White Wine Quality Prediction and Analysis with Machine Learning Techniques

Xianghui Jiang[1, †], Xuanyu Liu[2, †, *], Yutong Wu[3, †] and Dehuai Yang[4, †]

[1] School of Science/School of Big-Data Science, Zhejiang University of Science and Technology, Hangzhou, Zhejiang, China

[2] The Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA, United States

[3] Maple Leaf International school-Chongqing, Chongqing, China

[4] Transportation Engineering school, Tongji University, Shanghai, China

* Corresponding author email: Xuanyul5@uci.edu

[†] These authors contributed equally.

Abstract. The wine originated 7,500 years ago, and it is one of the most popular alcoholic beverages in the world today. With modern advances in production processes, the quality of wine has been significantly improved. White wines can be graded by measuring various physical and chemical quantities. In reality, it is very rare for white wines to have dramatic quality differences, so precisely capturing these important characteristics will be difficult. In addition, the available dataset for wine quality is relatively small and imbalanced, so it is not easy to train a machine learning model for wine quality prediction. This paper explores the applicability of applying logistic regression machine learning models for wine quality prediction. The Synthetic Minority Oversampling Technique (SMOTE) algorithm is applied to address the imbalanced data issue. The key features in wine production are analyzed to the physical or chemical quantities that are most likely to influence the quality of white wines. In several tests, the random forest model gives good results, and free sulfur dioxide, chlorides, and fixed acidity are the most likely characteristics to influence the quality of white wines.

Keywords: White Wine; Quality Prediction; SMOTE; Random Forests; Multiple Logistic Regression.

1. Introduction

Wine is now the most popular banquet drink in the world, and we can see them on various occasions. The origin of wine can be traced back to 7500 years ago. At that time and in present-day Iran, wine was not distinguished by colors, and it might be red, white, or even other colors [1]. However, the earliest use of white wine was recorded in the ancient Greek era, when doctors used white wine to treat patients [2]. These historical facts have proved how amazing it is that wine is still people's favorite alcoholic beverage after so many years.

In addition, the production process of white wine has become an art. Every step, from planting grapes to selling to consumers, is full of delicate and complex steps. The modern process of making wine is as follows: 1. Firstly, plant suitable grapes and then wait for maturity; 2. Then extract the juice from grapes; 3. Remove sediments in juice by precipitation; 4. Alcoholic fermentation by adding sugar and yeast; 5. Oaked the wine in the barrel to complete malolactic conversion; 6. Finally, the wine is filled, corked, and bottled to finish the production [3]. These are a few important steps summarized by omitting many details. In modern industry, a small change in any step will lead to a great difference in taste and quality after the final white wine is produced.

How to test whether a glass of white wine is good or bad? The Wine & Spirit Education Trust (WSET) gave the tasting grid to evaluate different wines. The first is to judge the appearance of wine, whether the color is pure, and whether the impact of changes over time is serious. The second is the smell of wine. The oxygen content in wine is increased by rotating the wine glass, and then the wine is graded by the smell and intensity. The final step is to evaluate the taste of wine, in which there are

specific instructions and standards for the sweetness, acid, tannin, alcohol, body, and intensity of the wine. In addition, the taste of wine can also be compared with the taste of fruits, portraits, herbs, or spices to obtain the specific attributes of different wines [4]. After these steps, the wine taster will give a number value as the final quality level of the wine.

However, if such a delicate and complex wine tasting process is widely used, it will be very difficult and waste resources. Each substance is composed of different ingredients. If the external environment is roughly the same, the quality and taste of wine with the same composition must be very similar. In order to help improve the quality of white wine in production, this study aims to explore the impact of different ingredients on the quality of white wine through data analysis. It then predicts the content of ingredients and chemical characteristics of high-quality white wine to help improve the quality and taste of industrial-produced wine and even optimize the detection process of white wine.

2. Dataset

We use the Wine Quality Dataset [5] from the UCI Datasets[6], which contains two sub-datasets for local red and white wine samples from northern Portugal. The white wine quality dataset is a multivariate and real dataset, which contains 4898 instances and 12 attributes. It was used in fields like business and was donated on 2009-10-17. Wine certification is usually evaluated by physical, chemical and sensory tests[7]. Thus, the data cover the ranges that could affect the quality of the white wine. It includes the following sections: volatile acidity, free sulfur dioxide, citric acid, fixed acidity, chlorides, pH, total sulfur dioxide, residual sugar, density, sulfates, alcohol, and quality. All of this stuff is because of privacy and logistic issues. Only physical and chemical, and sensory variables are included in that dataset. We used box plots to show the features of this dataset in Fig. 1.

3. Methods

3.1 Data Balancing

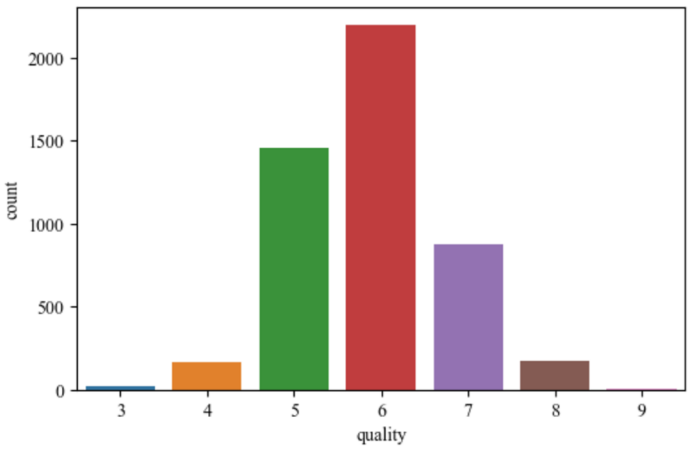

Fig 1. Category statistics bar chart

Fig. 1 demonstrates that the number of categories in this dataset is very imbalanced. The number of categories 3, 4, 8, and 9 is very small, while categories 5, 6, and 7 account for a very large amount of data. This can be explained by the very low probability of very good quality wines (categories 8 and 9) and very poor quality wines (categories 3 and 4) in normal wine production. However, in model training, such data can bias the results. Due to the imbalance in the data, the results of the model will be more biased towards more observed categories.

A method of oversampling is used to solve this problem. The SMOTE algorithm[8], the idea of which is to oversample the minority classes by obtaining a sample of each minority class x_i and

introducing synthetic examples along a line segment connecting any/all k-nearest neighbor minority classes $\hat{x_i}$ by

$$x_{new-sample} = x_i + rand(0,1) \times (\hat{x_i} - x_i) \quad (1)$$

By using the SMOTE algorithm, an oversampled sample data bar chart is obtained after trying different parameters, as shown in Fig. 2.

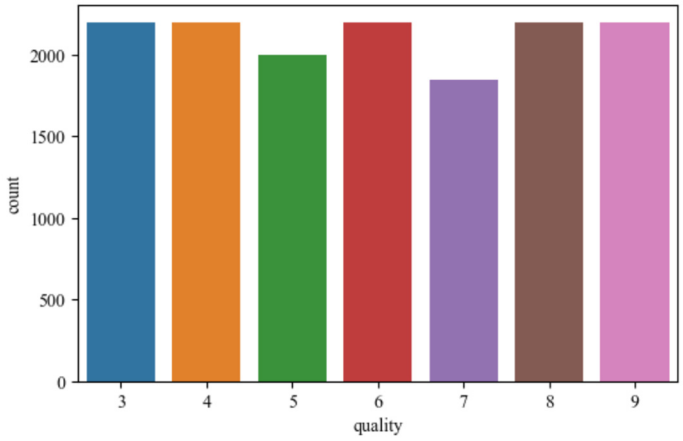

Fig 2. Bar chart of sample statistics after oversampling

3.2 Models

3.2.1 Random Forest

Random Forest [9] is a classical integrated learning algorithm with decision trees as the base learner. The main steps are to first perform M random sampling operations with put-backs from a dataset containing N samples, and one can obtain T randomly sampled datasets containing M. Then, a subset containing k attributes will be randomly selected from all attributes of each base decision tree, and in general, k takes the value (where f is all the attributes of the sample) by:

$$k = log_2 f \quad (2)$$

Finally, a forest is obtained, and when a new sample enters, each decision tree makes a judgment and produces T predictions. In the classification task, these T predictions are voted on to produce the final prediction.

The random forest classification model will be used in this study, and we will compare it with other models.

3.2.2 Multinomial Logistic Regression

Multinomial logistic regression is a classification method that applies logistic regression to multiclass classification problems that have at least two classes [10]. The main idea of these methods is that one can construct a linear function to predict a sample's probability for each class using a set of weights. We can use the linear combination of the weights with the explanatory variables from the observed samples using a dot product:

$$score(X_i, k) = \beta_k \cdot X_i \quad (3)$$

Where X_i represents the explanatory variables' vector of the observed sample i, β_k is the vector of weights, regression coefficients of the output k, and the score is given by $\text{score}(X_i, k)$, which is the probability that the sample i belongs to category k. For the sample i, the class with the highest prediction values is chosen as the predicted result.

4. Results and Discussion

In this section, we first conduct a descriptive analysis of the white wine quality and its factors, then we predict the quality using the two models using the scikit-learn library in Python. For the random forest, we set the number of decision trees, i.e., the number of classifiers, to 450 and the

maximum depth of the decision tree to 25. In addition, we set the maximum number of features and class weights to achieve the best classification results. For the multinomial logistic regression, we set 7 categories for the quality of the wine. Then we use this model to train 70% of the whole data. Finally, we use 30% for testing with this model.

4.1 Wine Quality Analysis

The initial processing method for data sets is to conduct descriptive analysis, through which one can judge the distribution of data and the general trend. The results of the descriptive analysis will also serve as the basis for further data processing and analysis. There are 12 reference factors in this white wine data set, of which the first 11 are independent variables (composition and chemical content), and the last factor - quality is the dependent variable. In other words, the result of the previous factor is the quality of the result. In addition to macro descriptive statistics and distribution analysis of data sets, this study still needs a more detailed and systematic descriptive analysis and statistical model for quality and different influencing factors. This new descriptive statistical method will analyze the distribution of these 11 independent variables at different quality levels and then get a trend according to the distribution. This reflects the distribution of these components and chemical characteristics after quality improvement. In order to more intuitively see the results from the analysis charts, this descriptive analysis uses box plot charts and line charts to display. The box diagram is used to represent the mass distribution of various factors under different quality levels. The line chart is to obtain the trend by connecting the average number of different influencing molecules under different mass levels to observe the numerical changes of various factors after the mass rises. The results obtained by this method will be more helpful for later experiments.

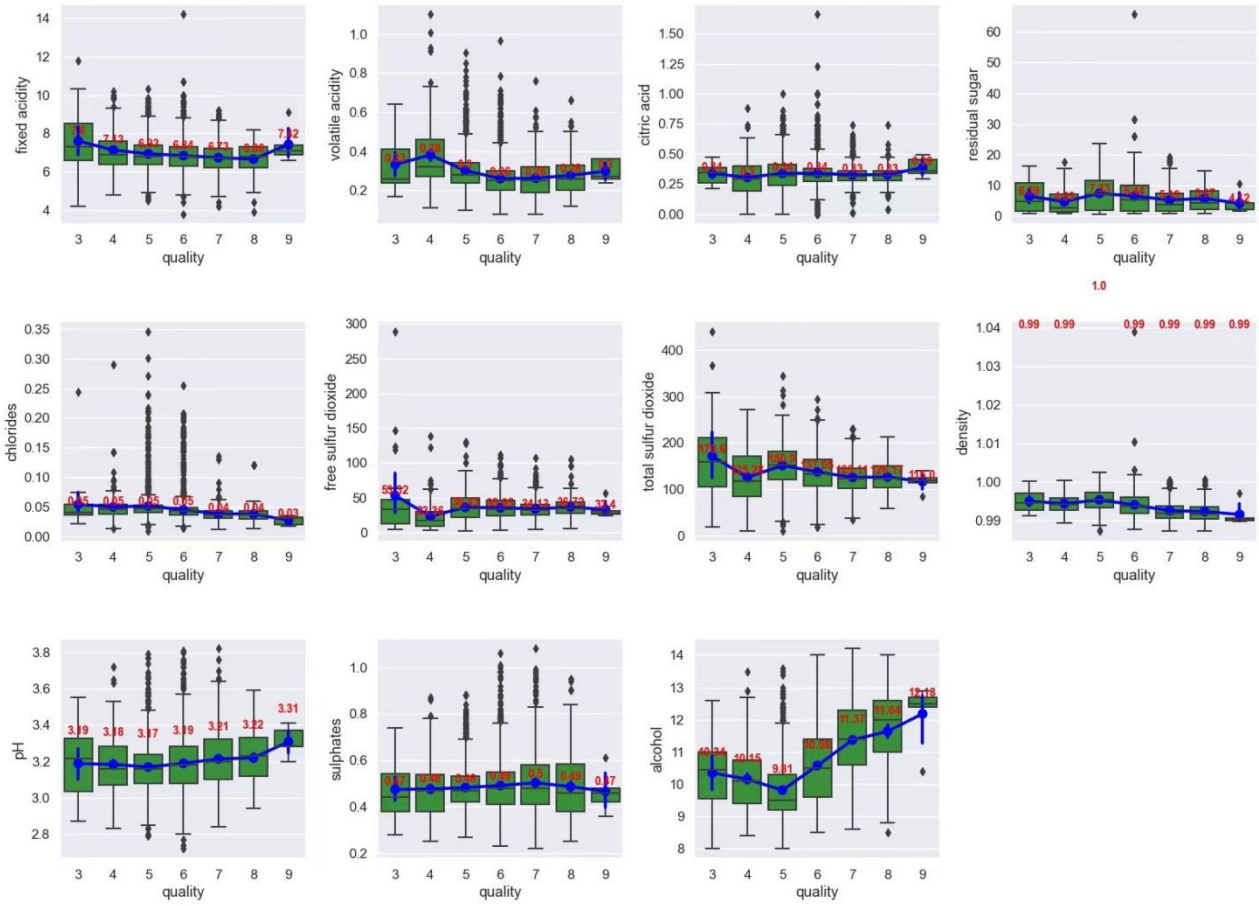

Fig 3. White Wine Quality Dataset Descriptive Analysis by Boxplot & Line Charts

Fig. 3 is the result of the descriptive analysis on the distribution of 11 components or chemical characteristics through different quality levels. However, only a part of the influence factors will have a large response due to the change in quality level. The fixed acidity of lower and higher quality levels is slightly higher than the average of other quality in the fixed acid, and the fixed acid data is more stable in the case of high quality. Furthermore, chlorides are more stable and tend to decrease with the increase in quality. However, with the gradual increase of the quality, the pH value tends to rise. The last important factor is alcohol. The higher the quality level of white wine is, the higher the alcohol. In addition, other influencing factors have no obvious trend, but this descriptive analysis also shows the distribution of data and provides a reference for further research.

4.2 Wine Quality Prediction

In our problem, the wine quality ranges from 1 to 10, so we divided the quality into ten categories, and each of the categories represents a different quality of the wine. We evaluated the random forest model and the multinomial logistic regression on our dataset. Compared with the random forests model, multinomial logistic regression does not have any parameters for us to adjust. Therefore, the only thing that we could do to try to increase the accuracy is to change the value of the independent variables. After balancing the data and using $log(x)$ function to change the raw output, it turned out that the accuracy did not improve apparently, with the best performance at 52%.

However, the Random Forest model's accuracy is much higher. Conclusively, we think it will be better to use the Random Forest model to predict the quality of the quality. In the random forests model, we can rank the importance of each feature weight.

Table 1. Feature importance ranking

Feature	Importance weights
free sulfur dioxide	0.122811
chlorides	0.109765
fixed acidity	0.102393
total sulfur dioxide	0.101026
alcohol	0.100458
pH	0.089324
volatile acidity	0.085991
density	0.079583
citric acid	0.079145
residual sugar	0.068460
sulphates	0.061045

From Table 1, we can see that the importance weights of the first five of these eleven features do not vary greatly. We can then analyze these features according to their impact on the prediction results, thus giving recommendations. For example, the "free sulfur dioxide" feature has the greatest impact on the sample results; our recommendation, combined with the previous data analysis, is to place a null value of 30-50 in the range for the feature "free sulfur dioxide". Values that are too high or too low may significantly reduce the quality of the wine.

We can explain how giving more attention to these first five attributes in the industrial production of white wines may improve the quality of the resulting white wines.

5. Conclusion

Wine quality prediction is a meaningful task considering the great popularity of wine worldwide. The main difficulty of applying machine learning techniques for wind quality prediction is the limited availability of training data and imbalanced data. This work is focused on exploring the factors that might influence the quality of the wine. Wine with great quality could not only appeal to more customers and even enjoy a worldwide reputation. Therefore, if one desires to explore the impact

factors, one needs to use math models to do data analysis. This work experimented with the Random Forest and Multinomial Logistic Regression models for wine quality prediction. The accuracy of the Multinomial Logistic Regression model is about 50%, whereas the Random Forest Model achieved over 90% accuracy. The advantage of the Random Forest model is that the model is more flexible in adjusting the parameters. Therefore, this work can analyze these features according to their impact on the prediction results and give recommendations. Conclusively, "free sulfur dioxide" has the greatest impact on the sample results. The experiment results suggest that the wine producer should pay more attention to these first five attributes in the industrial production of white wines might improve the quality of the white wines.

References

[1] Philippe Testard-Vaillant, A nectar for 7500 years, CNRS, consulted on 4 June 2010 (in French).

[2] Hugh Johnson, A World History of Wine from Antiquity to Modern Times, Hachette, 1990, ISBN 2-01-015867-9, 464 pages, p46.

[3] jellederoeck. "How White Wine Is Made from Grapes to Glass." Wine Folly, https://winefolly.com/deep-dive/how-is-white-wine-made/.

[4] Tilden III, Marshall. "How to Taste Wine like a Pro." Wine Enthusiast, 16 Mar. 2021, https:// www. winemag. com/2018/07/10/wine-tasting-grids/.

[5] 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, 2009, 47(4):547-553.

[6] A. Asuncion, D. Newman UCI Machine Learning Repository, University of California, Irvine(2007) http:// www.ics.uci.edu/~mlearn/MLRepository.html.

[7] S. Ebeler. Flavor Chemistry --Thirty Years of Progress, Kluwer Academic Publishers,1999:409-422. chapter Linking flavour chemistry to sensory analysis of wine.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, Vol. 16 (2002).

[9] Gerard Biau. Analysis of a Random Forests Model, Journal of Machine Learning Research 13 (2012) 1063-1095.

[10] Prasad, K. D. V., & Vaidya, R. (2018). Causes and effect of occupational stress and coping on performance with special reference to length of service: An empirical study using multinomial logistic regression approach. Psychology, 9(10), 2457-2470.