# AI Algorithms: Final Project

# Statement of Work V2

Name: Sravan Pingali

Student ID: 100809884

## Rational and Problem Statement

'The older a wine is, the better its quality.' This is a statement that is commonly used for wine. While it is true there are many other factors that are responsible for the quality analysis of wines.

The wine market all over the world is around 355 billion US dollars. Global wine market has seen steady growth over the last decade and the global wine market is expected to worth around 429 billion US dollars in three years. That is a projected growth of 21%. This market isn't going decline in the coming years, if not decades and crafting quality wine is vital in maintaining this steady growth. Much like artwork, quality wines can be extremely rewarding. Wine making is a time-consuming process. Many commercially successful wine companies are very meticulous in how they make wine. This includes using the accurate measurements of ingredients for fermentation. Until a few decades ago, making quality wines would have taken excruciating amount of time, as people had to heavily depend on trial and error approaches to create new flavors. Wine companies had to make batches with different ingredient ratios to figure out the best possible combination to make good wine. This has been the traditional way in which new wine flavors were found. But now, with the advent of data analytics and AI, this process is made simpler. Using data analytics to predict the quality of wine, can be used as a preliminary way of discarding ingredient ratios that may now work. This will help in reducing the number of trail sessions required to find new flavors. Hence, predicting the quality of red wine can help save resources of a wine making company. Essentially, the computer algorithm will select which combination of the features constitute for a wine to be of good quality.

This problem can be solved using classification. The goal of this project is to predict wine quality based on certain characteristics of the wine.

## Data Sources and Data Requirements

The dataset that will be used here is obtained from Kaggle. This data was initially obtained by UCI Machine Learning repository. The complete list of features is provided below.

| | |
|---|---|
| 1. | Fixed acidity |
| 2. | Volatile acidity |
| 3. | Citric Acid |
| 4. | Residual Sugar |
| 5. | Chlorides |
| 6. | Free Sulphur dioxide |
| 7. | Total Sulphur dioxide |
| 8. | Density |
| 9. | pH |
| 10. | Sulphates |
| 11. | Alcohol |
| 12. | Quality |

Of these, the first eleven are the features that will be used to predict the quality of wine. The data set will be divided into test and training set to check the accuracy of the method applied.

Though there may be other factors influencing the quality of wine, it is assumption here that these 11 features will be enough to get an accurate prediction for the quality of a wine. It is also assumed that this dataset is a representative of most wines.

Some of the limitations are:

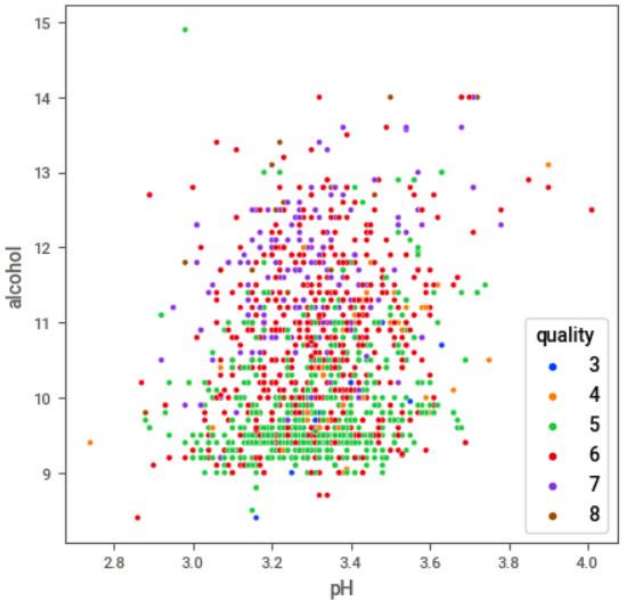
i) Incomplete data
ii) Inconsistent data (outliers)
iii) Duplicate data

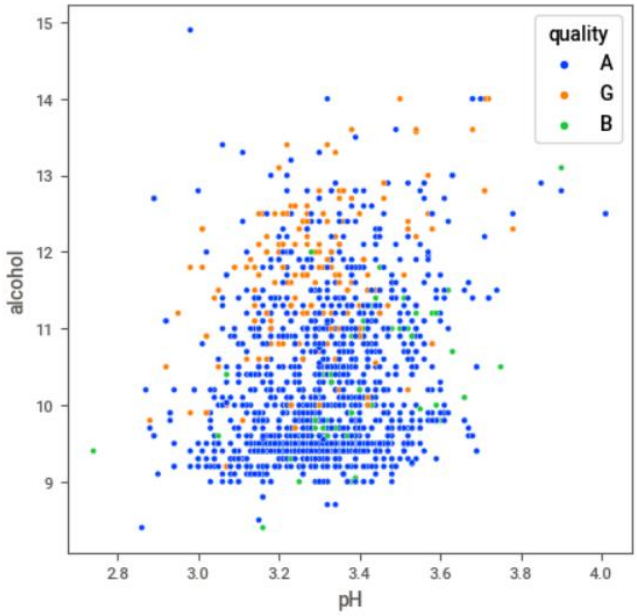
## Test Process to guarantee quality of work

Since, this is a classification problem. Accuracy will be used as the primary factor in determining the quality of the work. Accuracy will be used to before and after feature selection to check if there is an improvement or a decrement in the performance of the prediction. Training set will be used to train the classification models. After training, test sets will be used to check the success of a given model. The prediction accuracy over these test sets will determine the quality of work. Higher accuracy means good quality of work and lower accuracy means poor quality of work. Cross validation will be used to check which classification models have performed the best.

## Exploratory Data Analysis

Upon performing some rudimentary EDA, we can see that the almost all features seem to have independent of other features. Further visualizing of it is required. EDA was also after changing quality feature from a numerical variable to a categorical variable. They can be seen in the notebook file.




More EDA is in the notebook file.

## Data Manipulation and Cleaning

The data downloaded from Kaggle for this project is very clean. Nonetheless, the data was checked, and any 'Not a Number' entries have been removed. Further, the quality values *may* be divided into categorical values to see if any new insights can be found.

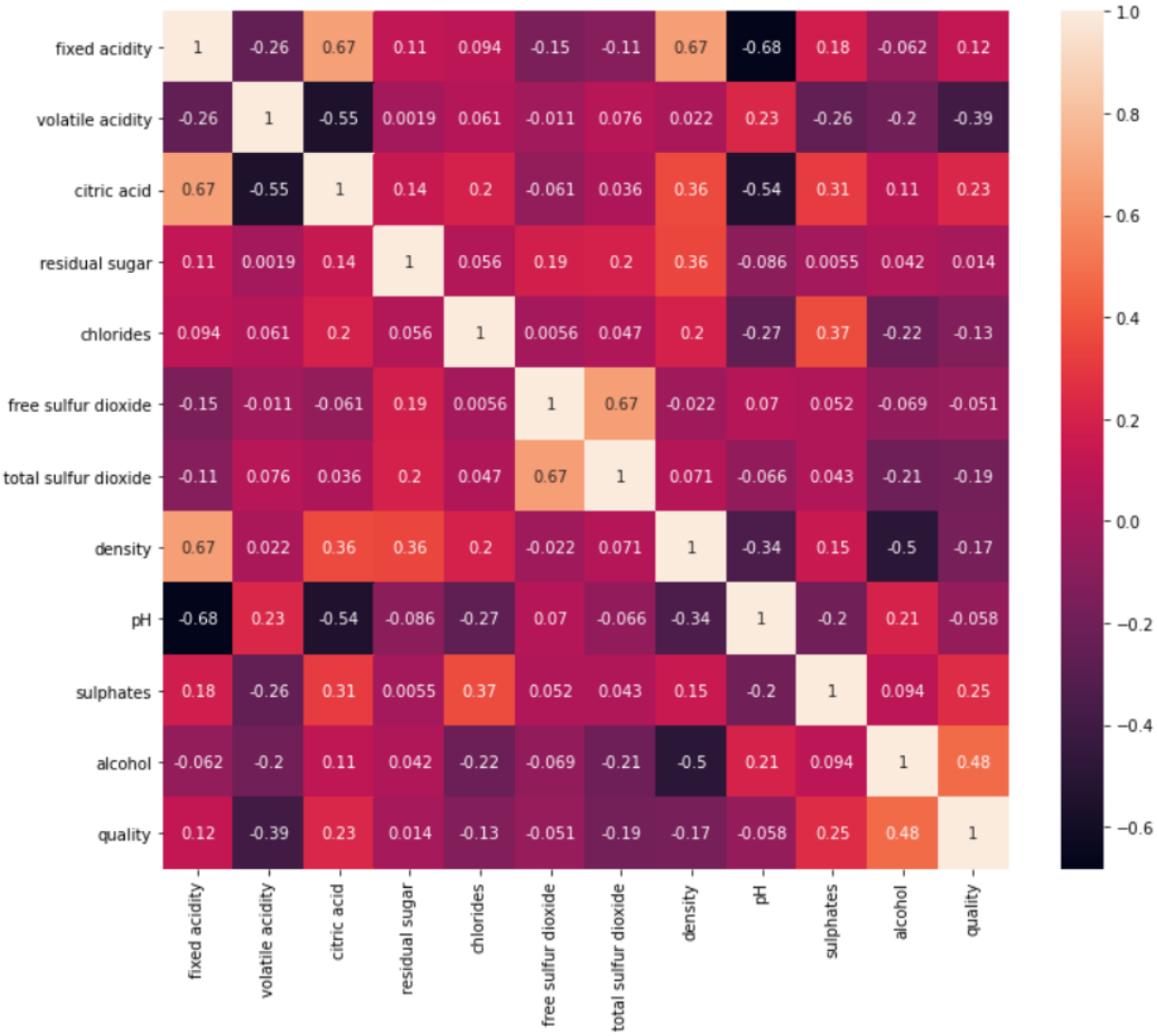
The categorical values will be B (Bad), A(Average), G(Good). Currently the data has 5 numerical values – 3, 4, 5, 6, 7 & 8. The distribution to categorical values will be equally distributed among these numerical values. They can be seen below.

3, 4 – Bad
5, 6 – Average
7, 8 – Good

## Correlations

Heat maps and correlation functions have been used to find some correlations.

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **fixed acidity** | 1.000000 | -0.256131 | 0.671703 | 0.114777 | 0.093705 | -0.153794 | -0.113181 | 0.668047 | -0.682978 | 0.183006 | -0.061668 | 0.124052 |
| **volatile acidity** | -0.256131 | 1.000000 | -0.552496 | 0.001918 | 0.061298 | -0.010504 | 0.076470 | 0.022026 | 0.234937 | -0.260987 | -0.202288 | -0.390558 |
| **citric acid** | 0.671703 | -0.552496 | 1.000000 | 0.143577 | 0.203823 | -0.060978 | 0.035533 | 0.364947 | -0.541904 | 0.312770 | 0.109903 | 0.226373 |
| **residual sugar** | 0.114777 | 0.001918 | 0.143577 | 1.000000 | 0.055610 | 0.187049 | 0.203028 | 0.355283 | -0.085652 | 0.005527 | 0.042075 | 0.013732 |
| **chlorides** | 0.093705 | 0.061298 | 0.203823 | 0.055610 | 1.000000 | 0.005562 | 0.047400 | 0.200632 | -0.265026 | 0.371260 | -0.221141 | -0.128907 |
| **free sulfur dioxide** | -0.153794 | -0.010504 | -0.060978 | 0.187049 | 0.005562 | 1.000000 | 0.667666 | -0.021946 | 0.070377 | 0.051658 | -0.069408 | -0.050656 |
| **total sulfur dioxide** | -0.113181 | 0.076470 | 0.035533 | 0.203028 | 0.047400 | 0.667666 | 1.000000 | 0.071269 | -0.066495 | 0.042947 | -0.205654 | -0.185100 |
| **density** | 0.668047 | 0.022026 | 0.364947 | 0.355283 | 0.200632 | -0.021946 | 0.071269 | 1.000000 | -0.341699 | 0.148506 | -0.496180 | -0.174919 |
| **pH** | -0.682978 | 0.234937 | -0.541904 | -0.085652 | -0.265026 | 0.070377 | -0.066495 | -0.341699 | 1.000000 | -0.196648 | 0.205633 | -0.057731 |
| **sulphates** | 0.183006 | -0.260987 | 0.312770 | 0.005527 | 0.371260 | 0.051658 | 0.042947 | 0.148506 | -0.196648 | 1.000000 | 0.093595 | 0.251397 |
| **alcohol** | -0.061668 | -0.202288 | 0.109903 | 0.042075 | -0.221141 | -0.069408 | -0.205654 | -0.496180 | 0.205633 | 0.093595 | 1.000000 | 0.476166 |
| **quality** | 0.124052 | -0.390558 | 0.226373 | 0.013732 | -0.128907 | -0.050656 | -0.185100 | -0.174919 | -0.057731 | 0.251397 | 0.476166 | 1.000000 |




From the heap map and the correlation matrix, we can see the correlation among quality and other features in the data.

We can see that quality has the most positive correlation with alcohol, followed by sulphates and citric acid features. Whereas it has the most negative correlation with volatile acidity and total sulphur di oxide.

This means it is likely that if volatile acidity is high in the wine, then its quality is likely going to be bad (here, meaning a low value). Please note that this is just an approximation and could be used as an accessor of quality if none of the other feature information is not available.

Similarly, higher alcohol means good quality wines. Again, this is a very vague way of determining quality.