**Red Wine Quality Prediction Project Blog**

**Author Name: Vaibhav Tayade**



**PROBLEM DEFINITION & INTRODUCTION:**

You must have heard this famous saying “Beer is made by men, wine by God”. Indeed this saying sounds so true, there would be barely anyone on this planet who doesn’t like wine. Wine also has special place in western society. They say wine is Jesus’s Blood in Christianity. Obviously, there are various types of wines present in the world, but here we’ll be talking about Red Wine. The actual color of the red wine can range from dark violet through brick red to brown red. Red wine is typical wine made from dark-colored grapes. The red color to the wine comes from anthocyan pigments present in the skin of the grapes. Red wine is a delicacy around the world. Here in this project we’ll be majorly talking about the Quality of red wine. In our problem statement, dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

**DETAILS OF DATASET:**

The dataset which we have got contains data about fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, Quality. These datasets are based on physicochemical tests. We are going to discuss about these parameters below.

**Fixed acidity:-** The predominant fixed acids found in wines are tartaric, malic, citric, and succinic. Their respective levels found in wine can vary greatly but in general one would expect to see 1,000 to 4,000 mg/L tartaric acid, 0 to 8,000 mg/L malic acid, 0 to 500 mg/L citric acid, and 500 to 2,000 mg/L succinic acid.

**Volatile acidity:-** Volatile acidity is a measure of the low molecular weight (or steam distillable) fatty acids in wine and is generally perceived as the odour of vinegar. Winemakers are usually most concerned with acetic acid, which accounts for more than 93% of steam distillable acids in wine.

**Citric acid:-** Citric acid is often added to wines to increase acidity, complement a specific flavor or prevent ferric hazes. It can be added to finished wines to increase acidity and give a “fresh” flavor.

**Residual sugar:-** Sweetness in wine is called residual sugar and is usually measured in grams per litre (g/L). Basically, residual sugar or 'RS' is the sugar from the grapes that's left over after fermentation. The more residual sugar remaining in a wine, the sweeter the wine is.

**Chlorides:-** It significantly contributes the wine's sensory characteristics, affecting color, clearness, flavor and aroma. Therefore, moderate to large concentrations of chlorides and sodium might give the wine a salty flavor which may turn way potential consumers.

**Free sulphur dioxide:-** SO2 is used for prevention of wine by oxidation and microbial spoilage.

**Total Sulphur dioxide:** Total Sulfur Dioxide (TSO2) is the portion of SO2 that is free in the wine plus the portion that is bound to other chemicals in the wine such as aldehydes, pigments, or sugars.

**Density:-** The perceived quality of both wines decline as the amount of chlorides , or salt, increase. As density increases, perceived quality declines. In the case of winemaking, a hydrometer is used to measure must or wine density, which is increased by fermentable sugars and other must/wine substances.

**pH:-** Typically, the pH level of a wine ranges from 3 to 4. Red wines with higher acidity are more likely to be a bright ruby color, as the lower pH gives them a red hue

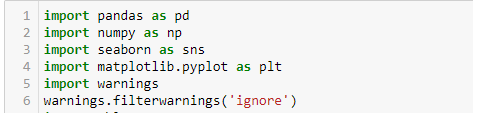
**Sulphates:-** Sulphates are a food preservative widely used in winemaking, thanks to their ability to maintain the flavor and freshness of wine.

**Alcohol:-** There is little percent of alcohol is also present in wine.

**Quality:-** Here in our dataset, wine quality values ranges from 0 to 10. We been asked in the problem statement to categorize the Quality of the Wine below 7 as not-good wine while 7 or above as good wine.

The biggest question which comes in one’s mind is about its quality. How can one know the quality of the wine he or she is consuming. Quality of the wine can be differentiated according to smell, flavor and color. But we are not the expert to find quality of the wine given to us on our own, we can do one thing we can make a machine learning model using various algorithm on the dataset which we have got and find out the best model having good accuracy, less overfitting, and less errors. The very first thing which we need to do while staring with the model building is importing various libraries inorder to analyse the dataset.

**Importing Various Libraries:**



Let’s see brief details about these libraries, pandas is the library used to for data analysis, numpy is another library used for numerical data. seaborn and matplotlib are the library used for data visualization.

**Importing data containing csv file:**



We have imported our csv file using using pandas library and created a dataframe named ‘ds’.

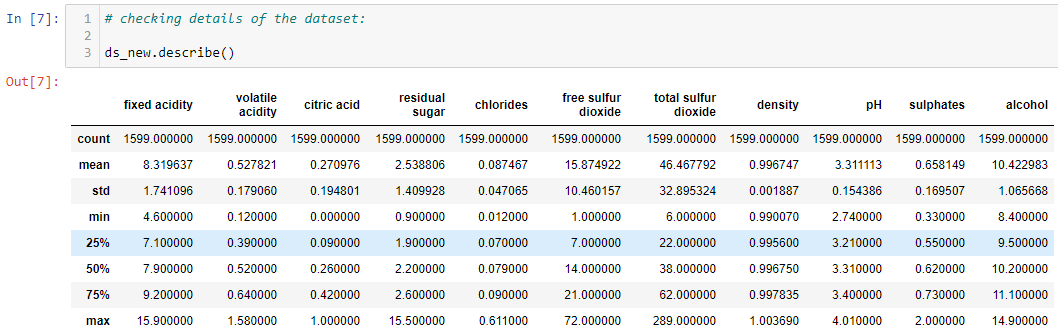
As per our problem statement we need to convert the Quality column into Good Wine and Not good wine according the value of the Quality present there. We have to categorize value 7 or above as ‘good’ wine and value below 7 as ‘not good’ wine. We can do that by replacing 7 and 8 values by ‘good’ and values 3, 4, 5, 6 by ‘not good’.



Now, we have a categorical problem type of dataset containing a target column having two categories ‘good’ and ‘not good’.

**DATA ANALYSIS:**

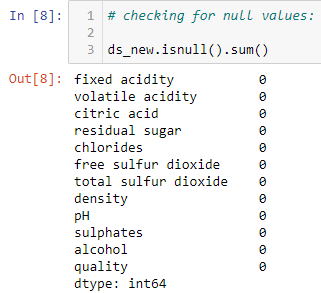
Let’s check the details of the dataset.



By checking the details of the dataset, we can see that the Mean is higher than the Median in 'fixed acidity' and 'total sulfur dioxide' columns, which indicates presence of outliers and skewness in these columns. We can confirm this when we do univariate analysis of the columns.

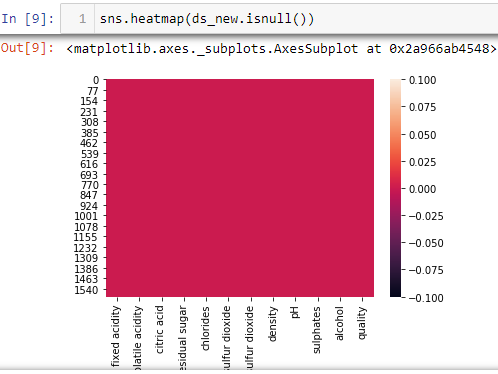
**Exploratory Data Analysis:**

The most important thing which we need to check before proceeding to further analysis is checking the **presence** **of Null data** because machine learning algorithms need numerical data, it doesn’t work on string data-types.



We can see that there are no Null values in our dataset.

We can confirm this by plotting the heatmap of the dataset, if any null values present in the dataset we can spot it easily.

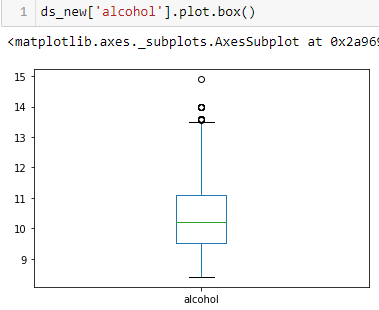
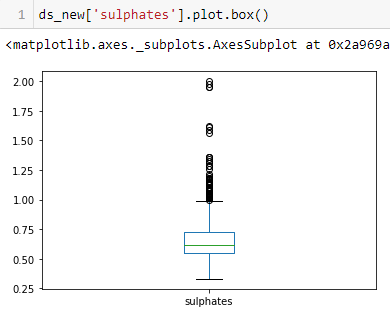
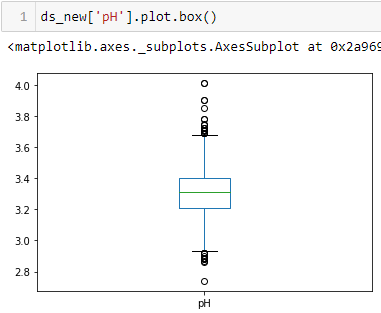
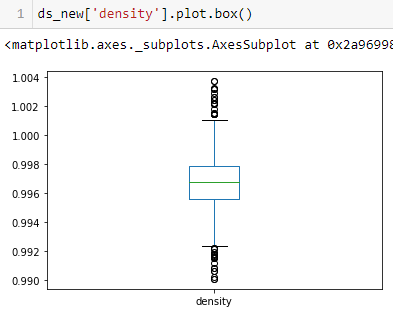
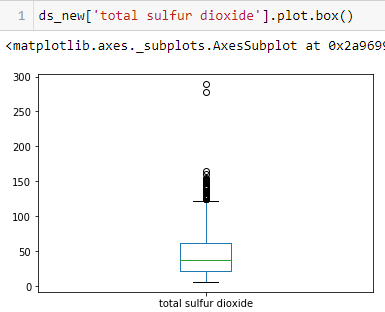
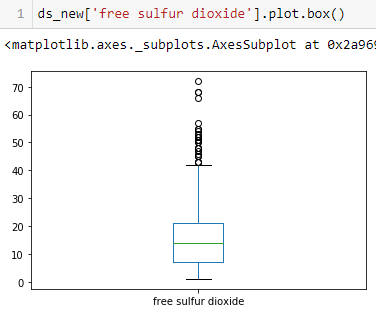
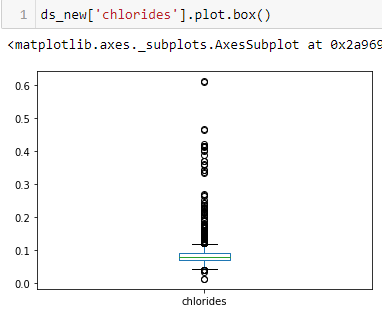
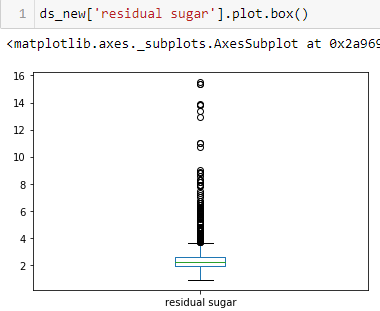
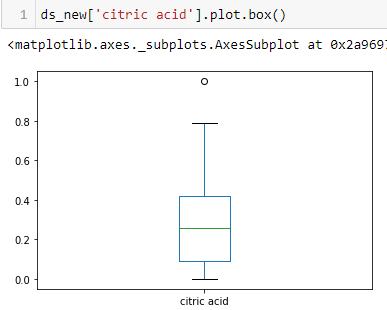
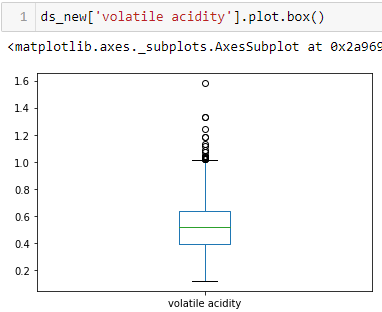
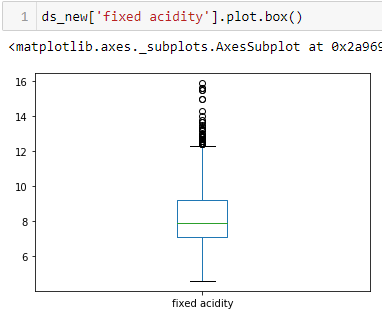


Here, in heatmap of dataset also we could not find any presence of null values.

**UNIVARIATE ANALYSIS:** The second most important thing in doing machine learning model is understanding the content of the dataset. There should not be presence of outliers in the dataset, outliers is nothing but the extraordinary values present in the dataset which must be present due to some external factors or wrong mentioning of data while building the dataset or maybe due to some other reasons.

In other words they are data records that differ dramatically from all others, they distinguish themselves in one or more characteristics. In other words, an outlier is a value that escapes normality and can (and probably will) cause anomalies in the results obtained through algorithms and analytical systems. We need to identify the range of the values presence in the dataset, if little percent of data values falls outside the range of the 90% values, we can identify such values as outliers. We can identify outlier values in any column by plotting box plot of that column.

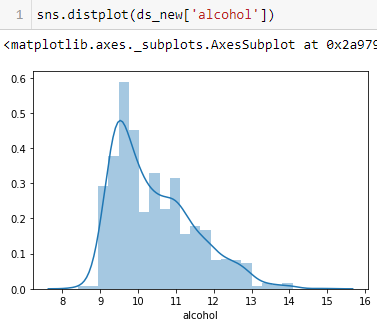
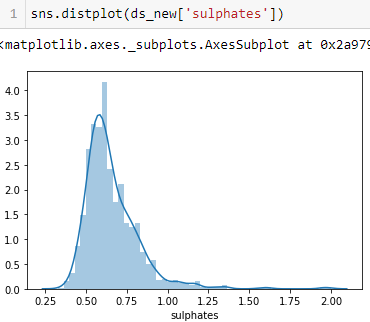
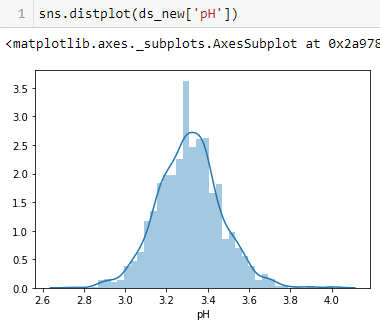
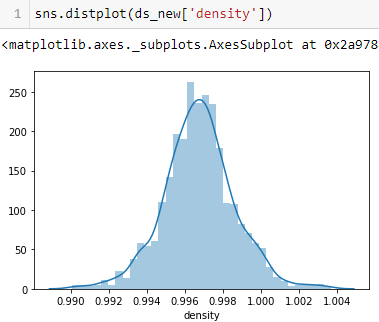
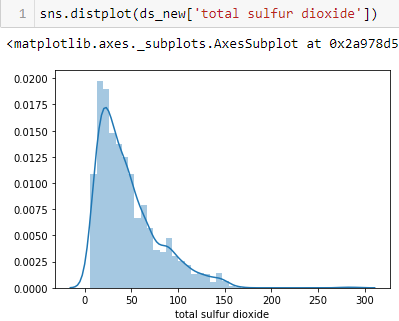
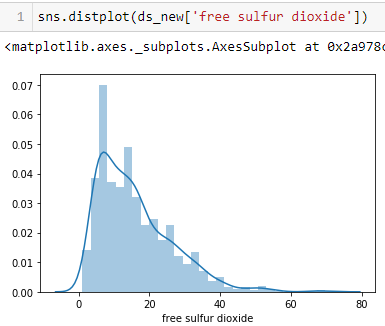
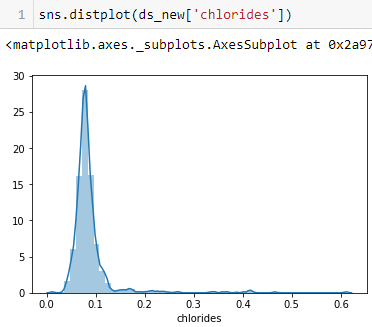
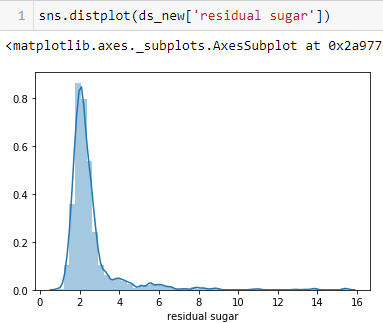
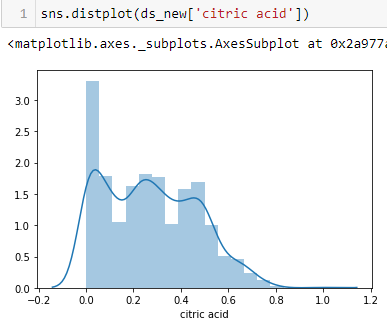
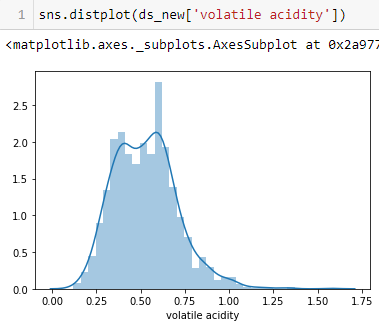
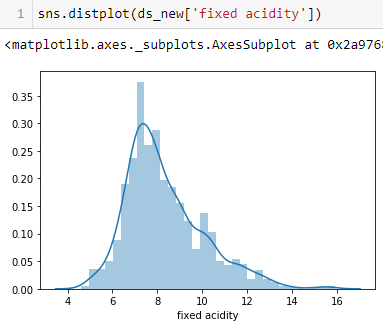
When we plotted box plot of all the columns in the dataset, we identified columns like ‘fixed acidity’, ‘volatile acidity’, ‘citric acid’, ‘residual sugar’, ‘chlorides’, ‘free sulfur dioxide’, ‘total sulfur dioxide’, ‘density’, ‘pH’, ‘sulphates’ and ‘alcohol’ contain outliers in them.



**Observation:**

**We observed by plotting boxplot of various columns that, around 11 columns contains outliers in them.**

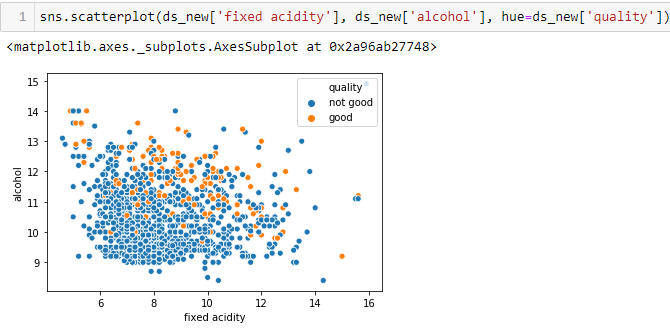
The another most important thing in **Exploratory Data Analysis** is checking skewness of the datasets. The skewness is nothing but a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side). In machine learning model making we check skewness of dataset with the help of distplot using seaborn library. If we pass the full two-dimensional dataset to kdeplot, we will get a two-dimensional visualization of the data showing data distribution. We made distplot of all 11 columns containing outliers found in earlier analysis of EDA.



**Observation:**

In distplot analysis we found out that, columns like ‘fixed acidity’, ‘volatile acidity’, ‘citric acid’, ‘residual sugar’, ‘chlorides’, ‘free sulphur dioxide’, ‘total sulphur dioxide’, ‘sulphates’, ‘alcohol’ contains right side skewness while columns like ‘density’ and ‘pH’ columns contains both sided skewness.

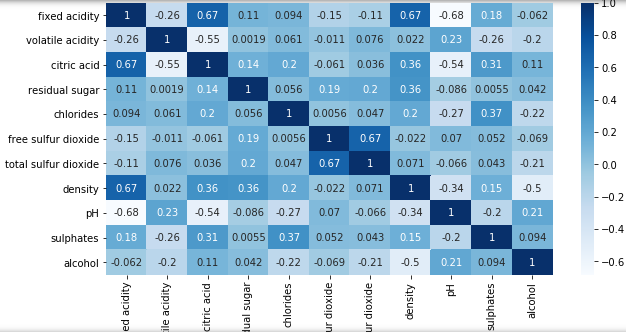
**BIVARIATE ANALYSIS:** Bivariate analysis is the type of analysis in which we analyse the dataset using two columns, here we did analysis using ‘fixed acidity’ and ‘alcohol’ column and taking hue as ‘quality’ column.



**Observation:**

Here, in this scatter plot of ‘fixed acidity’ and ‘alcohol’ columns we observed that, with increase in fixed acidity and alcohol content, quality of the alcohol increases.

**Correlation Matrix Analysis:** A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

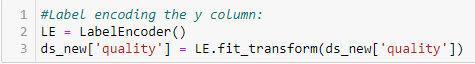


**Observation:** We observed here in this **heatmap** of **correlation matrix** that, ‘density’ column is highly correlated with alcohol with -0.5 values, ‘pH’ column is positively correlated with alcohol with 0.21 values.

**PRE PROCESSING PIPELINE:**

**LABEL ENCODING THE CATEGORICAL COLUMNS:** Before proceeding further, we need to confirm that all columns are numeric in nature, but we found out that all columns except ‘quality’ column which is of ‘object’ data type in nature. We need to convert it into numeric data types, we can do that using **Label Encoding technique** for converting this column in to nuemeric. Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

Below are codes used in order to do so.



**REMOVING OUTLIERS AND SKEWNESS:** Here in this dataset, we have used z-score method of removing outliers. A **z-score** (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is. A z-score can be placed on a normal distribution curve. We can choose a threshold value, Here, in this problem we have choosen 3 as a threshold value above which all are outliers. So we finally made a new dataframe named ‘df\_new’ containing values below 3 of z-score values. Codes are given below:

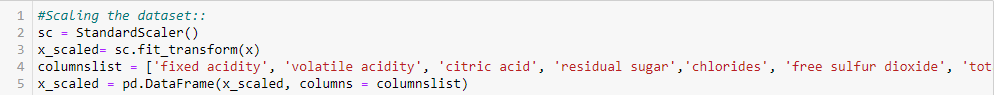


**Observation:** We are losing almost 8.81% of datavalues by removing outliers from the dataset which is very much fine.

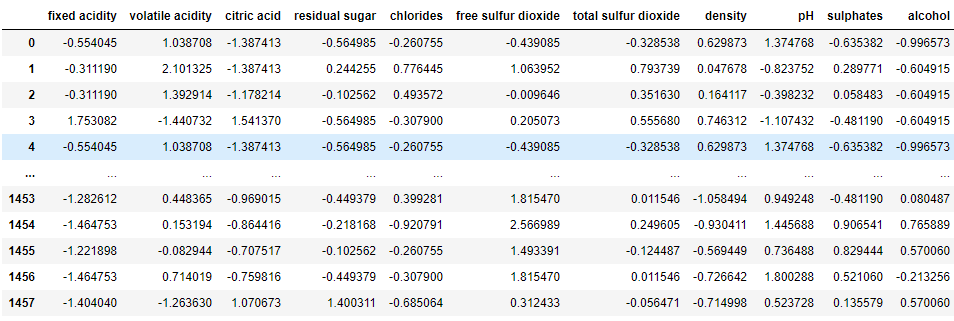
After removing outliers we need to extract ‘x’, ‘y’ data frames from ‘df\_new’ data frame in order to proceed further with the model making. ‘x’ data frame contains all columns except the ‘quality’ column while ‘y’ data frame exclusively contains ‘quality’ column only.



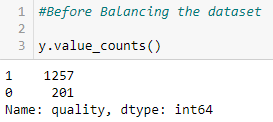
**SCALING THE DATASET:** The most important step in machine learning model making is to ensure all data columns are in same scale of values, we can bring them in same scale by scaling them using Standard Scaler scaling library. Standard Scaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. StandardScaler can be influenced by outliers (if they exist in the dataset, but we have already removed outliers) since it involves the estimation of the empirical mean and standard deviation of each feature.



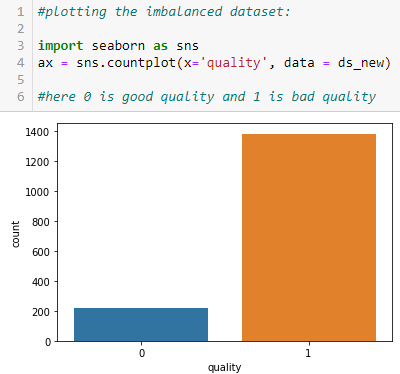
After scaling the dataset by using Standard Scaler as shown by above method, we got new dataset which is now scaled as shown below.



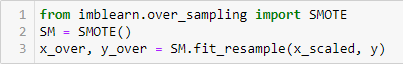
**CHECKING TARGET COLUMN DATA BALANCING:** Imbalanced data typically refers to a classification problem where the number of observations per class is not equally distributed; often you'll have a large amount of data/observations for one class (referred to as the majority class), and much fewer observations for one or more other classes (referred to as the minority class), that’s why checking balancing of the target column is must.



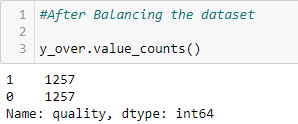
Plotting the imbalanced datset.



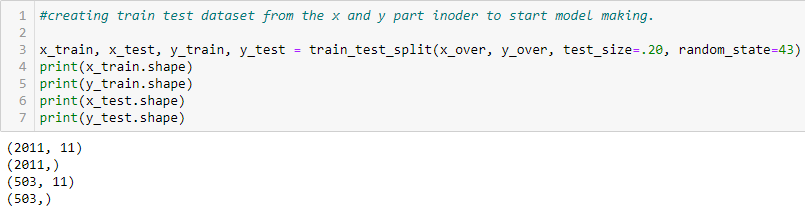
We can see that in the above plot, where 1 is ‘good’ and 0 is ‘not good’ type of wine values. We have datasets containing ‘good’ wine is around 1257 while datasets containing ‘not good’ wine is around 201, which clearly indicates imbalanced dataset. When dataset is imbalanced, it causes the machine learning model to be more biased towards majority class. It causes poor classification of minority classes. Thus balancing of datasets is very important. We can handle this issue by using SMOTE Library functions. In this type of library there are techniques by which we can over sample the under sample datasets.



After balancing over sampling the dataset, we can see that we have successfully balanced the datasets as shown below:



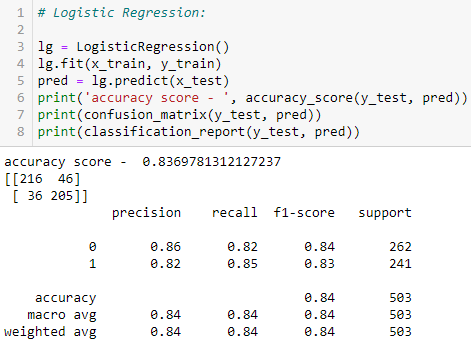
**TRAIN-TEST DATASET BUILING:** The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets Here we have kept 20% of dataset for testing the model while 80% of dataset for training the model.



**BUILDING MACHINE LEARING MODELS:**

**Machine learning model making:** Here in our problem we will try making machine learning model with six different types of algorithms i.e. Logistic Regression, GaussianNB, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier, Support Vector Classifier methods. We will try to find out each model’s Accuracy score, Confusion matrix, Classification report as well. Accuracy score will give us the percentage accuracy of the model in predicting the test datasets. Confusion matrix is a 26 by 26 matrix with the probability of each reaction to each stimulus. The classification report is used to measure the quality of predictions from a classification algorithm. The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives.

**Logistic Regression Method:** Logistic regression is basically a classification type of method, a statistical analysis method used to predict a data value based on prior observations of a data set. Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

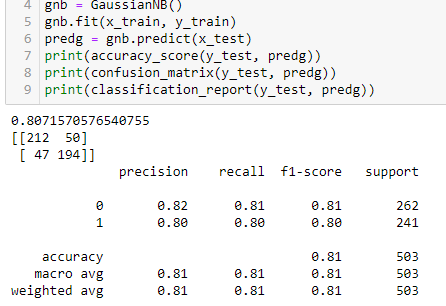


By using logistic regression method, we got accuracy score of **83.69%** which is very good. But this accuracy score may be due to over fitting of the data, we can check for over fitting by checking cross validation score of the model. It is basically a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Overfitting is a modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points. Overfitting the model generally takes the form of making an overly complex model to explain idiosyncrasies in the data under study. Here we’ll check cross validation score on 5 subsets of the datasets.



We can see that the cv score of logistic regression model is 80.78% which is very much close to Accuracy score of model i.e. 83.69%. We’ll try to make machine learning model on various other algorithms as well and we’ll check their accuracy score and cross validation score. The model which has least difference in their accuracy and cv score, we’ll select that model as our final model.

**GaussianNB Method:** A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e, normal distribution. Gaussian Naive Bayes method works on two categorical target datasets only, as our dataset contains two target categories only. We can very well use this method.

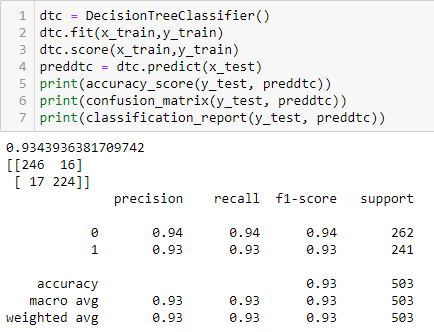


The accuracy score of the GaussianNB method is 80.71% which is again very good, but we need to check cv score as well.



Cv score of GaussianNB model is 77.52% which is again very close to our model’s accuracy score indicating less overfitting.

**Decision Tree Classifier Method:** A Decision Tree is an algorithm used for supervised learning problems such as classification or regression. Each leaf of the tree is labeled with a class or a probability distribution over the classes. A tree can be "learned" by splitting the source set into subsets based on an attribute value test.

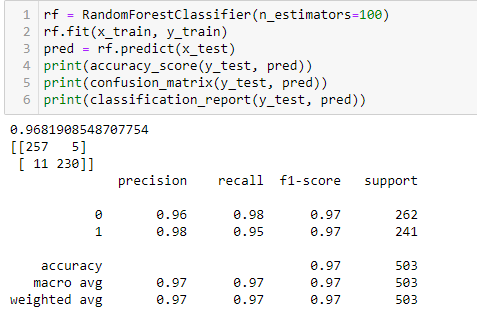


The accuracy score of the decision tree classifier method is 93.43% which is superb but again this might be due to overfitting so we need to check cross validation score as well.



Cross validation score of decision tree classifier model is 86.31% which is very much below decision tree classifier model’s accuracy score indicating overfitting of the datasets in the model making.

**Random Forest Classifier Method:** Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

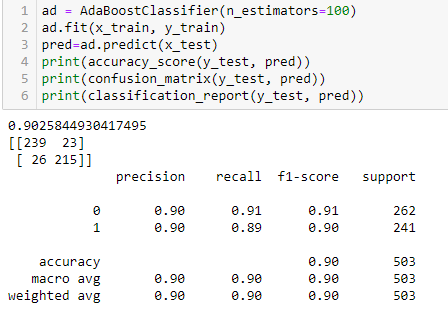


Accuracy of random forest classifier model is 96.81% which is again very good. Now we’ll check the cross validation score of the same.



Cross validation score of the random forest classifier model is 91.64% which is again very much below the accuracy score indicating overfitting of the values.

**Ada Boost Classifier Method:** Ada-boost or Adaptive Boosting is one of ensemble boosting classifier It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. Adaboost helps you combine multiple “weak classifiers” into a single “strong classifier”. The weak learners in AdaBoost are decision trees with a single split, called decision stumps. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well.

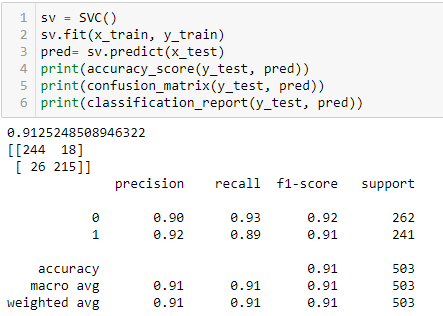


Ada Boost Classifier has an accuracy of 90.25%. but we need to check its cross validation score as well.



Cross validation score of ada boost classifier model is 85.59% which is very much below the accuracy score indication presence of overfitting in the model.

**Support Vector Classifier Method:** A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. So you're working on a text classification problem. The classifier separates data points using a hyperplane with the largest amount of margin. That's why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points.



Accuracy score of support vector classifier is 91.25% which is again good, but we need to check the cross validation score as well.

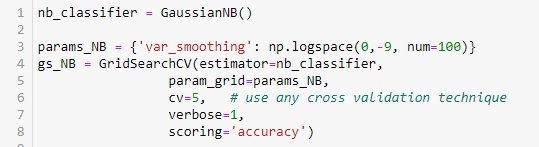


Cross validation score of the support vector classifier model is 86.03% which is very much below the accuracy score of the model indicating presence of overfitting values in the model.

**Observation:** After analyzing each model’s accuracy score and cross validation score, we found out that difference between accuracy score and cross validation score of GaussianNB model is least also and of good percentage also. we can say that GaussianNB is our best fit model. we have confirmed with cross checking the cross validation score of GaussianNB that overfitting is very less in the model. Thus we can conclude here that GaussianNB model is our final model.

**HYPER PARAMETER TUNING OF MODEL:** Hyperparameter tuning is basically choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyperparameter tuning.

Grid search is basically the most used hyperparameter tuning method. With this technique, we simply build a model for each possible combination of all of the hyperparameter values provided, evaluating each model, and selecting the architecture which produces the best results.

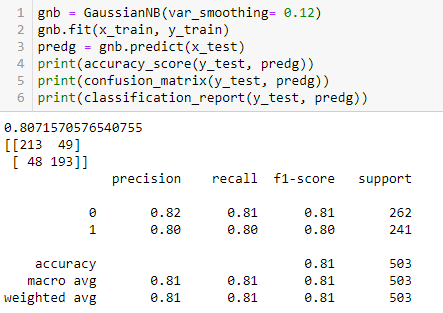


While using grid search cv, for finding best fit parameters, we used a range of values for var\_smoothing parameter of GaussianNB method and then we’ll check what best value we get in order have maximum accuracy of the model. Passing the train test values and checking the best parameters of grid search value.



We found that ‘var\_smoothing’ value of 0.12 is the best parameter for final model making.

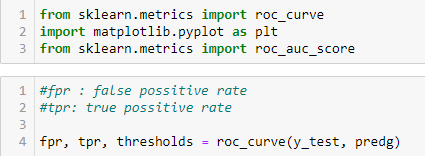
**Final Model Making:** We used this value of var\_smoothing i.e. 0.12 as a parameter in making final model as follows.



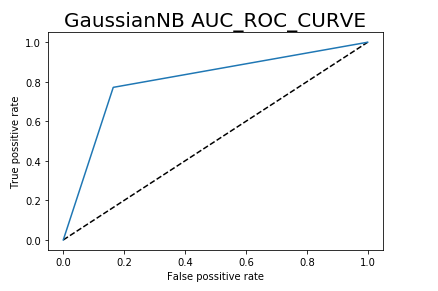
After using the best parameter of grid search cv also we found that we have got accuracy of 80.71% which is similar to earlier GaussianNB model, so we can say that’s the maximum accuracy we can get in GaussianNB model.

We can conclude that now GaussianNB model is our final model with the accuracy score of 80.71%.

**AUC-ROC CURVE MAKING:** The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. We can say that the Higher the AUC, the better the model is at distinguishing between wine with the good quality and with bad quality or not good quality. For auc-roc curve making also we need to import various other libraries as shown below.



Fpr is nothing but false positive rate and Tpr is nothing but true positive rate.



**Observation:** From the above Auc-Roc Curve, we can say that the curve auc-roc is having a very sharp curvature indicating very good model building. And the area under curvature is also good.

**SAVING THE MODEL:** We have saved our final model having accuracy of 80.71% by model name ‘Vaibhav\_Red\_Wine\_Project\_Model.pkl’.

**CONCLUDING REMARKS:**

**Conclusion:** Our model is very well built and having accuracy of 80.71%.

**File path:** File is saved on github.

https://github.com/vaibhav903174/datasci/blob/9b5843a6c5b2ac043703c8ceee827c60c4ed1956/Vaibhav\_Red\_Wine\_Quality\_Prediction\_26\_06\_2021.ipynb

**Thank You**