RED WINE PROJECT

Project Flow :

1.Problem Statement

2.Data Analysis

3.EDA Conclusions

4.Pre-Processing Pipeline

5.Model Building

6.Concluding remarks

By :

Naveen

Problem Statement

The 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.).

This dataset can be viewed as classification task. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Attribute Information

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

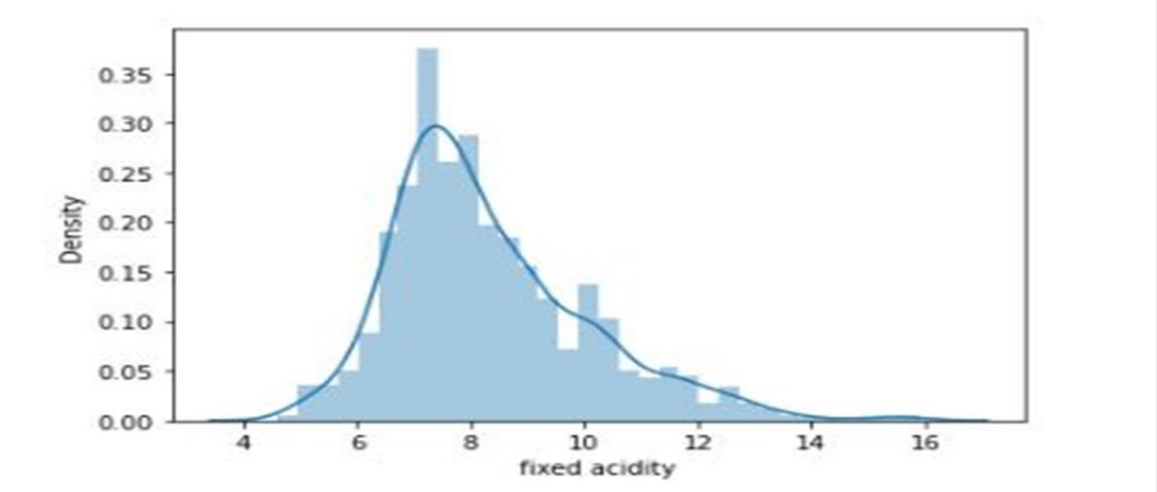
8 - density

Data Analysis

We have analysed this dataset in three stages

1.Univariate Analysis :

In this analysis we have taken each column one by one and we thoroughly analysed it, depending upon the data it holds we used suitable technique to extract maximum information out of it,

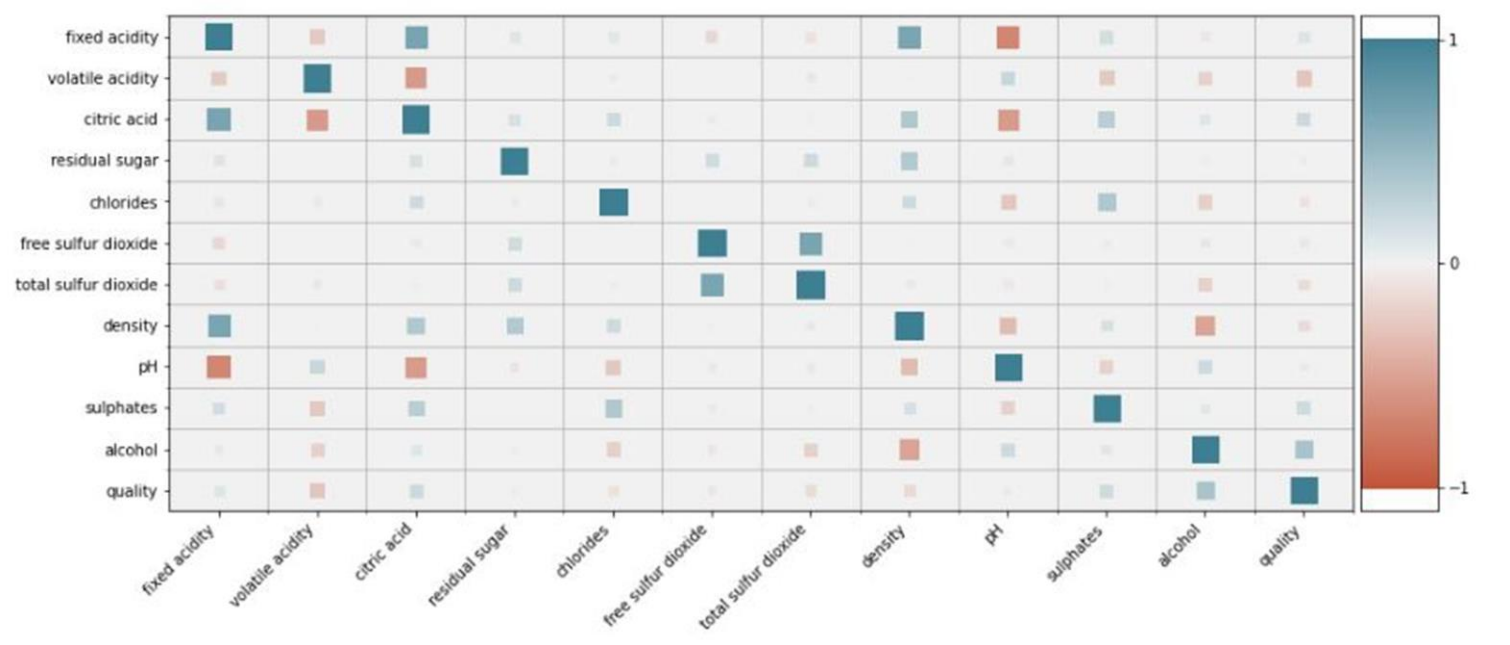


Above is an example of Fixed acidity column analysis through distplot showing how the fixed acidity are being distributed, We can see that the fixed acidity is slightly are right skewed, Univaiate analysis mainly infers that analysing only one variable.

Through Univariate analysis we will get to know how the data is been distributed like whether its uniformly distributed or skewed, or any outliers are there etc.

2.Multivariate Analysis :

In case of the multivariate analysis we analyse more than two variables simultaneously, We will get to know the relationship between differerent variables, From this technique we can extract maximum informations which are hidden.



Above is an example of multivariate analysis, the above plot is corrplot depicting the relationship between all the elements simultaneously.

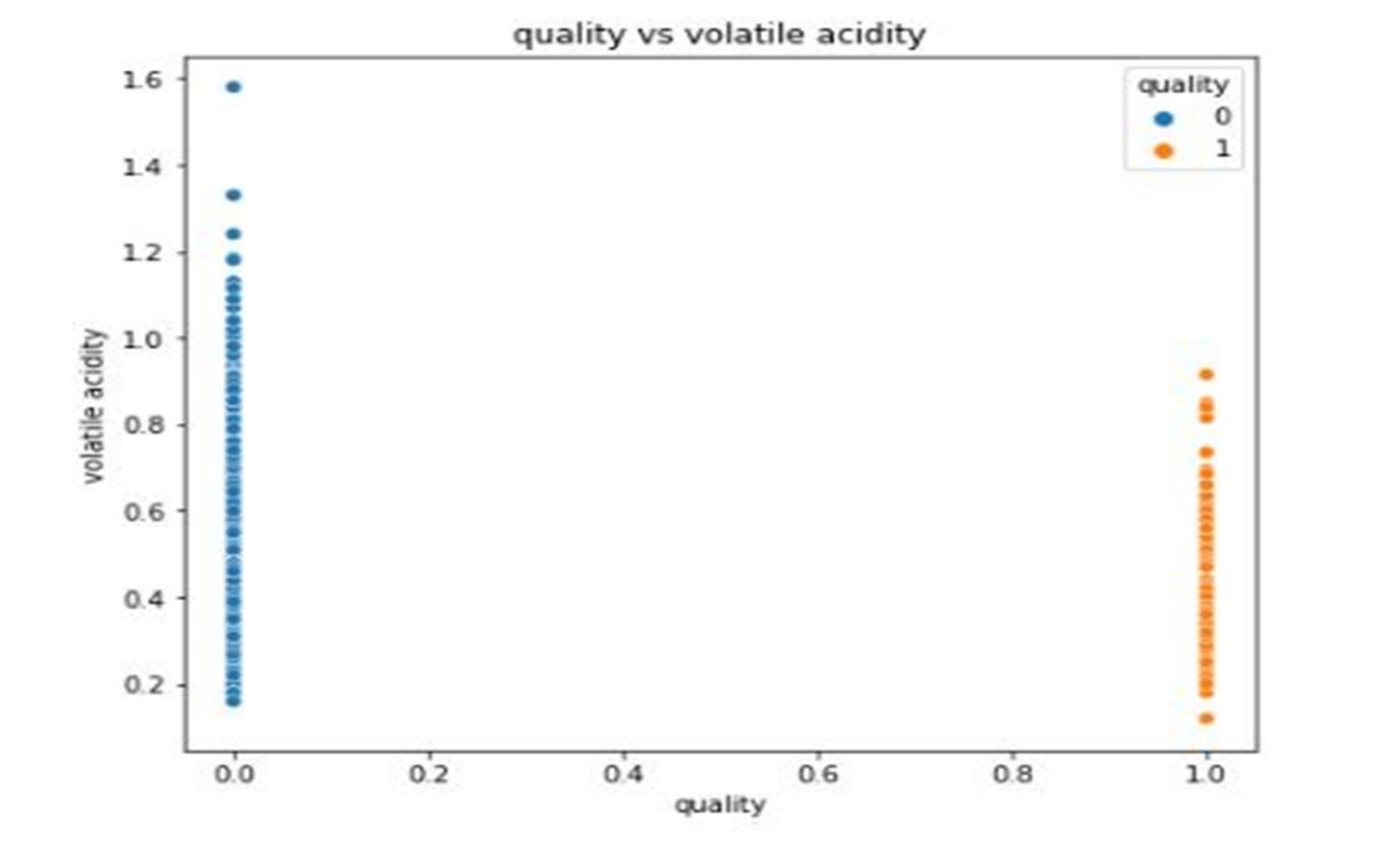
We have different plots for carrying out the multivariate analysis, Heatmap and corrplot are the major ones through which we will get to know how each variables are related to each other.

Grouping the variable and analysing them is also a type of multivariate analysis, from which we can extract maximum information because we can group the categories like yes or no and we can compare the features.from heatmap we will get to know whether an variable is positively correlated or negatively correlated.

3.Bivariate Analysis :

From multivariate analysis we will get to know which are the independent variables which are significantly correlated to each other we will analyse those pair of variables in Bi vaiate analysis.

From multivariate analysis we will get to know which are those independent variables which are significantly correlated to the response variable we can analyse those pairs in the bi variate analysis.



Above is an example of bi variate analysis in which we are analysing volatile acidity and quality

From bi variate analysis we will get know how an independent variable is realated to response variable, and we can prescribe required precautions to be effective.

Conclusions from EDA

1.The value of volatile acidity value wont go higher than 1 for good class wine.

2.For producing good wine dont keep the value of citric acid value more than 0.8.

3.To produce good wine keep the sulphates value less than 1.15

4.Dont keep the value of alcohol less than 9 for producing good quality wine.

5.Dont keep the value of residual sugar more than 9 for producing good wine.

6.Keep the value of chlorides less than 0.2 to produce good quality wine.

7.keep the value of free sulphur dioxide less than 40 to produce good wine.

8.Keep the value of total sulphur dioxide less than 120 for producing good wine.

#keep the values in following range for producing high rated wines

1.The fixed acidity value should be in the range of 8.56 to 10.22

2.volatile acidity value should be in the range of 0.42 to 0.85

3.citric acid value should be in the range of 0.39 to 0.53

4.residual sugar value should be in 2.57 to 2.6

5.chlorides value should be in the range of 0.06 to 0.075

6.free sulphur dioxide should be in the range of 13.27 to 16.5

7.total sulphur dioxide should be in the range of 33.44 to 43

8.density should be in the range of 0.995 to 0.998

9.PH should be in the range of 3.26 to 3.72

10.sulphates should be in range of 0.76 to 9.82

11.alcohol should be in the range of 12.09 to 12.87

Cleaning and preparing an precise input data:

1.Cleaning the datatset: This phase includes finding null values or unexpected characters in the datatset and treating them,We can treat them by various techniques like replacing, deleting, or by removing the rows.

2.**Removing the outliers :** This phase includes removing of the outliers, There are various methods for removing the outliers 1.zscore method 2.IQR method,3.Standard deviation method, We can use convinient method depending upon the situation, we have to ensure that the data loss shouldnt be more than 7%.

3.**Removal of the skewness :** Except categorical columns we have to remove the skewness of every other column.

4.We have to scale the independent variables if required so that the model will adapt high accuracy.

Building the models

We have ample number of models for analysing and building a solution for the classfication type of problem, In this particular scenario we gonna look at the best accuracy score and f1 score of each model and select the best model based on the result.

1.**Logistic Regression :**

```
accuracy score of LogisticRegression()
0.926
              precision    recall  f1-score   support

           0       0.94      0.98      0.96       321
           1       0.59      0.34      0.43        29

    accuracy                           0.93       350
   macro avg       0.77      0.66      0.70       350
weighted avg       0.91      0.93      0.92       350
```

From the above results we can see that Logistic regression has produced an accuracy of 0.926 and f1 score as 0.96/0.43

2.GaussianNB :

```
accuracy score of GaussianNB()
0.886
              precision    recall  f1-score   support

           0       0.98      0.89      0.93       321
           1       0.40      0.79      0.53        29

    accuracy                           0.89       350
   macro avg       0.69      0.84      0.73       350
weighted avg       0.93      0.89      0.90       350
```

From the above results we can see that the accuracy score is 0.886 and f1 score is 0.93/0.53

3.Support vector classifier :

```
accuracy score of SVC()
0.917
              precision    recall  f1-score   support

           0       0.92      1.00      0.96       321
           1       0.00      0.00      0.00        29

    accuracy                           0.92       350
   macro avg       0.46      0.50      0.48       350
weighted avg       0.84      0.92      0.88       350
```

Through svc we got an accuracy of 0.917 and f1 score of 0.96/0.

Decision tree classifier :

From decision tree classifier we got an accuracy score of 0.889 and f1 score of 0.94/0.47

```
accuracy score of DecisionTreeClassifier()
0.889
              precision    recall  f1-score   support

           0       0.96      0.92      0.94       321
           1       0.39      0.59      0.47        29

    accuracy                           0.89       350
   macro avg       0.67      0.75      0.70       350
weighted avg       0.91      0.89      0.90       350
```

KnearestNeighborsclassifier:

```
accuracy score of KNeighborsClassifier()
0.917
              precision    recall  f1-score   support

           0       0.96      0.95      0.95       321
           1       0.50      0.55      0.52        29

    accuracy                           0.92       350
   macro avg       0.73      0.75      0.74       350
weighted avg       0.92      0.92      0.92       350
```

From KNN classifier we got an accuracy score of 0.917 and f1 score of 0.95/0.52

Adaboost Classifier :

```
accuracy score of AdaBoostClassifier()
0.923
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       321
           1       0.53      0.55      0.54        29

    accuracy                           0.92       350
   macro avg       0.75      0.75      0.75       350
weighted avg       0.92      0.92      0.92       350
```

From this classifier we got an accuracy score of 0.923 and f1 score of 0.96/0.54.

RandomForestClassifier :

```
accuracy score of RandomForestClassifier()
0.951
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       321
           1       0.77      0.59      0.67        29

    accuracy                           0.95       350
   macro avg       0.87      0.79      0.82       350
weighted avg       0.95      0.95      0.95       350
```

From Random forest classifier we got an accuracy of 0.951 and f1 score of 0.97/0.67

Based upon the least difference between accuracy and cross val score, and based upon the f1 score its better to choose Random Forest Classifier as the best model.

Conclusion remarks:

The model has been built for best accuracy optimising all loopholes in the datatset.The model has secured an auc_roc score of 0.951 and It scored very well in all its test.

.....Thank you!!