White Wine Quality Prediction and Feature Importance Analysis Based on Chemical Composition and Machine Learning Models

Jialiang Yan*

Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0002

*Corresponding author email: jialiang@vt.edu

Abstract. Consumers tends to purchase white wines based only on their taste and price due to the difficulty of studying the composition of white wine. Popular classifications of white wines are usually depending on some easily understandable aspects such as carbon dioxide pressure and grape harvest time. A detailed way to classify the quality of the white wines is needed for both the consumers and the market regulators. In this research, a white wine dataset with 11 parameters and a final quality value is being used to train the machine learning models for future prediction. To avoid extreme values influencing the dataset, this paper used the Interquartile Range method to remove the outliers. After processing the data, six machine learning models were applied to the dataset to test the initial accuracies of the models. The Random Forest had the best accuracy among all the models. Then the focus of the research turned into the feature importance of the Decision Tree and Random Forest methods. The project found out that it is possible to remove one of the parameters from two parameters that have similar importance while maintaining almost the same model accuracy. Both models' parameter number were reduced to nine instead of 11 at the cost of less than 3% of accuracy. This provides people a useful way to make their analyzing processes easier in machine learning research.

Keywords: white wine, machine learning, random forest, data set

1. Introduction

With the improvement of the quality of human life, wine has become a daily consumable rather than scarcity and luxury. After reaching the legal age, everyone could buy all kinds of wines easily from grocery stores, restaurants, and retail shops. According to the U.S. Wine Institute, 1.1 billion gallons of wine were consumed in 2021 with over 3 gallons per resident per year [1]. The total sale of the wines reached an all-time high of $45.6 billion in the same year [2]. Lay consumers mostly only care about the price and the taste of the wine when purchasing while neglecting many other compositions of wine. Thus, it is tough for the citizens to identify the quality of the wine. People could only rely on the government and related organizations to help monitor the whole marker. Due to the complex compositions of wines, categorizing and classification of wines become very important for organizations to monitor their quality. Rough classifications based only on alcohol concentration, or the color of wine are definitely not feasible at a high level.

According to Paul and Isak's research on wine history, winemaking starts over 7000 years ago [3]. After a long period of development and evolution, hundreds of different wines were invented and gradually became a part of human life. Prior to this point, most wines were categorized based on their taste, color, sugar content, ethanol concentration, and so on for easy marketing purposes. For example, based on the carbon dioxide pressure, wines are separated into still and sparkling wines; based on the grape harvest time, they are divided into ordinary, late harvest, and noble rot wines [4]. All these classifications were made for the lay consumers to understand the features of the wines easily. People usually do not have the time and knowledge to fully understand the components of wines even if they bother to look carefully at the bottles. Suppose there is some kind of classification method that can calculate the quality level of the wine using the values of specific compositions in the wines. In that case, it will be much easier for the market to supervise the industry and for the consumers to make

their choice when they are facing hundreds of brands of wines. Others have done some similar research about the classification of wines. Yogesh Gupta posted a paper about predicting wine quality with some important features in 2017. He used statistical analysis instead of the feature importance of the data during the research, which could lead to a less accurate result [5]. Paulo Cortez and his team also did research on predicting the quality value using some parameters. However, they did not include the feature importance analysis to have a more comprehensive look at the dataset [6].

In the absence of a comprehensive approach to evaluating how good a wine is, people cannot take every aspect into account when choosing the wine. In this research, the detailed composition values of white wines will be investigated and analysed to give each wine an evaluation of its quality. After displaying the values, it will be easy to detect the wine with some extreme chemical composition contents. These are the unqualified products or products that had some unexpected errors in machining. Since the goal of this research is to train some machine models to predict the quality value of the wine, these outliers will be removed from the dataset for a more accurate result. With the quality value predicted for each wine, producers can price their products accordingly and see what quality of wine consumers prefer to buy; Consumers can easily avoid wine with bad quality and pick their choices based on their standards and economic capabilities; The relevant institutions could monitor the market conveniently by looking at the unqualified rate of each wine brand and if the products' prices match their quality.

The content of the paper is organized as follows: Section 2 will introduce the dataset being used in this research and what models will be applied. Then the paper is going to focus on pre-processing the dataset and using different models to predict the quality value of white wines in Section 3. Section 4 summarizes the whole paper and presents some final findings.

2. Method

2.1. Project Workflow

In this project, the raw data will be processed using the Interquartile Range method to get a better distribution. Then, the machine learning models could be applied to the processed dataset to have an initial investigation on the performance of each model. After that, the focus of the research will be moved to the Decision Tree and Random Forest methods to see if removing some parameters based on feature importance is a valid method to make the research process easier. Figure 1 shows the complete workflow of this research

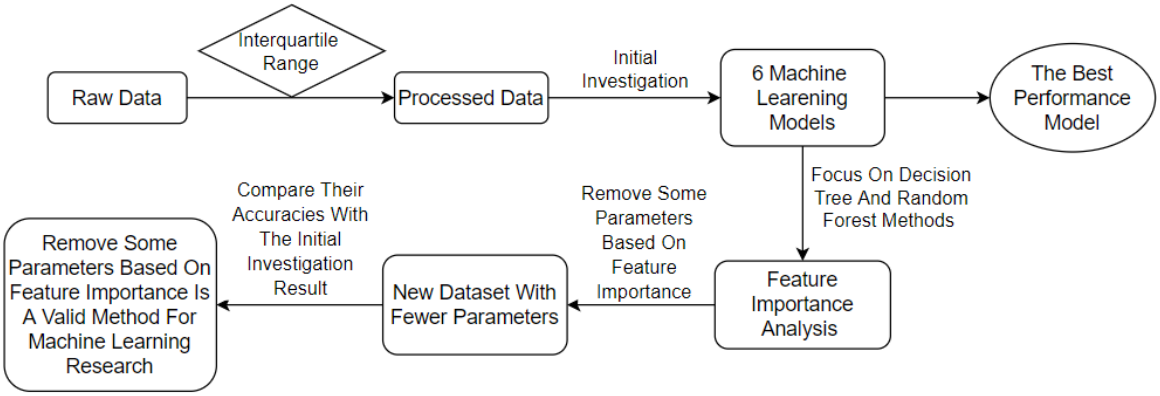

Figure 1. Workflow of the Complete Method

2.2. Dataset Description and preprocessing

The dataset used in this research is called "White Wine Quality" from Kaggle [7]. It contains over 4800 different white wine data from the market. For each data, the values of 11 chemical data of the white wine were recorded as the parameters for the research. The parameters are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH,

sulphates, and alcohol. There is a final quality value calculated based on the parameters for each white wine. The range of quality lies between 0 and 10. Then people could know the quality of the white wines based on the detailed chemical compositions easily from this value. Several models will be applied to the dataset to predict the quality value for other white wine data. Table 1 presents the parameters and some sample data from the dataset.

Table 1. Dataset Variables and Data Examples

	Example 1	**Example 2**	**Example 3**	**Example 4**
Fixed Acidity	7.0	6.3	8.1	7.2
Volatile Acidity	0.27	0.30	0.28	0.23
Vitric Acid	0.36	0.34	0.40	0.32
Residual Sugar	20.7	1.6	6.9	8.5
Chlorides	0.045	0.049	0.050	0.058
Free Sulfur Dioxide	45.0	14.0	30.0	47.0
Total Sulfur Dioxide	170.0	132.0	97.0	186.0
Density	1.0010	0.9940	0.9951	0.9956
pH	3.00	3.30	3.26	3.19
Sulphates	0.45	0.49	0.44	0.40
Alcohol	8.8	9.5	10.1	9.9
Quality	6	6	6	6

After acquiring the dataset, it was displayed using Histplot to reflect its data distribution. As shown in the plots, there are some extreme values in several categories that only appear very few times (Figure 2). Data like those will influence the classification a lot while holding little to no value. This paper used the Interquartile Range method to remove the extreme values. The lower bound is 1.5 times quartile 1 and the upper bound is 1.5 times quartile 3. This range includes all the data that contributes the most and needs to be focused on. Neglecting those extreme values can make the models only pay attention to the data that are important and avoid errors being made during data collection. 925 data that are outside of this boundary were removed as outliers. The comparison of the data after the processing is shown in Figure 3.

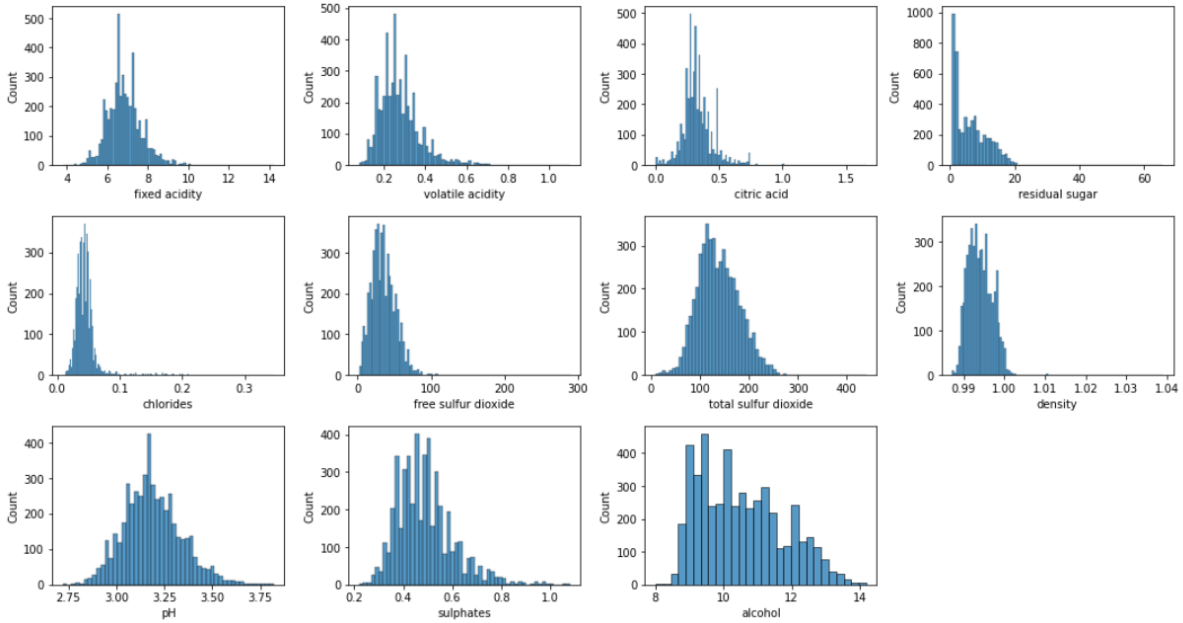

Figure 2. Raw Data Distribution

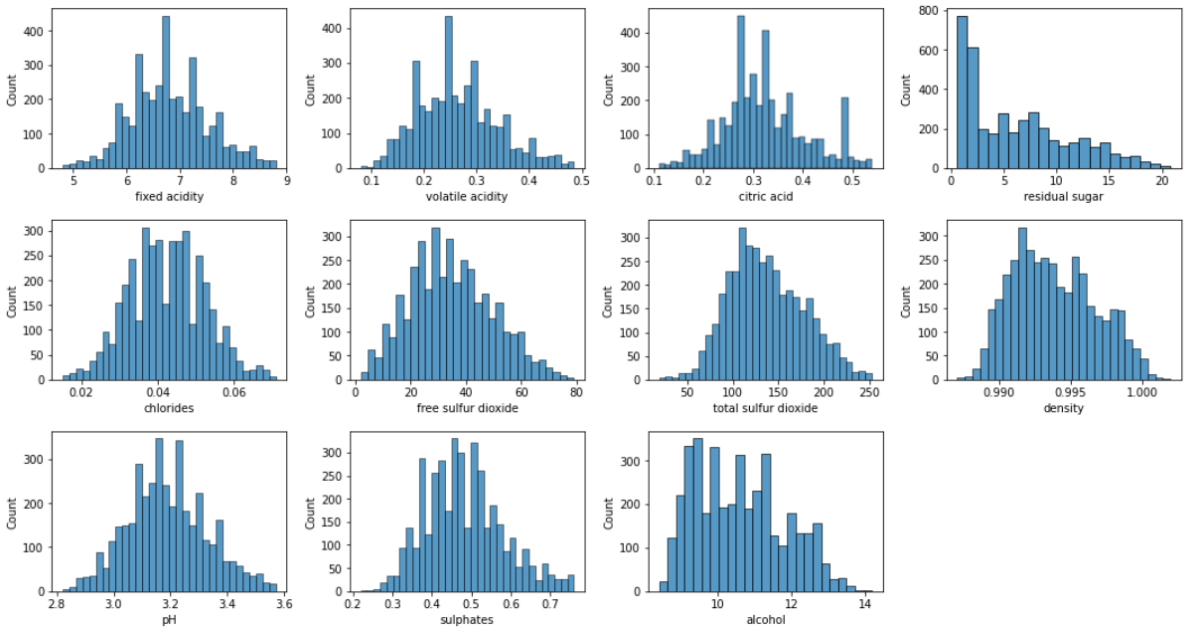

Figure 3. Processed Data Distribution

2.3. Machine Learning Models

In this project, multiple machine learning models will be applied to the dataset to make predictions about the quality value of other white wines. This research will primarily focus on the Decision Tree model and the Random Forest Classifier, but it will also include Gaussian Naive Bayes, Support Vector Machines, Logistic Regression Classifier, and K Neighbors Classifier. After applying all these six models to the dataset, the comparison of the accuracies of the models will be displayed to decide which model performs the best. All the models used in this project are imported from the scikit-learn library.

The Decision Tree model is a non-parametric supervised learning method that is used for classification. It takes the features of the dataset as input and infers decision rules to predict a target value. The decision tree is a very easy-to-use model that can be used in machine learning problems [8]. It has a parameter called criterion, which is set to “entropy” in this research. Random Forest

Classifier is an estimator that uses many decision tree classifiers on different subsamples of the dataset. It then uses averaging to get better overall accuracy. This model has a parameter n_estimators, it is set to 100 on default [9]. To ensure accuracy and convenience, this research set the n_estimators equal to 100.

The Naive Bayes method is a supervised learning algorithm that predicts target values based on Bayes' theorem. The Support Vector Machines model is very good at performing multi-class classifications on a dataset. The Logistic Regression Classifier can assign classifications to the dataset based on the observation and return the probability using the logistic function. The K Neighbors Classifier is an algorithm that predicts values by focusing on storing the data instead of trying to build a model to fit the dataset [10].

This research will apply all these models to the processed white wine data mentioned earlier to predict the quality value. The dataset will be split into a training set and a testing set by using the "train_test_split" function from the sklearn library. 70% of the data is used for training while leaving the other 30% to test the accuracies of the models. After getting an initial look at the accuracies, the focus will be switched to the Decision Tree model and the Random Forest Classifier. The paper decided to investigate the feature importance of different parameters used to train the models. After calculating and visualizing the importance of every parameter for each model, the project will focus on if it is possible to remove specific parameters while maintaining similar accuracy.

3. Results and Discussion

3.1. Result

First, six machine learning models were applied to the processed dataset to have an initial look at their performance. All of them used the 11 parameters to predict the quality value for the test dataset. The initial accuracies are as follows: Decision Tree 58.1%, Random Forest 67.9%, Naive Bayes 45.8%, Support Vector Machines 47.5%, Logistic Regression 47.4%, and K Neighbors Classifier 49.9%. The Random Forest classifier is the best model for the dataset. Figure 4 shows the comparison of the accuracies of the models.

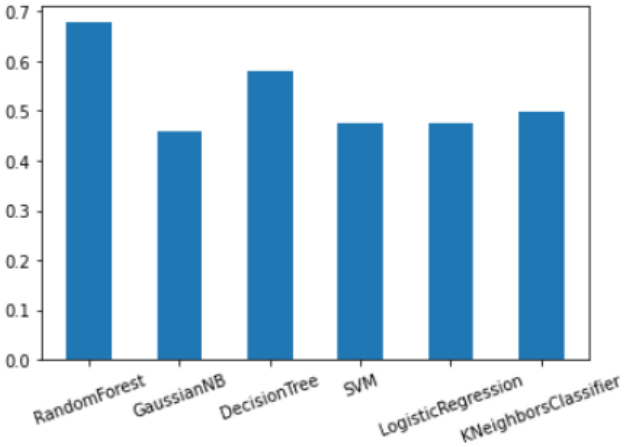

Figure 4. Initial Accuracy Comparison

Second, the research focused on the feature importance of Decision Tree and Random Forest methods to see the contribution of different parameters in each model. Figure 5 shows the feature importance of the Decision Tree model and Figure 6 shows the feature importance of the Random Forest model.

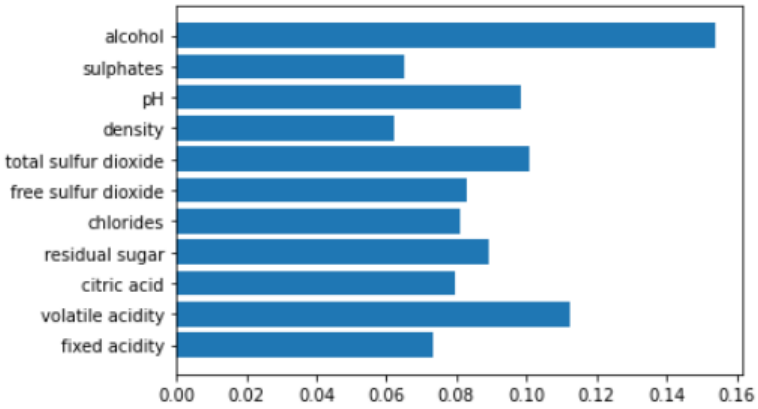

Figure 5. Decision Tree Feature Importance

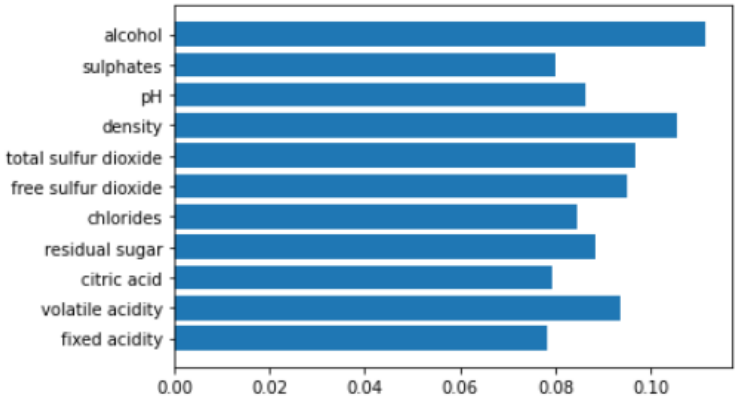

Figure 6. Random Forest Feature Importance

As shown in the figures, alcohol value played the most important role in both models. However, its importance does not stand out that much in the Random Forest model compared to the Decision Tree. All other parameters in the Random Forest model made a similar contribution to the quality value with a difference of less than 2%. In the Decision Tree model, the importance of alcohol is very apparent with a lead over other parameters of more than 4%.

After figuring out the feature importance of these two models, the research moved its focus to find out if it is possible to maintain similar accuracies of the models while removing some of the parameters based on their feature importance.

For the Decision Tree model, citric acid and fixed acidity, sulphates and total sulfur dioxide are two pairs of parameters with almost the same feature importance. Thus, citric acid and sulphates were dropped from the dataset. After applying the same model, the accuracy was 57.8% compared to the initial 58.1%. For the Random Forest model, sulphates and citric acidity, residual sugar and chlorides are two pairs of parameters with similar feature importance. Sulphates and residual sugar were dropped to get a model accuracy of 65.6% compared to the initial 67.9%.

3.2. Discussion

From the initial accuracies of the six models, it is not hard to tell that these models did not perform that well in predicting the quality value for this dataset. None of the models has an accuracy higher than 70%.

The study about the feature importance of the Decision Tree and Random Forest models proved a very effective research method and may benefit future machine learning projects. When having many different parameters in the dataset, it may take a while for the models and the machines to process and predict the data. One possible way to save some time without losing much accuracy is to remove some of the parameters that have similar feature importance as other remaining parameters. In this

research, both models decreased the number of parameters from 11 to 9 at the cost of less than 3% accuracy.

4. Conclusion

In this project, six machine learning models were applied to the processed white wine dataset to predict the quality value. Then the paper focused on the feature importance of the Decision Tree and Random Forest models to find the degree of influence of various variables on the results. The research found that it is possible to reduce the number of parameters used without losing much accuracy based on similar feature importance. In the future, more in-depth research is needed to find better models that fit this white wine dataset or better ways to adjust the parameters. On the other hand, processing the dataset with some better methods before applying the models may also boost the accuracy.

References

[1] Wine institute 2022 US Wine Consumption https://wineinstitute.org/our-industry/statistics/us-wine-consumption/

[2] Wine institute 2022 California US Wine Sales https://wineinstitute.org/our-industry/statistics/california-us-wine-sales/

[3] Chambers P J et al. 2010 Fermenting knowledge: the history of winemaking, science and yeast research. EMBO Rep 2010 Dec;11(12):914-20. doi: 10.1038/embor.2010.179

[4] Olavin 2022 The Ultimate Guide: How Wine is Classified? https://olavin.com/how-wine-is-classified.html

[5] Yogesh G 2018 Selection of important features and predicting wine quality using machine learning techniques Procedia Computer Science Volume 125 Pages 305-312

[6] Paulo C et al. Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems Volume 47 Issue 4 2009 Pages 547-553

[7] Kaggle 2020 White Wine Quality https://www.kaggle.com/datasets/piyushagni5/white-wine-quality

[8] Sklearn Tree 2022 https://scikit-learn.org/stable/modules/tree.html

[9] Sklearn 2022 Random forest classifier https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[10] Guo G 2003 et al. KNN model-based approach in classification. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems. Springer, Berlin, Heidelberg