Wine Quality Classification

**Bhushan Suryawanshi**[**DSC680-T302 Applied Data Science (2215-1)**](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_500232_1)

**Abstract:**

Today, if you walk into a wine shop there are about 100 different wine bottles on display. How to choose the one with good quality. To get quality wine either you need to be an expert wine taster or you need to follow one of them to get the list of good quality wine.   
Having data in hand and using that data for knowing the quality of wine will help democratize this space. Hence, conducted a study learning different chemicals available in wine to decide the quality of the wine.   
The study involves 1600 different observations across 12 features. The dataset was made available at both Kaggle and UCI machine learning repository. Having limited knowledge of different chemicals other than alcohol made it challenging. However, techniques such as finding feature importance and feature selection helped in identifying top influencing features for the model. Also, the target feature had imbalanced data and had to use the re-sampling technique to have uniform sampling across different categories of wine.   
Since this is a multi-class classification problem, used algorithms like logistic regression, KNN and also studied the impact of using neural network. The study shows that KNN worked the best with N=5 and balanced classes giving an accuracy of 82% on the test data.

**Keywords:** KNN, Logistic Regression, Neural Network, Wine, Quality, Deep Learning, Classification.

**Introduction/Background**

Wine drinkers are increasing day by day and hence you see there are more wineries around. The main reason behind the consumption of wine is, it is adapted in many cultures at different events. Wines are available in different colors, red, white, pale red, etc. Color is also one of the quality-defying factors in wine. Some of the other factors that define wine quality are weather, the soil in which grapes are cultivated, the location of the winery, at what temperature wine is stored and fermented, etc.

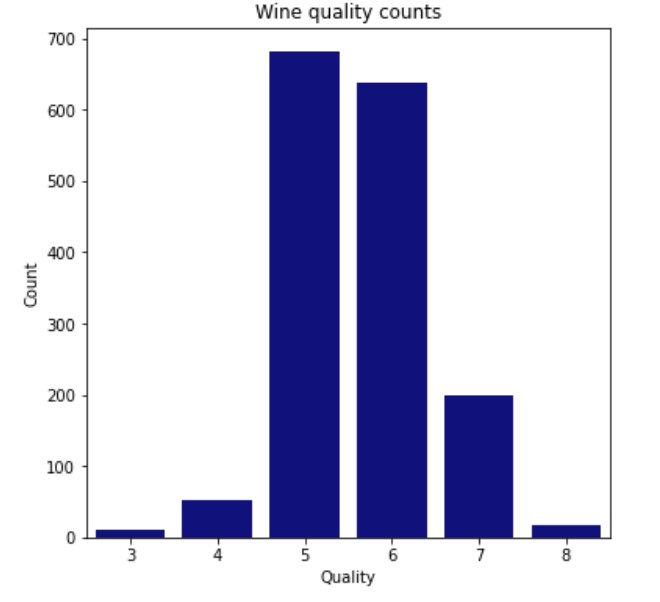
Similar to production and environment factor there are chemical factors those affect the quality of the wine. pH value is one of the factors in identifying the quality of the wine. Another well-known factor is the amount of alcohol. So if these chemical contents are known then finding the quality of wine can be automated.   
With the advancement in statistical analysis, machine learning techniques can be used to find the quality of the wine. Today only expert wine tasters can define the quality of wine and the rest follow blindly. If this tool can be generated statistically this big wine community can understand the quality of wine they want to drink tomorrow. This will not only help end consumers but also wine shops that can sell quality wines.

**Approach:**

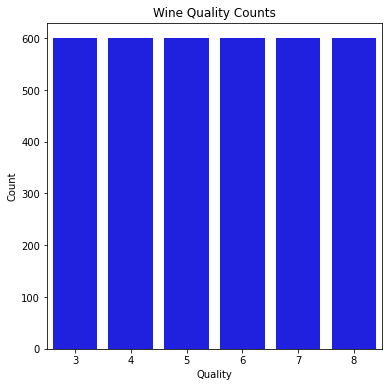
In this study, the standard CRISP-DM steps are followed to explore and find insights from the data provided. Following are details on each step performed -

Data Preparation -

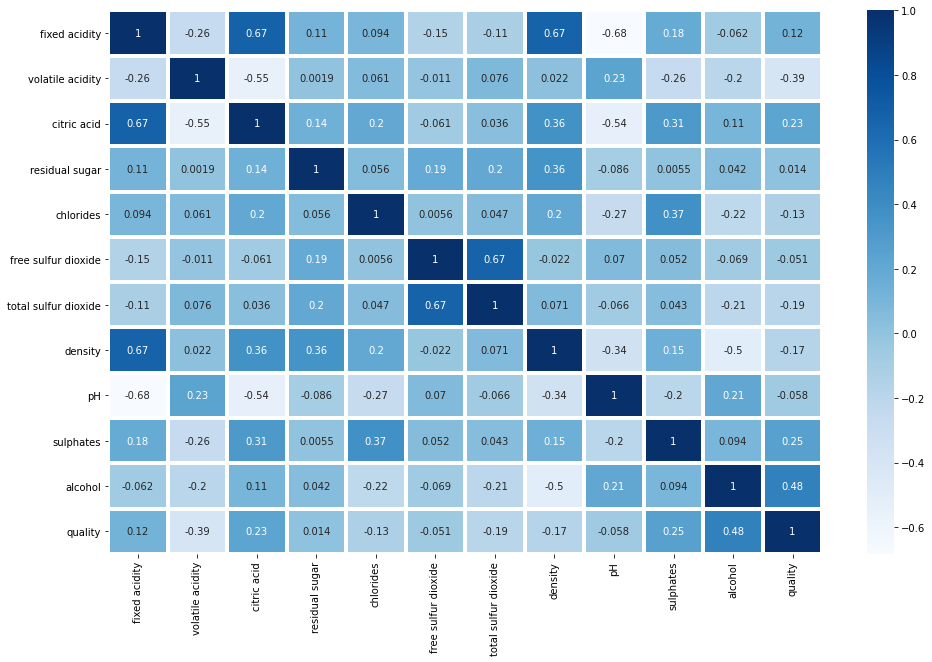
Data is provided in the CSV format hence data collection did not require much effort. The data has no missing or null values. The histogram of variables shows that some of the variables like pH, fixed\_acidity, and density are close to the normal distribution and others like citric acid are bimodal whereas residual sugar, free sulfur dioxide, chlorides are skewed. Since the dataset contains a smaller sample set outliers are not removed.   
When plotted the histogram of target feature - “quality” it was observed that the values are spread across 6 different categories from 3 to 8. Also, it can be seen from the graph that those are not uniformly distributed leaving an unbalanced dataset. However, kept the dataset as it is for comparing the difference between results from an unbalanced and balanced dataset.



For balancing dataset used re-sampling method “resample” from sklearn.utils. This helped to have another dataset with balanced categories.

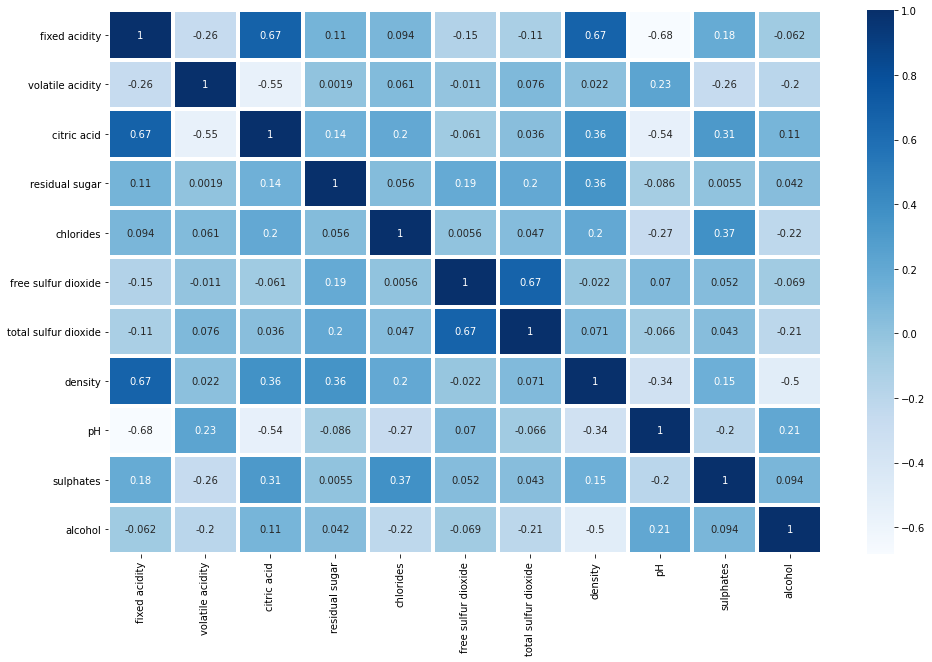


Then looked at the correlation of the data and it shows high collinearity between independent variables.



Feature transformation -

StandarScalar from sklearn.preprocessing is used to transform highly correlated variables. This will help to transform feature values in such a way that 68% of values will lie between 0 and 1. Here is the correlation after the transformation of data.



Kept the standard scalar transformation for unbalanced class dataset. Whereas for the dataset with balanced data for different quality categories are normalized using log-transformation to normally distribute all features.

Feature selection -

When you have multiple features it is important to choose those impact more and useful for the model. This will avoid “Garbage In Garbage Out” scenario. Feature selection will help to train the model faster with lower complexity.

sklearn.feature\_selection provides a function mutual\_info\_classif which helps to identify the importance of the feature. Higher the value for the feature means higher dependency. Following are the values after features are fit using mutual\_info\_classif.

alcohol 0.174967

volatile acidity 0.131487

sulphates 0.103866

density 0.077403

total sulfur dioxide 0.069870

citric acid 0.060860

fixed acidity 0.042183

chlorides 0.034264

pH 0.029401

residual sugar 0.027817

free sulfur dioxide 0.021509

Used mutual\_info\_classif with function from feature selection technique called SelectKBest to get top 7 independent features from the dataframe.

Model Building -

During model building, multiple classification models are used for comparison. A common function is written for model training and performance metrics to allow the reusability of code.   
The dataset is split into 70% training and 30% testing data. Use multi-class classification models like logistic regression, K-Nearest Neighbor (KNN), RandomForestClassifier, DecisionTreeClassifier, etc. The models are trained both with the balanced and imbalanced datasets. Also, did hyperparameter tuning for the selected best model depending on their accuracy?   
The neural network is another area that can be used for multi-class classification. Used sequential model to train with both balanced and imbalanced dataset.   
Following is a comparison of results found after training and hyperparameter tuning.   
Model, Training Accuracy, Precision, Recall, F1-Score

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Dataset Type | Training Accuracy | Testing Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | Unbalanced | 60 | 59 | 56 | 59 | 56 |
| KNN | Unbalanced | 70 | 62 | 60 | 62 | 60 |
| Random Forest | Unbalanced | 100 | 67 | 64 | 67 | 65 |
| Decision Tree | Unbalanced | 100 | 57 | 57 | 57 | 57 |
| Neural Network | Unbalanced | 70 | 62 | - | - | - |
| KNN | Balanced | 100 | 85 | 84 | 85 | 84 |
| Random Forest | Balanced | 100 | 66 | 66 | 66 | 64 |
| Decision Tree | Balanced | 100 | 58 | 58 | 58 | 58 |
| Neural Network | Balanced | 71 | 64 | - | - | - |

Based on the results above, it can be seen that for the problem in hand KNN is working the best. Again this is a wine quality problem and someone is really interested in knowing a set of wines in the category. Hence KNN works best to cluster wines of similar quality. KNN worked better on both balanced and imbalanced data with a testing accuracy of 85% and 62% respectively.

**CONCLUSION**

Wine consumption is increasing day by day. As per the Wine Institute in the United States, 2.99 gallons of wine is consumed per person which accounts for 966 million gallons of wine every year. People who drink wine frequently prefer quality wine. Knowing the quality of wine is dependent on human tasters. The wine quality is decided by the sensory mechanism by a bunch of experts.   
 Having a statistical tool for knowing the quality of wine will not only help sellers but also the consumer of wine. Having this tool handy will ensure good quality wine demand. The increased demand for wine will definitely help wineries but also farmers that grow raw materials for the wine. This shows the outcome of this study indirectly contributes to the economy.   
 The rich dataset provided at UCI and Kaggle repository helped to derive insights. Different multi-class classification models are trained and compared for their outcomes. Also, sequential neural networks are attempted with the given dataset. The multiple iterations of training and hyperparameter tuning shown great results. It has been observed that K-Nearest Neighbor(KNN) performed much better with 85% and 62% testing accuracy with balanced and imbalanced classes respectively.

**ACKNOWLEDGMENT**

I would like to acknowledge my family for allowing me with much needed time for study. My classmate were great with the feedback required to improve on the project. Finally I would like to thank Prof. Fadi Alsaleem for his guidance throughout the course and project work.

**References -**

1. Yogesh Gupta, Selection of important features and predicting wine quality using machine learning techniques, 2018. (https://www.sciencedirect.com/science/article/pii/S1877050917328053)
2. Nikita Sharma, Quality Prediction of Red Wine based on Different Feature Sets Using Machine Learning Techniques, 2018. (https://www.ijsr.net/archive/v9i7/SR20718002904.pdf)
3. [US Wine Consumption | Wine Institute](https://wineinstitute.org/our-industry/statistics/us-wine-consumption)
4. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

# The Key Chemicals in Red Wine – Colour, Flavour, and Potential Health Benefits (https://www.compoundchem.com/2014/05/28/redwinechemicals/)

1. Wikipedia, Acids in Wine (<https://en.wikipedia.org/wiki/Acids_in_wine>)
2. [10 Ingredients You Probably Didn't Know Were in Your Wine (winecoolerdirect.com)](https://learn.winecoolerdirect.com/surprising-wine-ingredients/)
3. [Understanding Acidity in Wine | Wine Folly](https://winefolly.com/deep-dive/understanding-acidity-in-wine/)
4. [How Alcohol Content Affects Wine | Ayza (ayzanyc.com)](https://ayzanyc.com/blog/how-alcohol-content-affects-wine/)
5. [How Much Sugar in Wine | A Wine Sugar Content Guide (binwise.com)](https://home.binwise.com/blog/how-much-sugar-in-wine)