Machine Learning on Wine Quality: Prediction and Feature Importance Analysis

Quanyue Xie

Department of Computer Science, College of Engineering, Oregon State University, Covallis, United States

xieq@oregontstate.edu

Abstract. Recently, wine has become a common drink in most people's homes, but most people have different opinions on the evaluation of wine quality. Artificial intelligence can provide a relatively fair assessment and help practitioners focus on certain features to improve wine quality. This study uses decision trees and random forests to learn and predict on wine datasets and investigate feature importance to derive the features that have the greatest impact on wine quality. First of all, this study deals with the original data reasonably, and uses the IQR method to remove some outliers, specifically the data of the first 0.09 and the last 0.09. Second, since the correlation between the two features of density and residual sugar is as high as 0.84, this study removes density to improve the final prediction accuracy. When using both the decision tree and random forest models, the parameters are debugged multiple times in this study, and the three results are retained in this paper. Finally, on the basis of random forest, this study analyses feature importance, and draws a bar graph and the ranking order of different feature importance. In the final result, the prediction accuracy of random forest is relatively higher than that of decision tree, because the random forest model optimizes the decision tree to some extent. In the study on feature importance, alcohol has the greatest impact on the quality of white wine, while the smallest feature is citric acid. This study adjusts the original data set and compares the accuracy of different models, focusing on the importance of features based on the random forest model.

Keywords: Wine quality, Machine learning, Feature importance.

1. Introduction

Currently, wine is a staple in many households, and the high quality of wine has become the goal pursued by the majority of people. Numerous elements, including but not limited to grape variety, place of origin, brewing technique, sweetness, aroma, etc., can be used to categorize the quality of wine into distinct grades. Although everyone has different feelings about the quality of wine, higher quality wines are still favoured by most consumers. In some countries, the grading of wine quality is very significant. For instance, France, the world's largest wine producer, has incredibly tight rules. Based on these regulations and years of experience, the French wine quality classification is very accurate, and the grades are directly proportional to the price. On the other hand, the quality level of the wine also constrains the manufacturer's production process in disguise. The raw materials for higher-quality wines are often produced in specific regions, and the production is scarce, and the process is complicated. Therefore, the quality of wine also reflects the process of production of raw materials. While years of experience and regulations can give some degree of accurate wine quality, manual measurements results are often varying from person to person. With the rapid development of technology today, artificial intelligence can be considered as an effective alternative in this case. Artificial intelligence does not have personal tastes and sensory differences. It will only be classified by different characteristics, and will make a final judgment after learning from tens of thousands of data. Compared with manual measurement, the application of the artificial intelligence in this field may bring greater potential.

In the past, the quality of wine was determined by many factors, including but not limited to grape origin, variety, yield, viticulture method, wine brewing method, ethanol content, etc. However, manual measurement and judgment are relatively inaccurate, and everyone may have a certain bias in the taste and perception of wine, and artificial intelligence happens to be able to solve this problem.

In 2021, Dahal et al. predicted the quality of wine in the field of machine learning using many algorithms [1], e.g. Ridge [2], SVM [3], Gradient Boosting Regressor [4], and neural network algorithms [5]. In 2022, Bhardwaj et al. used random forests, SMOTE, etc. to make quality predictions on smaller wine datasets [6]. In 2020, Gupta et al. used algorithms such as random forest and KNN to predict the quality of wine [7]. However, in the above paper, although the accuracy of the results is very high, the question of which features can have a greater impact on the results is not mentioned. Feature importance is a non-negligible issue to identify the most important features for further research, or to focus on those features.

To solve this problem, this study downloaded about 5, 000 white wine data on kaggle, this data set shows 11 features that may affect the quality of white wine, and divides the quality into 11 grades from 0-10. First, random forest was employed to classify and predict the data, and 70% of the data are used in the training process. After this, the remaining 30% of the data was utilized to make predictions to see whether the learned accuracy matched the data. In this process, the accuracy rate is 0.667. To test which feature have the greatest impact on the quality of white wine, feature importance obtained from the random forest is also investigated. If there are features with higher weights, that feature will affect the quality of white wine the most.

2. Method

To find a faster and more accurate solution, this paper uses the dataset on Kaggle called White Wine Quality [8]. According to the context of this dataset, the privacy-related part of this dataset is not included, and all its features and results are physical. This dataset has a total of 11 features, namely fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. These 11 features will affect the quality of white wine which is the output variable in this dataset called quality. The quality is divided into 11 grades from 0 to 10. The lower the grade, the worse the quality of the wine; correspondingly, the higher the grade, the better the quality. Table 1 presents a part of the sample data.

Table 1. The sample data in the collected wine dataset

Feature name	Sample 1	Sample 2	Sample 3
fixed acidity	7.0	6.3	8.1
volatile acidity	0.27	0.30	0.28
citric acid	0.36	0.34	0.40
residual sugar	20.7	1.6	6.9
chlorides	0.045	0.049	0.050
free sulfur dioxide	45.0	14.0	30.0
total sulfur dioxide	170.0	132.0	97.0
density	1.0010	0.9940	0.9951
pH	3.00	3.30	3.26
sulphates	0.45	0.49	0.44
alcohol	8.8	9.5	10.1

This dataset has 4, 898 data and all data are numbers without any blank entries and strings. However, during the initial processing of the data, each feature of this dataset has a small number of outliers. These outliers are too small or too large. In order to avoid outliers from affecting the final results, this paper makes the following treatments. First, making a copy of the dataset to avoid corruption of the original data, and store it in a new variable. In addition, the drop method is used, combined with the Interquartile Range (IQR) algorithm, to remove about the first 0.09 and the last 0.09 of all features. The total number of discarded data is 925, leaving 3, 973 data. However, such data processing only solves the most basic problems. In this paper, after calculating the correlation between 11 features, the correlation coefficient between density and residual sugar reached 0.84. This is a number very close to 1, and two features with higher correlation may lead to the final result is inaccurate. Therefore, on the basis of drop outliers, the density feature is also discarded using the drop method. After two data processing, there are a total of 10 features and 3, 973 data in the current dataset.

In machine learning, decision trees are a very common and efficient algorithm. As the name suggests, a decision tree is a tree-like flow chart with nodes and branches, where each internal node represents an attribute judgment, each branch represents a different judgment possibility, and each leaf node represents a different classification result. According to research investigated by Navada et al., the logic of the decision tree is very intuitive, the patterns of its input data form internal nodes, and the categories of these patterns form the leaf nodes of the decision tree [9]. It can be seen from this, decision tree algorithm is very effective in this dataset. After learning, it can quickly classify wines into different qualities based on their features. In addition, random forest is also one of the very common algorithms in machine learning. Its essence is to combine many decision trees and finally obtain a single judgment result. According to research investigated by Liu et al., random forest combines a certain number of decision trees, and votes based on the results, and finally calculates the result with the most votes [10]. So, random forests will have higher accuracy than decision trees to a certain extent. In decision trees, sklearn is a very common and efficient tool. In this paper, the first 70% of the dataset will be used for learning, the decision tree will categorize and remember how different features lead to different outcomes, i.e. which branch should be chosen at each internal node. After training, the remaining 30% of the data will be used for testing and prediction. The decision tree determines which branch to enter by examining the values of the sample features. Upon reaching the leaf node, the decision tree produces the final conclusion, which in this dataset is the quality of white wine. In random forest, the model will select random samples from the dataset and build a decision tree for each distinct sample. In this process, the learning process and prediction process of the model are the same as the decision tree. After each decision tree predicts the result, the random forest model summarizes all the results and generates the majority of the results as the final result. On the basis of decision trees and random forests, feature importance appears as a very important algorithm in order to obtain the characteristics that can most affect the final quality. Feature importance can calculate a different score for each feature, and the score represents the degree of influence of each feature on the result, that is, importance. The higher the score, the greater the influence of the feature, and vice versa.

3. Result and Discussion

3.1. Performance of decision tree and random forest

Table 2. The accuracy of decision tree with three different parameters

Test size	Random state	Max depth	accuracy
0.3	42	24	0.5973
0.2	35	30	0.5825
0.4	50	20	0.5791

After processing the raw data, this study starts to train the machine learning model. First of all, Table 2 is the different parameters and accuracy of the data after using the decision tree model. Here, three different parameters can play a certain role in the data, namely test size, random state and max depth. In the case of test size=0.3, random state=42 and max depth of 24, the accuracy is higher, which is 0.5973. In addition, this study makes certain adjustments to these three parameters, and tries to increase or decrease each parameter. After adjusting the parameters, although the accuracy is not significantly different from the original results, it still shows a certain downward trend, which are 0.5825 and 0.5791.

Table 3. The accuracy of random forest with two different parameters

Test size	n_estimators	accuracy
0.3	120	0.6661
0.2	100	0.6579
0.4	150	0.6458

Table 3 is the result of using random forest to learn the data. This study uses two parameters to adjust the final accuracy, namely test size and n_estimators. When test size=0.3 and n_estimators=120, the accuracy reaches 0.6661, which is the best result so far. Also, this study adjusts for test size and n_estimators, but also has a very small downward trend in accuracy, which are 0.6479 and 0.6458 respectively.

3.2. Feature importance of Random Forest

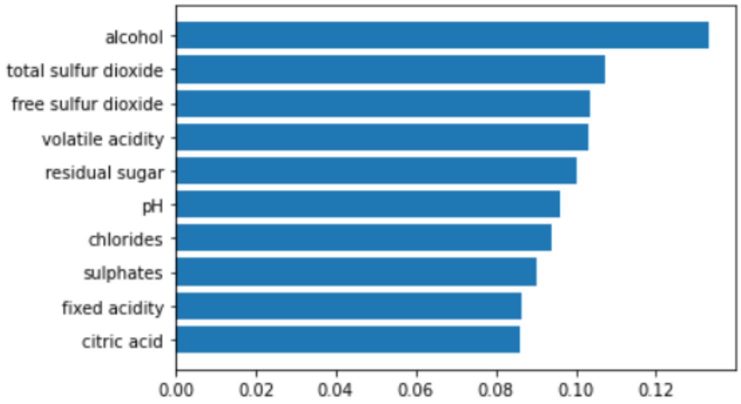

Figure 1. The feature importance of the random forest

Figure 1 shows the importance of different features for wine quality. According to the chart, the importance of these features to wine quality is ranked in descending order: alcohol, total sulfur dioxide, free sulfur dioxide, volatile acidity, residual sugar, pH, chlorides, sulphates, fixed acidity, citric acid. The importance of alcohol goes well beyond the other features in this data, which may have the biggest and most intuitive impact on a wine's quality. As the alcohol ratio increases, the quality of the wine may be higher, and vice versa. The least important is citric acid, a feature that has little impact on wine quality. Compared with alcohol, citric acid will have much less impact on wine quality. Feature importance can give the coefficients of different features that have an impact on the model results. The larger the coefficient, the greater the impact. For example, the Gini coefficient in feature importance, which is based on the importance of each node, and the number of times the data arrives at the node when training the model, the impurity value of the node, and other nodes around the node to get the final value. This value is the coefficient in the feature importance.

3.3. Discussion

Both decision tree and random forest are very effective models for data learning and prediction. However, the accuracy of the decision tree is lower than that of the random forest, because the random forest combines the results of multiple decision trees, and improves the accuracy again on this basis. However, decision tree and random forest can only predict the data, and cannot reflect the influence of different features on the results, and feature importance can solve this problem well. For this data, feature importance shows the different importance of the 10 processed features to the results, which also provides direction and value for follow-up research. For the most important feature of alcohol, follow-up research can try to increase a certain percentage of alcohol in the wine making process, or focus on this feature, which can improve the quality of wine to a certain extent, while for less important alcohol citric acid, the research value is lower than alcohol.

4. Conclusion

This paper uses different models to study the relationship between the features and wine quality, judges the wine quality according to ratios of different features, and studies the influence of different features on wine quality. This paper takes the white wine quality data set on kaggle as an example. After reasonable processing of the data, the decision tree and random forest are used to learn part of the data, and the quality of wine is successfully predicted according to the learning results. In addition, this paper also studies feature importance on the basis of random forest. This method reflects the importance of different features to wine quality and ranks them. In the decision tree, the result with the highest prediction accuracy is 0.5973, and due to the multiple learning of the model by random forest, the accuracy is higher, which is 0.6661. After ranking the feature importance, alcohol has the greatest impact on white wine, and the feature with the least impact is citric acid. This paper did not obtain a higher prediction accuracy at the moment. In the future, this paper will further analyse and process the data to improve the prediction accuracy.

References

[1] Dahal K R et al. 2021 Prediction of wine quality using machine learning algorithms Open Journal of Statistics 11.2 278-289.

[2] Dobriban E et al. 2018 High-dimensional asymptotics of prediction: Ridge regression and classification The Annals of Statistics 46.1 247-279.

[3] Yue S et al. 2003 SVM classification: Its contents and challenges Applied Mathematics-A Journal of Chinese Universities 18.3 332-342.

[4] Prettenhofer P et al. 2014 Gradient boosted regression trees in scikit-learn PyData 2014.

[5] Yu Q et al. 2019 Semantic segmentation of intracranial hemorrhages in head CT scans 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE.

[6] Bhardwaj P et al. 2022 A machine learning application in wine quality prediction Machine Learning with Applications 8 100261.

[7] Gupta U et al. 2020 Wine quality analysis using machine learning algorithms Micro-Electronics and Telecommunication Engineering Springer Singapore 11-18.

[8] Kaggle 2020 White Wine Quality https://www.kaggle.com/datasets/piyushagni5/white-wine-quality?select=winequality-white.csv.

[9] Navada A et al. 2011 Overview of use of decision tree algorithms in machine learning 2011 IEEE control and system graduate research colloquium IEEE.

[10] Liu Y et al. 2012 New machine learning algorithm: Random forest International Conference on Information Computing and Applications Springer Berlin Heidelberg.