White Wine Quality Analysis and Prediction

Data Mining and Knowledge Discovery

by

Amgad Khalil

Université Jean Monnet
Saint-Étienne

April 30, 2024

Contents

White Wine Quality Analysis and Prediction

Data Mining and Knowledge Discovery

Abstract

Main regional market leaders, Portuguese "Vinho Verde" wines, have a niche for their different, light, and fresh profile. In this study, an extensive dataset covering details of the chemical attributes of these white wines was used, which included measures of acidity, sugar, alcohol, and other important chemical properties. Using regression analysis, it is intended to develop a prediction model to estimate the quality of white 'Vinho Verde' wine from these attributes, trying to minimize human dependence on wine tasters. Such an approach will serve not only to streamline the quality assessment process of future wine batches but also to leverage the understanding of what chemical attributes bear a significant impact on wine quality. This model will have great value as a tool for vintners and businesses, enabling them to make a well-informed decision in advance about their production and marketing strategy for wine by providing them with a reliable prediction of quality. Initial results suggest a strong correlation between some chemical characteristics and the perceived quality of wine, indicating that this model could guide improvements in wine production.

Keywords: Vinho Verde, Chemical attributes, Regression analysis, Quality, Correlation, White wines

1. Introduction and Problem Understanding

The dataset used in this study was created in 2009 by a team of researchers and wine experts led by Paulo Cortez, who included Antonio Cerdeira, Fernando Almeida, Telmo Matos, and Jose Reis from the Viticulture Commission of the Vinho Verde Region (CVRVV). The dataset is designed such that it could be viewed as a classification as well as a regression problem. It is a particularly good tool for carrying out regression tasks, like the one in this project. The data is developed to allow a quantitative-oriented analysis of wine to predict the quality, based on the physicochemical properties of wines rather than on human experts' subjective judgment. It facilitates not only the process of quality assessment but also provides insights into what factors could matter most in terms of wine quality. This project is intended to provide streamlined quality control. That means it will make the quality control process easier for producers and businesses that want to predict the quality of a wine batch they intend to purchase by reducing the need for professional tasters.

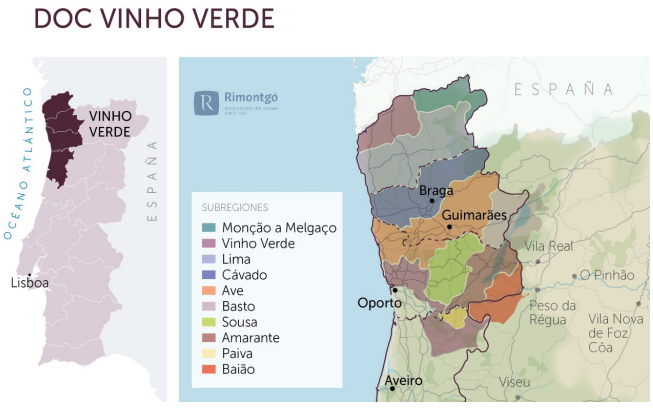

Figure 1: Vinho Verde Region

2. Data Understanding

This section gives an overview of the dataset used in the study, which includes several chemical properties of white 'Vinho Verde' wine. The dataset consists of 4898 instances and 12 features (11 characteristics plus one output variable). The following is a full summary of the characteristics of the dataset:

Feature	Description	Type
Fixed Acidity	Main acids involved in wine stability and color	Continuous
Volatile Acidity	Amount of acetic acid in wine, influencing aroma and taste	Continuous
Citric Acid	Found in small quantities, adds freshness and flavor to wines	Continuous
Residual Sugar	Sugar remaining after fermentation stops	Continuous
Chlorides	Amount of salt in the wine	Continuous
Free Sulfur Dioxide	Prevents microbial growth and oxidation	Continuous
Total Sulfur Dioxide	Amount of free and bound forms of S02	Continuous
Density	Density of the wine which is influenced by alcohol and sugar content	Continuous
pH	Describes how acidic or basic a wine is	Continuous
Sulphates	Wine additive which can contribute to sulfur dioxide gas (S02) levels	Continuous
Alcohol	Alcohol content of the wine	Continuous
Quality (output variable)	Score Ranges from 3 to 9, based on sensory data	Integer

Table 1: Features of the White Wine Quality Dataset

*2.1. **Missing Values***

The dataset was examined for missing values and 0 missing values were found.

*2.2. **Duplicate Records***

The dataset contains 937 duplicate entries. Later discussed in the 4.1 section.

*2.3. **Summary***

Understanding the attributes of the wine and the completeness of the data is made easier with the help of this explanation, which facilitates further analysis.

3. Data Visualization

Data visualization is very important in understanding patterns and relationships within the dataset where statistical numbers alone are not enough to show them. This section highlights the most informative plots that we found while analyzing the chemical attributes of the white wine and their relationship with the quality target variable.

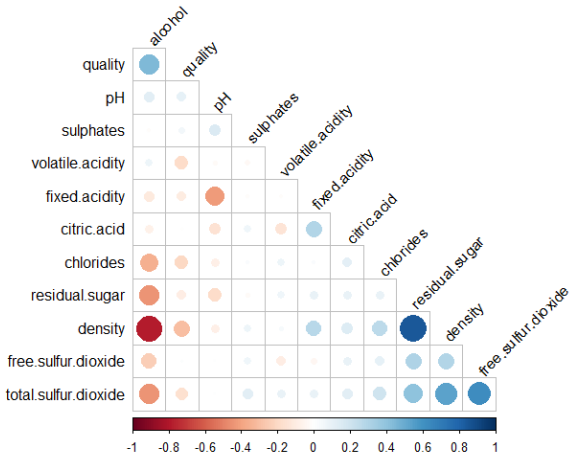

Figure 2: Correlation Between Features

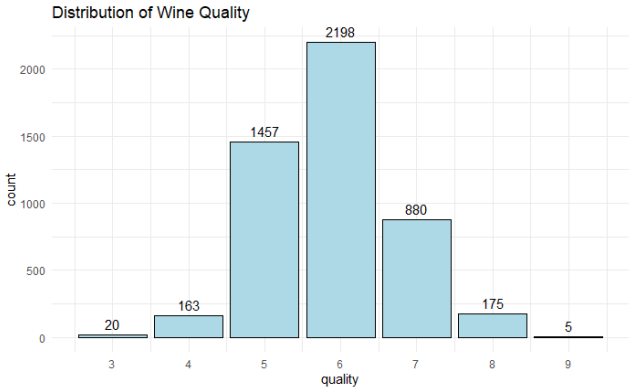

Figure 3: Class Distribution

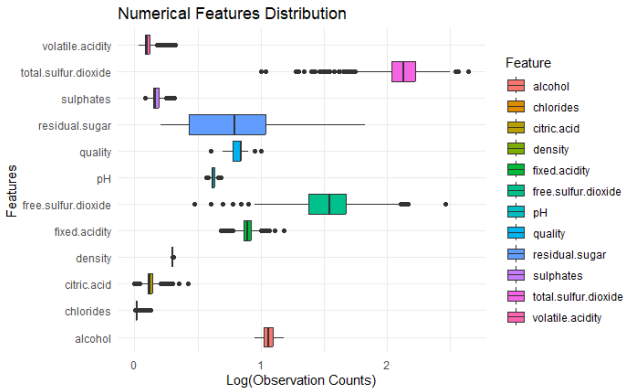

Figure 4: Features Box Plot Distribution

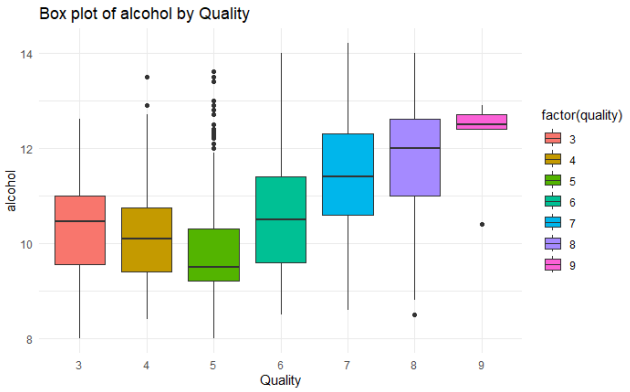

Figure 5: Box Plot of Alcohol By Quality

*3.1. **Correlations***

As evident in Figure 2 during analysis, several high correlations were observed, providing valuable insights into the physicochemical properties that influence wine characteristics:

- **Residual Sugar and Density**
  - **Correlation:** 0.84
  - **Interpretation:** Very high positive correlation indicating that as residual sugar increases, density tends to increase as well. This makes sense because sugar usually adds to the density of liquids.
- **Alcohol and Density**
  - **Correlation:** -0.78
  - **Interpretation:** Strong negative correlation. As alcohol increases, the density decreases. The correlation is logical because alcohol is less dense than water.
- **Quality and Alcohol**
  - **Correlation:** 0.44
  - **Interpretation:** Out of All the chemical attributes only Alcohol showed a mild positive correlation with the Quality of the wine. This suggests that wines with a higher amount of alcohol tend to have a better quality rating which could help a lot in the process of making the wine and marketing for it since it aligns with certain consumer preferences.

3.2. ***Class Distribution***

As we can see in Figure 3 it is obvious that our target variable is highly imbalanced. Most of the wines are concentrated around ranges 5-7 which means there are far more average wines with very few wines rated as low or high quality.

3.3. ***Features Distribution***

As shown in Figure 4, the features' ranges vary a lot showing different scales, meaning that during data preprocessing scaling is required.

The box plot in Figure 5 once again shows the slight trend of higher quality wines having higher ranges of alcohol. It also shows that some outliers are present in quality class 5 for example.

4. Data Pre-Processing

Several quality checks were performed on the data to guarantee its dependability for further analysis.

4.1. ***Duplicates***

To eliminate potential bias in the analysis, and to maintain data integrity we initially removed the 937 duplicates found in the dataset. Shortly after modeling and examining an increase of error results on test sets, we found that these duplicated rows are likely due to several wine testers rating the wine similarly. Hence, it will be relevant to keep all the observations as it can add more information.

4.2. ***Missing Values***

Checked and confirmed that the dataset doesn't contain any missing values.

4.3. ***Feature Selection***

Initially, we decided to remove the density feature because, as shown in Figure 2 before, it has high correlations with the two features (Alcohol and Residual Sugar) which introduces redundancy and multicollinearity. We also found that some features have very low correlations with the target variable close to 0 like free sulfur dioxide; which could indicate that it does not influence the quality of the alcohol. After deeper analysis, we found that the Density variable has a significantly high correlation with the target quality variable falling second in order right after alcohol so we decided to run the dataset through Random Forest, on the cloud (google colab) in Python due to long computations locally, to benefit from the feature importance property before deciding which features to remove. As shown in Figure 6 we see that density and free sulfur dioxide are in the top 5 features for predicting the quality. Due to the reasons above, the higher errors examined when removing those features, and the fact that we can't contact an expert in this domain; we decided to retain all the features as statistical information is not sufficient enough to capture the relationship between the features and the prediction of the quality variable.

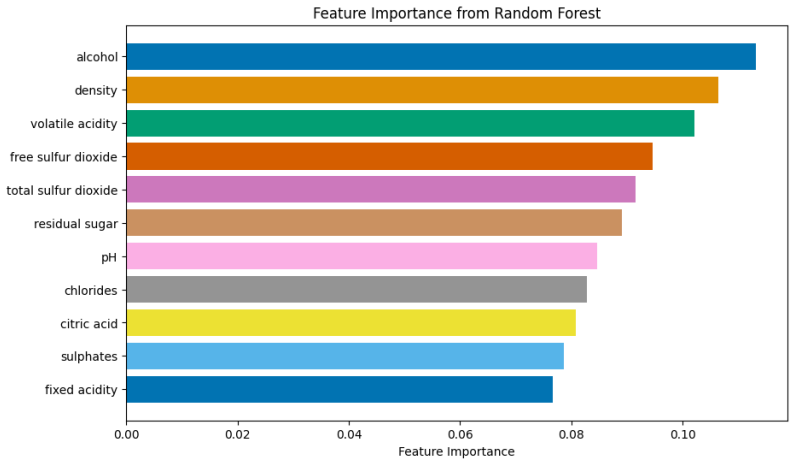

Figure 6: Random Forest Feature Importance

4.4. ***Feature Scaling***

Scaling is crucial in helping machine learning models to be more accurate because most of the models, especially the ones that require distance calculations, are sensitive to the scale of the input features. Features that have higher scales tend to influence the model which leads to a biased model. We preprocessed the numerical features to have a mean of 0 and a standard deviation of 1.

4.5. ***Principal Component Analysis (PCA)***

In this project, PCA was used to analyze the physicochemical attributes of wines and their impact on quality scores in lower dimensions.

The PCA plot shown in Figure 7 shows that the data is mainly concentrated in the center indicating that many wines have similar quality scores regarding the first two components. Alcohol and Density were the best represented in the plots by having the largest magnitudes implying these features are influential in the dataset.

Despite the data reduction done by PCA, it was not possible to directly evaluate the quality of wine in this two-dimensional space because no clusters or separations were observed. This suggests that quality is a more complex trait that may require more features to be understood properly.

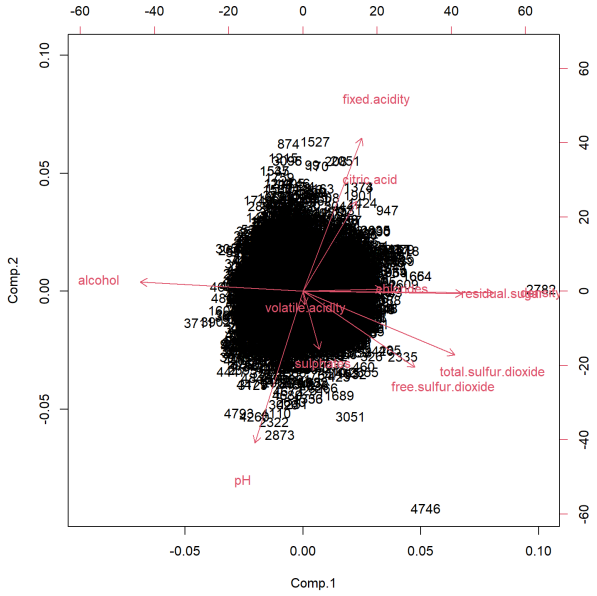

Figure 7: PCA PLot

4.6. ***Handling Class Imbalance***

High-imbalanced datasets like this always introduce challenges with machine learning models because they introduce bias that leads the model to be more in favor of the majority class. In such a scenario, resampling techniques could be used to either up-sample the minority classes or down-sample the majority class. Since the minority class (9) had only 5 observations in this dataset, up-sampling was more the approach to go with but this induced other challenges on us during this project because it increased the number of observations radically which made the computation process very hard for us. Due to our computational limitations, we decided to work with the imbalanced dataset.

5. Modeling

The wine quality dataset can be looked at as either a regression or classification problem. The target variable, 'quality' seems to have, the characteristics of a categorical variable yet also represents an ordinal nature where a higher score represents higher quality. Initially, the intuition was to consider both regression and classification tasks by categorizing quality into three classes: poor, normal, and high quality using predefined ranges. However, computational limitations and the excessive runtime of classification algorithms during local execution in R forced us to do regression analysis only. Regression was also a better choice because it takes into consideration the ordinal nature of our output variable.

5.1. ***Model Selection***

To select a model, we began with training multiple regression models on the dataset such as Linear Regression, Lasso Regression, Ridge Regression, and SVM with a radial basis function kernel (SVMRADIAL). We evaluated the Root Mean Square Error of each model as a measure of accuracy, where lower RMSE indicates better results. As shown in the table below 2, SVM showed the best results out of the four models.

Model	RMSE
Linear Regression	0.7632294
Ridge Regression	0.7632374
Lasso Regression	0.763556
SVMRadial	0.7026423

Table 2: Initial Modeling Results

Regularization introduced by 'Ridge' and 'Lasso' suggests that penalizing large coefficients by 'L2 Regularization' in 'Ridge doesn't change the predictions, and the 'L1 Regularization' in Lasso that promotes sparsity (some coefficients become zero) also doesn't change the predictions much. 'SVM' performed 'best' by having a *RMSE* of '0.70' showing a stronger performance. This could be due to the SVM's ability to handle 'non-linear' relationships.

5.2. ***SVM Hyperparameter tuning***

To tune the hyperparameters of the selected model we used grid search and cross-validation to maximize the performance of the Support Vector Machine (SVMRadial).

5.2.1. ***Grid Search and Cross Validation***

To conduct Grid Search, a parameter grid (svm_grid) with values for "C," "sigma," and "kernel" had to be created. This grid included the kernel functions to be "rbf," along with a variety of hyperparameter combinations. To find the best hyperparameters for the SVM model, Grid Search was carried out using the method grid in R, which included trying all possible combinations in the grid provided and a 5-fold cross-validation.

C	sigma	RMSE	Rsquared	MAE
0.1	0.01	0.75	0.30	0.58
0.1	0.10	0.72	0.35	0.55
0.1	1.00	0.81	0.26	0.60
1.0	0.01	0.72	0.34	0.56
1.0	0.10	0.69	0.39	0.52
1.0	1.00	0.69	0.40	0.48
10.0	0.01	0.71	0.36	0.54
10.0	0.10	0.70	0.39	0.52
10.0	1.00	0.68	0.42	0.46

Figure 8: Model Results

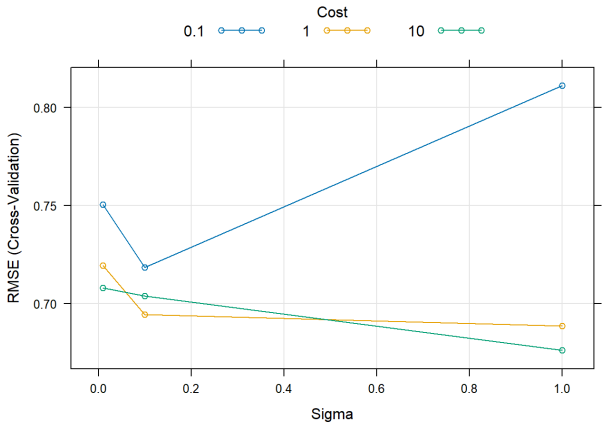

Figure 9: Grid Search Results

RMSE was used to select the optimal model using the smallest value. This model showed a good error reduction where the RMSE value was reduced from 0.7 to 0.65 on the test set. The final values used for the model were sigma set to 1 and C (cost) set to 10.

6. Conclusion and Deployment

The analysis of the prediction of wine quality has shown significant results. The alcohol content was shown to be the most important feature in predicting the wine quality, showing an up trend where higher quality wines have higher alcohol content.

After testing and evaluation, the Support Vector Machine (SVM) with a radial basis function kernel performed best. It achieved the lowest Root Mean Square Error (RMSE) of 0.65. This was an improvement from the 0.70 RMSE introduced initially before tuning, The cost parameter (C) was set to 10 and the sigma parameter was set to 1.

The Tuned SVM model was saved for deployment since it showed the best results out of all the models that we tried. A script, named `deployment.R`, shows an example for the deployment pipeline: it loads new data, performs the necessary preprocessing, and uses the trained SVM model to predict the quality of the wine based on its chemical attributes. This deployment process shows an easy way to implement this model in real-world applications, which will help with the production of wine and with quality prediction.

7. Acknowledgements

We would like to express our sincere gratitude to Professor Fabrice Muhlenbach for his supervision throughout Data Mining with R. We would also like to thank Professor Jeudi Baptiste for his teaching of the theoretical part of the course.

References

[Cor+09a] Paulo Cortez et al. "Using Data Mining for Wine Quality Assessment". In: Oct. 2009, pp. 66–79. ISBN: 978-3-642-04746-6. DOI: 10.1007/978-3-642-04747-3_8.

[Cor+09b] Paulo Cortez et al. *Wine Quality*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T. 2009.

[Pen] Penn State University. *Analysis of the Wine Quality Data*. https://online.stat.psu.edu/stat508/lesson/analysis-wine-quality-data.

[Wik22] Wikipedia contributors. *Vinho Verde*. 2022. URL: https://en.wikipedia.org/wiki/Vinho_Verde.