Project Name: Wine Quality Prediction

EDA Summary

- There are no missing values in the White wine dataset
- The quality rating 5,6, and 7 holds the major proportion of the data.
- Residual Sugar and Density, Free Sulfur Dioxide and Total Sulfur Dioxide, Total Sulfur Dioxide and Density, Alcohol and Quality are strongly correlated.
- Alcohol is weakly correlated with Density, Residual Sugar and Total Sulfur Dioxide. Also, pH is weakly correlated with Fixed Acidity.
- 80% of wines are of bad quality and only 20% of wines are of good quality in the dataset.
- 1833 rows are removed due to the presence of outliers.

Preliminary Data Manipulations

- There were outliers present in the dataset which were detected by Interquartile Range (IQR) and removed to provide better accuracy.
- The quality label has been classified as Good and Bad. If the quality is 6 or less than 6 then the wine will be considered of bad quality and if the wine quality is greater than 6 then it will be considered of good quality.
- The good and bad wine quality has been encoded to 0 and 1 with the Scikit-learn Label Encoder.

Statistical Analysis

- There is a large difference between 75% (75th percentile) and max values in Residual Sugar, Free Sulfur Dioxide and Total Sulfur Dioxide.
- From the statistical analysis, it is clear that there were outliers present in the dataset.

Key Candidate Features

- The key candidate features in the dataset are Quality, Alcohol, Residual Sugar and Density.

Rational Statement

Wine which is considered a luxury product is enjoyed by most people around the world. Portugal is the largest producer and exporter of premium quality wines and one such white wine is Vinho Verde which is appreciated by people all over the world. To produce such wines, certification and assessment of wine quality are necessary which is done by grading the wine on the basis of taste and vintage. This process of testing the quality of wines is in-efficient and time-consuming. Moreover, it becomes costly for the producer to get the wine quality checked by the experts when there is a huge demand for these wines.

To overcome this problem, I will develop a machine-learning algorithm to predict the quality of the wine based on the various parameters which play an essential role in determining the quality of wines. These parameters are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol which are obtained by the physiochemical test of wines. The output of this model is to determine the quality of wine on a scale of 1 to 10. This model will not only make the whole process of quality checking efficient but also cheaper with less human interaction and it will benefit both the certification bodies and producers who will improve the quality of a wine by modifying various physiochemical properties.

Data

In this project, I will use 11 physiochemical parameters as independent variables and 1 dependent variable to score the quality of wines from 1 to 10. The data must be split into training and testing data.

White wine Dataset

Source: https://www.kaggle.com/christiankobayashi/winequalitywhite

Dataset Size: 258.21 KB

This dataset contains 4898 entries of Portuguese Vinho Verde white wine. The data is structured in CSV format.

Data Description

1. **Fixed Acidity:** are non-volatile acids that do not evaporate readily

2. **Volatile Acidity:** are high acetic acid in wine which leads to an unpleasant vinegar taste

3. **Citric Acid:** acts as a preservative to increase acidity. When in small quantities, adds freshness and flavor to wines

4. **Residual Sugar:** is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between sweetness and sourness. It is important to note that wines > 45g/ltrs are sweet

5. **Chlorides:** the amount of salt in the wine

6. **Free Sulfur Dioxide:** it prevents microbial growth and the oxidation of wine

7. **Total Sulfur Dioxide:** is the amount of free + bound forms of SO2

8. **Density:** sweeter wines have a higher density

9. **pH:** describes the level of acidity on a scale of 0–14. Most wines are always between 3–4 on the pH scale

10. **Sulphates:** a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant

11. **Alcohol:** available in small quantities in wines makes the drinkers sociable

12. **Quality:** which is the output variable/predictor

Limitations

This project is limited to exploring the parameters related to the composition of white wine. The other two factors limiting the current research are time and cost. The data has been collected from the open-source which is freely available on the internet.

Testing Process

I will divide the dataset into two parts i,e. 75% training dataset and 25% test dataset. The model will be trained with the training dataset and then it will be tested for accuracy with the test data.