Wine Quality Prediction

Data Wrangling

OVERVIEW

This document describes the dataset for Wine Quality prediction and lists any data wrangling steps performed.

DATA

The dataset has been obtained from Kaggle and it consists of a comprehensive list of red and white wines features which include volatile acidity, pH, alcohol content, citric acid content, residual sugars, chlorides along with the corresponding quality rating which ranges between 1-10.

But what do these features mean in real life?

1. fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2. volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines
4. residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
5. chlorides: the amount of salt in the wine
6. free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. density: the density of water is close to that of water depending on the percent alcohol and sugar content
9. pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11. alcohol: the percent alcohol content of the wine.

All features are measured in g/dm^3 except for total + free sulfur dioxide and alcohol which is in mg/dm^3 and % by volume respectively.

DATA WRANGLING

Procedure

- The rows and columns of the dataset were examined. It has 6497 rows and 13 columns.
- Dataset was checked for any missing or null values. It did have null values which were replaced by the median value of that particular column. Medians are more robust/ less sensitive to outliers when compared to means which is why I chose to replace the nulls with medians.
- The IQR was used to detect outliers. There were a few outliers in almost all the feature columns. But for this project, the outliers were kept as it is and sklearn’s RobustScaler which is not sensitive to outliers, will be used for feature scaling in the later sections of this project.

CONCLUSION

After checking the data for missing or null values and outliers, and taking the necessary data wrangling steps to deal with them, the dataset is now ready for Exploratory Data Analysis.