UTILIZING MACHINE LEARNING CLASSIFICATION MODELS TO PREDICT WINE QUALITY

Heather Daries

Northwestern University, Practical Machine Learning

August 25, 2024

1. Executive Summary

In the highly competitive wine industry, quality has become a crucial factor for winemakers to consider when creating and distributing their products. The ability to predict wine quality based on its chemical features can be vital in maintaining and improving superiority of the product, ensuring it continues to be poured into the glasses of wine enthusiasts across the globe. In this study, we explore a dataset containing information on various physicochemical characteristics of wine in an effort to accurately predict quality.

The objectives of this study include developing predictive models to forecast wine quality as well as analyzing feature importance to understand the key drivers of quality. In the process, evaluation and comparison will be conducted on different machine learning classification algorithms and model interpretability will be prioritized for stakeholders.

The analysis intends to find the physicochemical properties that have a significant impact on wine quality and leverage machine learning techniques to develop a model that will accurately predict quality based on the analyzed features. Appropriate metrics will be utilized to evaluate the model and optimization will be conducted where opportunities present.

Overall, this study will provide valuable insights for winemakers and stakeholders within the wine industry. By understanding the factors that most heavily influence wine quality and leveraging predictive modeling, winemakers will be able to make informed decisions to produce high quality wines that meet consumer expectations and maintain a competitive edge in the market.

2. Problem Statement & Research Objective

The problem addressed in this research revolves around the winemaker's ability to predict product quality based on physicochemical properties of wine. The dataset used for analysis contains various physicochemical features, including information about fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. The primary research objective is to build machine learning classification models that can accurately predict the quality of wine based on these features and help inform winemaking companies on how to best create their products in an effort to keep sales and profits strong.

3. Exploratory Data Analysis

The data being used in our analysis comes from the UC Irvine Machine Learning Repository wine quality data. The data consists of 4,898 red and white wine vinho verde samples produced in the north of Portugal. For this analysis, only the white wine dataset will be utilized, which contains 1,599 samples. Data was collected on the 12 properties listed above. The quality feature is based on sensory data, while the remaining features are chemical properties of the wines. All chemical properties are continuous, while quality is ordinal with 1 being the worst and 10 being the best.

To start data exploration, distribution was examined for each property. Results show a right skew for most features with alcohol and pH being the only normally distributed variables. The most common quality ranks were 5, 6, and 7.

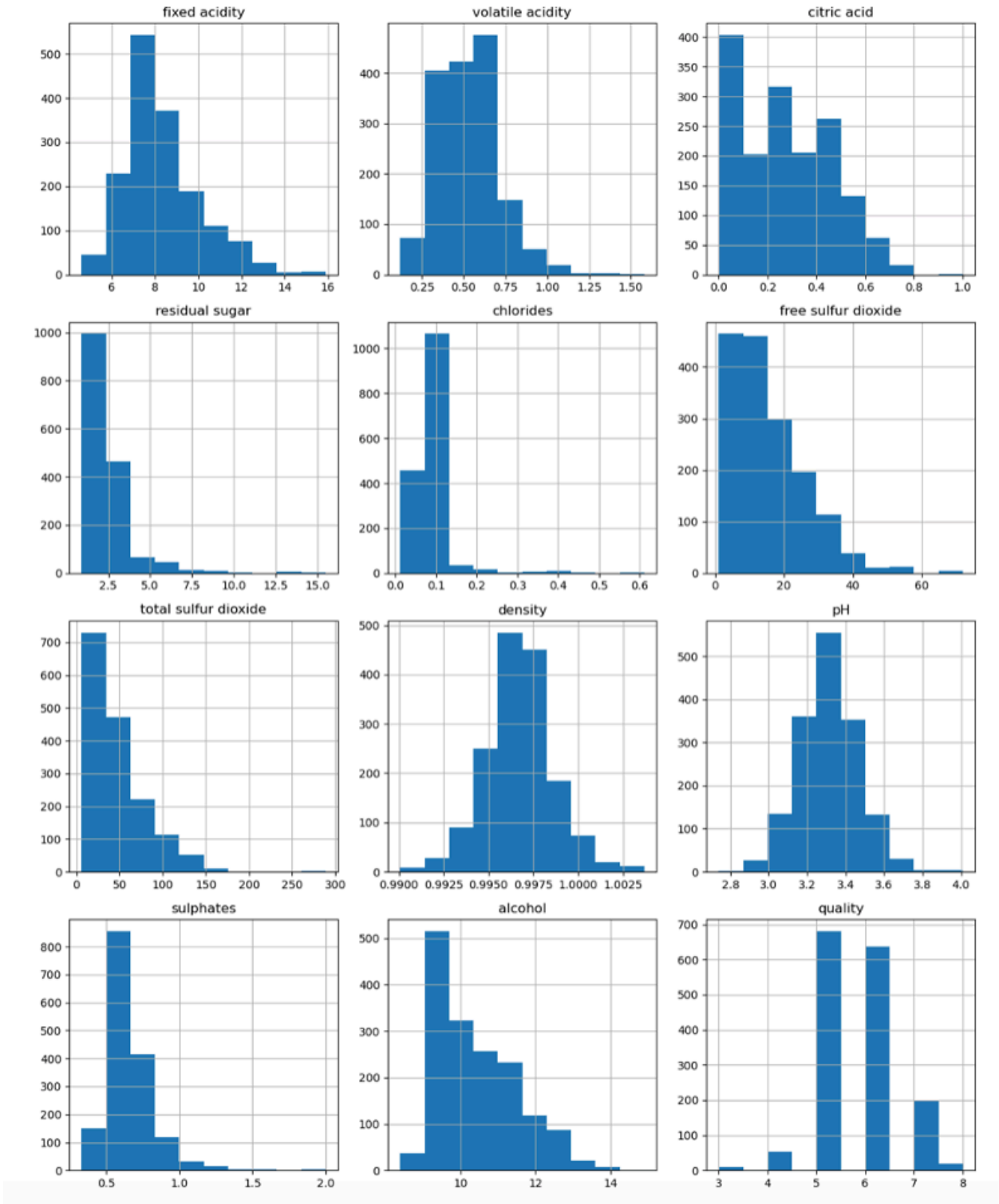

Figure 3.1: Distribution of wine features

Boxplots were then created for all properties to examine outliers, as well as descriptive statistics confirming the distribution of data shown in the histograms above. It can be observed

from the boxplots that there were extreme outliers present in many of the properties that required consideration in the following data preparation steps.

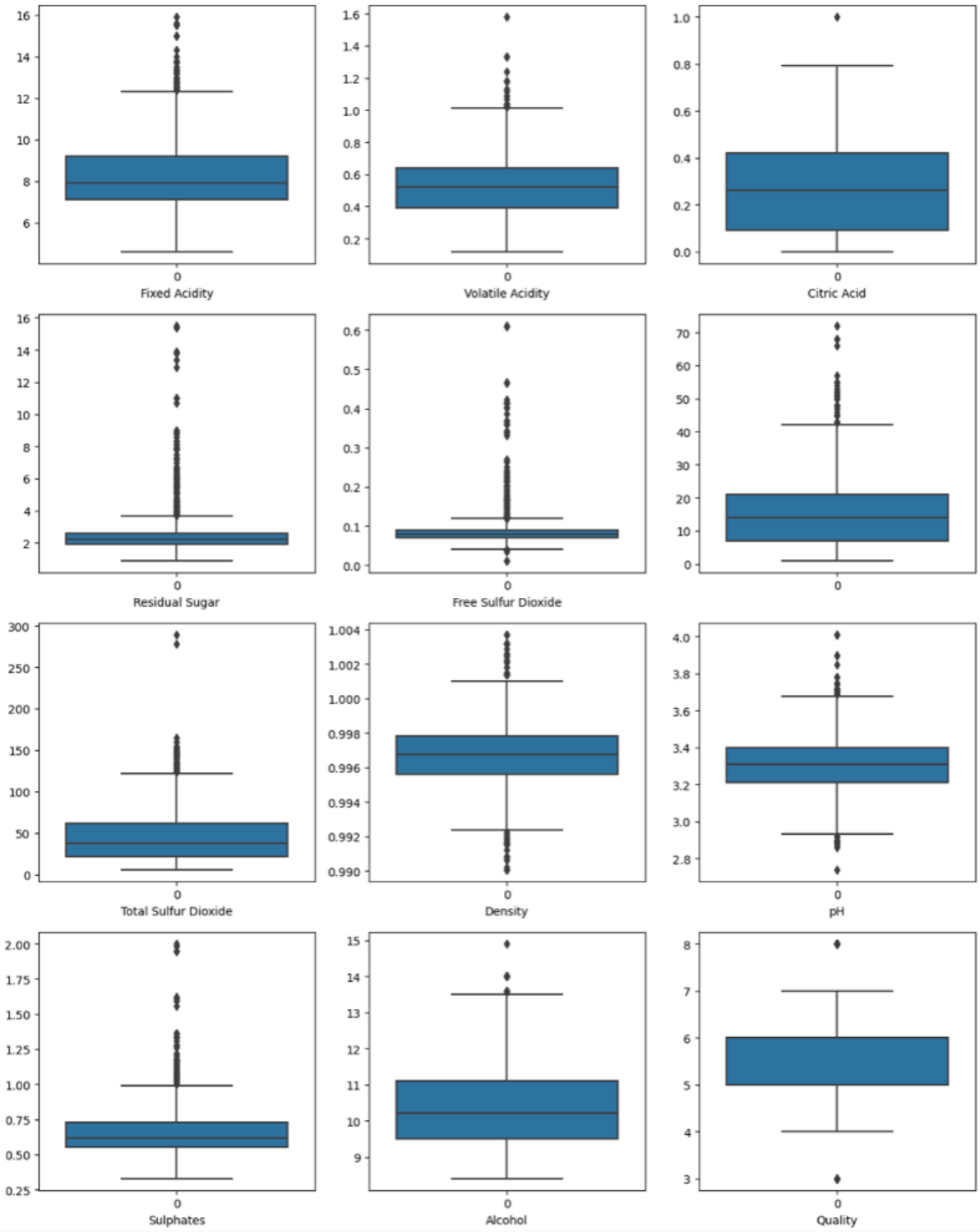

Figure 3.2: Boxplot showing outliers present in wine features

Finally, a correlation matrix was completed to see the relationships that existed between the variables, specifically those correlated with quality. The matrix showed alcohol, sulphates, citric acid, and fixed acidity to be most positively correlated with quality scores. Total sulfur dioxide, density, chlorides, and volatile acidity showed to be most negatively correlated with quality scores.

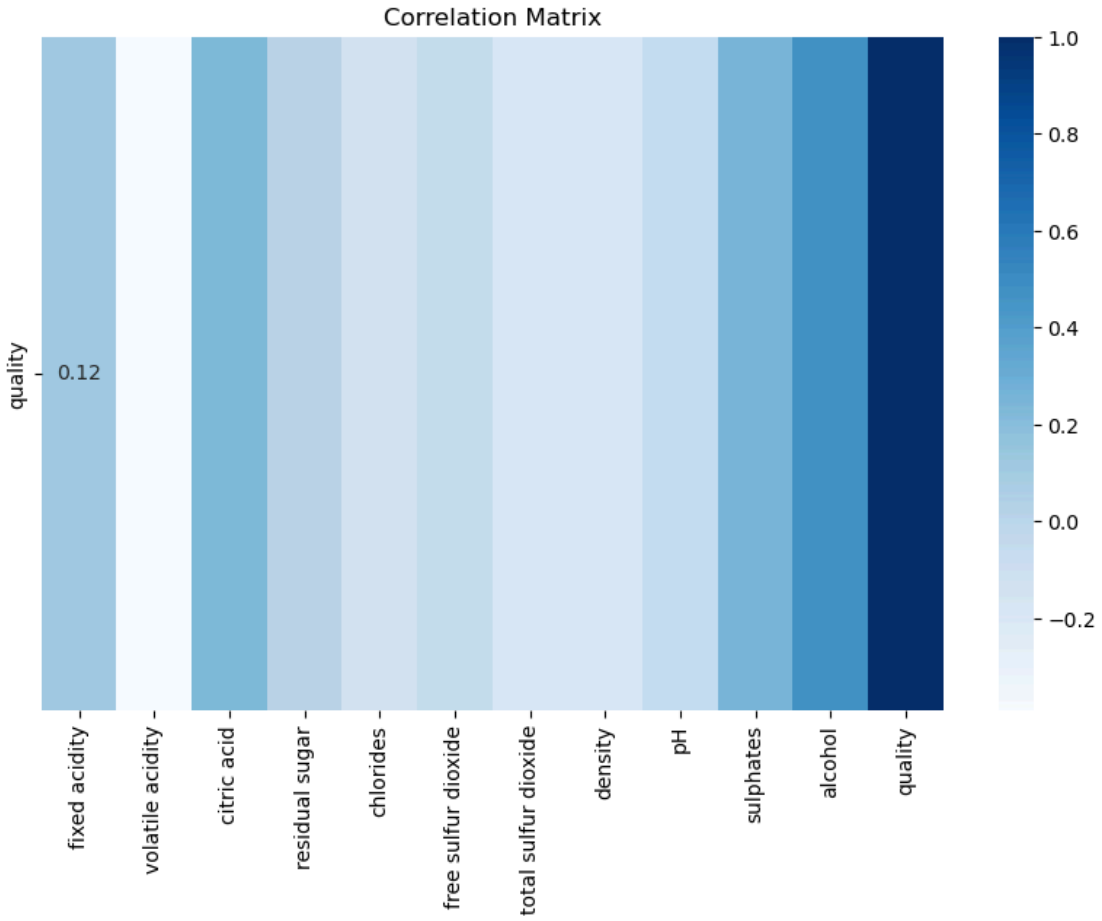

Figure 3.3: Correlation matrix showing correlations between wine features and quality

These steps helped to gain sufficient insight into the distribution of data and prepared for the next steps of data processing.

4. Data Preparation and Feature Engineering

The dataset presented with all numerical values and no null fields. This eliminated any need to encode data or fill in missing values. As mentioned above, there were several features of the data set that were right skewed along with extreme outliers present. Outliers were only removed from the quality feature in an effort to improve model accuracy during training. This step removed 28 wines from the dataset resulting in a total of 1,571 white wines used for analysis. Following this, wine quality ratings were classified into a binary variable to further improve model accuracy. Wines with a rating between 1 and 5 are considered to be poor wines, while wines with a rating of 6 and over are considered to be good wines. This step resulted in 734 poor wines and 837 good wines.

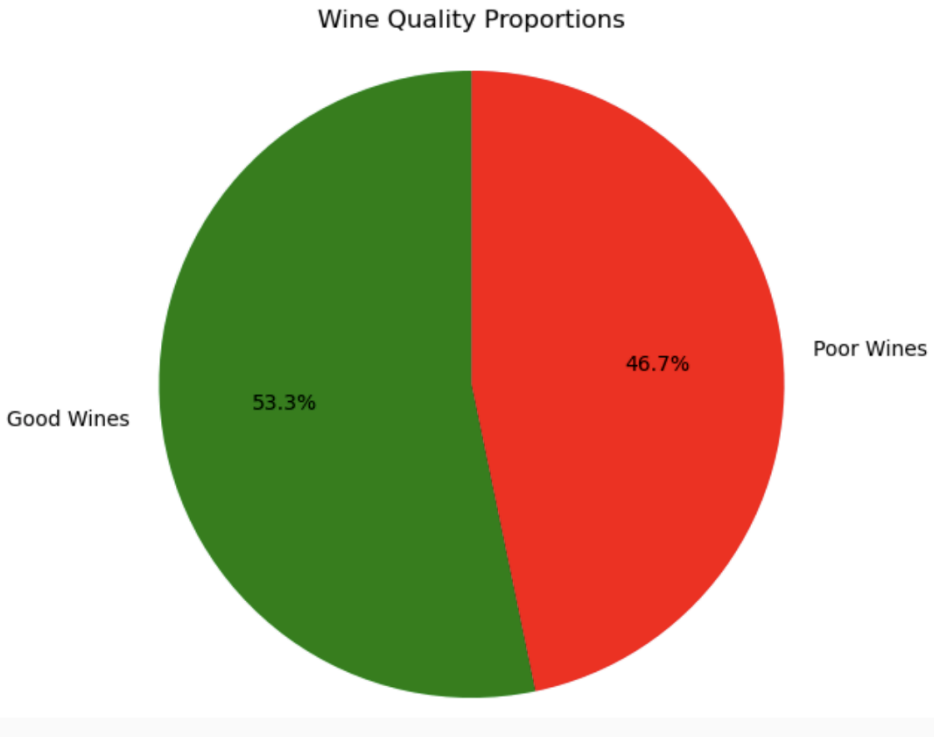

Figure 4.1: Pie chart of good/poor wine quality proportions in the dataset

Additionally, three new features were created from the existing properties. This was mainly done to capture interactions between variables in an effort to help the model learn more

complex relationships in the data. The three features that were created were alcohol sulphate interaction, acidity alcohol interaction, and sugar acidity balance. Alcohol and sulphates were two of the most positively correlated variables with quality, so the alcohol acidity interaction feature represents the relationship between alcohol content and sulphates in the wine. Fixed acidity was another significant positively correlated variable, so the acidity alcohol interaction feature was created to represent the relationship between acidity level and alcohol content in determining wine quality. Lastly, the sugar acidity balance feature was created to represent the balance between acidity and sweetness in the wine to see how that impacts the quality.

An additional correlation matrix was created to see how the new features relate to the quality variable. Unsurprisingly alcohol sulphate interaction and acidity alcohol interaction were the most positively correlated features with quality. This makes sense since they were created from three of the most highly correlated variables. The sugar acidity balance feature does not show to be as highly correlated, but could still reveal insights on wine quality.

5. Methodology

For this analysis, four classification models were built and deployed using the sklearn library in Python. The four models selected were Random Forest Classifier, Logistic Regression, Gradient Boosting Trees, and Linear Discriminant Analysis. Random Forest Classifier was chosen for its ensemble learning approach that combines multiple decision trees to improve predictive accuracy and reduce overfitting. Logistic Regression was chosen for its simplicity and interpretable classification algorithm that is well suited for binary classification problems, such as this. Gradient Boosting Trees was chosen for similar reasons as the Random Forest Classifier in addition to its ability to capture complex interactions between the input features and the target

variable. Lastly, Linear Discriminant Analysis was chosen for its dimensionality reduction and computational efficiency that makes it suitable for high-dimensional datasets with linear decision boundaries.

All four classification models were deployed on two different feature sets - the initial 12 features in the dataset and the initial features with the addition of the 3 added features mentioned previously. Additionally, models were trained on the white wine data using a training-testing split. The builds were created with a k-fold cross-validation approach where data was divided into k subsets or folds, with one fold held out as the validation set, and the remaining folds used for training. This process was repeated k times, with each fold serving as the validation set exactly once. Accuracy was used as the performance evaluation metric and it was calculated for each fold and averaged to provide an overall assessment of the model's predictive power. This validation process ultimately provided a comprehensive assessment of the machine learning models' performance and their ability to predict wine quality.

6. Findings and Conclusions

When evaluating all models based on overall accuracy scores, it can be observed that the Random Forest Classifier and the Gradient Boosting Trees models performed much better compared to the Logistic Regression and Linear Discriminant Analysis models. Additionally, the Random Forest Classifier had a slightly higher accuracy score with the added features compared to the initial features alone, although this difference is quite negligible. On the other hand, the Gradient Boosting Trees model had a higher accuracy score with the initial features only. The Logistic Regression model performed slightly better with the addition of the added features

while the Linear Discriminant Analysis also had a negligible difference in accuracy scores between feature sets. All accuracy scores can be observed below.

Model Accuracy Comparison		
Model	**Accuracy Score with Initial Features**	**Accuracy Score with Initial Features + Added Features**
Random Forest Classifier	82.17%	82.48%
Gradient Boosting Trees	81.21%	79.94%
Logistic Regression	75.80%	77.07%
Linear Discriminant Analysis	76.11%	76.43%

Figure 6.1: Table of model accuracy score comparisons

This analysis demonstrates the power of ensemble learning approaches when it comes to the classification of wine quality based on its physicochemical features. Furthermore, it can also be observed that the initial 12 features of the dataset produced an accurate result for both the Random Forest Classifier and the Gradient Boosting Trees models, showing that the additional features that were created based on high correlations with quality had little to no impact on the models' predictive power. This suggests that more experimentation and exploration is needed to create feature interactions that will better improve the models' accuracy.

7. Lessons Learned & Recommendations

In terms of recommendations based on this specific research, it can be concluded that winemakers should focus heavily on the 12 initial chemical properties when producing wines as these appear to have the most impact when predicting quality. Ensuring the utmost attention to these features can help to ensure that the highest quality wines are being produced and keep winemakers competitive within the market.

In evaluating these results, it is also important to consider the limitations of the research. The data used only contained a limited amount of information on white wines from the northern region of Portugal. Using a more robust dataset that includes more wine varieties from other regions could potentially further improve the accuracy of the classification algorithms. Further feature engineering that better exposes the relationships between physicochemical properties in wine could also have a positive impact on future research. Experimentation with different models, specifically other models that utilize an ensemble learning approach, could uncover new methods with stronger predictive power. Additionally, experimenting with other models outside of classification also has the potential to further improve the work that has been conducted thus far. Lastly, accuracy is the only evaluation metric used in this research. Adding additional evaluation metrics could uncover new conclusions when comparing different models.

Overall, this study shows promise in the use of predictive modeling to improve the quality of wine production and ensure that winemaking companies are focused on the right aspects to improve the quality of their products. This will help to ensure continued sales and profitability as they grow and hold their ground in the competitive wine industry.

8. References

When reviewing the literature related to this study, previous research shows that feature selection is an important component when applying machine learning techniques to wine quality prediction. In an article titled "*A machine learning application in wine quality prediction*" authors acknowledge that wine quality is one of the most significant issues in the wine industry, hence the value that comes with predicting it (Bhardwaj, Olejar, Parr, Tiwari 2022). In the study, similar machine learning techniques, such as Gradient Boosting Trees and Random Forest

Classifier among other models were utilized to predict wine quality as a binary classification problem (Bhardwaj, Olejar, Parr, Tiwari 2022). The results of the study found that feature selection was key as the performance of all classifiers improved when the models were trained and tested with essential variables (Bhardwaj, Olejar, Parr, Tiwari 2022). This further solidifies the importance of feature selection when attempting to address the objective herein with predictive modeling.

In an earlier research article titled "*Selection of important features and predicting wine quality using machine learning techniques*" the author uses alternative machine learning models, specifically linear regression and neural networks to tackle the complex task of predicting wine quality (Gupta, 2018). Despite the differing approach in model selection, the results still show that these models predict the target variable more accurately when only important features are considered rather than considering all features (Gupta, 2018).

This work, once again, confirms how important feature selection is. In this current body of research, the main area for future work rests in the feature selection. While the ensemble learning methods show the best path forward, it is observed that perfecting feature selection will further boost performance and assist winemakers in perfecting their products.

References

Bhardwaj, P., Kulasiri, D., Olejar, K., Parr, W., Tiwari, P.,. 2022. "*A machine learning application in wine quality prediction*." Machine Learning with Applications 8: 1-11. https://doi.org/10.1016/j.mlwa.2022.100261.

Gupta, Y. 2018. "*Selection of important features and predicting wine quality using machine learning techniques*." Procedia Computer Science 125: 305-312. https://doi.org/10.1016/j.procs.2017.12.041.

URL to wine dataset used: https://archive.ics.uci.edu/dataset/186/wine+quality