**Project：Analysis of Red Wine Quality Data Set**



Date**：30/11/2023**

1. **Introduction**

We used the data source：<https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>. Our goal was to classify whether a wine bottle is of high or low quality based on the following 11 attributes: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. All attributes are numerical and continuous, and the quality is categorical.

The dataset had 1599 records with no missing values, but 240 records were duplicated and therefore removed.

Quality had values of between 3 and 8, but due to class imbalance we grouped it into low (3-5) and high (6-8).

1. **Methodology and Results**

**Preliminary Data Analysis**

We started by importing and cleaning the data before starting preliminary exploration.

The use of *ggpairs* suggested there are some outliers and a possible class imbalance. As there are 11 attributes and over 1000 records, it was difficult to get much more. *Summary* also suggested there were outliers. Using *plot3d* to show pH, alcohol, density, and salts, with different colors to the different levels, also suggested class imbalance. This suggested we might need to use fewer level, rather than staying with the detailed classification of 6 levels.

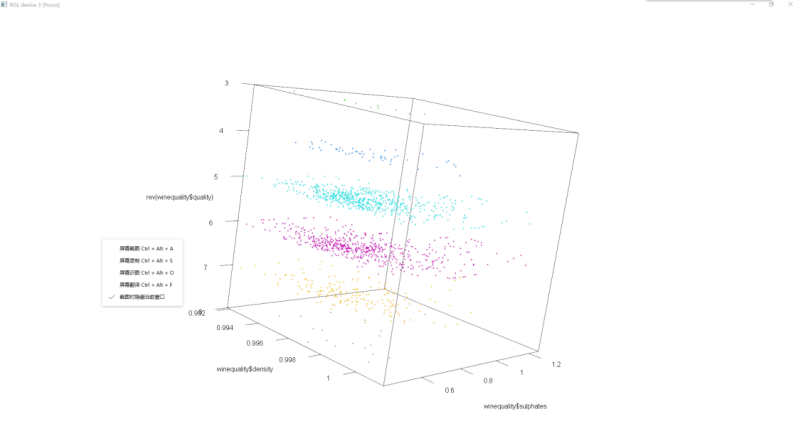
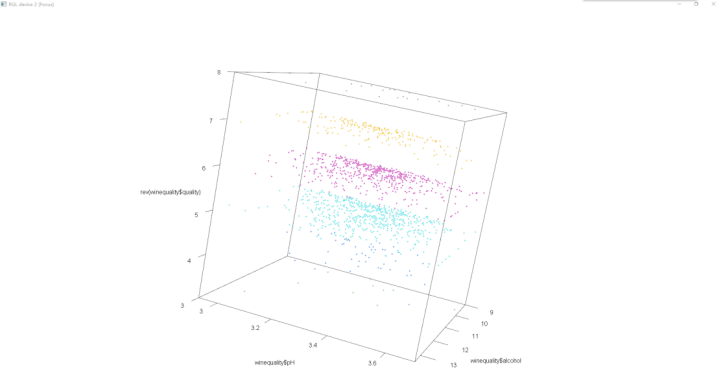
The use of histograms showed both outliers and that the data was mostly well spread across the range. A few exceptions were residual sugar, chlorides, and sulphates that were more centered. It also showed class imbalanced, where most data were of levels 5 and 6, while levels 3, 4, and 8 had fewer than 100 records. Finally, the use of boxplots confirmed the data had many outliers.

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

To resolve the outliers’ issue, we first limited the data to 1.5 times the IQR but that resulted in predictions that were less accurate than the original data. The use of R to set a new value to the outliers as equal to the 99 percentile was also less accurate. We therefore used python to remove the duplicates and set a new maximum value equal to the 99 percentile which improved the results. The Python code and new csv file are attached.

Correlation matrix revealed some correlation between some of the attributes. For example, between fixed acidity and citric acid, density, and pH (0.67, 0.67, and -0.69 respectively) and between free sulfur dioxide and total sulfur dioxide (0.67).

A graph with numbers and symbols

Description automatically generated

**Data Mining**

We started by fixing the class imbalanced by grouping levels 3, 4, and 5 into “low”, and levels 6, 7, and 8 into “high”. This enabled both classes to have enough records for proper analysis. “high” accounted for 52.91% and “low” to 47.09% (719 and 640 records respectively). We then scaled the data as there were variations in size between the different attributes. For example, “total sulfur dioxide” ranged from 6 to 113 while “citric acid” ranged from 0 to 0.7042.

We split the data into 70% training data with 504 “high” records and 448 “low” records (52.94% and 47.06% respectively), and 30% training data with 215 “high” records and 192 “low” records (52.83% and 47.17% respectively). Best practices range from 80/20 to 70/30 splits, for our data 70/30 gave the best results (need reference).

We trained the training data on the following four models and compared the results: Conditional Inference Tree (Decision Tree), K-Nearest Neighbors, Naïve Bayes Classifiers, and Linear Support Vector Machines. These models fit the classification task we’re trying to solve, and each have advantages and disadvantages.

Cross validation was created using a fixed, 10-fold, sampling scheme with the goal of getting a reliable estimate of the model’s performance.

The comparison found the KNN had the highest accuracy and Kappa out of all the models.

A screenshot of a computer

Description automatically generated A diagram of a graph

Description automatically generated with medium confidence

z

KNN tried k ranges from 1 to 10, best results were when k=1, as shown below.

A screenshot of a computer

Description automatically generated

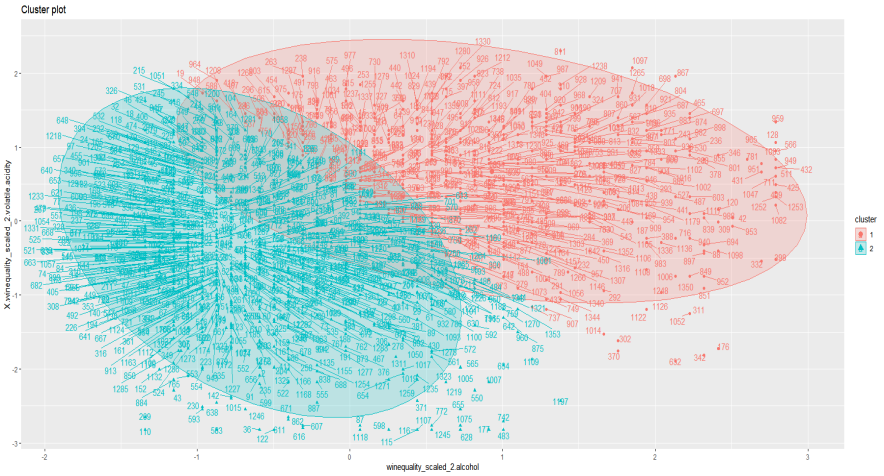
However, when using KNN on the test data accuracy was reduced to 0.7592 with a Kappa of 0.5158, even though we fixed the imbalance in the training and test data.

We further explored the two-class data to see if we could reduce the number of attributes to improve the results, as KNN preforms better with lower dimensionality. We trained the same models again with alcohol and volatile acidity, with alcohol, volatile acidity, and sulfates, and after removing pH, citric acid, density, and total sulfur dioxide. The first three attributes appeared to be somewhat well separated in parallel coordinates plot, and the last option had attributes with higher correlation removed. The best result was for the option of alcohol and volatile acidity, with an accuracy of 0.742 and a Kappa of 0.4817.

A graph showing a variety of lines

Description automatically generated with medium confidence

If we look at the below graph, we can see that the two are mostly well separated. It is possible that some of the other attributes assist in correctly identifying the overlapping area, which is why the full model is slightly more accuracy with a higher Kappa.



Mean, median, variance

1. **Discussion and Conclusion**

The data had very few records for wines ranking 3, 4, and 8. While it is reasonable to assume that most whines are “ok” - not extremely good nor bad, it does present difficulties when trying to classify. Adding more wines to these levels could improve our classifications abilities, prediction capabilities, and recommendations.

With the current data, we can suggest that wines with higher content of alcohol and lower levels of volatile acidity, appear to be of better quality. This is supported by Ribéreau-Gayon et al. (2006), which indicate that the flavor begins to deteriorate at volatile acidity levels higher than 0.72 g/L (as brough in AWRI, 2018). Higher alcohol levels are generally considered positive as it brings out flavors. However, balance is important and it’s possible that this is why the outliers of wines of level 5, that have similar alcohol content to higher levels, don’t belong to a higher quality group (King et al, 2013).

1. **References**

AWRI. (2018). Volatile Acidity. *Grapegrower & Winemaker*, 648. <https://www.awri.com.au/wp-content/uploads/2018/03/s1982.pdf>.

King, E.S., Dunn, R.L., & Heymann, H. (2013). The influence of alcohol on the sensory perception of red wines, *Food Quality and Preference*, Volume 28, Issue 1, 235-243. <https://www.sciencedirect.com/science/article/abs/pii/S0950329312001796>.

1. **Veriguide Report**

### Appendix A: Need reference

### Input variables (based on physicochemical tests):

### Fixed acidity：most acids involved with wine or fixed or nonvolatile (do not evaporate readily).

### Volatile acidity：the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.

### Citric acid：found in small quantities, citric acid can add 'freshness' and flavor to wines.

### Residual sugar：the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.

### Chlorides：the amount of salt in the wine.

### Free sulfur dioxide：the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.

### Total sulfur dioxide：amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.

### Density：the density of water is close to that of water depending on the percent alcohol and sugar content.

### pH：describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.

### Sulphates：a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.

### Alcohol

### Output variable (based on sensory data):

### Quality (score between 0 and 10)