Wine Quality Prediction Project

1. Introduction

The Wine Quality dataset from the UCI Machine Learning Repository was used for this project. It contains information about red and white wines, with various attributes such as pH, alcohol, and sugar content, alongside a quality score rated between 0 and 10. The primary objective was to predict wine quality using a combination of machine learning models, including Random Forest Regressor and Gradient Boosting Regressor.

2. Data Exploration & Preprocessing

The dataset comprised 1599 samples of red wine and 4898 samples of white wine. Key insights gained from exploratory data analysis included the identification of outliers in attributes such as 'residual sugar', 'free sulfur dioxide', and 'total sulfur dioxide'. These were handled using IQR-based capping. Features were normalized to the [0, 1] range using Min-Max scaling, and the categorical variable 'wine_type' was one-hot encoded.

3. Model Selection & Training

Random Forest Regressor and Gradient Boosting Regressor were selected for their strong predictive capabilities in regression tasks. The models were trained on an 80-20 train-test split, and hyperparameter tuning was performed using Grid Search to optimize their performance. Both models were evaluated on their mean absolute error (MAE) and mean squared error (MSE).

4. Performance Evaluation

The Random Forest Regressor achieved lower MAE and MSE compared to the Gradient Boosting Regressor. Hyperparameter tuning improved both models' performances, with the Random Forest

MAE improving from 0.0729 to 0.0726 and Gradient Boosting from 0.0886 to 0.0699. Combining the models into a stacked regressor further reduced errors, achieving an MAE of 0.0664.

5. Challenges & Future Work

Key challenges included handling outliers, ensuring proper feature scaling, and identifying the best model. These were addressed through systematic preprocessing, hyperparameter tuning, and feature importance analysis. Future work could involve using SHAP for deeper interpretability and exploring alternative optimization techniques like Bayesian Optimization for faster hyperparameter tuning.