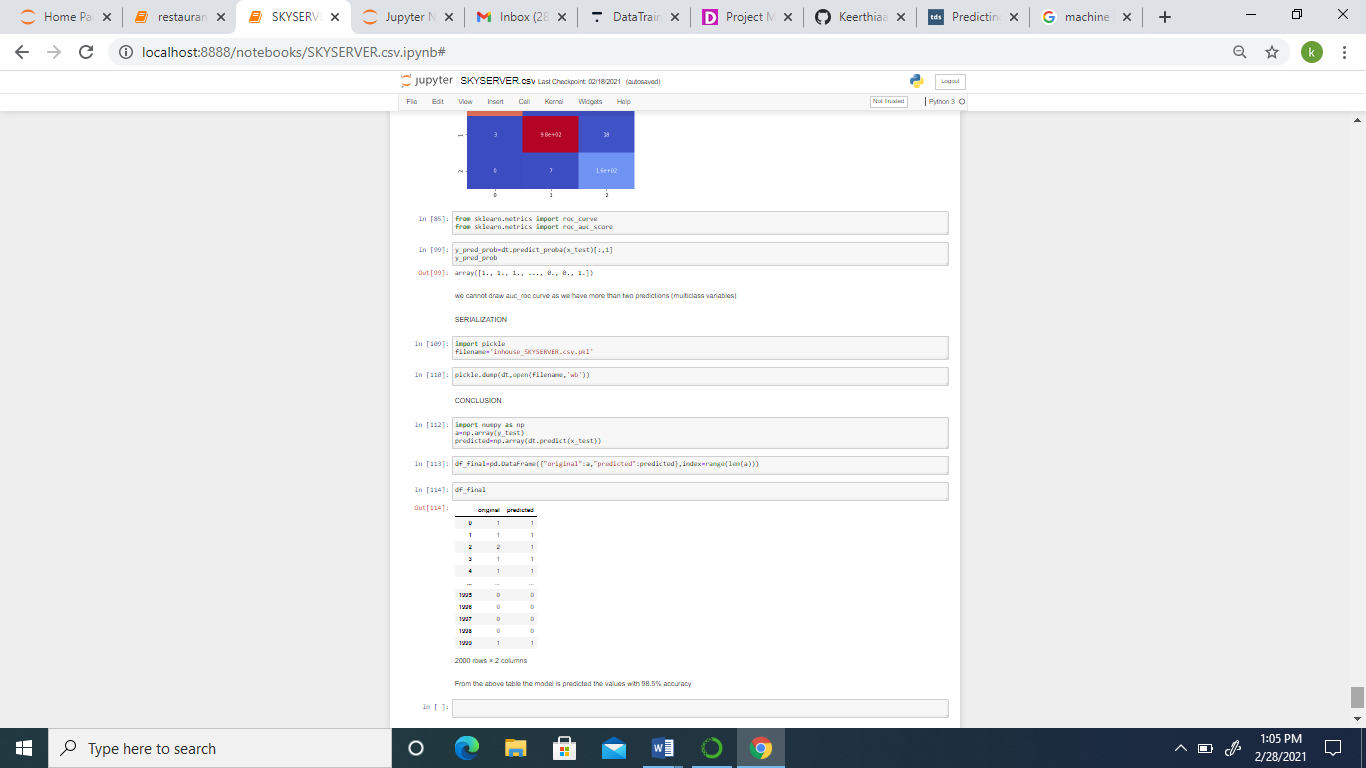
Same way we will do Support vector classifier, GaussianNB classifier, Decision Tree Classifier, random forest classifier and K-Nearest Neighbors



|  |  |
| --- | --- |
| K Nearest Neighbors | 92% |
| Decision Tree | 98.55% |
| Gaussion NB | 88% |
| Support Vector Classifier | 96% |
| Random Forest Classifier | 98.5% |
| Logistic Regression | 94% |

These scores are after making Grid Search CV for each Classifier we will be choosing Decision Tree Classifier as our best classifier for this dataset and plot confusion matrix for the same, AUC ROC curve is not plotted for this dataset since we have more than two predictable items.

Lastly through serialization process we are importing pickle file and dumping Decision Tree Classifier as a best fitted model in pickle file and finally concluding with the 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 98.55%.