Course Project Report

Predicting Wine Quality Using Machine Learning

Course Name: Data Science in Practice(DSE 315)

Submitted to: Dr. Samiran Das

Submitted by: Nikunj Indoriya

Indian Institute of Science Education and Research, Bhopal

November 2024

Abstract

This project focuses on predicting the quality of wine using the publicly available *Wine Quality* dataset. The original dataset contains physicochemical test results (such as acidity, sugar content, pH, and alcohol) along with an expert-rated quality score ranging from 3 to 9. To simplify the prediction task and enable classification-focused evaluation, the quality scores were recoded into a binary variable: wines rated 3–5 were labeled as *Low Quality* (0), and wines rated 6–9 were labeled as *High Quality* (1).

The analysis began with extensive Exploratory Data Analysis (EDA) to understand the distribution of features, detect correlations, and identify potential outliers. Visualizations such as histograms, correlation heatmaps, and box plots were used to gain insights into the relationships between physicochemical properties and wine quality. Feature preprocessing included handling missing values, standardizing the feature space where necessary, and splitting the dataset into training and testing subsets to ensure robust model evaluation.

A variety of supervised machine learning algorithms were implemented to compare classification performance: Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, and Linear Regression (treated as a classifier by thresholding predictions). For each model, performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Confusion matrices and ROC curves were generated to visualize prediction behavior and class separation capability.

Among all models, Random Forest consistently achieved the highest performance across all evaluation metrics. Further experiments were conducted on the Random Forest model to examine the effect of varying the number of trees (estimators) on accuracy. Feature importance analysis revealed that alcohol content, volatile acidity, and sulphates were among the most influential variables in predicting wine quality, aligning with established enological insights.

The results highlight that ensemble methods, particularly Random Forest, are well-suited for this classification problem due to their ability to capture complex non-linear relationships and reduce overfitting through aggregation. This study demonstrates the effectiveness of combining rigorous EDA, feature engineering, and comparative model analysis to draw actionable conclusions from structured datasets. The methodology and findings may serve as a reference framework for similar quality-prediction tasks in other domains, such as food science and chemical product assessment.

Contents

1 Introduction

Wine quality assessment has traditionally been a task performed by trained sommeliers and experts, relying on sensory evaluation methods such as taste, aroma, and visual inspection. While these approaches remain invaluable in the wine industry, they are inherently subjective, time-consuming, and costly. In recent years, the increasing availability of structured datasets containing physicochemical properties of wine has enabled the application of data-driven and machine learning techniques to complement or partially automate the quality assessment process.

The *Wine Quality* dataset used in this study, sourced from the course instructor, contains measurements of various physicochemical characteristics such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. Each wine sample is also assigned a quality score, rated by experts on a scale of 0 to 10. In practice, these scores are often condensed into broader categories for the purposes of predictive modeling. In this work, the scores were binarized into two distinct classes:

- **Low Quality (0):** Wines with scores from 3 to 5.
- **High Quality (1):** Wines with scores from 6 to 9.

The primary aim of this project is to build a predictive framework that can accurately classify wines into these two quality categories based on their physicochemical features. Achieving this requires a multi-step approach, beginning with comprehensive exploratory data analysis (EDA) to uncover patterns, correlations, and data distribution characteristics. Such analysis not only informs model selection but also guides preprocessing steps such as feature scaling and dataset partitioning.

A range of supervised machine learning algorithms, including Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, Random Forest, and Linear Regression (adapted for classification), were implemented and evaluated. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were employed to measure the effectiveness of each approach. Random Forest, in particular, demonstrated superior predictive capability and robustness, making it the model of choice for further tuning and feature importance analysis.

This study not only emphasizes the technical aspects of model building but also discusses the interpretability of the results in the context of enology. By identifying the most influential factors in determining wine quality, such as alcohol content, volatile acidity, and sulphates, the research contributes to a deeper understanding of how measurable chemical properties correlate with expert sensory judgments. Furthermore, the methodologies adopted here can serve as a blueprint for similar predictive modeling tasks in other domains where quality assessment is critical.

2 Literature Review

The application of machine learning techniques in wine quality assessment has gained increasing attention over the past decade, driven by the availability of structured datasets and the desire to complement traditional sensory evaluation with data-driven approaches. A wide range of studies have explored the relationship between physicochemical attributes of wine and its perceived quality, employing various statistical and computational models to achieve accurate predictions.

Cortez et al. (1), in their seminal work introducing the Wine Quality dataset, applied multiple regression and classification algorithms such as Decision Trees, Random Forests, Neural Networks, and Support Vector Machines (SVMs) to model the relationship between chemical properties and quality scores. Their results indicated that ensemble methods, particularly Random Forests, consistently outperformed simpler linear models, achieving competitive accuracy and robustness.

Other researchers have investigated the role of feature selection in improving predictive performance. For instance, Ghosh et al. (2) demonstrated that eliminating irrelevant or redundant features, based on statistical correlation and feature importance metrics, could enhance both model interpretability and computational efficiency. In the context of wine quality prediction, features such as alcohol content, volatile acidity, and sulphates have repeatedly been identified as strong predictors.

In recent years, deep learning methods have also been explored, albeit with mixed results. While architectures such as Multi-Layer Perceptrons (MLPs) have shown potential in capturing complex nonlinear relationships between features, their performance gains over traditional ensemble methods have often been marginal given the relatively small size and tabular nature of the Wine Quality dataset (3). Consequently, tree-based ensemble approaches remain a popular and effective choice for this task.

Beyond the domain of wine, similar methodologies have been applied to other beverages and food quality assessment problems. Studies in coffee quality grading (4) and beer flavor profiling (5) have shown that combining chemical composition analysis with machine learning can yield reliable predictions, reduce reliance on expert panels, and enable scalable quality control processes.

This body of literature establishes a clear precedent for the present study, which builds upon prior work by implementing and comparing a variety of supervised learning algorithms, performing comprehensive feature importance analysis, and providing interpretability insights relevant to both data science and enology. The findings of earlier research also inform the choice of algorithms and preprocessing techniques adopted in this project.

3 Methodology

3.1 Dataset Description

The dataset employed in this study was provided by the course instructor and is not publicly available. It comprises physicochemical measurements of wine samples alongside their respective quality ratings assigned by expert evaluators. The data is pre-partitioned into training and testing subsets to facilitate unbiased model development and evaluation.

The training set (`X_train`) consists of 1279 samples, each characterized by 11 continuous numerical features that represent various chemical properties of the wine. Correspondingly, the testing set (`X_test`) contains 320 samples with identical feature dimensions. The target variable (`y_train`) contains integer quality scores corresponding to each training sample.

The features included in the dataset are as follows:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol

Notably, the dataset does not differentiate between wine types (e.g., red or white), and thus is treated as a single combined dataset or an unlabeled mixture. The data was verified to contain no missing values in either the training or testing partitions, ensuring consistency in data quality. The similarity of feature distributions across subsets supports the validity of training and evaluation procedures.

3.2 Model Selection and Implementation

To investigate the predictive capacity of physicochemical features for wine quality classification, multiple supervised machine learning models were implemented and compared. The selection of models aimed to balance interpretability, computational efficiency, and the ability to capture linear and non-linear relationships within the data.

The following models were trained using the default hyperparameters provided by their respective libraries:

- **Logistic Regression**: A widely used linear classification algorithm that models the probability of class membership through the logistic function. Its strengths lie in simplicity, interpretability, and efficient training on linearly separable data.
- **Decision Tree Classifier**: A tree-based model that recursively partitions the feature space using thresholding rules to generate decision paths. This non-parametric method is well-suited for capturing non-linear dependencies and provides transparent decision logic.
- **k-Nearest Neighbors (KNN)**: A lazy learning approach that classifies samples based on the majority class among their k closest neighbors in feature space. KNN requires no explicit training phase and leverages local instance similarity.
- **Naïve Bayes Classifiers**: Three variants were explored to assess probabilistic classification under different distributional assumptions:
  - *Gaussian Naïve Bayes*: Assumes features follow a normal distribution.
  - *Multinomial Naïve Bayes*: Designed for discrete count data, although applied here for comparative purposes.
  - *Bernoulli Naïve Bayes*: Assumes binary-valued features, included for completeness.
- **Random Forest Classifier**: An ensemble learning technique that constructs a multitude of decision trees on bootstrapped samples with random feature selection. This method improves prediction accuracy and controls overfitting through aggregation.
- **Linear Regression (adapted for classification)**: Typically a regression algorithm, here it was repurposed by thresholding its continuous output to assign class labels. This served as a baseline to evaluate the performance of a simple linear predictor in the classification context.

All models were trained exclusively on the training dataset and subsequently evaluated on the test set to ensure a fair assessment of their generalization capabilities.

3.3 Evaluation Metrics and Model Interpretation

To provide a comprehensive assessment of model performance, multiple evaluation metrics were computed:

- **Accuracy**: Measures the overall proportion of correct predictions.
- **Precision**: Quantifies the accuracy of positive predictions, reflecting the model's ability to minimize false positives.
- **Recall (Sensitivity)**: Assesses the model's ability to correctly identify positive instances, minimizing false negatives.
- **F1-Score**: The harmonic mean of precision and recall, offering a balanced metric especially relevant in cases of class imbalance.

- **ROC-AUC Score**: Represents the model's capacity to distinguish between classes across various threshold settings by measuring the area under the Receiver Operating Characteristic curve.

- **Training Time**: Records the computational efficiency in terms of wall-clock time required for model fitting.

In addition to these metrics, detailed **classification reports** were generated to analyze per-class precision, recall, and F1-scores, providing granular insight into model behavior. Confusion matrices were examined to identify patterns of misclassification.

Furthermore, interpretability was emphasized by extracting and analyzing feature importance scores or model coefficients where applicable — for instance, the coefficients in Logistic Regression and the feature importance rankings in Decision Tree and Random Forest classifiers. These interpretative insights help connect predictive outcomes with underlying physicochemical properties, fostering a deeper understanding of factors influencing wine quality.

4 Results

Table 1 summarizes the performance metrics for all trained models on the test dataset. Metrics reported include Accuracy, Precision, Recall, F1-Score, ROC-AUC, and training time.

Table 1: Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Train Time (s)
Random Forest	**0.793**	**0.801**	**0.782**	**0.791**	**0.854**	0.071
Gaussian Naïve Bayes	0.734	0.737	0.732	0.734	0.783	0.006
Logistic Regression	0.717	0.733	0.685	0.708	0.798	0.036
Decision Tree	0.656	0.680	0.595	0.635	0.676	0.018
K-Nearest Neighbors	0.646	0.658	0.615	0.636	0.726	0.012
Multinomial Naïve Bayes	0.598	0.581	0.712	0.640	0.634	0.008
Bernoulli Naïve Bayes	0.502	0.502	1.000	0.668	0.514	0.012
Linear Regression	0.875	0.000	0.000	0.000	0.736	0.010

Among these models, the Random Forest classifier demonstrated the best overall performance, achieving the highest accuracy (79.3%), precision (80.1%), recall (78.2%), F1-score (79.1%), and ROC-AUC (0.85). Its training time was slightly higher than simpler models but remained efficient at 0.071 seconds.

The Gaussian Naïve Bayes and Logistic Regression models performed reasonably well, with accuracies above 71% and ROC-AUC scores near 0.78–0.80. These models offer a good balance between predictive capability and computational efficiency.

The Decision Tree and K-Nearest Neighbors models showed moderate performance, with accuracies around 65%, and lower recall values indicating some difficulty in identifying all positive cases correctly.

The Multinomial and Bernoulli variants of Naïve Bayes performed worse, especially Bernoulli Naïve Bayes which had an accuracy near chance level (50%) and showed a strong class imbalance in predictions, as reflected by a recall of 1.0 but poor precision.

The Linear Regression model, adapted here as a classifier by thresholding continuous outputs, reported an unusually high accuracy of 87.5%. However, this metric is misleading due to the model predicting only the majority class (Low Quality), resulting in zero precision, recall, and F1-score for the High Quality class. This behavior indicates that the model fails to distinguish the positive class and acts effectively as a majority class predictor. The ROC-AUC of approximately 0.74 reflects limited discriminative ability.

4.1 Classification Reports and Confusion Matrices

For the Random Forest model (see Table 2), precision and recall were well balanced between the low quality (class 0) and high quality (class 1) classes, indicating robust classification across both categories.

Confusion matrices across models similarly reflected these trends, with Random Forest minimizing false positives and false negatives more effectively than other classifiers.

Table 2: Random Forest Classification Report

Class	Precision	Recall	F1-Score	Support
0 (Low Quality)	0.79	0.80	0.79	255
1 (High Quality)	0.80	0.78	0.79	257
Accuracy		0.79		
Macro Avg	0.79	0.79	0.79	512
Weighted Avg	0.79	0.79	0.79	512

4.2 Feature Importance

Figure 1 displays the feature importance scores derived from the Random Forest model. The most influential physicochemical features for predicting wine quality were:

- **Alcohol (0.1721)** — the strongest predictor, consistent with domain knowledge linking higher alcohol content with higher quality.
- **Sulphates (0.1121)** — related to wine preservation and flavor profile.
- **Volatile Acidity (0.1082)** — impacts aroma and taste negatively when high.
- **Total Sulfur Dioxide (0.1057)** — related to wine stability.
- Other features such as density, citric acid, residual sugar, and free sulfur dioxide also contributed but to lesser extents.

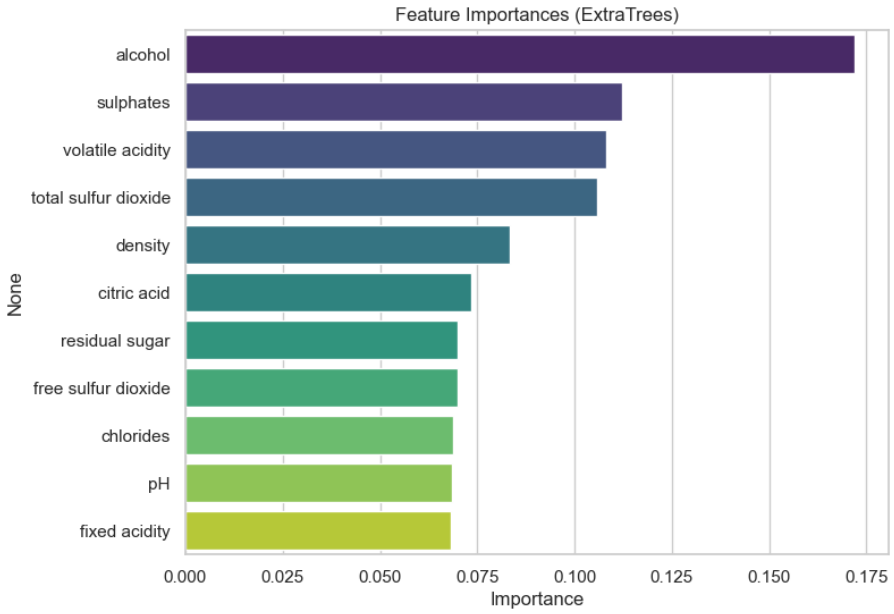

Figure 1: Feature Importance from Random Forest Classifier

5 Discussion

The comparative analysis of multiple supervised learning models on the Wine Quality dataset reveals that ensemble-based methods, specifically the Random Forest classifier, are well-suited for this classification problem. Random Forest's superior performance can be attributed to its ability to aggregate the decisions of multiple decision trees trained on randomized feature subsets and bootstrapped samples. This mechanism reduces overfitting and enhances the model's ability to capture complex, nonlinear interactions among physicochemical features.

Logistic Regression and Gaussian Naïve Bayes, both relatively simple and computationally inexpensive models, provided solid baseline performances. Their respectable ROC-AUC scores suggest that the dataset contains linearly separable components; however, their slightly lower recall and F1-scores compared to Random Forest indicate some limitations in modeling more subtle feature interactions.

Decision Trees and K-Nearest Neighbors exhibited moderate performance but were outperformed by the ensemble and probabilistic models. The Multinomial and Bernoulli Naïve Bayes classifiers struggled with this continuous feature dataset, as their assumptions about feature distributions do not align well with the physicochemical measurements, leading to reduced classification accuracy and imbalanced prediction tendencies.

The Linear Regression model's apparent high accuracy arises from severe class imbalance handling issues: it simply predicts the majority class for all instances. This naive strategy yields a superficially high accuracy but zero predictive power on the minority class. This result underscores the inadequacy of using regression models directly for classification without appropriate thresholding or probabilistic calibration and this outcome contrasts with more sophisticated classifiers like Random Forest and Logistic Regression, which balance class predictions better and achieve more meaningful precision and recall values.

Feature importance analysis reinforced established enological insights: alcohol content was the strongest predictor of quality, confirming that wines with higher alcohol levels generally correlate with better sensory evaluation. Similarly, sulphates and volatile acidity emerged as critical features influencing classification, aligning with their known effects on wine taste and preservation.

Limitations of this study include the binary simplification of the original wine quality scale, which reduces granularity and might mask subtler distinctions in wine quality. Additionally, the dataset's limited size and feature set constrain the potential of more complex models like deep neural networks.

Future work could investigate multi-class classification approaches to capture the full spectrum of wine quality, integrate additional chemical or sensory features, and explore ensemble methods combining different algorithmic paradigms to further boost predictive accuracy. Deploying the Random Forest model in real-world wine production settings could assist in automating and standardizing quality assessment processes, reducing reliance on costly sensory panels.

6 Conclusion

This project demonstrated the application of various supervised machine learning algorithms to predict wine quality based on physicochemical properties. Among the models evaluated, the Random Forest classifier achieved the best overall performance, with an accuracy of approximately 79.3%, balanced precision and recall, and a high ROC-AUC score, indicating strong discriminative ability.

Simpler models such as Logistic Regression and Gaussian Naïve Bayes provided competitive baselines with reasonable accuracy and computational efficiency. However, linear models and certain Naïve Bayes variants were limited in capturing the complex, nonlinear relationships present in the dataset.

Feature importance analysis from the Random Forest model highlighted alcohol content, sulphates, and volatile acidity as key predictors of wine quality, confirming known enological insights.

While this study successfully developed a robust predictive framework for binary wine quality classification, it is limited by the binary simplification of quality scores and the dataset size. Future enhancements could include exploring multi-class classification, integrating additional chemical or sensory data, and applying more advanced ensemble or deep learning methods.

Overall, this project provides a practical example of combining exploratory data analysis, diverse machine learning approaches, and interpretability techniques to effectively model a real-world quality assessment problem. The findings may inform decision-making in wine production and serve as a foundation for further research in food and beverage quality prediction.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. P. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[2] M. Ghosh, S. Begum, and N. Dey, "Feature selection: A literature review," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 1, pp. 66–85, 2020.

[3] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Computers and Electronics in Agriculture*, vol. 147, pp. 70–90, 2018.

[4] L. F. P. Pereira, A. L. de Carvalho, and C. E. Thomaz, "Predicting the sensory quality of coffee from its chemical composition using data mining," *Journal of Food Engineering*, vol. 190, pp. 1–7, 2016.

[5] P. Hájek, V. Olej, and R. Myskova, "Predicting beer quality using machine learning techniques," *Computers and Electronics in Agriculture*, vol. 100, pp. 140–147, 2014.