

Comparison of the red wine quality prediction accuracy using 5 Machine Learning Model

Zhanbei Liu*

GSAS, Columbia University, New York, NY USA 11101

* Corresponding author: zl3070@columbia.edu

Abstract. Wine quality is rather important for alcohol industry, especially for some unique product (e.g., Vinho Verde). Since the quality evaluation of wine is often part of the certification process and is helpful for setting prices, it is necessary to find a model with the highest model accuracy to predict the quality. The paper collected the red wine data from a certain database and changed it to a classification problem with quality as 0 to bad and 1 to good. This dataset is like a normal distribution with the quality of 5 and 6 having the highest ratio. In order to find the approach with the best performance, five machine learning approaches are adopted and compared in terms of various metrics and judgments. According to the analysis, the Support-vector machine (SVM) model predicts preforms the best, i.e., the model has the highest accuracy. This article is a guide for the wine market for the model selection and shed light on guiding further exploration regarding to wine quality prediction.

Keywords: Vinho Verde, red wine, wine quality, accuracy

1. Introduction

Contemporarily, red wine quality plays a crucial role in marketing and sales for the wine industry [1-4]. Among various types of wine, the Vinho Verde is well-known product of Portugal, the largest producing region in the country. This wine is unaged and medium in alcohol, where about 15% of them are red and Rose style. Vinho Verde is characterized by high acidity, low alcohol, and a light spray [6]. To the current knowledge, there is no definitive red variety for it, although several studies on Vinho wines have been published. Exports to Vinho Verde rose 36 percent between 1997 and 2007. Because of its rapid growth, the quality of the wine market must be guaranteed. Quality assessment is able to help to determine prices. As a result, it is necessary to find the best model for predicting the quality of red wine.

Wine quality assessments are usually carried out by "wine experts". Though this issue still has problems, it can be solved and addressed at least in some extents, especially when repeatedly evaluating the same wine. In many cases, quality results are best communicated through categories that represent the range of scores rather than the scores themselves. In these cases, the use of correlation to estimate is not ideal within the same medal range. From a wine show judging point of view, scoring wine in repeated tastings, if both are within the same rating range, is actually the best result [7-10].

Through this article, the research uses the dataset with quality as the dependent variable. The goal of the paper is first to change the quality to a Boolean with good or bad and find the best machine learning model to calculate the accuracy of the quality from Naive Bayes, Decision Tree, K-nearest neighbor algorithm (KNN), Random Forest, and Support-vector machine (SVM).

This article has six parts with first the introduction for the background and literature review for both the red wine and the importance of the wine quality and also the summary of the research. The second part is a brief description of the dataset for this article and the model used in this research. The third section is the result of the article through the boxplot, correlation table, and accuracy table. The fourth section is the limitation of this article and the approximate way to improve it. The last section is the reference cited for this research.

2. Data & methods

2.1. Data

The red wine data of this study is collected from UCI machine learning repository [11]. The original dataset includes two datasets, related to red and white vinho Verde wine samples, from the north of Portugal. In this dataset, the 11 input variables are based on physicochemical tests. The variables are as follows: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol. The output variable is the quality (score between 0 and 10).

2.2. Models

In this project, after finding the dataset, this study investigates the background and do the literature review in the first 2 days to determine my topic. Then, the exploratory data analysis (EDA) is carried out for this dataset, including the barplot for the dependent variables and the boxplot between the 11 independent variables to the dependent variable, and the correlation matrix for the 12 variables. The goal is to turn the 10-point quality scale into a quantitative variable to make the dataset into a classification problem by using quality > 6.5 to good and the others are bad. Subsequently, this study offers the code to calculate the accuracy using 5 machine learning methods that are used normally: Naive Bayes, Decision Tree, K-nearest neighbors algorithm(KNN), Random Forest, and Support-vector machine(SVM). Since this study calculates the accuracy for the 5 models using the default parameters, this paper then optimizes to improve the accuracy by writing a loop to choose the best tuning parameters for KNN, Random Forest, and SVM models. By comparing the accuracy score, the result can find the best model to predict the quality of red wine for this project. To change the variable quality to a categorical variable, the model will classify quality >6.5 to good and bad otherwise.

$$\text{Number of good quality} = \frac{\text{Number of data with quality greater than 6.5}}{\text{Number of whole dataset}} \quad (1)$$

$$\text{Number of bad quality} = \text{Number of whole datasets} - \text{Number of good quality} \quad (2)$$

To calculate the accuracy for the machine learning models, following expression is used:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (3)$$

The Euclidean of KNN model is

$$d(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \quad (4)$$

3. Results & Discussion

Based on the EDA, this research will count the number and also make a barplot for the response variable. As shown in Fig. 1, the dataset is almost a normal distribution. The quality of 5 and 6 have the highest ratio in the dataset. On the contrary, the quality of 3 and 8 only contains a small amount. Then, this study will make the boxplot between quality and each of the predictors.

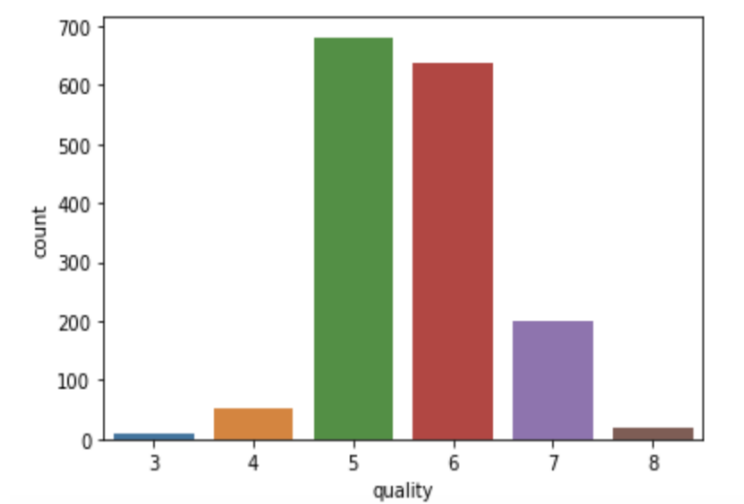
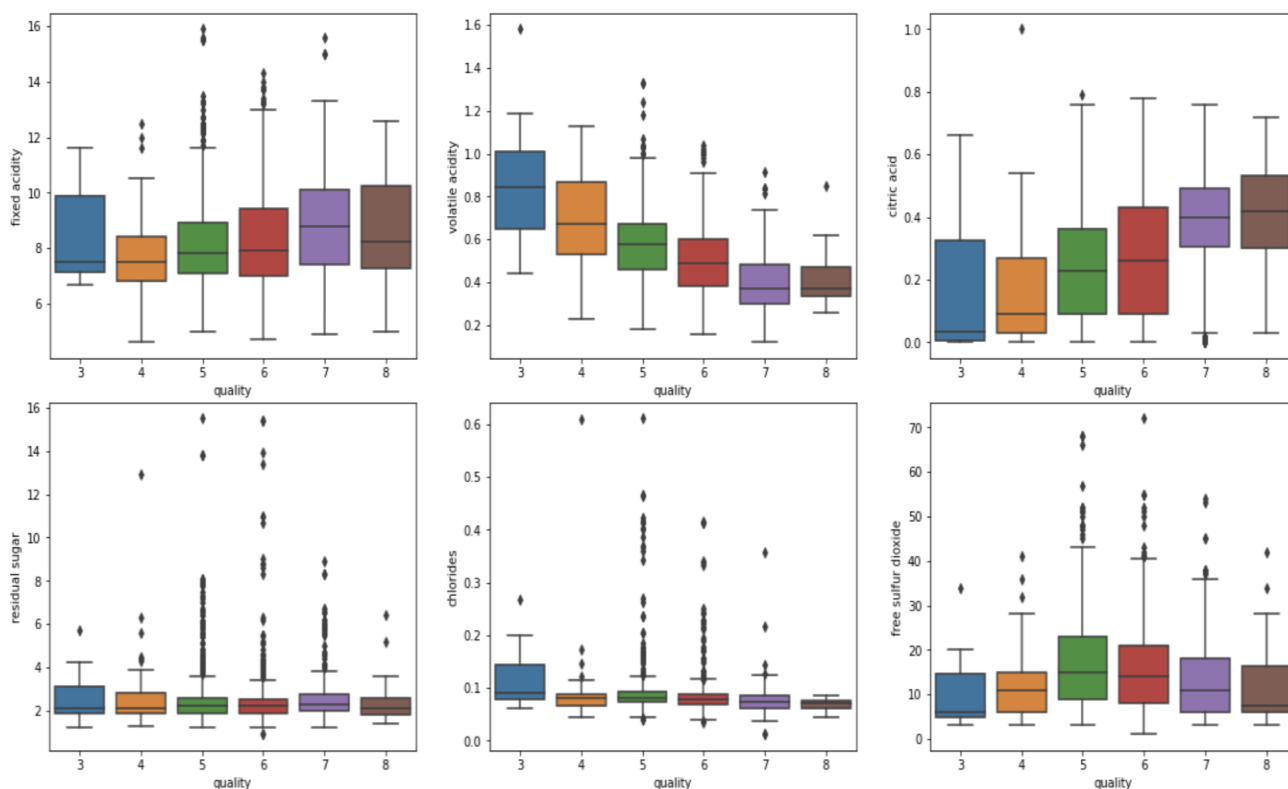


Figure 1. Count of each Quality

From Fig. 2 of the boxplot for each predictor, the fixed acidity, residual sugar, chlorides, and total sulfur dioxide variables, the median of each quality does not have a big difference. However, there are several outliers for variable chlorides and free sulfur dioxide at a quality equal to 5 and 6. For the predictor of volatile acidity, density, and pH, as the quality level increase, which has a descending trend from the boxplot. On the contrary, for the variable citric acid, sulfates, and alcohol, as the quality level increase, which has an increasing trend from the boxplot. Afterwards, this research gives a correlation matrix among each of the variables in the following. According to the correlation matrix in Fig. 3, it does not lead to a multicollinearity problem in this project. It shows that there is a moderate correlation between fixed acidity density, pH, and citric acid and a moderate correlation between free sulfur dioxide and total sulfur dioxide.



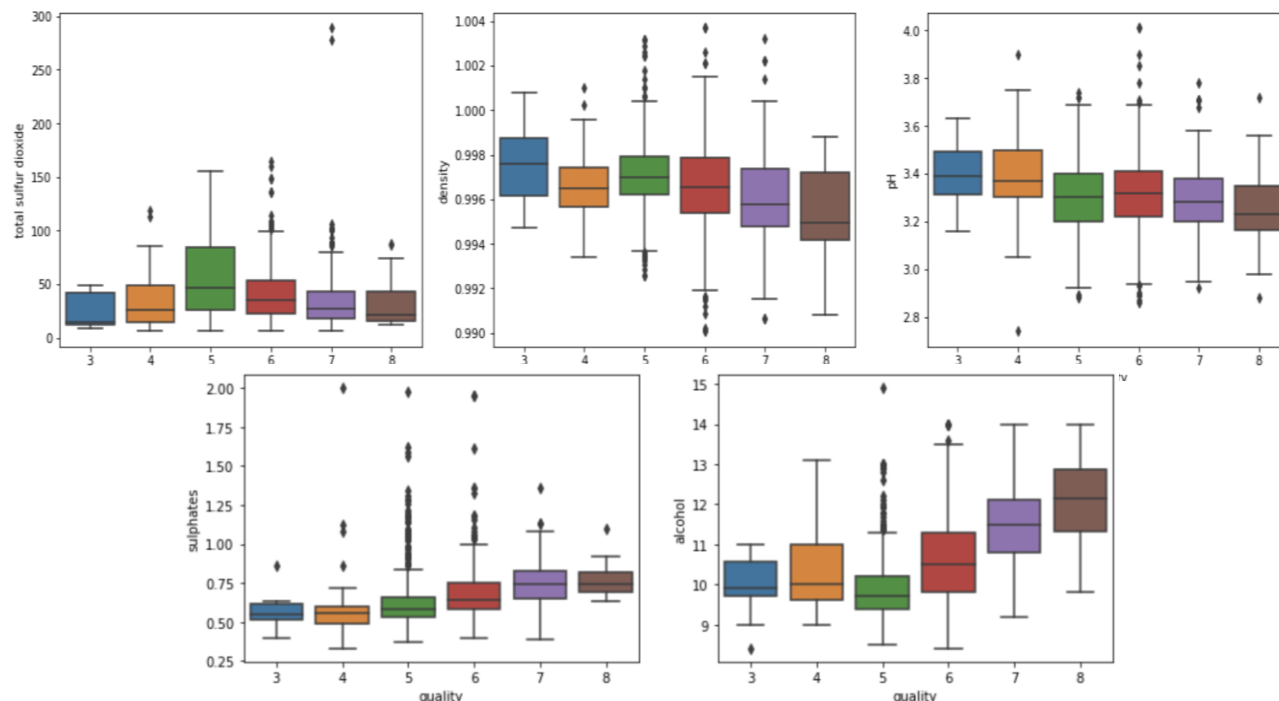


Figure 2. Boxplot for each Predictor to Quality.

Since the goal is to change the quality scale into a classification problem. This study writes a code to make a quantitative variable by using quality > 6.5 to good and the others are bad like the code in the following and change it to a boolean variable with 0 to bad and 1 to good. This paper also makes a barplot in Fig. 4 to present the result. From the barplot and the number count, it shows that the bad quality is far more than the good quality.

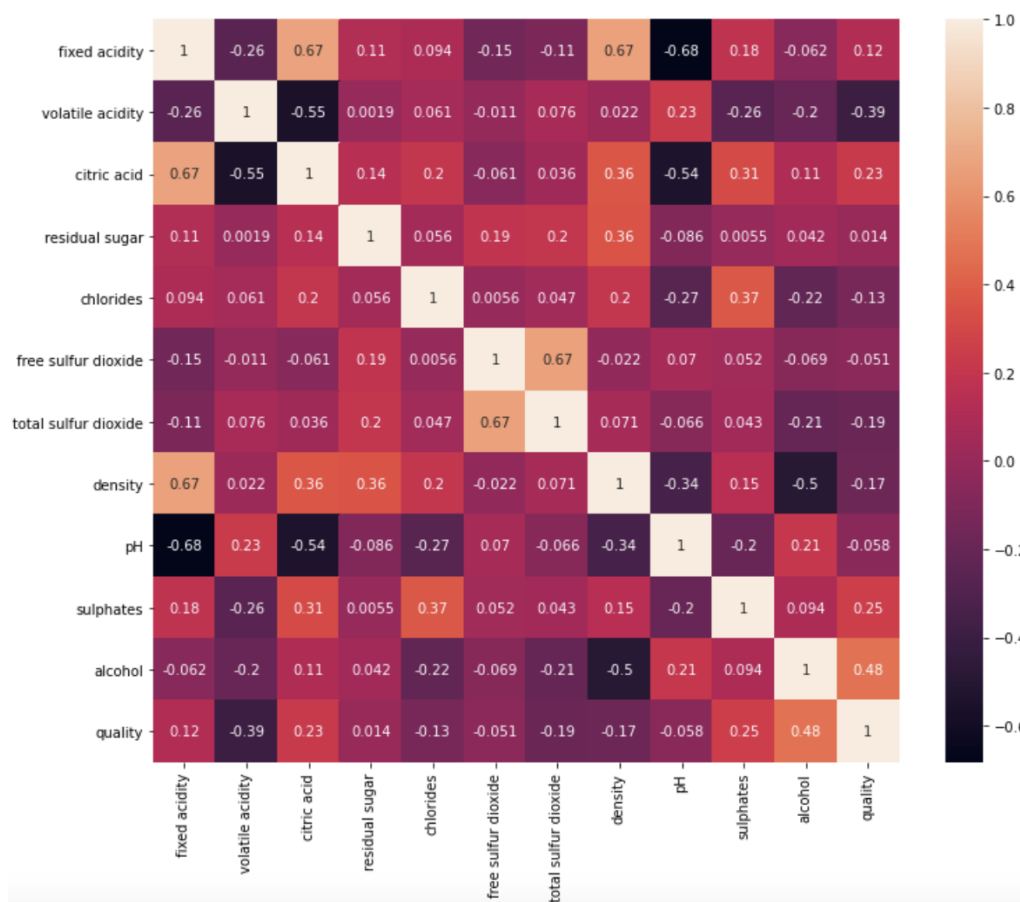


Figure 3. Correlation table for each Variable.

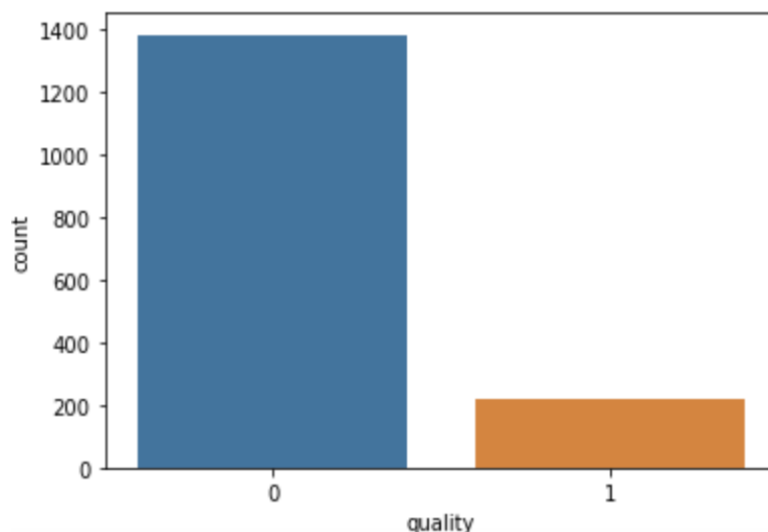


Figure 4. Counts for quality with bad or good.

After doing the EDA, this study standardizes the independent variables and divides the data into a testing set and training set with 30% and 70%. Then, this paper calculates the model accuracy like the example code for the KNN model for the 5 machine learning models using the default parameters: Naive Bayes, Decision Tree, K-nearest neighbors algorithm (KNN), Random Forest, and Support-vector machine (SVM). The results are summarized in Table 1. Based on the results the SVM has the highest accuracy compared to the 5 machine learning methods.

Table 1. Margins and print area specifications.

Method	Accuracy using default parameters
Naive Bayes	82.29%
Decision Tree	84.16%
KNN	88.54%
Random Forest	91.04%
SVM	91.25%

The model is fitted using all the default parameters, this study then does model optimization by writing a loop and using the cross-validation method to choose the best parameters. Table 2 shows the new accuracy as well as the parameter used to fit the new models. This research changed the number of neighbors in KNN method, the maximum depth and the number of trees in the random forest method and the regularization parameter C and the kernel coefficient in SVM model. It is easy to find an increase in the model accuracy of these 3 models. According to the accuracy score, it presents the best model to predict the quality of red wine is the Support-vector machine (SVM) model.

Table 2. The accuracy with the best tuning parameters in each machine learning model.

Method	Accuracy using default parameters	Choosing best parameters	Accuracy using choosing parameters
Naive Bayes	82.29%		
Decision Tree	84.16%		
KNN	88.54%	n_neighbors=43	91.04%
Random Forest	91.04%	max_depth=20, n_estimators=50	91.25%
SVM	91.25%	C = 3, gamma = 1	92.5%

4. Limitation & Prospects

This project still can have lots of improvements. To be specific, people can use R first to do the forward and backward selection to find the best model to predict the quality and then use Python to do the above. Besides, people can calculate not only the accuracy score to compare the model but the MLE and MAP. Since this paper only uses the dataset of red wine, for future research, the author can also use the white wine data to do the same thing and compare the result of white wine and red wine to give more suggestions for marketing. In addition, the dataset used in this article only contains 1,599 observations. To make a more accurate prediction, it is better to enlarge the observations of the dataset. In addition, the dataset is chosen in the year 2009, the result may not suit the wine market in 2022 and should be changed to follow the latest era. Moreover, since the dataset only contains 11 predictors, there are still many variables that may influence the quality of red wine.

5. Conclusion

In conclusion, the quality of 5 and 6 have the highest ratio in the dataset, which means most of the red wine only touches the standard quality to use. After changing the dataset to categorical, the bad quality is 5 times more than that of the good quality. It is understandable since the low quality of the red wine is, the lower price. The median standard is suitable for most people to buy. In addition, for the fixed acidity, residual sugar, chlorides, and total sulfur dioxide variables, the median of each quality does not have a big difference. According to the highest accuracy score, the paper results in the best model to predict the quality of red wine being the Support-vector machine (SVM) model with an accuracy equal to 92.5% no matter if adjusting the tuning parameters in each model. However, adjusting the model with the most suitable tuning parameter can increase the model's accuracy.

References

- [1] Gupta, Y. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, 125, 305-312 (2018).
- [2] Dahal, K. R., Dahal, J. N., Banjade, H., and Gaire, S. Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics*, 11(2), 278-289 (2021).
- [3] Castillo-Sánchez, J. X., García-Falcón, M. S., Garrido, J., et al. Phenolic compounds and colour stability of Vinhao wines: Influence of wine-making protocol and fining agents. *Food Chemistry*, 106(1), 18-26 (2008).

- [4] Castillo-Sánchez, J. J., Mejuto, J. C., Garrido, J., and García-Falcón, S. Influence of wine-making protocol and fining agents on the evolution of the anthocyanin content, colour and general organoleptic quality of Vinhão wines. *Food Chemistry*, 97(1), 130-136 (2006)s.
- [5] Denise M. Gardner Former Penn State Extension Associate in Enology Pennsylvania State University. Volatile Acidity in Wine. Penn State Extension, 7 Aug. 2022, Retrieved from: <https://extension.psu.edu/volatile-acidity-in-wine#:~:text=What%20Is%20Volatile%20Acidity%3F,smell%20and%20taste%20of%20vinegar>.
- [6] Red Wine Definition & Meaning. Merriam-Webster, Merriam-Webster, Retrieved from: <https://www.merriam-webster.com/dictionary/red%20wine>.
- [7] Gawel, R., and Godden, P. W. Evaluation of the consistency of wine quality assessments from expert wine tasters. *Australian journal of grape and wine research*, 14(1), 1-8 (2008).
- [8] 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).
- [9] What Is Vinho Verde? Understanding the Many Wines from Portugal's Prominent Wine Region - 2022. MasterClass, Retrieved from: <https://www.masterclass.com/articles/what-is-vinho-verde>.
- [10] Comissão de viticultura da região dos Vinhos Verdes, Retrieved from: <http://www.vinhoverde.pt/>.
- [11] Learning, UCI Machine. "Red Wine Quality." Kaggle, 27 Nov. 2017, Retrieved from: <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>.