# **Predicting Vinho Verde wine quality using its chemical composition**

By Mohammad Bilal Iqbal

**Introduction**

Vinho Verde is a wine from Portugal and has been protected since 1908, which the Vinho Verde region was officially demarcated. The name is a reference to Vinho Verde being a young wine, typically released between 3 and 6 months of the grapes being harvested. It is meant to be consumed soon after bottling.

The dataset comes from P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, who kindly donated it to the UCI Machine Learning Repository. The dataset has already been separated into two CSV files, for red and white wines. This is essential as the attributes that make up a good white wine may be significantly different in values than those that make up a quality red wine. For this analysis, we will be focusing on the white wine dataset for two reasons: white wines make up 86% of all Vinho Verde wines and the white wine dataset is significantly bigger than the red wine dataset (4,898 observations vs 1,599 observations).

**Exploration**

This particular dataset was cleaned before it was donated to the UCI Machine Learning repository and basic exploration reveals that it does not have any missing values. Thus it will not require significant preparation ahead of testing.

Twelve attributes make up the dataset, 11 of them being input variables and one being the target variable. They are listed below:

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Role |
| Fixed Acidity | Real Number | Input |
| Volatile | Real Number | Input |
| Citric Acid | Real Number | Input |
| Residual Sugar | Real Number | Input |
| Chlorides | Real Number | Input |
| Free Sulphur Dioxide | Real Number | Input |
| Total Sulphur Dioxide | Real Number | Input |
| Density | Real Number | Input |
| pH Level | Real Number | Input |
| Sulphates | Real Number | Input |
| Alcohol | Real Number | Input |
| Quality | Real Number | Target |

Figure 1 below shows a snapshot of the data using R’s built-in summary function:

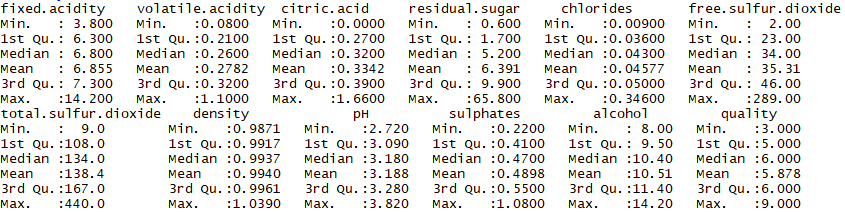


Figure : White Wine Data Snapshot

Immediately it is clear that the variables vary in magnitudes, which means that the data will need to be normalized for classification techniques that work on the magnitude of measurements, such as KNN.

The boxplot for variables in the white wine dataset is given in Figure 2. We see that some of the variables have significant number of outliers. They will not be removed or processed due to the assumption that the dataset was already significantly tested for the original experiment and outliers do not constitute mistakes in encoding.

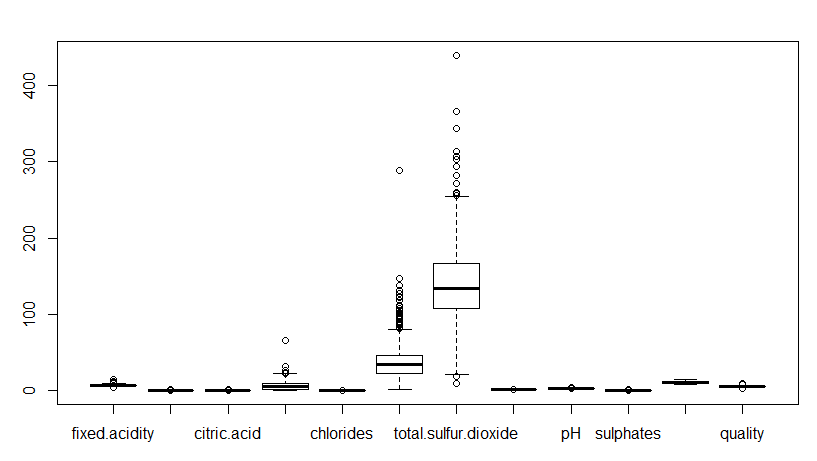


Figure : Box Plot for the Distribution of Variables in the Wine Dataset

Even though wine quality is rated from 1-10, checking the dataset reveals that the wines observed were ranked from 3-9. It is clear from Figure 3 that the observed samples are not uniformly distributed. There are far fewer wines on the fringes (only five were rated 9 and 20 were rated 3) than in the center. Thus it may be harder for classification data mining techniques to model characteristics of wines that are either very poor or high in quality.



Figure : Distribution of Wine Quality

**Modelling**

For each classification attempt, the dataset has been divided into a 75-25 split, i.e. 75% of data for training and the remaining for testing of the model.

**ctree**

Since this is a classification problem, we will start off with the most basic classification techniques i.e. trees. We begin with Caret’s ctree implementation. We get an accuracy of 54.01% on the test dataset. The confusion matrix is shown in Figure 4. Even though the accuracy is much better

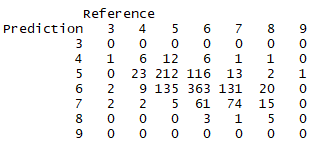


Figure : Confusion Matrix for Ctree

Even though the Accuracy of the model is much better than random assignment would suggest (100/7=14.28%), the model only performs reasonably well on wines of Quality 5 and 6, with a recall rate of 58.24% and 66.12%, respectively.

The distribution of classification as a result of this model is shown in Figure 5.



Figure : Distribution of Classification for Ctree

Overall, as it is, these results are not very useful to get meaningful insight into the quality of a wine using its chemical composition.

**rpart**

Rpart performs worse than ctree with a 50.74% accuracy on the test dataset. The model performs reasonably well on wines of Quality 5 and 6, showing an improvement in recall rates for wines of Quality 5 (66.48 and 62.30%, respectively). However, like the Ctree model, these results are not very useful to get meaningful insight into the quality of a wine using its chemical composition.

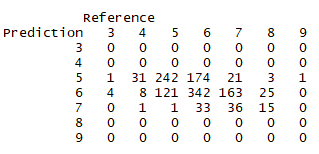


Figure : Confusion Matrix for Rpart

As seen in Figure 7, Rpart does not perform at all for wines outside of the middle pack, with zero predictions for wines of Quality of 3,4, 8 and 9.



Figure : Distribution of Classification for Rpart

**KNN**

Since KNN uses a distance metric, we would need to normalize the dataset before training. This is because the predictive variables have different scales and would influence the algorithm if they are not made comparable. We will use min-max normalization on the dataset combined with scaling, as this combination gave better predictive performance. The model picks k=31 due to a better Kappa value compared to other values of k.

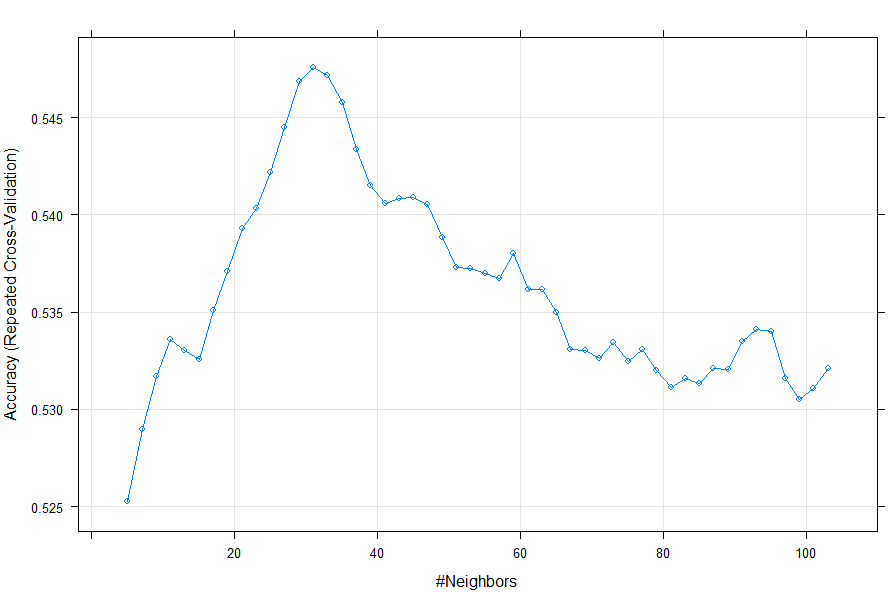


Figure : KNN Accuracy across different values of K

The overall accuracy for the KNN model is 55.4%, which is slightly better than the results achieved by ctree. This model performs well on Wines of Quality 6, getting a recall rate of 70.67%. The recall rate for Wines Quality 5 is slightly worse than ctree (57.14% vs 58.24%) and much worse than rpart (66.48%). The recall rate for Wines of Quality 7 using KNN see some improvements at 35.45% (versus 33.63% for ctree and an abysmal 16.36% for rpart).

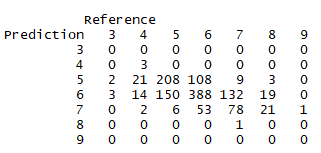


Figure : Confusion Matrix for KNN

**Random Forest**

Random Forest generates the best results of all the algorithms tested so far. Unprocessed data was fed to the model, as Random Forest does not benefit from normalization. Using Caret and setting tunelength to 5, the algorithm choose mtry=2 for the final model. This results in a significant improvement for the test dataset at 68.17%.

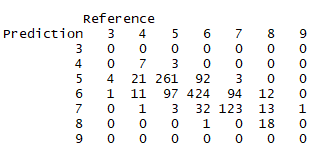


Figure : Confusion Matrix for Random Forest



Figure : Class Recall for Random Forest

Looking at figures 10 and 11, we see that Random Forest appears to be approaching usable levels for Classes 5 and 6, which Recall rates of 71.70% and 77.23%.

Gradient Boosted Trees were also used on the dataset but performed worse than Random Forest and have been omitted from this analysis.

Even though Random Forest generates some usable results for wines of quality 5 and 6, its results for other qualities are not very promising and usable. Classification predictions on Wine datasets, thus, will generate results that may be significantly off at time and thus not very helpful to any supermarket trying to make informed decisions based on the chemical compositions of the wines.

**Discussion**

A closer look at Figure 10, however, does indicate that if we divide the dataset into two categories, i.e. wines with below average quality and wines with above average quality, we may be able to achieve more useful results. Though not quite as desirable as being able to predict the exact quality of the wine, we could use this new classification approach to quickly whittle out the bad wines from the decent ones, thus being able to focus more expert energies on wines that are classified as above average.

Re-labelling the dataset into two categories and applying Random Forest to it (since it worked the best with the original dataset), we get an accuracy of 83.82%. This modified approach results in much better recall rate of 91.03% for wines classified as above average (quality 6 and above). The recall performance of the model for wines classified as below average (quality 5 and below) was 69.51%.



Figure : Confusion Matrix for Two Labels using Random Forest

Using this approach gives excellent results to identify wines that are of above average quality. Combining this information with the Random Forest model used on the original dataset can help wine experts speed up their classification process.