Predicting Wine Quality: A Comparative Analysis of Classification Methods

Jayesh Pamnani
University of Maryland, College Park
College Park, USA
jpamnani@umd.edu

***Abstract*—Wine quality assessment plays a pivotal role in both consumer satisfaction and the success of the viticulture industry. This paper explores the prediction of wine quality, classifying it as good or bad using various classification methods. The methods include Decision Tree, Logistic Regression, KNN, Naive Bayes, and Linear Discriminant Analysis, applied to a dataset containing the chemical properties of wines. Logistic Regression emerges as the most effective method, closely followed by Linear Discriminant Analysis. The selection was based on accuracy metrics, the area under the Receiver Operating Characteristic Curve, the number of false positives, and subsequently, the computing efficiency of the algorithm. Cross-validation is applied to ensure the robustness and generalization of the models. The paper discusses the practical implications of accurately assessing wine quality, potentially aiding both consumers and producers in making informed decisions. To elevate wine quality predictions, future enhancements may involve dataset expansion, and advanced algorithm experimentation, offering the potential for more accurate and reliable models in the viticulture industry.**

***Index Terms*—Wine-quality prediction, Decision Tree, Logistic Regression, kNN, Naïve Bayes, Linear Discriminant Analysis, Accuracy, K-Fold Cross-Validation, Receiver Operating Characteristic-Curve**

I. Introduction

The decision to predict wine quality holds practical importance in guiding consumers in their wine selection and assisting producers in quality control processes. This paper addresses the decision problem of predicting wine quality, a task with significant implications for both consumers and producers in the wine industry. Understanding the inherent quality of wine before purchase enables consumers to align their choices with personal preferences, ensuring a more enjoyable and satisfying experience. This is particularly relevant in a market where an extensive array of wine options is available, varying in taste profiles, aromas, and overall quality. On the producer's side, predicting and maintaining wine quality is a critical aspect of ensuring brand reputation and customer loyalty and it directly affects customer retention. Consistent delivery of high-quality products not only satisfies existing customers but also establishes a positive reputation in the competitive marketplace. The primary objective of this study is to develop and choose a reliable model that can accurately classify wines as good or bad quality based on their chemical properties, and provide satisfaction to both, the customers and the producers by minimizing the overall cost associated with misclassification of good-quality wine.

In this study, the primary aim is to assess the effectiveness of five machine learning models on a wine quality dataset, encompassing the chemical properties of wine. The selected algorithms for evaluation include Decision Tree, Logistic Regression, k-Nearest Neighbors, Naïve Bayes, and the Linear Discriminant Analysis classifier. These models will be trained on the wine quality dataset to predict whether the wine quality is good or bad. The resulting predictions will be validated and supported by cross-validation metrics and the Area Under the Receiver Operating Characteristic curve. The evaluation process will also consider the Confusion matrix for the two leading models, identified as potential candidates for being the most suitable for the given dataset. The algorithm yielding the best results will be adopted by the wine industry for predicting wine quality.

A. Dataset

This study leverages a wine quality dataset that was originally sourced from the UCI Machine Learning Repository [1], subsequently procured from Kaggle [2], and comprises approximately 1600 samples, to undertake predictive modeling for the assessment of wine quality. The dataset encompasses a set of crucial chemical properties, each contributing to the intricate profile of the wines under consideration. These properties include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. The target variable, which the machine learning models aim to predict, is the quality of the wine. The dataset encapsulates a diverse range of chemical attributes, providing a comprehensive foundation for evaluating the factors that contribute to the overall quality of wines.

B. Literature Review

In the realm of wine quality prediction, significant strides have been made, with two notable studies laying the groundwork for our exploration. In the first study [3], the authors employed a multifaceted approach, demonstrating a comprehensive methodology. However, the study in this paper distinguishes itself by contending that the intricate nature of predicting wine quality extends beyond the confines of data science alone.

A second pivotal study in the domain of Wine Quality Prediction [4] took an extensive approach, training their model on samples of data. The researchers successfully predicted wine quality and validated their predictions against industry standards. Notably, this study achieved a remarkable level of accuracy in its predictions, setting a benchmark for the predictive capabilities of machine learning models in the context of wine quality.

The paper is structured into four key sections to provide a coherent and comprehensive approach to predicting the quality of wine. The ’Introduction’ sets the stage by introducing the problem definition, the algorithms and dataset used, the paper’s objectives, and the literature survey. ’Methodology’ delves into the technical implementation of the five models employed in the study, along with the metrics chosen for model evaluation. Special attention is given to the validation of results through different techniques. In the ’Results’ section, the author presents the findings of the experiments, focusing on accuracy scores, cross-validation results, and the area under the curve for each of the five models. Finally, the ’Discussion’ critically analyzes the results, offering insights into the strengths and limitations of the classifiers showcasing the best results, and suggesting directions for potential future research.

II. Methodology

In this section, the methodological approach used to evaluate machine learning models for classifying wine quality as either good or bad based on the aforementioned dataset is outlined. The five selected models for this study are the Decision Tree [5], Logistic Regression [6], the k-Nearest Neighbor [7] model, the Naive Bayes Classifier, and the Linear Discriminant Analysis Classifier [8]. Model performance was compared based on accuracy, and model robustness was assessed through Cross-Validation. Furthermore, the most suitable model was determined using the Area under the Receiver Operating Characteristic Curve.

The following subsections will elaborate on the versions of software used, the steps involved in data pre-processing, data processing, and result validation.

A. Versions Used

The methodology employed in this study is grounded in Python, utilizing version 3.11.2, and relies on crucial libraries such as sklearn (version 1.3.0), NumPy (version 1.25.2), and Pandas (version 2.1.0) for data manipulation and analysis.

B. Data Preprocessing

The pre-processing steps applied to the wine Dataset to convert it into a pandas dataframe, started by reading the CSV file using the ’read_csv’ function present in the pandas’ library [9].

Listing 1. Reading the CSV file using pd.read_csv()

```
wineDataSet =
    pd.read_csv('wineQuality_dataset.csv')
```

Setting the ’x’ (containing the list of features) and ’y’ (target variable) values by extracting the features from the wine DataSet was the next step involved in the process. Moving ahead, mapping the values in the target column (’good’ and ’bad’) to numerical values (1 and 0) was done to convert it to a binary classification problem.

Listing 2. Mapping the values in the target column

```
wineDataSet = target_column_index = 11
wineDataSet.iloc[:, target_column_index] =
wineDataSet.iloc[:, target_column_index]
        .map({'good': 1, 'bad': 0})
```

The above listing is a combination of the integer-location-based indexing (’iloc’) used to select all rows of a specific column specified by ’target_column_index’ and the ’map’ function that is used to replace values in the selected column. In this case, it’s replacing the string values ’good’ with the numeric value 1 and ’bad’ with the numeric value 0.

In the subsequent step, the values in both variables (x and y) are converted to integers to ensure that the data is consistent and compatible with the expectations of the machine learning algorithms. This is also done as classification algorithms work more efficiently with integer data types.

Listing 3. Converting the data types of the variables y and x to integers

```
y=y.astype('int')
x=x.astype('int')
```

The models in the study are trained using 67% of the dataset and are tested on the remaining 33% of the dataset.

C. Data Processing

The study incorporates five machine learning models: Decision Tree, Logistic Regression, k-Nearest Neighbors (KNN), Naive Bayes, and Linear Discriminant Analysis, with their implementations detailed in [5], [6], [7], [8]. For the k-NN algorithm, the range of k values was set from 1 to 50, aiming to identify the optimal k value for maximizing accuracy, as outlined in [7].

The performance of the models is evaluated based on accuracy, calculated using the ’accuracy_score’ method from the ’metrics’ package in scikit-learn, as explained in [10].

To assess the model’s robustness and check against overfitting or underfitting, a 3-fold cross-validation was employed, following the methodology presented in [11].

The Receiver Operating Characteristic Area Under the Curve (ROC AUC) metric was employed to gain a comprehensive understanding of true positive and false positive rates as extensively described in [12], providing readers with a comprehensive guide on its description, initialization, usage, and plotting.

D. Result Validation

Apart from cross-validation and the ROC-AUC metric, to further validate the obtained results, a confusion matrix was generated for the two models with the highest accuracy to gain further insights into the false positives as in [10]. The matrix proved to be valuable in understanding the model's classification performance, revealing instances where predictions align with the ground truth (true positives) and where errors occur (false positives).

III. Results

A. Accuracy Summary

Table 1 below presents the accuracy scores of five models, along with their respective mean accuracy values derived from cross-validation.

TABLE I
Accuracy Scores of the Five Models and Their Mean Accuracy from Cross-Validation

Model ↓	Accuracy	Cross-Validation Accuracy
Decision Tree	69.12%	69.40%
Logistic Regression	73.48%	72.74%
KNN	66.85%	67.91%
Naive Bayes	53.21%	56.54%
Linear Discriminant Analysis	73.29%	72.84%

The K-nearest neighbors (KNN) model was tested with neighbors ranging from 1 to 50, revealing that the optimal k, yielding the highest accuracy, was 12 (66.85%). This finding is summarized in the accompanying table. The accuracy outcomes for each count of neighbors within the tested range are graphically represented in Figure 1 below, providing a visual depiction of the model's performance across different k values.

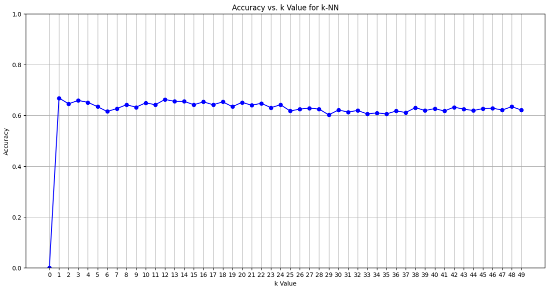

Fig. 1. Accuracy vs k value for KNN

B. Area under the ROC Curve

Table 2 below presents the area under the ROC (Receiver Operating Characteristic Curve) scores of five models.

TABLE II
AUC of the Five Models

Model ↓	Area Under the ROC-Curve
Decision Tree	0.694
Logistic Regression	0.731
KNN	0.668
Naive Bayes	0.515
Linear Discriminant Analysis	0.730

Figure 2 below illustrates the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for the five machine learning models employed in this study. The figure provides further insight into the true positive and the false positive values.

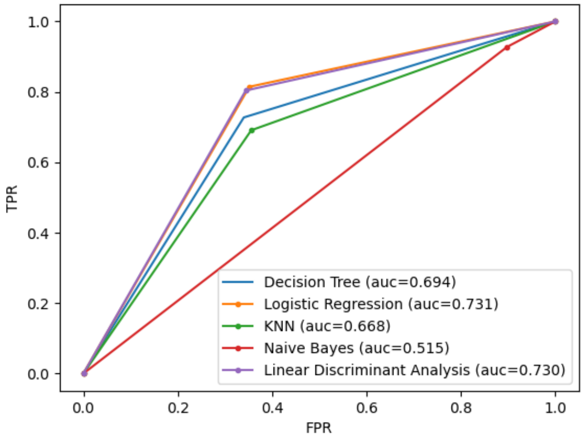

Fig. 2. AUC of the five models

C. Confusion Matrix for Logistic Regression and Linear Discriminant Analysis Models

After being tested on 33% of the dataset, the confusion matrix analysis for the Logistic Regression model revealed that the model accurately identified 164 instances of good-quality wine (true positives) and correctly classified 224 records as bad-quality wine (true negatives). However, the model missed out on 51 samples (false negatives) and erroneously classified 87 records as good-quality wine when they were not (false positives).

Similarly, for the Linear Discriminant Analysis model, the confusion matrix depicted that the model correctly predicted 166 instances of good-quality wine and 221 instances of bad-quality. However, it misclassified 54 instances of good quality as bad (false negatives) and 85 instances of bad quality as good (false positives).

D. Results Discussion

Given their comparable accuracy scores, both the Logistic Regression (LR) and Linear Discriminant Analysis (LDA) models emerge as front-runners, positioning them as the most suitable models for the given dataset. The alignment of their performance metrics beckons us to delve into the subtleties of second-order effects. Despite the Area Under the Curve (AUC) maintaining consistency across both models, a nuanced distinction emerges in the slope, revealing a slightly steeper inclination in Logistic Regression (LR) compared to Linear Discriminant Analysis (LDA).

A more in-depth analysis redirects our focus to the strategic goal of reducing false positives to enhance overall model efficacy. This strategic objective is firmly rooted in the collective satisfaction of consumers and producers, with the overarching goal of preserving customer retention. However, a noteworthy challenge emerges as both models exhibit a comparable number of false positives, introducing a nuanced dilemma in determining the most appropriate course of action.

This scenario prompts a deeper analysis to identify potential nuances or contextual considerations that might inform the selection between LR and LDA, thereby ensuring a more informed decision-making process

IV. Discussion

The observed performance of the machine learning models in predicting wine quality reveals distinct patterns. Logistic Regression (LR) and Linear Discriminant Analysis (LDA) outshine other models, displaying accuracies of 73.48% and 73.29%, respectively. The test for the robustness of models is checked as cross-validation yields similar accuracy results. This consistency is crucial in ensuring that the models generalize well to new, unseen data, a pivotal aspect in real-world applications. The consistency between accuracy scores and the Area Under the ROC-Curve (AUC) underscores the robust predictive capabilities of both LR and LDA. Their similar AUC values suggest a balanced trade-off between sensitivity and specificity.

The strategic decision to now either minimize the false positives or false negatives introduces a nuanced dilemma in the context of wine quality prediction. On one hand, reducing false positives is important for customer retention, ensuring that customers consistently receive wines of the expected quality. This approach aims to prevent dissatisfaction among consumers, preserving brand loyalty and potentially increasing revenue. However, on the other hand, minimizing false negatives is crucial to avoid incorrectly classifying high-quality wines as lower quality, which could result in undervaluing premium products. In such a scenario, companies may lean towards an algorithm based on their specific business priorities and risk tolerance. If customer satisfaction and retention are paramount, a model that minimizes false positives might be favored. Conversely, if protecting high-quality wines from being undervalued is of greater concern, a model prioritizing the reduction of false negatives may be preferred.

This dilemma becomes particularly pronounced when both Logistic Regression (LR) and Linear Discriminant Analysis (LDA) showcase similar results, as observed in this study. The comparable performance of these models poses a challenge for decision-makers in the wine industry, as they must carefully weigh the consequences of misclassifying good-quality wine.

Ultimately, the choice between minimizing false positives and false negatives hinges on the unique objectives and business strategies of each wine company. Delving deep into the tertiary impacts, organizations should also take into account the computational costs associated with deploying a predictive model. Considering the computational complexity aspect, Logistic Regression emerges as a more computationally efficient option compared to Linear Discriminant Analysis. Given this, a company may lean towards Logistic Regression as the preferred choice, considering not only its computational advantages but also the strategic fit with the overall business objectives.

This paper significantly contributes to wine quality prediction, playing a pivotal role in influencing consumer choices and maintaining industry quality standards. Leveraging machine learning models, the research offers companies a competitive advantage by accurately identifying good-quality wines. This not only enhances customer retention but also has the potential to boost revenue. Logistic Regression and Linear Discriminant Analysis emerge as promising models, showcasing their practical applicability in real-world scenarios. Overall, this research opens avenues for the wine industry to employ advanced analytics, providing a nuanced approach to quality prediction in tune with evolving consumer preferences and industry standards.

Moving forward, enhancing model robustness could involve feature engineering, dataset expansion, and experimentation with advanced algorithms. Future research might explore additional data sources, such as data from different vintages, regions, grape varieties, vineyard practices, or weather conditions, to provide a more comprehensive understanding of the factors impacting wine quality. Ensembling techniques and hyperparameter tuning could further elevate predictive performance, contributing to the development of more accurate models for the wine industry's benefit. By embracing these strategies, the wine industry can anticipate the development of more accurate, resilient, and adaptable models, ultimately enhancing the precision and reliability of wine quality assessments

References

[1] https://archive.ics.uci.edu/ml/datasets/wine+quality

[2] https://www.kaggle.com/datasets/nareshbhat/wine-quality-binary-classification

[3] M. Koranga, R. Pandey, M. Joshi, and M. Kumar, "Analysis of white wine using machine learning algorithms," 2021

[4] K. Jain, K. Kaushik, S. K. Gupta, S. Mahajan, and S. Kadry, "Machine learning-based predictive modelling for the enhancement of wine quality," 2022.

[5] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu, "Decision Tree Analysis on Iris Dataset."

[6] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu, "Classification using Logistic Regression."

[7] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu, "K-Nearest Neighbor: A Comprehensive Study on 'K' and Decision Metrics."

[8] J. Pamnani, "Comparative Analysis of Classifier Performance on the Iris Dataset: Naive Bayes and Linear Discriminant Analysis vs. Logistic Regression, Decision Tree, and k-Nearest Neighbors"

[9] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html

[10] J. Pamnani, "Maximizing Profit by Minimizing Misclassification Costs using Decision Tree, Logistic Regression, and KNN Models on the Iris Dataset."

[11] J. Pamnani, A. Ketkar, J. Hasija, and D. Lnu, "Comparing Decision Tree with Logistic Regression using Cross-Validation Technique."

[12] J. Pamnani, "Evaluating Models For Predicting Flight Delays Using ROC-AUC Metrics."