White Wine Quality Prediction

Sahil Khandelwal
University of Notre Dame
South Bend, Indiana, USA
skhandel@nd.edu

Tyler Berg
University of Notre Dame
South Bend, Indiana, USA
tberg3@nd.edu

Mrinal Sharma
University of Notre Dame
South Bend, Indiana, USA
msharma2@nd.edu

Haotian Wang
University of Notre Dame
South Bend, Indiana, USA
hwang35@nd.edu

Abstract

This study develops a machine learning model to predict white wine quality using physicochemical attributes. We analyze a dataset with 4,898 wine samples and 12 features, applying data preprocessing techniques such as correlation analysis, and Box–Cox transformations for skewness of data. Several regression models, including Random Forest, XGBoost, MLP, and hybrid models, are evaluated. Hyperparameters are optimized using Optuna, and performance is assessed with RMSE, R^2, MAE, and classification accuracy. The hybrid MLP+XGBoost model achieved the best results (RMSE = 0.559, R^2 = 0.561). Our findings show that ensemble and hybrid models outperform simpler regressors, offering an objective, data-driven alternative to traditional wine quality assessments. Future work will expand the model to red wine and explore interpretability and deployment.

ACM Reference Format:
Sahil Khandelwal, Tyler Berg, Mrinal Sharma, and Haotian Wang. 2025. White Wine Quality Prediction. In *Proceedings of* . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The global wine industry represents a significant economic sector, with the market valued at approximately 340.8 billion in 2020 and projected to reach 528.7 billion by 2028. Wine quality assessment traditionally relies on expert sensory evaluation, which is subject to human factors, time-consuming, and inconsistent across different tasters. By leveraging data analytics and machine learning, wineries can establish standardized quality metrics that reduce subjectivity, thereby enhancing consumer trust and potentially increasing market share in the industry[1]. The primary objective of this study is to develop a Machine Learning predictive model for white wine quality using data-driven techniques. The dataset contains physicochemical properties of white wines, with quality as the target variable. The task is formulated as a regression problem, predicting a continuous numerical quality score from wine characteristics. Alternatively, a classification approach can be applied by categorizing wine quality into discrete labels (e.g., low, medium, high). The dataset is structured with 12 numerical and ordinal attributes, and the study aims to evaluate different machine-learning models to identify the most effective approach for predicting wine quality. Feature selection is informed by correlation analysis, ensuring that only the most relevant attributes will contribute to the final model.

,

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Works

The prediction of wine quality using machine learning (ML) techniques has garnered significant research interest due to its potential to offer objective, consistent, and reproducible evaluations that traditional sensory analysis methods often cannot. Several recent studies have explored diverse ML approaches, underscoring the advantages of ensemble and hybrid models due to their robust handling of high-dimensional, nonlinear data.

Bhardwaj *et al.* [2] provided an extensive analysis on the prediction of New Zealand Pinot Noir wine quality using both experimental and synthetically augmented datasets. Recognizing the limitation posed by small and homogeneous datasets, the authors employed Synthetic Minority Over-sampling Technique (SMOTE) to augment the dataset size significantly. Furthermore, the authors systematically applied multiple feature selection methods, including Random Forest, Gradient Boosting, and Extra Trees classifiers, to pinpoint the most significant chemical features influencing wine quality, such as Ethyl octanoate and Geraniol. Their experimental evaluation across various classifiers, including AdaBoost, XGBoost, and Random Forest, revealed AdaBoost as particularly effective, consistently yielding exceptional accuracy approaching 100%. The research underscored the importance of feature selection and data augmentation to enhance predictive accuracy in ML models applied to limited datasets.

Shaw *et al.* [3] undertook a comparative study of various supervised learning algorithms—Support Vector Machines (SVM), Random Forest (RF), and Multilayer Perceptron (MLP)—using the widely referenced UCI red wine dataset. Their study demonstrated clearly that RF outperformed other algorithms by achieving approximately 82% accuracy, substantially higher than SVM and MLP models tested in parallel. They attributed RF's superior performance to its intrinsic ability to handle complex nonlinear relationships and its robustness to noisy, high-dimensional data. This reinforced the prevailing observation within the literature that ensemble methods like RF are particularly suitable for applications where multiple interacting variables contribute significantly to the output.

In related work, Trivedi and Sehrawat [4] explored binary classification (good versus bad quality) of wine samples utilizing Logistic Regression (LR) and Random Forest classifiers. Their experimental results clearly indicated RF as more robust, yielding an accuracy of 84%, compared to LR's 76%. Notably, this study emphasized the necessity of employing metrics beyond mere accuracy, such as precision, recall, F1-score, and specificity, due to inherent class imbalance in wine quality datasets. Their work highlighted how RF's ensemble nature provides a distinct advantage, particularly in scenarios involving imbalanced classification tasks, by mitigating the biases that simpler models often encounter.

Further emphasizing the robustness and versatility of ensemble methods, additional benchmarking conducted using the datasets from the UCI Machine Learning Repository [1] consistently demonstrated RF's superior performance over conventional models such as Decision Trees and SVM. This benchmarking demonstrated RF models achieving near-perfect accuracy (98%), underscoring the practical value of ensemble learning frameworks in predictive modeling of wine quality.

In a broader context, the reviewed literature collectively emphasizes several key insights that directly influence the methodological direction of our research:

- The critical importance of rigorous feature selection and dimensionality reduction to enhance model accuracy.
- The beneficial impact of data augmentation methods, such as Box-cox transformation, in handling data scarcity and imbalance.
- The consistent superiority of ensemble and hybrid modeling techniques in capturing complex, nonlinear interactions within wine datasets.
- The necessity of employing multiple evaluation metrics to comprehensively assess model performance, especially in cases of class imbalance.

Motivated by these extensive findings, our study seeks to integrate proven ensemble methodologies and advanced data preprocessing techniques to predict white wine quality accurately. By leveraging insights from previous works, we aim to establish a reliable, data-driven model that effectively addresses the inherent complexity and variability characteristic of wine quality prediction tasks.

3 Problem Definition

The primary objective of our research is to develop an accurate and reliable predictive model for white wine quality utilizing advanced machine learning techniques. Traditionally, wine quality assessments have heavily depended on sensory evaluations performed by expert tasters. However, these sensory evaluations often introduce subjectivity, inconsistency, and variability, thus highlighting the need for a more objective and consistent evaluation method.

To address this challenge, we frame our research problem as a regression task. Specifically, our aim is to predict the numerical quality scores assigned to white wines, which range from 3 to 9. These scores reflect comprehensive evaluations based on sensory analysis and chemical composition.

The input for our predictive model consists of several physicochemical attributes that are measurable and quantifiable. These include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. Each of these attributes contributes differently to the sensory perception and overall quality of the wine, interacting through complex, nonlinear relationships.

Recognizing the intricacy of these relationships and the complexity of the dataset, our research leverages advanced ensemble and hybrid machine learning approaches. These models are chosen for their proven effectiveness in managing highly dimensional, correlated, and nonlinear data structures commonly observed in wine datasets.

Furthermore, our research addresses critical data challenges, including class imbalance (unequal distribution of wine quality scores), feature selection (identifying and prioritizing the most influential attributes), and data skewness (normalizing distributions for effective modeling). To enhance predictive accuracy and reliability, extensive preprocessing steps such as normalization, standardization (z-score scaling), and Box-Cox transformations for handling class imbalance are incorporated.

By thoroughly applying these preprocessing and modeling strategies and employing comprehensive evaluation metrics (such as RMSE, R^2, MAE, and accuracy), our research aims to deliver a predictive model that not only demonstrates high accuracy but also generalizes effectively to unseen datasets. Ultimately, this data-driven approach promises an objective, reproducible, and efficient alternative to conventional wine quality assessment methods, providing significant value to the wine industry in improving quality control and consumer satisfaction.

4 Data Analysis

For this study, we employed the white wine quality dataset sourced from the UCI Machine Learning Repository, comprising 4898 instances and 12 attributes, including one target variable (wine quality). Our analysis began with comprehensive exploratory data analysis (EDA) to understand the underlying patterns and characteristics of the dataset.

Initially, descriptive statistical analyses were conducted, summarizing key statistics such as mean, median, standard deviation, and quantile distributions of each attribute. The dataset exhibited a variety of distributional characteristics, notably highlighting skewness in attributes like residual sugar and total sulfur dioxide. To address these skewness issues and improve model performance, we applied the Box-Cox transformation, effectively normalizing the distributions of these attributes.

Distance and proximity analyses were performed using pairwise Euclidean distances and cosine similarities between data points. These measures revealed significant variability in chemical compositions (average Euclidean distance = 57.14), indicating that the wines varied widely across the dataset. High average cosine similarity (0.9891) suggested that despite variations in magnitude, many wine samples shared proportional similarities in their physicochemical profiles.

Correlation analyses were crucial for identifying the most significant predictors of wine quality. Alcohol content demonstrated the strongest positive correlation (0.44), while density (-0.31) and volatile acidity (-0.19) showed notable negative correlations. These

insights were integral in guiding feature selection, confirming the importance of these attributes in predicting wine quality accurately.

Finally, feature standardization (z-score scaling) was applied uniformly across all attributes to ensure comparability and enhance the stability of subsequent modeling processes. This meticulous data analysis and preprocessing established a solid foundation for developing robust predictive models, crucial for achieving accurate, reliable, and generalizable results.

5 Validation and Evaluation Methods

To ensure robust model evaluation, we partitioned the dataset into 70/15/15 train/validation/test sets. As this is a regression task predicting continuous quality scores, no resampling techniques (oversampling/undersampling) were employed. We used 5-fold cross-validation on the training set for hyperparameter tuning with Optuna, optimizing for RMSE (MLP) and R^2 (XGBoost), with early stopping to mitigate overfitting. Final model selection was based on validation set performance, followed by a single evaluation on the held-out test set.

Regression accuracy was assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2. To evaluate discrete predictive consistency, we rounded predictions to integer quality scores and calculated "rounded accuracy" (percentage of exact matches). While our primary focus was regression, we analyzed classification behavior through confusion matrices, observing error distributions and class-boundary challenges without formal precision/recall reporting. This dual approach ensured comprehensive assessment of both continuous prediction accuracy and practical grading utility.

6 Data-Driven Methods

Throughout the development of our wine quality prediction model, we systematically refined our approach based on both exploratory analysis and iterative experimentation.

6.1 Feature Engineering and Preprocessing

We retained all physicochemical features from the original dataset after initial correlation and variance analysis. Alcohol content, density, and volatile acidity were identified as key predictive features. Standardization was applied to all features using z-score scaling to ensure comparability and stabilize model training. A Box–Cox transformation was applied to the target variable (wine quality) to normalize its distribution and improve model convergence.

6.2 Distance and Proximity Analysis

To better understand the structure of the feature space, we analyzed pairwise Euclidean distances and cosine similarities among samples. Substantial variability in Euclidean distances and moderate-to-high cosine similarity indicated the need for flexible models capable of capturing both global and local patterns.

Figure 1: Correlation of Features

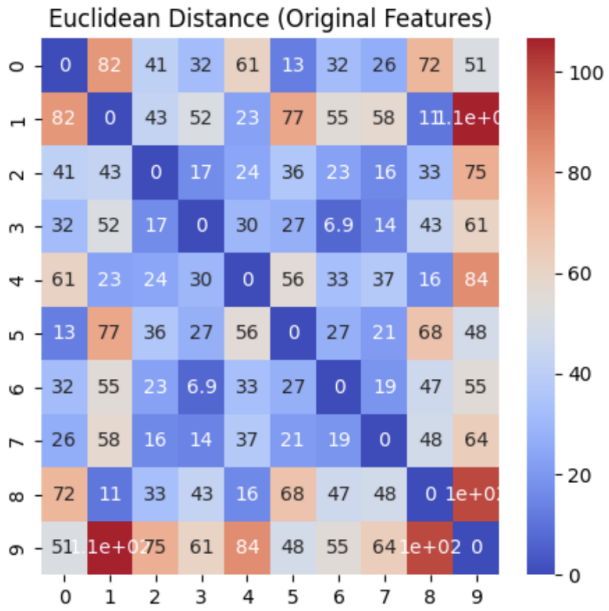

Figure 2: Euclidean Distance with Original Features

6.3 Model Development and Hyperparameter Settings

We evaluated a diverse set of regression models ranging from simple baselines to advanced ensembles and hybrid architectures. The key models and their final settings are summarized below:

- **Dummy Regressor**: Predicts the mean quality (default settings).
- **Decision Tree Regressor**: Default settings, with no maximum depth restriction.
- **Support Vector Regressor (SVR)**: RBF kernel, $C = 1.0$, $\epsilon = 0.1$.
- **Random Forest Regressor**:
  - $n\_estimators = 1500$

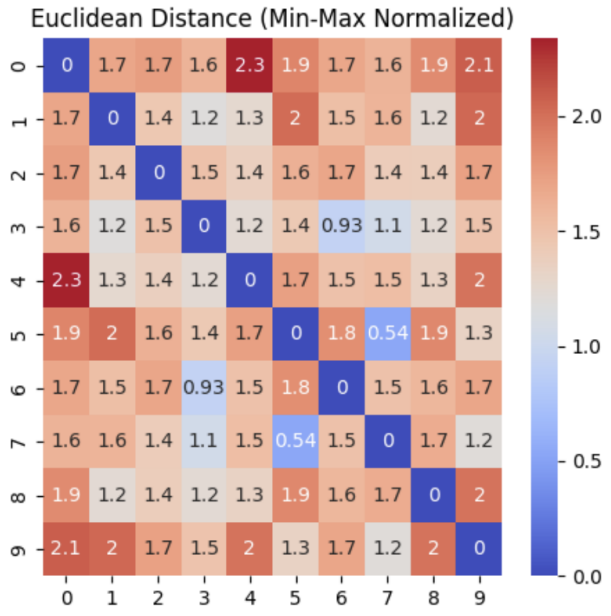

Figure 3: Euclidean Distance after Min-Max Normalization

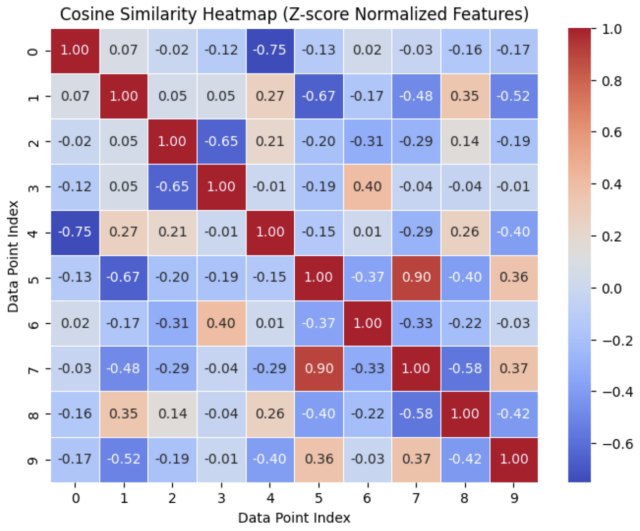

Figure 4: Cosine Similarity after Z-score Normalization

  - $max\_depth = 25$
  - $min\_samples\_split = 2$
  - $min\_samples\_leaf = 1$
- **XGBoost Regressor** (tuned with Optuna):
  - $n\_estimators = 670$
  - $max\_depth = 24$
  - $learning\_rate = 0.042$
  - $subsample = 0.8$
  - $colsample\_bytree = 0.7$
  - $reg\_alpha = 0.47$
  - $reg\_lambda = 0.71$
  - $objective = 'reg:squarederror'$
  - $random\_state = 42$
- **Stacked Regressor**: Combines XGBoost, Random Forest (1500 trees), and Ridge Regression ($\alpha = 1.0$) with Linear Regression as the meta-learner.
- **Multi-Layer Perceptron (MLP)** (tuned with Optuna):
  - 2 hidden layers, 192 units per layer
  - Dropout rate: 0.35
  - Batch size: 64
  - Optimizer: Adam
  - Learning rate: 0.0012
  - Batch normalization applied after each dense layer
- **Hybrid MLP + XGBoost**: Augmented the original features with inverse-transformed XGBoost predictions before training the MLP.

6.4 Hyperparameter Optimization Strategy

We employed *Optuna* for hyperparameter tuning. For XGBoost, the optimization objective was to maximize the cross-validated R^2 score, while for the MLP the objective was to minimize the validation RMSE. Early stopping was used to prevent overfitting in both cases. To understand how the number of estimators impacts

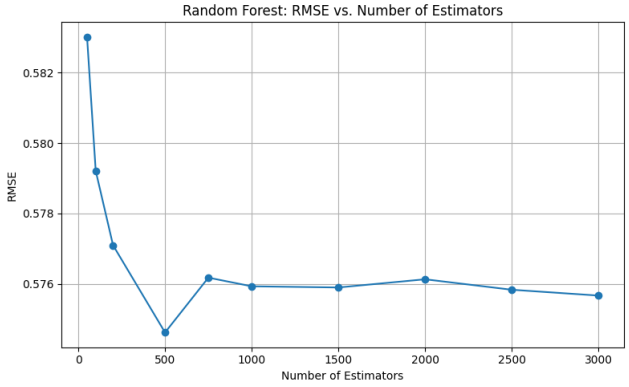

Figure 5: RMSE vs Number of Estimators in Random Forest

performance, we plotted RMSE against increasing estimator counts for both Random Forest and XGBoost (Figures 5 and 9, respectively).

For **XGBoost** (Figure 9), RMSE decreases significantly from approximately 0.655 to just over 0.560 as the number of estimators increases from 50 to 250. Beyond this point, the curve flattens, indicating diminishing returns in performance improvements beyond 500 estimators.

In the case of **Random Forest** (Figure 5), RMSE also declines initially—dropping from around 0.583 to 0.574 by 500 estimators—but fluctuates minimally thereafter. The improvements become marginal as the number of estimators exceeds 1000.

These trends suggest that while increasing the number of trees improves model performance early on, particularly for XGBoost, the benefit tapers off at higher estimator values. Thus, selecting an appropriate number of estimators involves balancing predictive performance with computational efficiency.

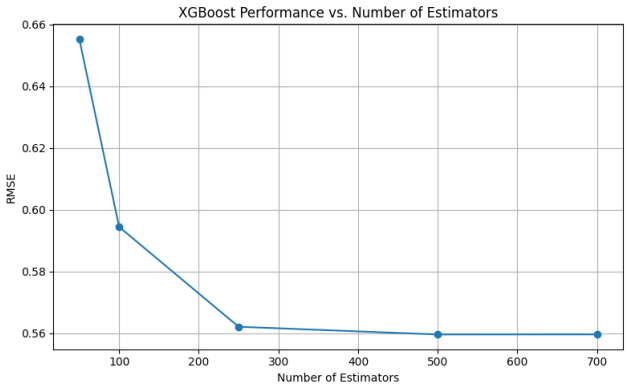

Figure 6: RMSE vs Number of Estimators in XGBoost

7 Experimental Results

7.1 Evaluation Metrics

Performance was assessed using:

- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of error.
- **R^2 Score**: Measures the proportion of variance explained.
- **Mean Absolute Error (MAE)**: Measures the average absolute error.
- **Rounded Accuracy**: Measures the percentage of integer-rounded predictions matching true labels.

7.2 Model Performance Summary

The results in Table 1 show a clear progression in model performance from simple to more advanced approaches. As expected, the Dummy Regressor, which predicts the mean of the target, performed the worst, establishing a baseline RMSE of 0.8553 and a negative R^2 score. Tree-based models such as Decision Trees and Random Forests showed marked improvements, with Random Forests achieving an RMSE of 0.5674 and R^2 of 0.548. Support Vector Regression performed moderately well but lagged behind ensemble methods. The XGBoost Regressor, tuned via Optuna, outperformed both Random Forest and SVR, validating the strength of gradient-boosted trees for this regression task. Neural networks also proved effective, with the standalone MLP achieving solid results. The best performance overall was achieved by the **Hybrid MLP + XGBoost** model, which leveraged the predictive structure of XGBoost and the nonlinear capacity of the MLP, resulting in the lowest RMSE (0.5588), highest R^2 (0.5612), and lowest MAE (0.3859). These results demonstrate that combining diverse model types through ensembling or feature augmentation can lead to superior predictive accuracy in structured tabular datasets like wine quality.

7.3 Key Observations

7.4 Confusion Matrix Analysis

To further evaluate model performance from a classification perspective, we converted regression outputs into discrete class predictions by rounding them to the nearest integer wine quality score. We then visualized and analyzed confusion matrices for the final three models: **Random Forest** (Figure 10), **XGBoost** (Figure 11), and the **Hybrid MLP + XGBoost** model (Figure 12).

Model	RMSE	R^2	MAE
Dummy Regressor	0.8553	-0.002	0.6257
Decision Tree Regressor	0.7148	0.286	0.5656
Support Vector Regressor (SVR)	0.6585	0.394	0.5103
MLP	0.6000	0.522	0.4522
XGBoost Regressor	0.5777	0.557	0.3729
Random Forest Regressor	0.5674	0.548	0.4325
Stacked Regressor	0.5668	0.549	0.4024
Hybrid MLP + XGBoost	**0.5588**	**0.5612**	**0.3859**

Table 1: Performance metrics of all evaluated models.

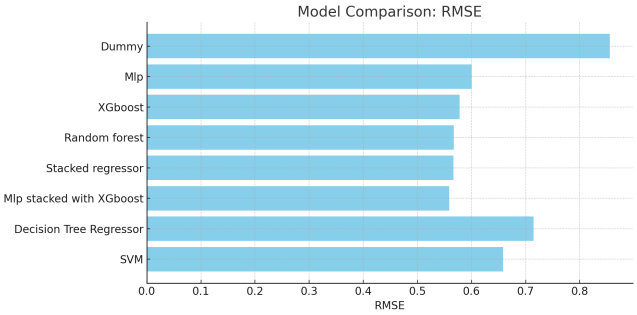

Figure 7: RMSE Comparison Across All Models

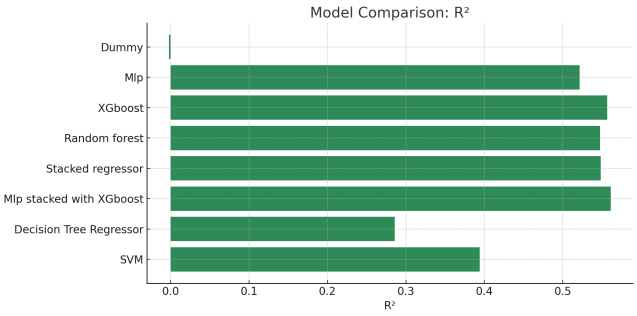

Figure 8: R^2 Comparison Across All Models

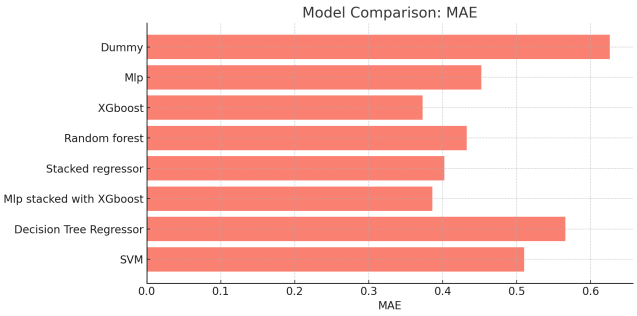

Figure 9: MAE Comparison Across All Models

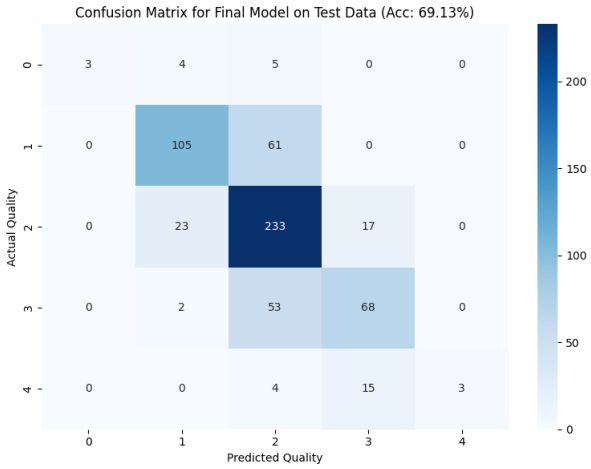

Figure 10: Confusion Matrix for Random Forest

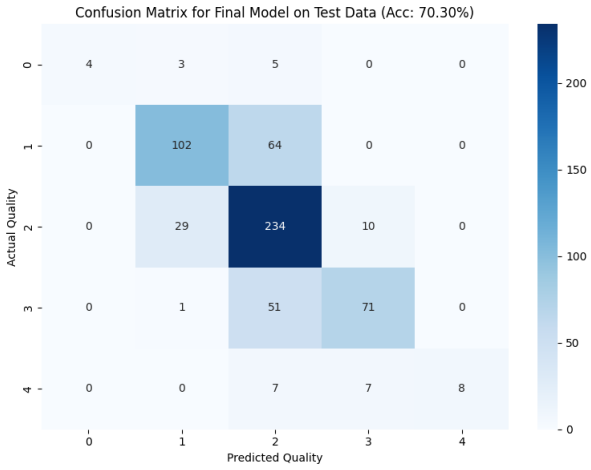

Figure 11: Confusion Matrix for XGBoost

Overall Accuracy. The rounded classification accuracies were as follows: Random Forest (70.30%), XGBoost (69.13%), and MLP + XGBoost (66.28%). Despite the hybrid model having the best regression metrics (RMSE and R^2), Random Forest achieved the highest classification accuracy. This is likely due to its robust handling of class boundaries and tendency to produce stable, central predictions.

Mid-Quality Class Performance. All models performed best on class 2 (corresponding to quality score 6), which also represents the most frequent class in the dataset. Misclassifications frequently occurred near class boundaries, especially between class 2 (quality 6) and class 3 (quality 7). Random Forest made the clearest distinction between these neighboring classes, correctly identifying 71 samples in class 3.

Edge Class Behavior. Performance on the lowest and highest quality classes was weak across all models. The hybrid model failed to correctly classify any instances in class 4 (quality 8), while Random Forest and XGBoost performed marginally better. This suggests that extreme classes, being underrepresented, are often confused with nearby central classes.

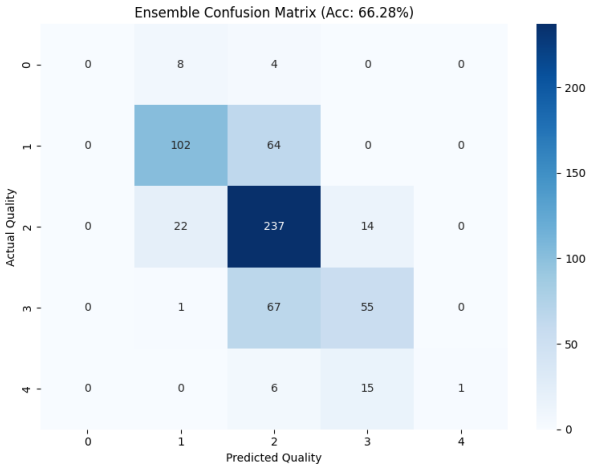

Figure 12: Confusion Matrix for MLP+XGBoost Ensemble

Error Magnitude. Most misclassifications were within ±1 of the true quality class. This indicates that while the models occasionally miss the exact class, they still capture relative quality fairly well. This consistency reinforces the usefulness of these models for practical wine scoring or quality ranking applications.

- **Random Forest (Figure 10)**: Best overall classification accuracy, strong class 3 detection.
- **XGBoost (Figure 11)**: Balanced performance, slightly better than hybrid model in classification.
- **Hybrid MLP + XGBoost (Figure 12)**: Most consistent regression, slightly blurred class boundaries.

These results suggest that while hybrid and boosting models offer superior numeric predictions, tree ensembles like Random Forest remain highly competitive when wine quality is treated as a discrete classification task.

- The **Hybrid MLP + XGBoost** achieved the best results, yielding the lowest RMSE (0.5588) and highest R^2 (0.5612).
- Tree-based ensemble models (Random Forest, XGBoost, and the Stacked Regressor) significantly outperformed simpler regressors like SVR and Decision Trees.
- The standalone MLP model performed competitively but showed enhanced predictive power when combined with XGBoost outputs.
- Applying the Box–Cox transformation helped normalize the target distribution, resulting in improved model stability and lower errors.

8 Conclusion

This project demonstrated a comprehensive and data-driven approach to predicting white wine quality using physicochemical attributes. Through iterative experimentation, we evaluated a range of regression models — from simple baselines like Dummy and

Decision Trees to more advanced architectures including Random Forest, XGBoost, and Multi-Layer Perceptrons (MLPs).

Key preprocessing steps, such as feature standardization and Box–Cox transformation of the target, helped stabilize model training and improve accuracy. We used Optuna to fine-tune hyperparameters for both XGBoost and MLP, which substantially boosted performance.

The best results were achieved by a hybrid model that combined inverse-transformed predictions from a tuned XGBoost regressor with an MLP. This approach achieved the lowest RMSE of 0.5588 and the highest R^2 of 0.5612, outperforming all other individual and ensemble models. These results highlight the strength of blending tree-based models with neural networks to capture both structured feature interactions and complex nonlinear patterns in tabular data.

8.1 Future Work

Several promising directions remain for expanding this work:

- **Extension to Classification**: While this project framed wine quality as a regression problem, future efforts could treat it as a classification task (e.g., low, medium, high quality) and evaluate performance using F1-score and precision-recall metrics.
- **Application to Red Wine Dataset**: The same methodology can be applied to the red wine dataset, allowing for comparison across wine types and potentially training a unified model.
- **Model Interpretability**: Tools like SHAP or LIME could be used to interpret feature importance and understand how individual chemical properties influence predictions.
- **Advanced Ensembles**: Future work could explore deeper stacking techniques, use model blending strategies, or incorporate LightGBM and CatBoost for additional gradient boosting alternatives.
- **Deployment and Real-World Testing**: Packaging the model into a lightweight web interface for winery use or quality control applications would demonstrate practical impact and usability.

Overall, this project provides a strong foundation for high-performing wine quality prediction and sets the stage for both methodological and real-world extensions.

References

[1] [n. d.]. Wine Quality Data Set. https://archive.ics.uci.edu/ml/datasets/Wine+Quality. Accessed: 2023-10-26.

[2] I. Bhardwaj, P. Tiwari, K. Jr Olejar, W. Parr, and D. Kulasiri. 2022. Machine learning application in wine quality prediction. *Machine Learning with Applications* 8 (2022), 100261. doi:10.1016/j.mlwa.2022.100261

[3] Bipul Shaw, Ankur Kumar Suman, and Biswarup Chakraborty. 2020. Wine Quality Analysis Using Machine Learning. In *Proceedings of the International Conference on Computational Intelligence and Data Science.*

[4] Akanksha Trivedi and Ruchi Sehrawat. 2018. Wine Quality Detection through Machine Learning Algorithms. In *2018 IEEE Conference on Computational Intelligence and Computing Research (ICCIC).* IEEE, 1–4.