# Prediction of wine quality (sensory data) from physicochemical properties using classification model

Alexander Witedja
Fahim Tahmeed

Royal Melbourne Institute of Technology

s3641837@student.rmit.edu.au
s3680881@student.rmit.edu.au

Date of report 28/05/2019

Table of content

## Abstract

The aim of the report was to investigate which physicochemical properties can affect the quality of wine and create a model that can predict wine quality with reasonable accuracy. Exploring the correlation between each features to the quality of wine and using K nearest neighbors and decision tree classifier as classification model. The report concludes that the features free sulfur dioxide and fixed acidity should be used as input for prediction and that KNN is a better model compare to decision tree in this problem. It is recommended that for a more accurate.

## 1 Introduction

Wine was a luxurious drink in the past where only few people get to drink, nowadays wine is available in different price ranges and is being enjoyed by many people all around the world. Due to the growth of popularity in wine, the industry is in great needs of new methods that will improve making wine and selling wine which can be achieved by having a better understanding of how wine quality evaluated.

Wine quality is usually evaluated by its physicochemical properties and sensory tests by wine experts, however this process proved to be difficult because human senses provides little information because describing taste is tricky. As a result of this it's going to be hard for wine makers to decide on what they should improve or what kind of physicochemical properties should they control after receiving feedback from wine experts, in other words there's a need to understand which physicochemical attributes affect wine quality.

Firstly, the report will discuss the correlation of each physicochemical features to the wine quality by using graphs and domain knowledge. Next, the report will discuss about how well two different classification models perform prediction. Lastly, the report will discuss which features are selected by our model that'll improve the prediction of wine quality.

## 2 Methodology

We used Jupyter notebook which is an application that Python 2.7 and it's relevant libraries to preprocess the data, visualize the data and finally train classification models.

### 2.1 Red wine data

The red wine data is provided by UCI repository. The dataset contains Vinho Verde wine which refers to wines that are produced in the northernmost part of Portugal. Vinho Verde wine is well established, it's wine making history dates back to 2000 years ago making it one of the oldest wine regions in Portugal. It's popularity is growing and is considered one of the best valued wine in the market.  According to the relevant paper mentioned by UCI, the data were collected between May 2004 and February 2007 and the data were collected by an organization known as CVRVV, they're an organization with the purpose to improving quality and marketing of Vinho Verde.

2.1.1 Attribute information

In total there are 12 attributes for this dataset, for every wine there are 11 physicochemical attributes to describe it and 1 attribute describing the quality of wine according to wine experts.

1. **Fixed acidity (g(tartaric acids)/L)** [2] in wine determines the sour and tart taste of a wine. A wine lacking in fixed acidity is described as a “flat” wine. Continuous min: 4.6 max: 15.9
2. **Volatile acidity (g(acetic acids)L)** [3] a different kind of acidity. The volatile acids in wine are primarily acetic wine which is a main component for vinegar. High amounts of volatile acids produces unpleasant tastes. Continuous min: 0.1 max: 1.6
3. **Citric acid (g/L)** [4] small quantities, adds sourness, freshness. Flavours the wine. Continuous min: 0 max: 1
4. **Residual sugar (g/L)** [5] is the remaining sugar left from fermenting grapes. Gives sweetness to wine. Continuous min: 0.9 max: 15.5
5. **Chlorides (g(sodium chloride)/L)** [6] refers to amount of salt in wine. Gives saltiness in wine. Continuous min: 0.01 max: 0.61
6. **Free sulfur dioxide (mg/L)** [1] prevents microbial growth and oxidation of wine it’s an effective preservative for wine. Continuous min: 1 max: 72
7. **Total sulfur dioxide (mg/L)**. [1] Sulfur dioxide comes in two forms free and bounded, the total sulfur dioxide is the sum of these two values. Gives unpleasant taste if above a 20-30 mg/l for experience tasters. Continuous min: 6 max: 289
8. **Density (g/L)** of wine is close to that of water. Can’t find any sources that can confirm density affects quality/taste. Continuous min: 0.990 max: 1.004
9. **pH** [8] affects the amount of free sulfur dioxide needed in wine, wines with lower pH levels age slower and less chance of unwanted bacteria to form. Doesn’t affect taste. Continuous min: 2.7 max: 4.0
10. **Sulphates (g(potassium sulphate)/L)** [12]is an additive to control sulfur dioxide levels which acts as preservative for wine. Continuous min: 0.3 max: 2.0
11. **Alcohol (% vol)** [9] contributes to taste. It tastes bitter, spicy, sweet and oily. Continuous min: 8.4 max: 14.9
12. **Quality.** A wine’s quality is given by wine experts according to their taste sense. Ordinal value between 1-10.

2.2 Data pre processing

Firstly the data set is downloaded from UCI and imported into Jupyter notebook environment using the read_csv() method. Next, we checked if there are any missing values by calling the isnull().values.any(). Then, we checked for any values that are out of bounds for each column, we found that there are 2 rows containing pH value which is bigger than 4 (max value stated in the previous section) but it’s only bigger by 0.01, since this is a negligible difference and would most likely not affect the modelling we chose to ignore it. Other than the pH value everything else seems to be in order. Finally from data exploration we noticed that our data is very uneven, there are a lot of wines labelled as 5 and 6 but the data is severely lacking for every other labels. To make up for this distribution we decided to

change the wine's label into two, all wine quality ranging from 1-5 will be labelled as 0 and 6-10 will be labelled as 1.

2.3 Data exploration

The graphs that were used for exploration is created by python matplotlib library. This section will be split into two. The first one shows the frequency distribution of each columns in histogram for continuous columns and bar graph for ordinal columns. The next section shows the correlation of each attributes to quality using boxplot. Boxplot is the appropriate chart because it's suitable for finding correlation between categorical vs numerical data.

2.4 Training classification models

The classification models used for this project are imported from Python's Scikit-learn library. In this section we will discuss how parameter is tuned for each model, how features are selected and finally.

2.4.1 K nearest neighbors

For K nearest neighbors we decide to tune 3 parameters which are the number of neighbors, weight value and the minkowski power. To tune the parameters we make use of for loops, as an example we iterate from 1 to 100 number of neighbors and at the end of iteration print out the best number of neighbors alongside the score, we got 36 as the best number of neighbors. Next, we compare the score between weights='distance' and weights='uniform', we create two different loops and find the best P value for each weights.

We also implemented hill climbing which is a feature selection technique, before training the model we dropped two features that were not selected by hill climbing which are fixed acidity and free sulfur dioxide. Then we train KNN model 3 times each with different test sizes. After that, we tried training the model with different sets of features and observing how it affects the model's performance

2.4.2 Decision tree classifier

For the decision tree classifier we decided to tune 5 parameters which are max_depth, min_samples_split, max_features, min_samples_leaf and max_leaf _nodes. We used the criterion as 'gini'. To tune the parameters, we used for loop to calculate a range of values and in the end printed out the best value according to then score it received. The best parameters we found were accordingly max_depth= 16, min_samples_split= 5, min_samples_leaf = 5, max_features=2 and max_leaf_nodes=44. We implemented hill climbing algorithm and dropped features that were noise. Features that were noise were volatile acidity, free sulfer dioxide and pH. We trained the model three times with different test sizes.  we tried training the model with different sets of features and observing how it affects the model's performance.

## 3 Results

3.1 Data visualization

3.1.1 Frequency distribution of each column

- **Fig 1.1** histogram for fixed acidity
- **Fig 1.2** histogram for volatile acidity
- **Fig 1.3** histogram for citric acid
- **Fig 1.4** histogram for residual sugar
- **Fig 1.5** histogram for chlorides
- **Fig 1.6** histogram for free sulfur dioxide
- **Fig 1.7** histogram for total sulfur dioxide
- **Fig 1.8** histogram for density
- **Fig 1.9** histogram for ph
- **Fig 1.10** histogram for sulphate
- **Fig 1.11** histogram for Alcohol
- **Fig 1.12** quality distribution of quality
- **Fig 1.13** quality distribution after re label

3.1.2 Correlation of each feature to quality

- **Fig 2.1** a boxplot of fixed acidity by quality
- **Fig 2.2** a boxplot of volatile acidity by quality
- **Fig 2.3** a boxplot of citric acid by quality
- **Fig 2.4** a boxplot of residual sugar by quality
- **Fig 2.5** a boxplot of chlorides by quality
- **Fig 2.6** a boxplot of free sulfur dioxide by quality
- **Fig 2.7** a boxplot of total sulfur dioxide by quality
- **Fig 2.8** a boxplot of density by quality
- **Fig 2.9** a boxplot of pH by quality
- **Fig 2.10** a boxplot of sulphate by quality
- **Fig 2.11** a boxplot of Alcohol by quality

3.1.3 Analysis of frequency histogram

Fig 1.12 shows that there are much more wines that are labelled 5 and 6 while the other labels are very lacking. Fig 1.4 and 1.5 shows that every wine in the dataset have similar amount of chlorides and residual sugar we could expect that these attributes have no correlation to quality. Fig 1.6 and 1.7 shows that both free and total sulfur dioxide are right skewed, this is because if amount of sulfur dioxide in wine exceeds a certain amount it could give off unpleasant taste so winemakers will reduce the amount more information about sulfur dioxide is explained in section 2.1.1.

3.1.4 Analysis of boxplot correlation

There are 4 boxplot graphs that really stood out from the rest, they are volatile acidity, citric acid, sulphates and alcohol depicted by fig 2.2, 2.3, 2.10, 2.11 respectively. Fig 2.2 suggests that wine of higher quality have less volatile acidity this makes sense since high amounts of volatile acidity will give wine a vinegar like taste which would be unpleasant. Fig 2.3 suggests that wine of higher quality have more citric acid, this makes sense since citric acid gives wine 'freshness' and flavours the wine in a good way. Fig 2.10 suggests that increasing sulphates quantity in wine will improve the wine's quality, since sulphates preserves the wine then it is possible that lower quality wines aren't as well preserved. Finally fig 2.11 shows there was an increase in quality as alcohol increases but this doesn't seem to apply to wines of quality 3, 4, 5, this was unexpected because the trend only applies to higher quality wines.

The rest 7 boxplot graphs shows a lack of correlation to quality. Fig 2.1 is boxplot for fixed acidity, the fact that there's a lack of correlation shown here was unexpected because fixed acidity (tartaric acid) is closely related to citric acid both are acids of the same type and citric acid does have a slight correlation to quality. Fig 2.8 and 2.9 are boxplot for residual density and pH, it seems that the levels of these attributes are similar across all wine quality. The lack of correlation of pH and density to wine quality was expected because we couldn't find any sources that mentioned that density and pH affects wine taste. Fig 2.4 and 2.5 are boxplot for residual sugar and chlorides, wines in this dataset have similar levels of residual sugar and chlorides, the lack of correlation is unexpected since residual sugar and chlorides affect how sweet and salty a wine is these attributes should affect quality. Fig 2.6 is boxplot for free sulfur dioxide which acts as preservatives, the lack of correlation is unexpected because sulphate is also another form of preservative yet we see a correlation there. Finally Fig 2.7 is total sulfur dioxide, there is no correlation to wine quality, however it is interesting to note that most wine in the data set no matter the quality doesn't exceed 50 mg/L.

## 3.2 Classification model outcomes

### 3.2.1 K nearest neighbours

After parameter tuning and feature climbing, the best model for KNN is where k=36 weights='distance' p=1. The features selected by hill climbing were citric acid, residual sugar, chlorides, sulphates, volatile acidity, total sulfur dioxide, alcohol, ph and density. There are also 3 different sets of features that we will be comparing the performance against hill climbing. Fig 3.1 shows the classification report of the model which is trained using features that are selected by hill climbing, the micro average for precision is 0.81 which is the highest we could get to at this point for KNN. Fig 3.2 shows another classification report but this time we tried removing an additional two features that were selected by hill climbing which are pH and density under the basis that these attributes shouldn't affect quality since both attributes doesn't contribute to taste at all and quality is based on human taste sensory. This model has a similar micro average for precision to the previous model however it has a slightly lower precision for wine quality 0. This suggests that there's a tiny correlation between those features and wine quality. Fig 3.3 shows a different classification report, this time we select our own features with only the help of data exploration for this instance we selected citric acid, volatile acidity, alcohol and sulphates as our features. The micro average

of precision for this model is lower than previous models by 0.01. This shows that these 4 attributes does a decent job of determining quality.

3.2.2 Decision tree classifier

After tuning the parameters the best parameters were accordingly max_depth= 16, min_samples_split= 5, min_samples_leaf = 5, max_features=2 and max_leaf_nodes=44. The features selected by hill climbing were fixed acidity, citric acid, residual sugar, chlorides, total sulfer dioxide, density, sulphates and alcohol. The classification report put out the scores of precision, recall, f1-score. I used weighted average of precision for comparison between the 3 test sets. In fig 4.2 we can see that for 20% of test data the weighted average score was .71 while we see in the figure 4.3 and 4.4 the score is same at .70. As for feature selection we used hill climbing and just removed the features which were determined as noise and the features are mentioned above. From the decision tree we can see that alcohol becomes the determining factor when it comes to wine quality.

3.2.3 Comparison between KNN and decision tree and recommendation

KNN has a better performance compared to decision tree this is because KNN has a better micro, macro and weighted average across all three measures (precision, recall and f1 score). We applied hill climbing to both of the models and both selected different sets of features. KNN didn't select free sulfur dioxide and fixed acidity while decision tree didn't select ph volatile acidity and free sulfur dioxide. KNN hill climbing clearly selects a better feature because for decision tree we were able to come up with sets of features that'll result in a better performance.

By these two factors it is quite decisive that KNN is the better model compared to decision tree for analysing this dataset.

## Discussion

From the results of the model we learnt that not all of the features are useful when determining quality of wine. Some of the features selected were very unexpected such as pH and density but upon further inspection it is possible that these features were selected by hill climbing by coincidence because we demonstrated that removing these two features resulted in a small drop of precision, not only that it wouldn't make sense to include these features since we couldn't find any sources stating that density and pH levels could affect taste. Some other features that were selected such as chlorides, residual sugar, total sulfur dioxide also surprised us although upon further testing we found out that with only 4 features citric acid, alcohol, volatile acidity and sulphates we could create a model that performs with acceptable accuracy.

The KNN model suggests that wine makers should focus on controlling these 4 attributes if they wish to create good wine although that doesn't mean that attributes such as chlorides, residual sugar, total sulfur dioxide should be completely neglected since as already explained before total sulfur dioxide gives unpleasant taste when above a certain

threshold which varies between tasters while chlorides and residual sugar contribute to the sweetness and saltiness of wine which according to [7] contributes to evaluating wine.

## Conclusion

In conclusion, we found out that there are 4 most significant physicochemical attributes that contributes to how wine is evaluated by the experts which are citric acid, volatile acidity, alcohol and sulphates. Wine makers must focus to control these 4 different attributes but not forget about other relevant attributes that could possibly affect wine quality but not shown by our study which are residual sugar, chlorides and total sulfur dioxide.

There are still knowledge gaps, the dataset doesn't have all the attributes that can be used to evaluate a wine. Wine tasting remains a very difficult field to fully understand and the physicochemical representation of wine shown in this study is lacking in many ways. For instance according to [11] wine are judged by so many factors like sight, smell and not just taste. Another example in [10] shows a long list of wine components that are untouched by our dataset. It is possible to improve the study even further by including chemical components in [10] when collecting data and collecting wines from different wine makers.

## Visualisations and figures

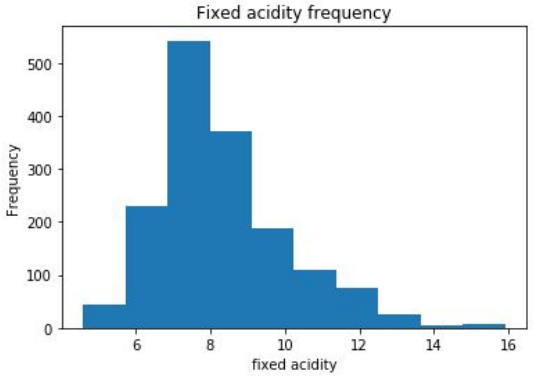



Fig 1.1
Histogram of fixed acidity

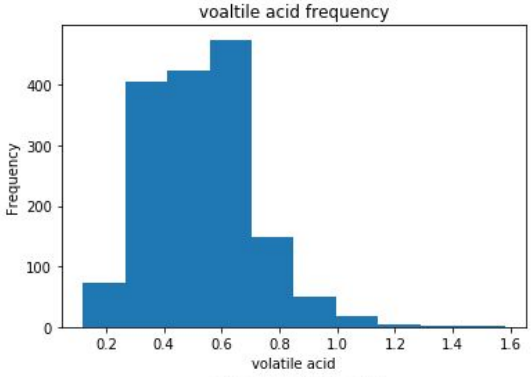



Fig 1.2
Histogram of volatile acidity

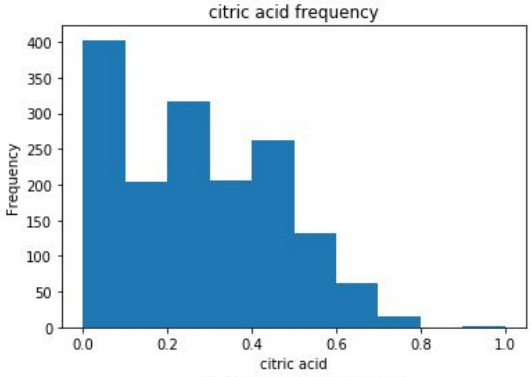



Fig 1.3
Histogram of citric acid

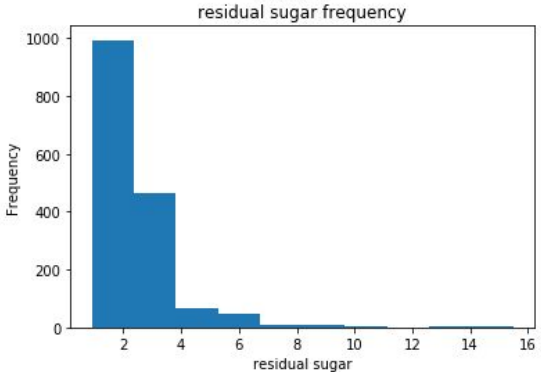



Fig 1.4
Histogram of residual sugar

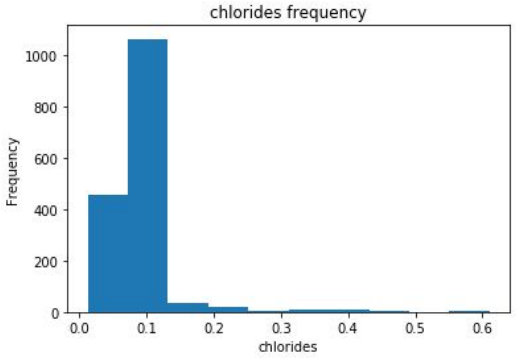



Fig 1.5
Histogram of chlorides

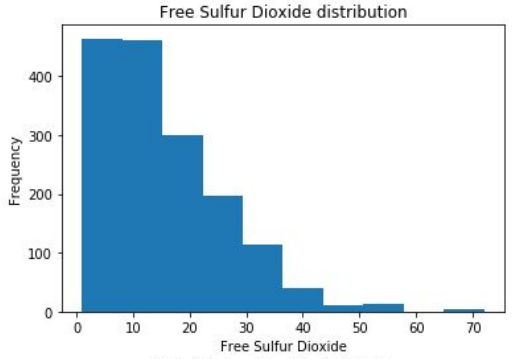



Fig 1.6
Histogram of free sulfur dioxide

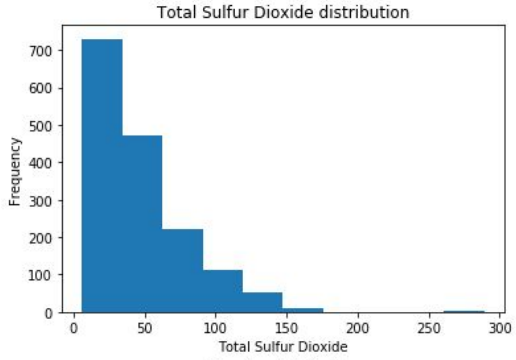



Fig 1.7
Histogram of total sulfur dioxide

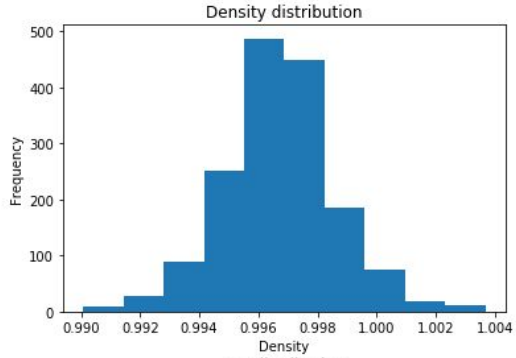



Fig 1.8
Histogram of density

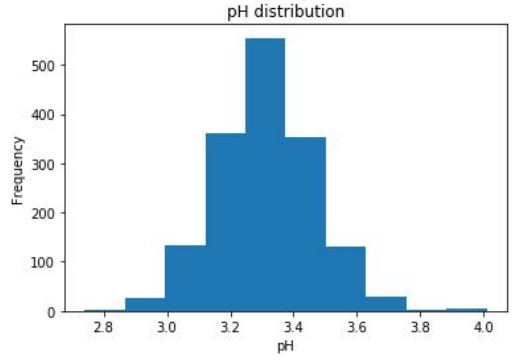



Fig 1.9
Histogram of pH

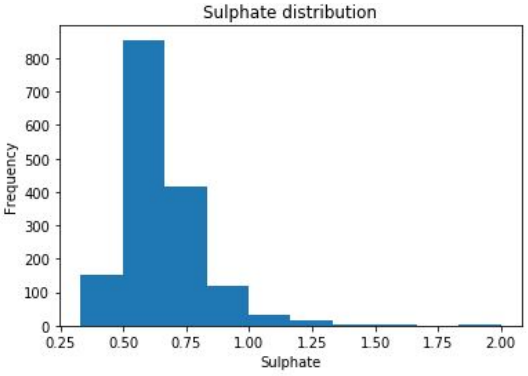



Fig 1.10
Histogram of sulphate

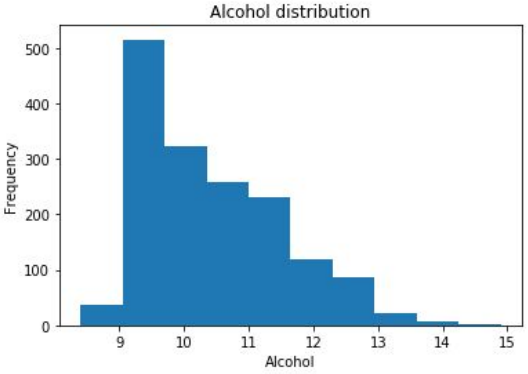



Fig 1.11
Histogram of alcohol

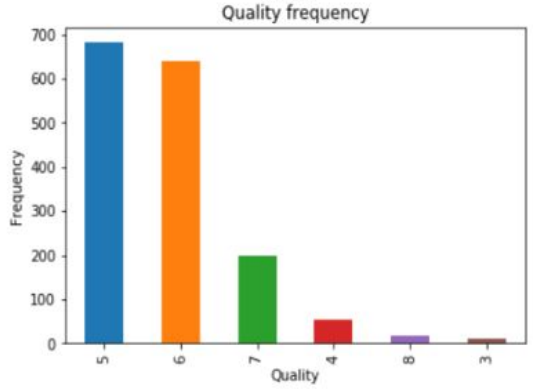



Fig 1.12
Bar chart of quality frequency

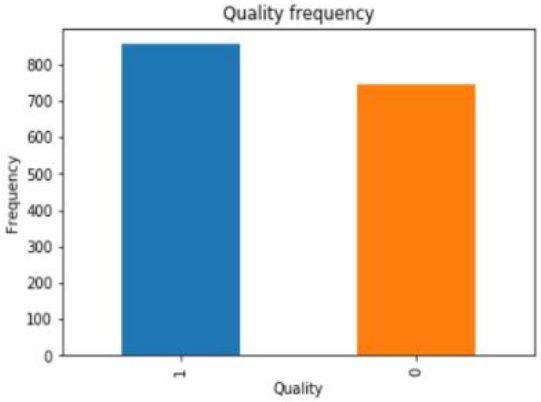



Fig 1.13
Bar chart of re labelled quality

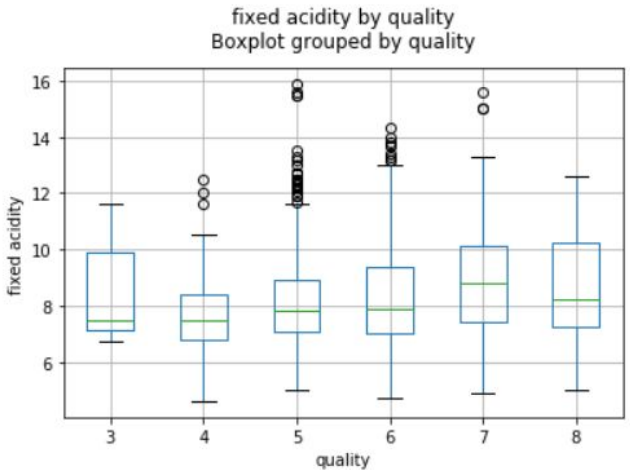



Fig 2.1
boxplot of fixed acidity by quality

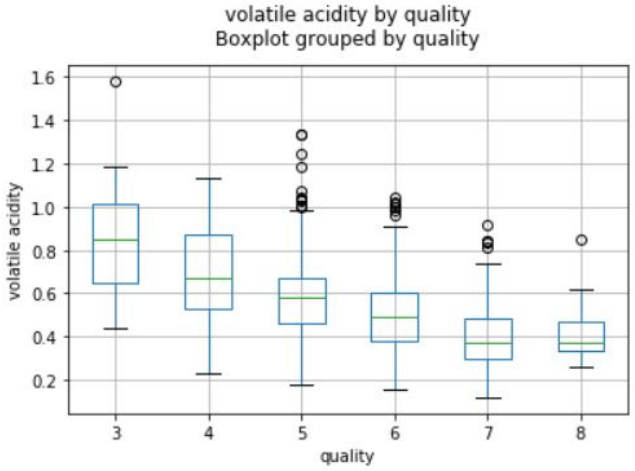



Fig 2.2
boxplot of volatile acidity by quality

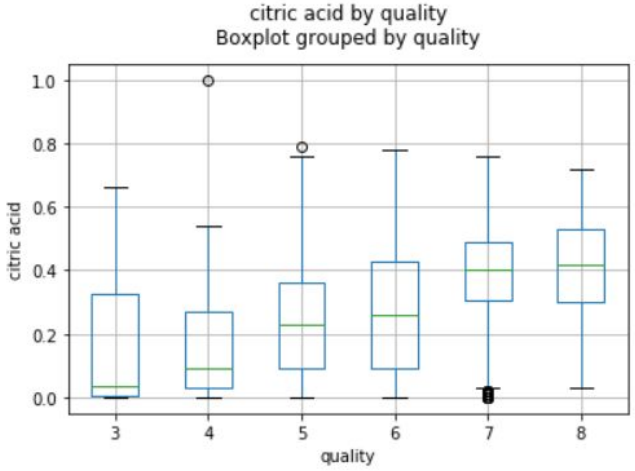



Fig 2.3
boxplot of citric acid by quality

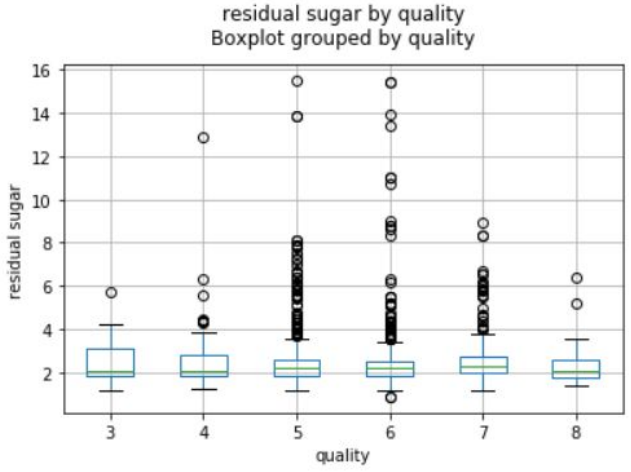



Fig 2.4
boxplot of residual sugar by quality

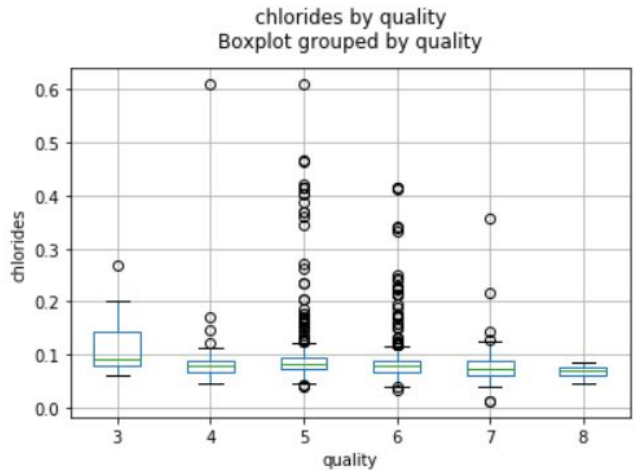



Fig 2.5
boxplot of
chlorides by quality

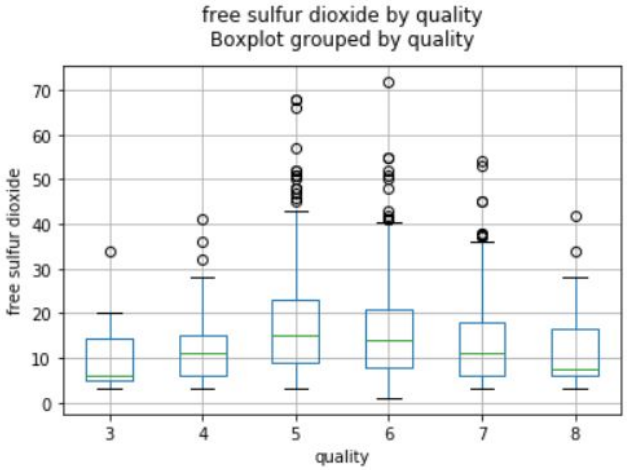



Fig 2.6
boxplot of free
sulfur dioxide by
quality

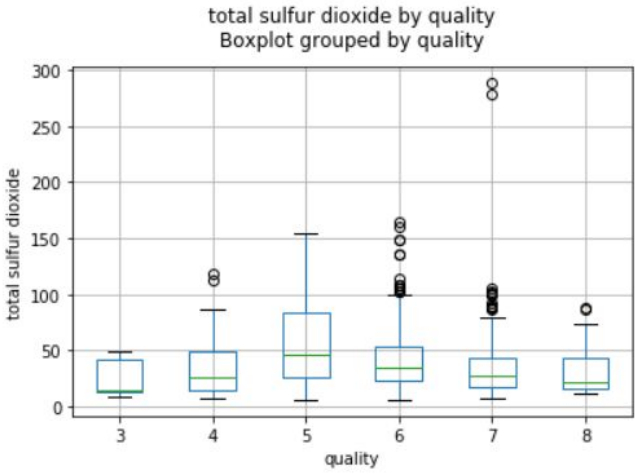



Fig 2.7
boxplot of total
sulfur dioxide by
quality

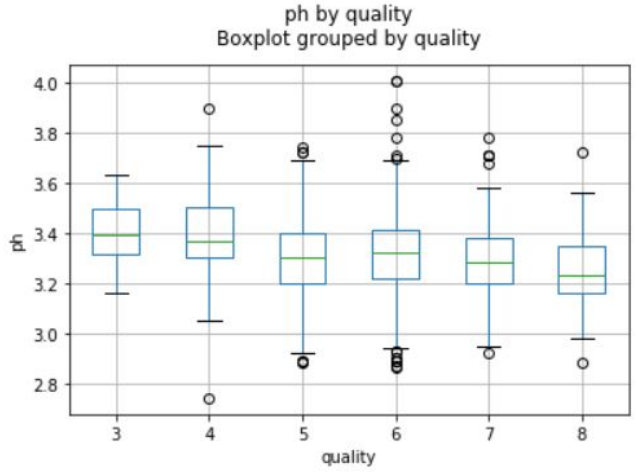



Fig 2.9
boxplot of pH by quality

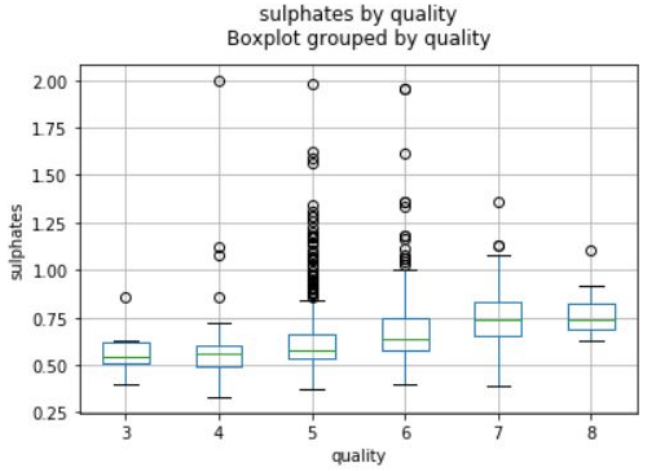



Fig 2.10
boxplot of sulphate by quality

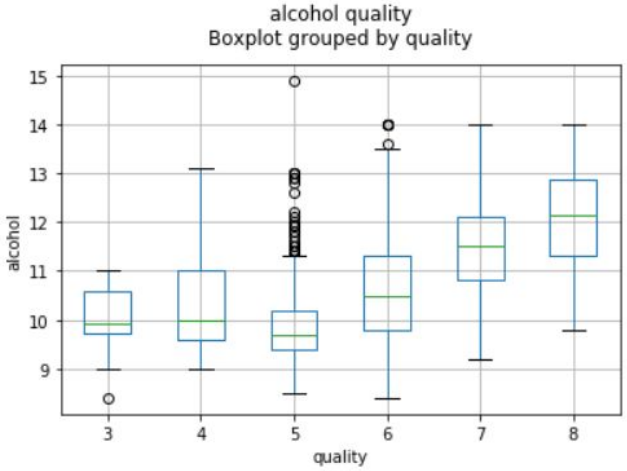



Fig 2.11
boxplot of alcohol by quality

```
[[127  27]
 [ 33 133]]
              precision    recall  f1-score   support

           0       0.79      0.82      0.81       154
           1       0.83      0.80      0.82       166

   micro avg       0.81      0.81      0.81       320
   macro avg       0.81      0.81      0.81       320
weighted avg       0.81      0.81      0.81       320
```

Fig 3.1
Features selected: Hill climbing
citric acid, residual sugar, chlorides, sulphates, volatile acidity, total sulfur dioxide, ph and density

```
[[127  27]
 [ 35 131]]
              precision    recall  f1-score   support

           0       0.78      0.82      0.80       154
           1       0.83      0.79      0.81       166

   micro avg       0.81      0.81      0.81       320
   macro avg       0.81      0.81      0.81       320
weighted avg       0.81      0.81      0.81       320
```

Fig 3.2 Hill climbing without ph and density
Features selected:
citric acid, residual sugar, chlorides, sulphates, volatile acidity, total sulfur dioxide

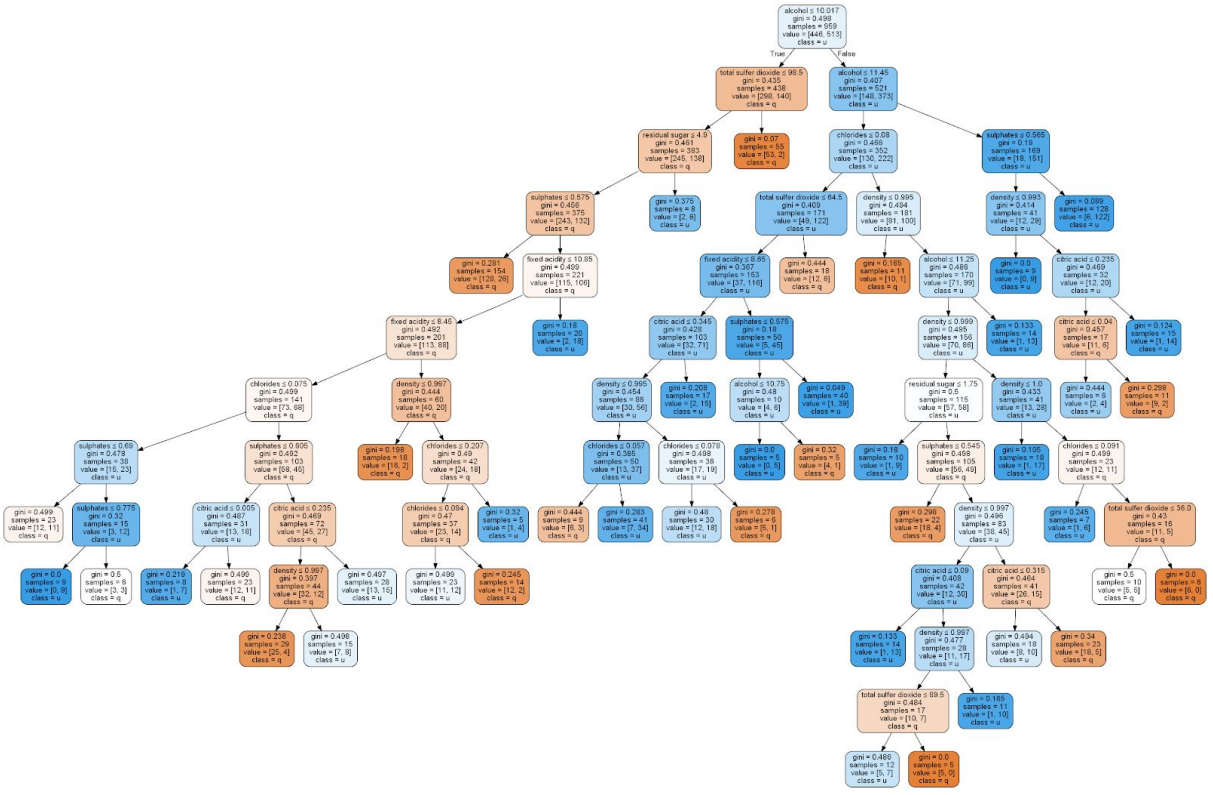
```
[[122  32]
 [ 33 133]]
              precision    recall  f1-score   support

           0       0.79      0.79      0.79       154
           1       0.81      0.80      0.80       166

   micro avg       0.80      0.80      0.80       320
   macro avg       0.80      0.80      0.80       320
weighted avg       0.80      0.80      0.80       320
```

Fig 3.3
Features selected: Data exploration
citric acid, alcohol, sulphates, volatile acidity

Figure 4.1 decision tree classifier (submitted png in zip)



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.64 | 0.81 | 0.72 | 154 |
| 1 | 0.77 | 0.58 | 0.66 | 166 |
| micro avg | 0.69 | 0.69 | 0.69 | 320 |
| macro avg | 0.70 | 0.70 | 0.69 | 320 |
| weighted avg | 0.71 | 0.69 | 0.69 | 320 |

Fig 4.2(Classification report 20%)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.69 | 0.65 | 0.67 | 298 |
| 1 | 0.71 | 0.74 | 0.73 | 342 |
| micro avg | 0.70 | 0.70 | 0.70 | 640 |
| macro avg | 0.70 | 0.70 | 0.70 | 640 |
| weighted avg | 0.70 | 0.70 | 0.70 | 640 |

Fig 4.3(Classification report 40%)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.68 | 0.69 | 0.69 | 375 |
| 1 | 0.72 | 0.72 | 0.72 | 425 |
| micro avg | 0.70 | 0.70 | 0.70 | 800 |
| macro avg | 0.70 | 0.70 | 0.70 | 800 |
| weighted avg | 0.70 | 0.70 | 0.70 | 800 |

Fig 4.4(Classification report 50%)

**References :**


1. http://www.morethanorganic.com/sulphur-in-the-bottle (total and free sulfur dioxide)
2. https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity (fixed acidity)
3. https://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity (volatile acidity)
4. https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid (citric acid)
5. https://whicherridge.com.au/what-is-residual-sugar-in-wine/ (residual sugar)
6. https://www.aromadictionary.com/articles/salt_article.html (chlorides)
7. http://thirstmag.com/wine/win-features/How-wines-are-judged (evaluating wines)
8. https://thegrapevinemagazine.net/article/the-importance-of-ph-in-winemaking/ (pH levels)
9. https://winefolly.com/review/wine-characteristics/ (alcohol)
10. https://waterhouse.ucdavis.edu/whats-in-wine/ (whats in wine)
11. https://www.winemag.com/2015/08/25/how-to-taste-wine/ (evaluating wine)
12. https://www.thekitchn.com/the-truth-about-sulfites-in-wine-myths-of-red-wine-headaches-100878 (sulphates)