Red wine Quality Prediction by Supervised Learning

Peini Li*

Wuhan NO.11 middle school, Wuhan, China

*Corresponding author: zhulan@tjh.tjmu.edu.cn

Abstract. In this essay, various machine learning techniques are carried out to model the red wine dataset, which contains 1599 samples and 11 physical features, and predict its quality. Data processing techniques including dropping duplicate values and removing outliers are first applied to normalize the features and bring them to a similar scale. Five algorithms, KNN, Multinominal Logistic Regression, Bagging, Random Forest, and Gradient Boosting are then applied. Cross-validation is performed to identify the best parameters for each model. Additionally, feature selection is also done in Logistic Regression based on the calculated p-values. All five models perform well, with test accuracies roughly around 60%, indicating that machine learning has enormous potential on red wine quality prediction to facilitate the whole industry. Among them, Logistic Regression and Random Forest have slightly higher accuracies. This paper highlights the importance of machine learning methods of Logistic Regression and Random Forest in red wine quality prediction.

Keywords: Red wine; quality; supervised learning.

1. Introduction

Red wine has existed for more than thousands of years in human's life, holding cultural and symbolic significance in various cultures. Its health benefits, such as cardiovascular benefits and antioxidant properties, have contributed to its popularity and importance in the world of wine. The quality of red wine is traditionally assessed based on sensory evaluations conducted by experts. During the process, panelists follow standardized criteria to assess different sensory aspects such as appearance, aroma, taste, mouthfeel, and overall liking. They evaluate attributes like color, clarity, intensity of aromas, complexity, acidity, sweetness, tannins, body, and balance, to determine its quality and overall sensory experience. However, the manual evaluation process is time-consuming, subjective, and prone to human errors [1].

With the rapid development of technology, machine learning can be applied to the field of red wine quality prediction today. Machine learning accurate predictions compared to conventional manual procedure by learning patterns and relationships in the data. It enables efficient prediction of red wine quality. Once the model is trained, it can quickly process large amounts of data and make predictions in real-time, saving time and effort compared to manual analysis. So a more subjective and accurate approach of examing red wine quality can be developed. This will make significant contributions to the wine market.

Previous researches have attempted to apply various supervised learning techniques for classification to the prediction of red wine quality [2-5]. However, relevant investigations are still limited. Thus, this paper focuses on this field, so as to benefit winemakers.

2. Methodology

2.1. Data Collection

The dataset is obtained from kaggle, containing 1599 samples of the Portuguese "Vinho Verde" red wine. For each sample, eleven features based on physicochemical tests are included, together with the quality based on sensory data. The physical features are taken as input variables and the quality is taken as output vairable. Python 3 is used for data processing and machine learning modelling.

2.2. Data Processing

To ensure all the data being used is valid and unique, null values and duplicate values are dropped at first. Then, outliers for each feature are removed according to statistical rules, so that the model can better fit the whole dataset. The number of samples becomes 985 after cleaning the data.

The target variable is classified into six groups from quality score 3 to 8, which is a relatively large number in machine learning. Besides, the data is highly imbalanced, with quality 5 and 6 dominanting the dataset, and few samples in quality 3 and 8. Fig. 1 shows the distribution of the target variable.

Standard Scaler is used to standardize the input features, making them comparabe and suitable for various machine learning algorithems. After that, the dataset is splitted into training set and testing set, with a ratio of 4:1.

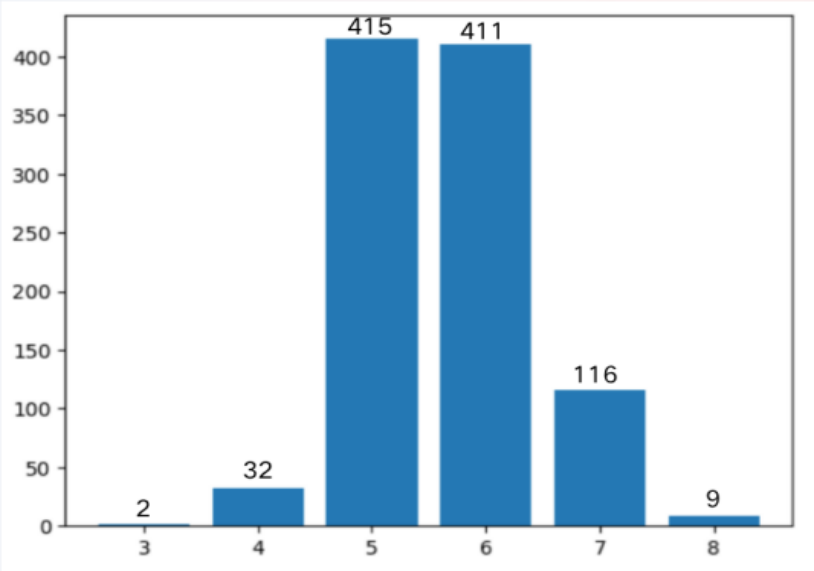

Fig. 1 Bar plot of the distribution of target variable after cleaning the data

2.3. Evaluate Indicators

Train accuracy: The prediction accuracy on the training set. It reflects how well the model fits the training dataset.

Validation accuracy: The prediction accuracy on the validation set, which is seperated from the training set in cross-validation, a method used to determine the best parameters of a model.

Test accuracy: The prediction accuracy on the testing set. It reflects how well the model makes prediction on unseen data, and is the primary measure to be considered.

Pro accuracy: An measure the author defines in this experiment. Pro accuracy is the percentage of prediction where the difference between the predicted value of red wine quality and its true value is no bigger than 1.

2.4. KNN

KNN is a wide-used supervised learning algorithm that predicts the class of a new data point using the majority class of its K nearest neighbors in the feature space. The most significant parameter for this model is the value of K. Cross-validation is employed to identify the optimal K value within the range of 1 to 100. The result shows that when K=74, the validation score tops the list. So K=74 is used to train the model.

2.5. Multinomial logistic Regression

It uses a modified version of the softmax function, which can normalize the probabilities across all categories. The model estimates separate coefficients for each category, and these coefficients are then used to calculate the probability of each category for a given input. By comparing the calculated probabilites, the category with the dominant value is considered the predicted class. The most

significant parameter for this model is the value of C. In this experiment, cross-validation is applied to decide on the value of C that gives the best validation score among 0.01, 0.1, 1, 10, 100, and the result is 1. So it is used again to determine the best value from 0.5 to 1, and the result is 0.7. So C=0.7 is used to train the data.

2.6. Tree-based methods

Decision tree is a graphical representation of a series of decisions or choices that lead to a particular outcome. The construction of a decision tree involves dividing the data based on different attributes and selecting the most informative features to make decisions. This process is done recursively until a stopping criterion, such as reaching a maximum depth or purity, is met. Tree-based methods ensemble structures of decision trees to make predictions. Three of them are carried out in this experiment:

Bagging: It generates multiple subsets of the initial training data by randomly selecting samples from the training data, allowing for repetition, and forming new datasets that match the size of the original dataset. Decision tree is trained on each subset independently. The predictions from all the models are combined using a voting approach, where the class with the most votes is chosen.

Random Forest: It works the way as Bagging does, expect that only a randomly chosen subset of features is used to built the tree. It is a powerful algorithm that can handle large datasets with high dimensionality and can capture complex relationships between variables.

Gradient Boosting: ¸It is a type of ensemble method where multiple weak predictive models, typically decision trees, are combined to create a stronger overall model. Each new model is trained to predict the residuals, or errors, of the previous model. By repeatedly adding these new models to the ensemble, the overall model learns to correct and improve upon the weaknesses of the previous models. This way, the ensemble gradually learns to predict the target variable more accurately with each iteration.

Fig. 2 shows how the validation score of each method varies with the tree number, which is the crucial paramter for all three algorithms.

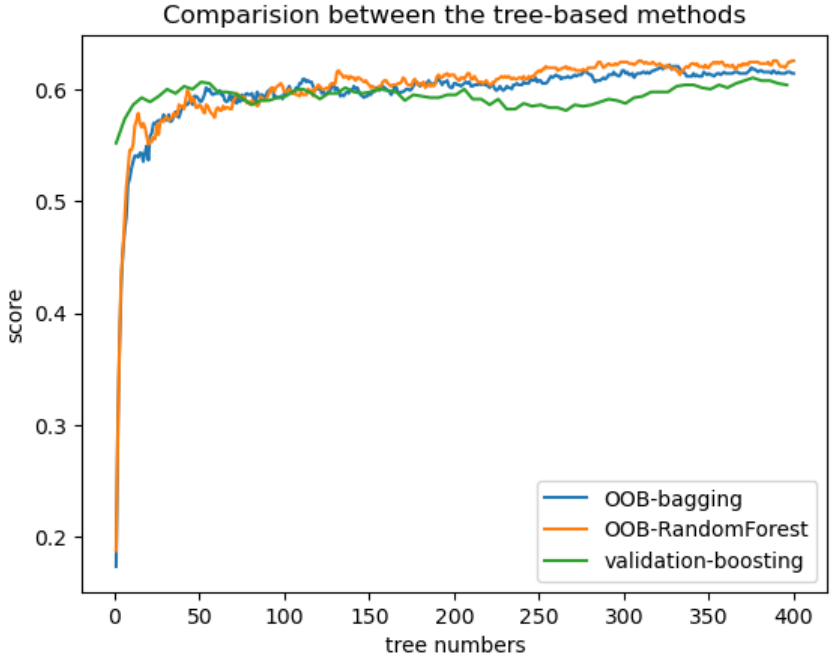

Fig. 2 Plot of validation score against the tree number

For Bagging and Boosting, OOB score is the validation socre calculated using the out-of-bag samples that were not used during its training as the validation set. But in Gradient Boosting, all the dataset is used for training, so cross-validation is used to obtain the validation score for each parameter.

From Fig. 2, the best value of tree number is determined, which is 320 for Bagging, 400 for Random Forest, and 380 for Gradient Boosting.

3. Results

Table 1 shows the result of the five models. All the test accuracies are just under 60%, with pro accuracies all quite considerable, illustrating that all five models perform pretty well on red wine quality prediction. Among them, Logistic Regression and Random Forest have slightly better performance.

Table 1. Table of accuracy

	Train Accuracy	Validation Accuracy	Test Accuracy	Pro Accuracy
KNN	1.000000	0.593905	0.558376	0.984772
Logistic Regression	0.613790	0.595131	0.583756	0.989848
Bagging	1.000000	0.618020	0.588832	0.969543
Random Forest	1.000000	0.625635	0.593909	0.974619
Boosting	0.382244	0.606508	0.578680	0.959391

4. Further Reflection

Although the result is good, compared to other studies, the test accuracies are ralatively low in general.

4.1. Imbalanced data

The extremely imbalanced data can lead to incorrect classification. In KNN, the k-nearest neighbors being considered for the test point's classification are more likely to be samples from the more prevelant classes, resulting in incorrect classifications of test cases from small classes. In Tree-based methods, in order to distinguish small classes, more branches are pruned. This caused overfitting, the increase in the error rate due to which is much more significant than the decrease due to correct classifications of small classes. Since the aim of pruning is to minimize the predicting error, these branches are removed, leading to the incorrect prediction of small classes [6]. However, the problem of imbalanced data can be solved by over-sampling methods like the Synthetic Minority Over-sampling technique (SMOTE), which synthetically generates samples in smaller classes using linear relationships within the classes, or by class decomposition schemes, including OVA and OVO, that simplify the problem into binary classification [7].

4.2. Droping outliers

Dropping outliers can not always lead to better results, only when there's strong confidence that outliers are measurement errors can this process make sense [8]. In the experiment, 250 samples are determined as outliers and dropped, which constitute more than 15% of the original dataset. So the process of dropping outliers possibly removes dozens of valid samples that could have contributed to the model training. Especially in case, where the data is highly imbalanced and the number of classes is large, the odds are that this makes the small classes even smaller and the data even more imbalanced, since samples in small classes tend to have more extreme physical properties and are more likely to be detected as outliers. So to improve accuracy, the process of dropping outliers should be reconsidered and taken in a more cautious way. The defition of outliers can be stricter, so that only those particularly extreme samples are dropped. Feature selection can also be done before dropping outliers, so that only the dominant features are processed, reducing the number of samples being removed.

4.3. Feature Selection

Feature is a process in machine learning and data analysis where a subset of relevant features or variables is selected from a larger set of available features [9]. By selecting relevant features, noise can be reduced and the most informative features can be more focused. At the same time, it prevents overfitting, where the model fits the training data too closely or captures noise and random fluctuations rather than the underlying patterns or general trends in the data, that results in poor test accuracy. So generally, feature selection has potiential to improve the model performance. Experiments also show that feature selection can also bring higher accuracy in red wine quality prediction [10].

5. Conclusion

In conclusion, this essay showcases a comprehensive approach to predicting red wine quality using a range of supervised learning techniques. It highlights the importance of data processing, parameter selection, and feature selection for achieving accurate predictions. The use of cross-validation throughout the process ensures the robustness and reliability of the models. For future work, some techniques like SMOTE can be tried to generate more artificial data to make the dataset more balanced, and the process of dropping outliers can be dealt with in a better way.

References

[1] Chen, J., Li, R., & Yang, J. A Prediction Model for Quality of Red Wine through Explainable Artificial Intelligence. In 2022 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control, 2022: 268-272.

[2] Kumar, S., Agrawal, K., & Mandan, N. Red Wine Quality Prediction Using Machine Learning Techniques. In 2020 International Conference on Computer Communication and Informatics, 2020: 1-6.

[3] Mohana, R., Sharma, P., & Sharma, A. Ensemble Framework for Red Wine Quality Prediction. Food Anal. Methods, 2023, 16: 30-44.

[4] Aich, S., Al-Absi, A. A., Hui, K. L., & Sain, M. Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques. In 2019 21st International Conference on Advanced Communication Technology, 2019: 1122-1127.

[5] Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques. In 2018 20th International Conference on Advanced Communication Technology, 2018: 139-143).

[6] Sun, Y., Wong, A. K. C., & Kamel, M. S. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(4): 687-719.

[7] Tanha, J., Abdi, Y., Samadi, N., et al. Boosting methods for multi-class imbalanced data classification: An experimental review. Journal of Big Data, 2020, 7: 70.

[8] Dahal, K. R., Dahal, J. N., Banjade, H., et al. Prediction of wine quality using machine learning algorithms. Open Journal of Statistics, 2021, 11(2): 278-289.

[9] Kumar, V., & Minz, S. Feature selection: A literature review. SmartCR, 2014, 4(3): 211-229.

[10] Gupta, Y. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 2018, 125: 305-312.