# Midterm project report

Yuchen Hua

## Introduction

**Dataset**: The datasets I am using for this project contain two main part, one is red wines' data and the other is red wine. The two datasets contain both physicochemical properties and sensory graded wine quality. White wine dataset has 4898 data points, and red wine has 1599 points.

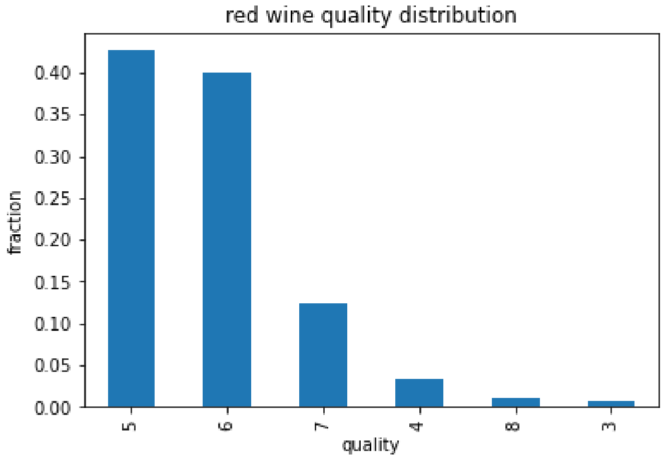
**Regression question:** The problem I want to investigate is that which physicochemical feature has important influence on the wine quality. The target variable is 'quality', which is sensory data from wine experts' evaluation. 0 means the quality is bad and 10 means excellent. This is originally a regression problem as I am interested in which feature or which combination of features give better quality of wine, but as this dataset could also be used for classification, I plan to start with regression method and then using classification method to see if it gives a better result.

**Interesting points:** This problem is interesting because when given the physicochemical features which can improve the quality of wine, then wine vinery can produce accordingly more tasty and popular wines to fit the market as well as improve wine production.

**Other publications:** The datasets are well described on UCI. The datasets are used from paper [Cortez et al., 2009]. In this paper, authors used multiple regression(MR), neural network(NN) and support vector machines(SVM) models, and the author find (according to SVM model which gives the better performance) for red wine, the most important factor impact wine quality are 'sulphates', 'pH', and 'total sulfur dioxide' .For white wine the most important three factors are 'sulphates', 'alcohol', 'residual sugar'. They found the most relevant factors tat effect wine quality and gain a model accuracy up to 89.0% (red) and 86.8% (white), and even above 90% for majority of classes.

## Exploratory Data Analysis

- Explore target variable

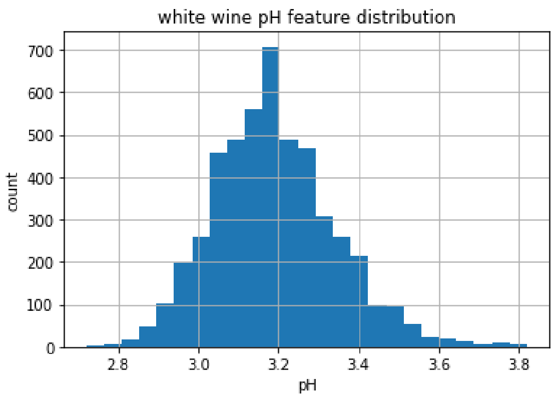


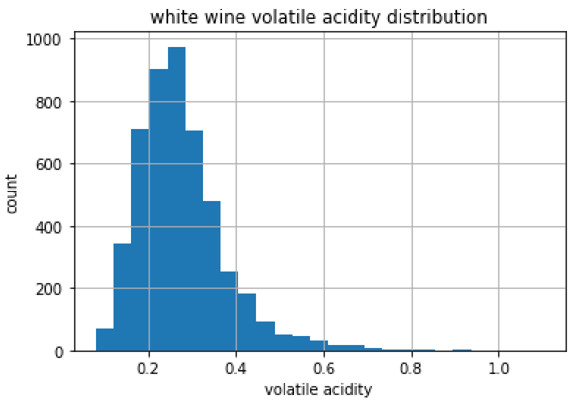
According to the figures, the data is imbalanced as most of the wine quality are centered around normal scale (around 5 and 6), just a few are extremely bad (around 3) and extremely good (around9). Also the data is already ordered (0 is bad and 10 is good). Red wine has 6 target variable values (3-8) and white wine has 7 (3-9). (Since red and white have same pattern so that I just include one figure, except white wine has quality class 9).

- Input variables distribution

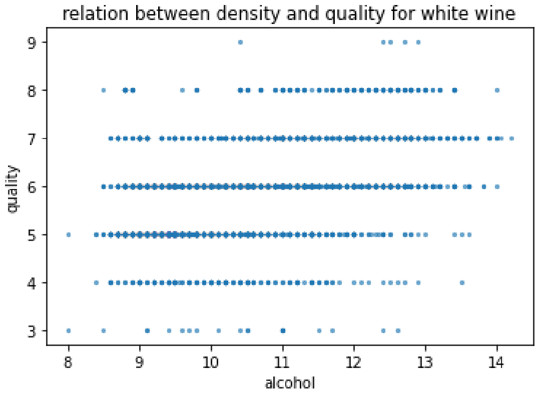
  As input 11 feature variables are all float type, it is appropriate to use MinMax or Standard Scaler. There are two typical kinds of distribution of those features. One is the feature has tail distribution and the other kind is more like a normal distribution. I include these two kinds as following (red wine and white wine has the same input feature patterns)

  Features 'alcohol', 'chlorides', 'critic acid', 'fixed acidity', 'free sulfur dioxide', 'residual sugar', 'sulphates', 'total sulfur dioxide', 'volatile acidity' are tailed distributed, as a result, I am using Standard Scaler on those features. 'density' and 'pH' are normal distributed, I applied MinMax Scaler on them in preprocessing.

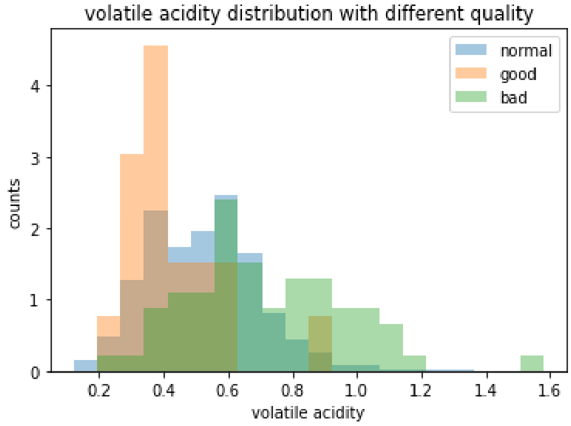




- Relations between each feature and target variable



The most linear related featured with target variable I found is alcohol. The data is ordered and rounded up to integers, but we could still see a linear relation between alcohol and quality, the larger percent the alcohol, the better quality of the wine.



For more direct visualization, I classify the quality as 3-4 for bad quality, 5-7 for normal and 8-9 for good. Except alcohol feature we have seen may have an influence on quality, another potential candidate is volatile acidity feature. From the figure we can see lower volatile acidity, better quality the wine is.

- **Data Preprocessing**

I am using stratified K-fold method splitting data, because the data is imbalanced. So that we can use stratified k-fold to put small fraction classes data (such as quality 8, 9 or 3, 4) into every set. The data is IID because it does not have group structure as well as time series generated data. There are 11 features in the data, and as stated above, all of them are numerical data having float type. I scaled two of them ('density' and 'pH') using MinMaxScaler, and the others suing StandardScaler as they have tailed property. As the label are already ordered encoded, there is no necessity to apply LabelEncoder on it.

- **References**

Cortez, Paulo, et al. "Modeling wine preferences by data mining from physicochemical properties." *Decision Support Systems* 47.4 (2009): 547-553.

**Github link:** https://github.com/yuchen996/data1030-final-project