**WINE QUALITY PROJECT**

*Hypothesis*

Using wine quality data with features such as citric acid levels, pH, alcohol, acidity, density, sulphates, residual sugars, chlorides, volatile acidity and fixed acidity, and a quality measure between 0 and 10, our team goals are to analyze two data sets, using Machine Learning to decide with various classification methods which yields the highest accuracy and to determine which features are most indicative of a high-quality wine.

*Reading Data*

After importing all the required libraries, we checked what technical information is contained in the data and what could be missing. There are 1599 rows and 12 columns in total for red wine data and 4898 rows and 12 columns in total in white wine data which is clean and has no missing values.

*Visualization*

This is the graphic representation of the data used to find useful information. The first plots are bar graphs of quality versus count to check the number of each type of wine that fall under a certain measure of quality. Both wines tend to concentrate in the midrange with very few lying in very good and very bad ranges. The histograms reveal how data is easily distributed on different wine features followed by descriptive statistics which measured position, dispersion, central tendency and frequency.

*Statistical test*

We used correlation matrix to find the bonding and relationship between two features. We found that total sulfur dioxide and free sulfur dioxide with a correlation of more than 0.7 is directly correlated therefore we dropped it off as an important determinant of quality.

*Most High-Quality Indicative features*

By comparing the three models that is; decision tree, random forest and support vector classifier, random forest model seems to yield the highest level of accuracy. The top two features in accuracy prediction are alcohol and density while the least is fixed acidity. Therefore, random forest model is the best applicable only to this project,