23012646 | UFCFMJ-15-M | 2,197 words excluding reference | LT2-SALAM/ML-assessment (github.com)

Wine Quality Predition: A Machine Learning Application

Latifah Salam
Department of Data Science
University of the West of England
Bristol, United Kingdom
latifah2.salam@live.uwe.ac.uk
LT2-SALAM/ML-assessment (github.com)

Abstract—This study investigates various machine learning approaches for predicting wine quality, focusing on the challenges posed by class imbalance in the dataset and this issue is common in real-world data which can hinder the learning system's ability to accurately predict the minority class. My analysis employs three different models—Random Forest, Support Vector Machine (SVM), and Logistic Regression—alongside a comprehensive experimental evaluation involving class balancing techniques and feature selection methods. These experiments reveal that while class imbalance can affect performance, it is not the sole factor. Other complicated elements, such as class overlapping, also play a significant role. To address these challenges, I implemented the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset, leading to more defined class clusters. The Random Forest model emerged as the best performer for wine quality prediction, offering a balanced and robust approach to handling class imbalance and achieving high accuracy, precision, recall, and F1-score metrics. The comparative results indicate that over-sampling methods, particularly Random Forest combined with SMOTE, yield more accurate predictions than other models, especially in terms of the area under the Receiver Operating Characteristic (ROC) curve (AUC).

Keywords—Wine; SMOTE; SVC; Logistic Regression; Random Forest; Accuracy; precision; recall; F1-score; confusion matrix; Cross validation; ROC-AUC

INTRODUCTION

The wine industry is a multifaceted and competitive market where quality is a significant determinant of consumer preference and market success. The wine business is actively investing in innovative technology to enhance both the wine production and sales operations. Wine certification and quality evaluation are crucial components in this context. Certification serves to prevent the unlawful adulteration of wines, therefore safeguarding human health, and ensures the quality of wines on the market. Quality assessment is frequently included in the wine certification procedure and can aid in enhancing wine production by identifying the most significant elements (1). These elements are often evaluated by physicochemical and sensory examinations (2). Physicochemical laboratory tests for wine characterization encompass the analysis of alcohol content, chlorides, density, and residual sugar among others while sensory tests heavily rely on the expertise of human experts (3).

The use of wine physicochemical components for prediction of wine quality can provide valuable insights for winemakers, allowing them to refine their production processes and enhance profit. However, predicting wine quality based on these factors poses several challenges, particularly when dealing with imbalanced datasets where certain quality classes are underrepresented.

In the context of machine learning, class imbalance can significantly impact model performance. When features in training data belonging to one class heavily outnumber those in other classes, learning systems may struggle to accurately learn the minority class concepts. This issue is prevalent in real-world data, where infrequent but important events, such as the production of high-quality wines, are often underrepresented. Addressing this challenge requires robust techniques to balance the dataset and improve the learning system's performance.

This study explores three machine learning models—Random Forest, Support Vector Machine (SVM), and Logistic Regression—to predict wine quality. I employ the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance by generating synthetic data for the minority class. Additionally, I conduct feature selection to identify the most relevant physicochemical properties influencing wine quality.

II. DATASET OVERVIEW

Data for this study was sourced from the UCI Machine Learning Repository, specifically designed to model wine preferences based on physicochemical tests. Originally compiled by Paulo Cortez and his team, the dataset features observations on Portuguese "Vinho Verde" wine variants, an exclusive product originating from the Minho area in the northwest of Portugal. It includes 1,599 samples of red wine and 4,898 samples of white wine, summing up to 6,497

instances. The dataset is publicly available and detailed by [Cortez et al., 2009], providing a comprehensive basis for analysis in decision support systems within the wine industry. Data attributes include:

- The basic physicochemical properties of wine are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content.
- The output variable is the wine "quality", rated on a scale from 0 (very poor) to 10 (excellent), based on sensory data from wine tastings.

This dataset overview sets the stage for further modeling and analysis, focusing on developing predictive models that can accurately assess wine quality based on measurable properties. The goal is to guide vintners in optimizing the quality of their wine production, leveraging data-driven insights to enhance consumer satisfaction and operational efficiency.

III. PROBLEM DEFINITION

The objective of this research is to address a supervised learning problem centered around predicting wine quality from physicochemical properties of wines, specifically Portuguese "Vinho Verde." Unlike typical classification tasks, the goal here is to classify wines into quality categories that range from 0 (very bad) to 10 (excellent), based on their inherent characteristics. Given the nature of the dataset, this task can be approached as both a classification and a regression problem, providing a unique challenge due to the ordered nature of the quality scores and the class imbalance present in the dataset.

Research Problem:

The primary challenge is to predict the sensory wine quality based on easily measurable physicochemical properties. This involves:

i. Predicting Wine Quality: Utilizing various machine learning models to predict the wine quality score as accurately as possible.

ii. Understanding Feature Influence: Identifying which physicochemical features most significantly impact wine quality to guide vintners in improving their product.

Data Complexity and Challenges:

- Class Imbalance: The dataset exhibits a significant imbalance with more normal wines than excellent or poor ones, complicating effective model training and performance evaluation.
- Outlier Sensitivity: The presence of outliers can affect model accuracy, especially in regression tasks where extreme values can disproportionately influence the model's loss calculations.
- Feature Selection: Not all physicochemical properties may be relevant or equally important for predicting wine quality, making feature selection a critical step.

Multivariate Analysis

Given that wine quality prediction involves multiple chemical properties and complex interactions between them, multivariate analysis is the appropriate approach. It allows for a more comprehensive understanding and accurate prediction by considering the combined effects of all relevant features.

Modeling Approaches:

To address these challenges, the study will employ various machine learning algorithms, including:

- Random Forest: It is an ensemble of decision trees that improves generalization through bagging and feature randomness, making it less likely to overfit and capable of handling class imbalance effectively. For classification tasks, the output of the random forest is the class selected by most trees (5).
- Support Vector Machines (SVM): Is a powerful and versatile machine learning model, capable of performing linear or nonlinear classification, regression, and even novelty detection. SVMs shine with small to medium-sized nonlinear datasets, especially for classification tasks (4).
- Logistic Regression: Is a type of supervised machine learning method that is specifically designed for classification problems. Its main objective is to estimate the likelihood that a certain instance belongs to a specific class or not.

Class Balancing Techniques:

SMOTE (Synthetic Minority Over-sampling Technique): Enhances the representation of minority classes by creating synthetic samples, thus balancing the dataset, and potentially improving model accuracy on less represented classes.

Evaluation Metrics:

- Accuracy: Is a fundamental assessment parameter for classification. It denotes the ratio of accurately predicted observations out of the overall number of observations. Put simply, it measures the frequency of the model's accuracy. The formula for accuracy is:

$$Accuracy = \frac{True\ Positives + True\ Negetives}{Total\ Sample\ Size}$$

I will also be using precision and recall because they are vital measures employed to assess the accuracy of a model. They provide useful Insight on the model's ability to accurately identify positive samples and minimize both false positives and false negatives. Precision is a measure that specifically evaluates the accuracy of positive prediction while recall quantifies the degree to which positive predictions are correctly detected. The formula for precision and recall is:

$$pecision = \frac{True\ Positives}{True\ Positves + False\ Positives}$$

$$recall = \frac{True\ Positives}{True\ Positves + False\ Negatives}$$

- F1-Score: Combines precision and recall into a single metric, balancing the trade-off between incorrectly predicting low quality as high (precision) and failing to identify high quality (recall). This study used the weighted-F1 score to select and evaluate the model. The weighted F1 score is a specific scenario in which we calculate and present the score for both the positive and negative classes. It is crucial to consider this factor when handling imbalanced classes. The formula for F1-score is:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

- ROC-AUC Score: Measures the model's ability to discriminate between classes at various threshold settings. The AUC (Area Under the Curve) provides a single value summary of the ROC curve, representing the likelihood that a model ranks a random positive example more highly than a random negative example. Provides a comprehensive measure of model performance across all classification thresholds, especially useful in imbalanced settings.

The trade-off between accuracy and recall may be illustrated by utilizing the Receiver Operating Characteristics (ROC) curve to compare the performance of different models. The comparative evaluation of the overall model performance is conducted by computing the area under the curve, also known as the ROC-AUC score.

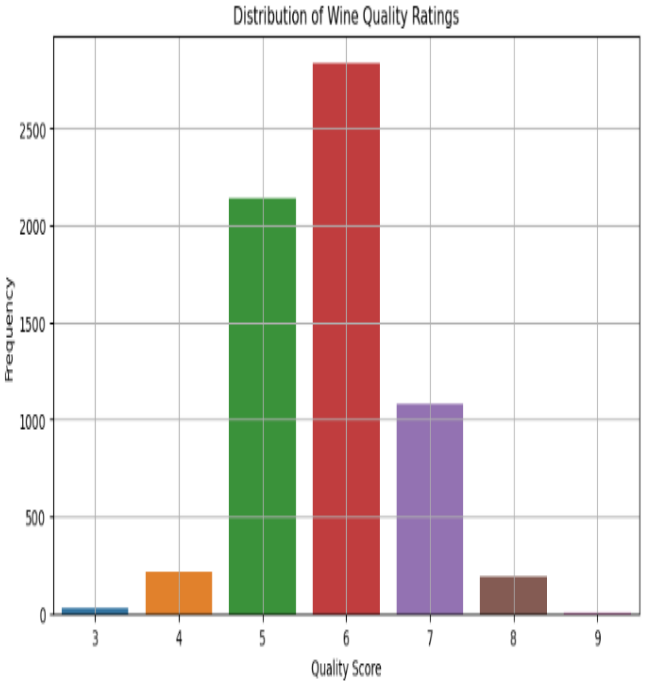
IV. ANALYSIS & EVALUATION

The wine quality prediction pipeline for the project sho

Figure 1: Modeling Pipeline for Wine Quality

Data Cleaning and Preprocessing	**Load Data**: Import red and white wine datasets. **Merge Datasets**: Combine red and white wine datasets to form a comprehensive dataset for analysis.
	Statistical Summary: Examine basic statistics of the data to understand distributions and central tendencies. **Initial Cleaning**: Inspect for missing values, handle anomalies, and remove duplicates.
	Correlation Analysis: Determine relationships between features using correlation matrices to identify highly correlated features that may need adjustment. **Visualization**: Use histograms, box plots, and scatter plots to visualize distributions and identify potential outliers.
Feature Selection and Class Balancing	**Feature Selection**: Applying SelectKBest with the ANOVA F-value method to select the top features. **Assess Imbalance**: Quantify the extent of class imbalance in the wine quality scores.
	Apply SMOTE: Use Synthetic Minority Over-sampling Technique to balance the dataset, enhancing the representation of minority classes in the training data.
Model Training	**Train Models**: Fit each model on the balanced and feature-engineered training set. **Cross-Validation**: Implement k-fold cross-validation to ensure the models generalize well to unseen data. using 5 numberd of folds.
Model Evaluation	**Performance Metrics**: Evaluate each model using accuracy, precision, recall, F1-score, and ROC-AUC (using 'ovr'-one vs rest approach for multi-class). **Confusion Matrix**: Generate confusion matrices for each model to visualize true positives, false positives, true negatives, and false negatives.
Model Interpretation and Reporting	**Feature Importance**: Analyze and report the importance of various features in the models, especially for tree-based methods like RandomForest.
	Coefficient Analysis: For linear models like Logistic Regression, examine the coefficients to interpret the impact of each feature on wine quality.
	Results Visualization: Present findings through ROC curves and other plots to compare model performances visually.
Model Interpretation and Reporting	**Summarize Insights**: Document the key findings, highlighting which models performed best and why. **Implications for Winemakers**: Provide actionable insights for winemakers on how to adjust wine-making processes based on the physicochemical properties that most influence wine quality.

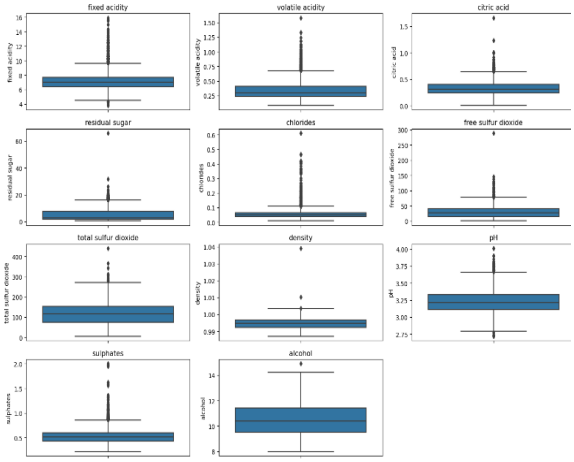
Fig 2: *Distribution of wine quality rating*



The bar plot represents the distribution of wine quality ratings in the wine dataset. The quality variable is mostly concentrated around scores of 5 and 6, indicating a skew towards average wines. Scores of 3 and 9 are rare, these could be the outliers (excellent or poor wines). Also, the distribution shows a clear class imbalance, with most wines rated around 5 and 6. This imbalance needs to be addressed for predictive modelling to ensure the model doesn't become biased towards the majority classes.

Treating outliers

Fig 3: *Displaying outliers using boxplot*



The presence of outliers in (Fig 3) indicates variability in wine production methods and ingredients. While some features like alcohol and density show tight control and consistency, others like citric acid and chlorides exhibit significant variation, which could be critical for certain types of wine.

Features Selection

Fig 4: Correlation matrix of wine attributes

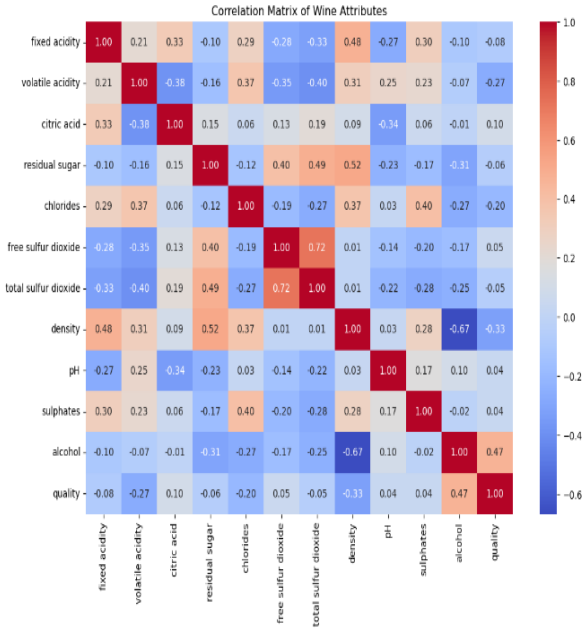


Fig 5: Features selection table

	Feature	Score
10	alcohol	307.977876
7	density	130.251647
1	volatile acidity	80.765877
4	chlorides	43.272754
5	free sulfur dioxide	13.204516
3	residual sugar	10.309397
2	citric acid	9.502731
6	total sulfur dioxide	8.128644
0	fixed acidity	7.147666
9	sulphates	3.567827
8	pH	3.451651

The correlation matrix shows a strong positive correlation that exists between wine quality and alcohol content where alcohol is a key feature in predicting wine quality. Ensuring optimal alcohol levels can significantly enhance wine quality. Volatile acidity and density have notable negative correlations with quality, indicating that controlling these factors can

improve wine quality. Lower volatile acidity and density are desirable. While features like sulphates, chlorides, citric acid, and residual sugar have weaker correlations, they still play a role in the overall quality and should not be neglected.

To determine which features are most relevant in predicting wine quality. For this stage, I use correlation matrix along with ANOVA F-value method to identify the most relevant features for predicting wine quality. The top 5 features with the highest ANOVA F-value score are used for model development.

Model Evaluation

TABLE I. SUMMARY OF MODEL EVALUATION

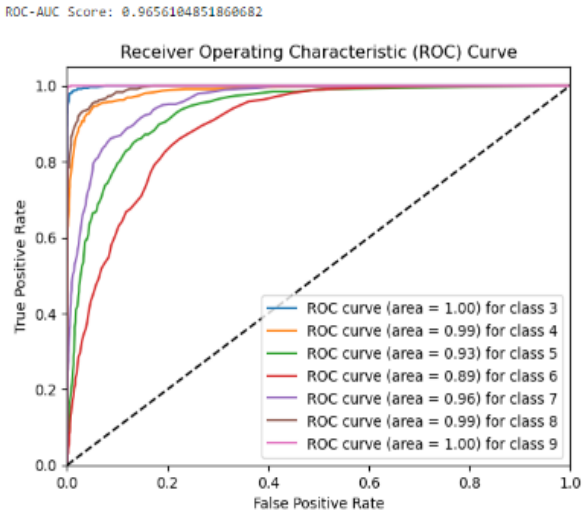
Models	**Trian /Test**	**Model Evaluation Table train and Test validation**			
		Accuracy	***Precision***	***Recall***	***F1-score***
Random Forest	Train	0.8137	0.8045	0.8137	0.8016
	Test	0.8096	0.7982	0.8097	
Linear SVM	Train	0.3690	0.3640	0.3690	0.3520
	Test	0.3619	0.3550	0.3619	0.3433
Logistic Regression	Train	0.4485	0.4273	0.4485	0.4312
	Test	0.4431	0.4202	0.4431	0.4258

Fig 6: *Model Comparison*

Metric	RandomForest	SVM	Logistic Regression
Accuracy	0.8097	0.3630	0.4470
Precision	0.8001	0.3583	0.4265
Recall	0.8097	0.3630	0.4470

The Random Forest model demonstrates the highest performance across all metrics with 80.97% average accuracy, indicating a strong ability to accurately classify wine quality. The SVM model shows low performance with 36.30% but Linear regression is slightly higher than SVN with 8.4% on average accuracy with made it still significantly lower than the Random Forest model. The Random Forest model is particularly effective at balancing precision and recall, making it reliable for predicting both positive and negative classes.

Fig 7: Receiver Operating Characteristic (ROC) Curve



The overall ROC-AUC score of 0.965 indicates that the random forest model performs exceptionally well in distinguishing between different wine quality classes. Most classes have very high AUC values, demonstrating that the model is highly effective in classifying wines correctly. The AUC values for classes 5 and 6 are slightly lower, indicating that the model has more difficulty distinguishing these middle-quality wines.

V. CONCLUSION

Random Forest model is the best model for wine quality prediction because the model achieved the highest accuracy (80.97%), precision (80.01%), recall (80.97%), and F1-score (80.49%) among the three models, indicating superior performance in classifying wine quality. The high F1-score of 80.49% suggests that Random Forest maintains a good balance between precision and recall, making it reliable for both identifying true positives and minimizing false positives. Random Forest is known for its robustness and ability to handle large datasets and complex interactions between features, which likely contributes to its superior performance on this wine quality dataset.

The comparison clearly indicates that Random Forest is the most effective model for wine quality prediction, making it the recommended choice for deployment.

Further Works

- Advanced Ensemble Methods: Explore more advanced ensemble methods like Gradient Boosting Machines (GBM), or XGBoost, which often outperform traditional methods.
- Independent Test Sets: Validate the models on independent test sets not used during training to ensure they generalize well to completely unseen data.
    - Treat red and white wine separately since the red and white tastes are quite different and compare their result.

Reference

[1] Cortez, P., António, C., Almeida, F., Matos, T. and Reis, J. (2009) Modeling Wine Preferences by Data Mining From Physicochemical Properties. Decision Support Systems [online]. 547–533 [Accessed 30 March 2024].

[2] Ebeler, S. and Flavor, C. (1999) Linking Flavour Chemistry to Sensory Analysis of Wine. Kluwer Academic [online]. 409-422 [Accessed 02 April 2024].

[3] Smith, D. and Margolskee, R. (2006) Making sense of taste. Scientific American [online]. 84–92 [Accessed 15 April 2024].

[4] Géron, A. (2022) Hands-on Machine Learning with Scikit-learn, Keras, and Tensorflow, 3rd Edition [online]. 3rd ed. : O'Reilly Media, Inc. [Accessed 01 May 2024].

[5] Tin Kan Ho, (1995) Random Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition [online]. 278–282, pp. 14-16. [Accessed 25 April 2024].

[6] María, M., Dupas De Matos, A., Azquez-araújo, L., Puente, V., Hernando, J. and Chaya, C. (2021) Exploring Young Consumers' Attitudes and Emotions to Sensory and Physicochemical Properties of Different Red Wines. Food Research International [online]. 143 (110303), pp. 14-16. [Accessed 28 April 2024].

[7] Hong-yue, Z., Si-yu, L., Xu Zhao, Yi-bin, L., Xin-ke, Z., Ying Shi, and Chang-qing, D. (2023) The Compositional Characteristics, Influencing Factors, Effects on Wine Quality and Relevant Analytical Methods of Wine Polysaccharides: A Review. Food Chemistry [online]. 403 (134467) [Accessed 30 April 2024].

[8] Qian Janice, W. and Charles, S. (2018) Wine Complexity: An Empirical Investigation. Food Quality and Preference [online]., pp. 238-244. [Accessed 05 May 2024].

Link to my repository: LT2-SALAM/ML-assessment (github.com)