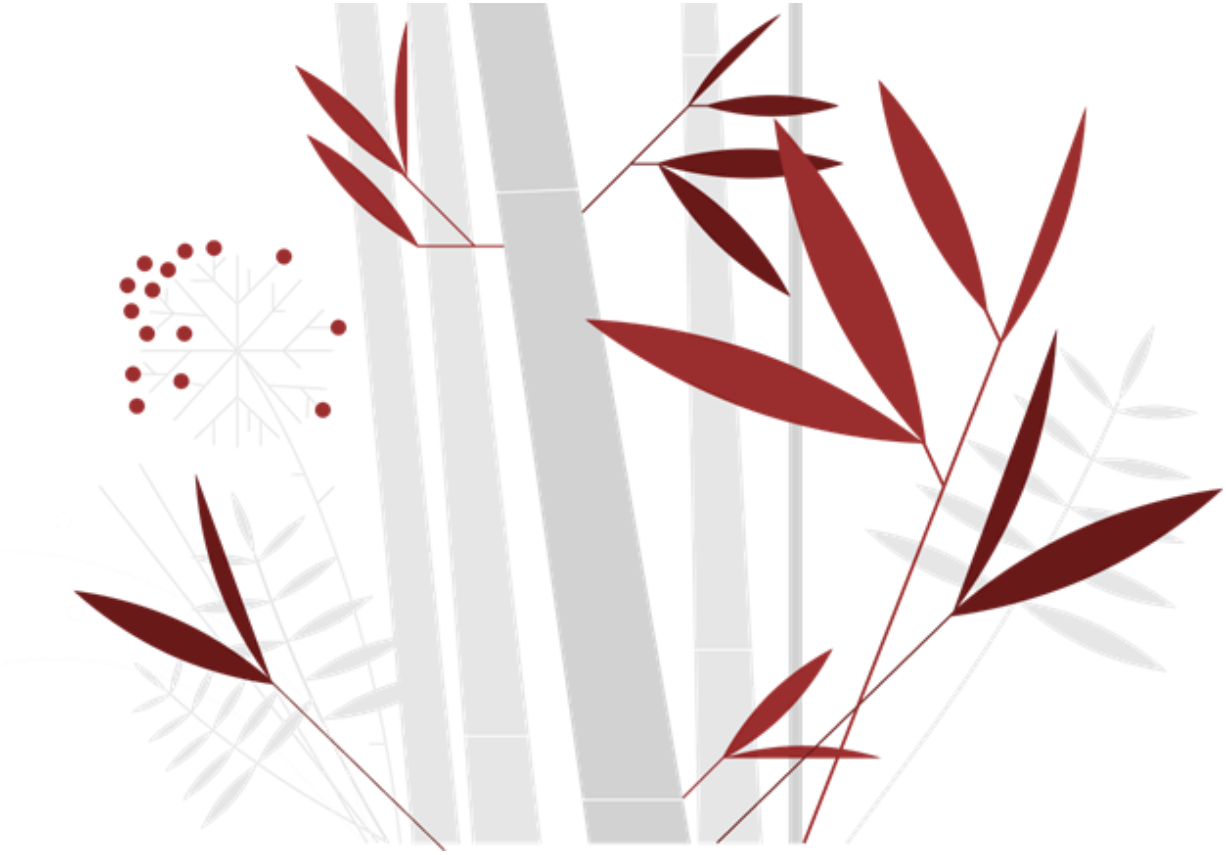
FINAL PROJECT REPORT

RED WINE QUALITY PREDICTION

Data Science & Machine Learning

Win, Mya Ei

4/5/2025



Abstract

This project explores the application of machine learning techniques to predict the quality of red wine based on its physicochemical properties. The analysis includes both regression and classification approaches. Regression is used to predict the exact quality score ranging from 0 to 10, while classification determines whether a wine is high quality or not based on quality label. The models selected include Linear Regression, Logistic Regression, MLP (Neural Network), and Gradient Boosting, with hyperparameter tuning applied to MLP and Gradient Boosting models using GridSearchCV. The project evaluates the performance of each model and explains why certain models outperformed others. Results show that Gradient Boosting performed the best overall in both tasks.

Introduction

The Red Wine Quality dataset contains physicochemical data of 1,599 wine samples. Each sample includes 11 numeric features 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 rated with a quality score ranging from 0 to 10, determined by human tasters.
This project aims to predict two outcomes:
1. The exact wine quality score for regression task.
2. Whether the wine is high quality (quality ≥ 7) or not for classification task.

Using both simple and complex models, this project provides insights into the prediction performance of different algorithms, including how hyperparameter tuning can enhance results.

Methodology and Models

Data Preprocessing

The wine quality dataset was loaded from a CSV file into a pandas Data Frame, and basic statistics were reviewed to understand the distribution of features. Any missing values were removed to ensure data quality, and the data was then split into training and test sets using an 80/20 split.

Model Selection and Explanation

The following models were chosen to analyze for this dataset:

Linear Regression: It was used as a baseline model in this project. Although it is easy to interpret and quick to run, it performed poorly due to the non-linear relationships in the wine data, with an R^2 score of only 0.40.
Logistic Regression: It is simple and interpretable. In this project, it served as a baseline model for classifying wine into high and low quality.
MLP (Multi-Layer Perceptron): A basic type of neural network capable of modeling non-linear relationships. Applied to both regression and classification. The MLPClassifier was used with hyperparameters.
Gradient Boosting: An ensemble method that builds models sequentially to minimize errors. Known for high performance in structured data tasks. It was used for both regression and classification, with hyperparameters tuned using GridSearchCV.

Model Evaluation and Tuning

The performance of models was evaluated using appropriate metrics: Mean Squared Error (MSE) and R^2 score for regression, and Accuracy, Precision, Recall for classification. GridSearchCV was applied to MLP and Gradient Boosting models to find optimal hyperparameters such as number of estimators, learning rate, max depth (for Gradient Boosting), and hidden layer size and activation function (for MLP).

Results and Comparison

Regression Models

Linear Regression provided a baseline with limited performance due to its assumption of linearity. MLP Regressor improved results by learning complex patterns. Gradient Boosting Regressor achieved the lowest MSE and highest R^2 score, proving to be the most accurate for predicting the wine score.

Model	MSE	R^2 Score
Linear Regression	0.39	0.40

MLP Regressor	0.37	0.43
Gradient Boosting	0.35	0.46

Classification Models

Logistic Regression performed reasonably well but had difficulty classifying the minority high-quality class. MLP Classifier improved accuracy slightly, while Gradient Boosting Classifier delivered the best results. The confusion matrix shows a better balance between true positives and true negatives with Gradient Boosting.

Model	Accuracy
Logistic Regression	0.85
MLP Classifier	0.86
Gradient Boosting	0.89

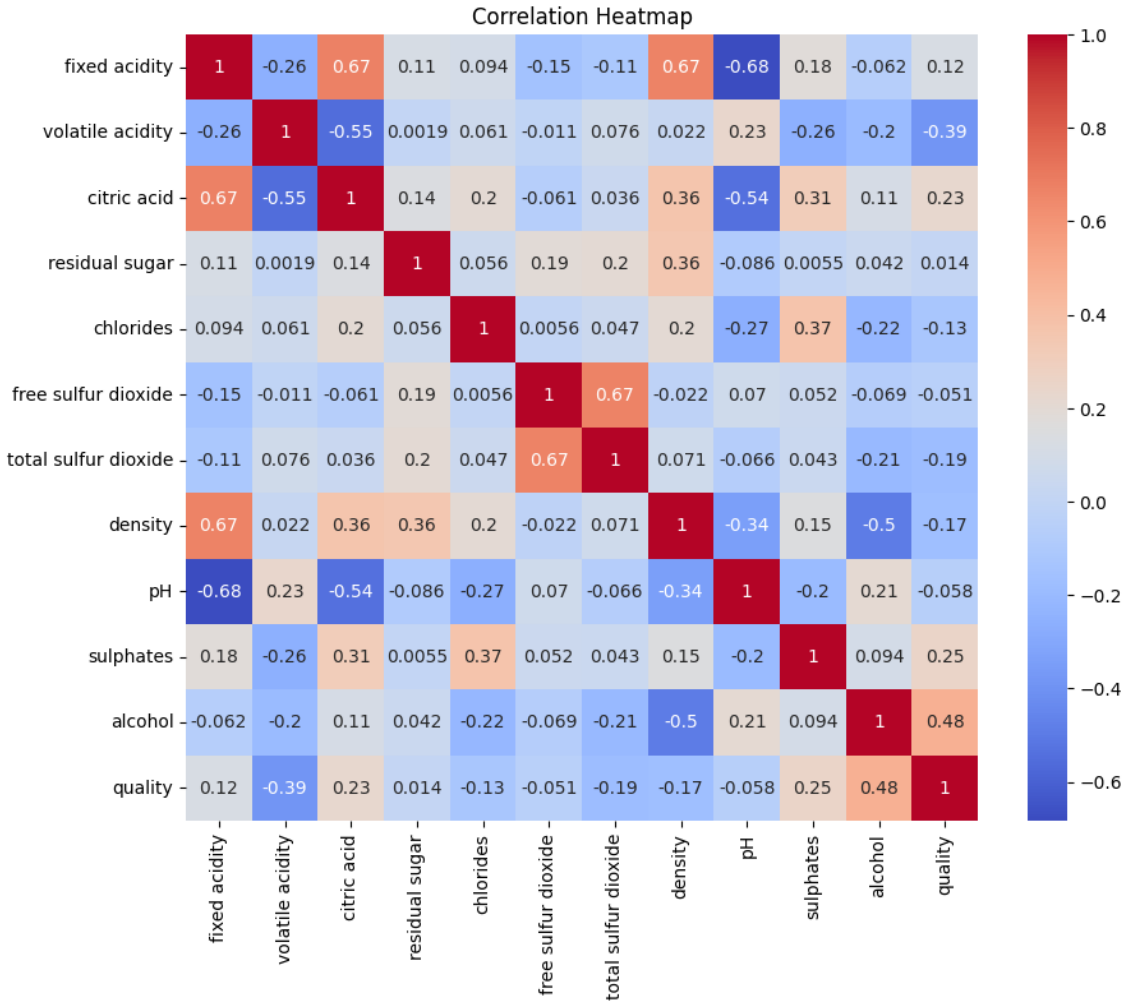
Visualizations and Interpretation

Several visualizations were used to support this analysis:

Correlation Heatmap

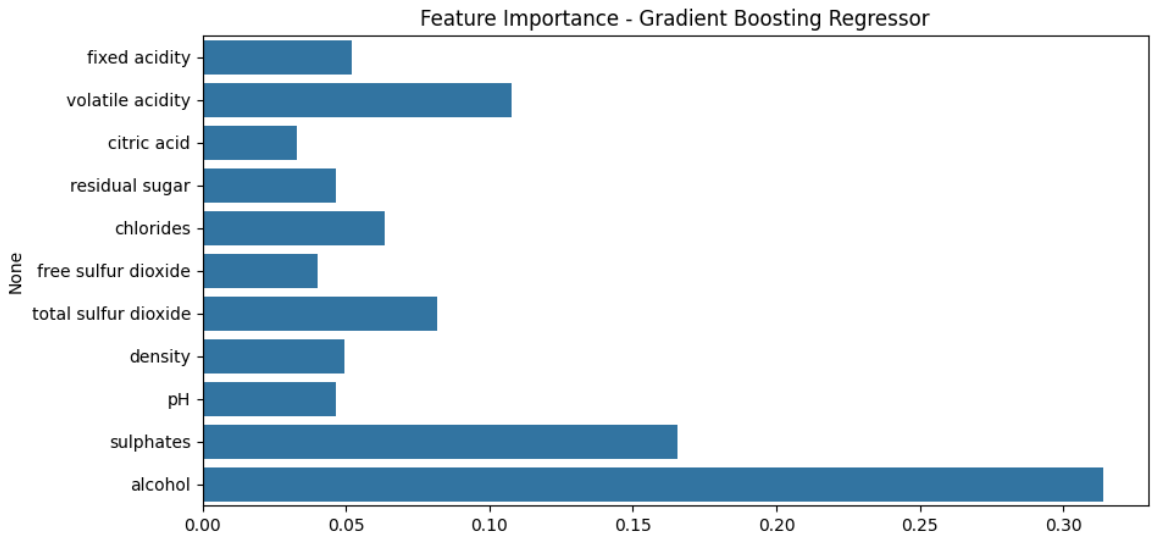
The correlation heatmap was used to explore the relationships between all the features in the dataset, including the target variable quality. From the heatmap, alcohol shows the strongest positive correlation with wine quality (0.48), meaning that wines with higher alcohol levels tend to receive better quality ratings. On the other hand, features like volatile

acidity have a negative correlation, which may lower the wine's quality. .



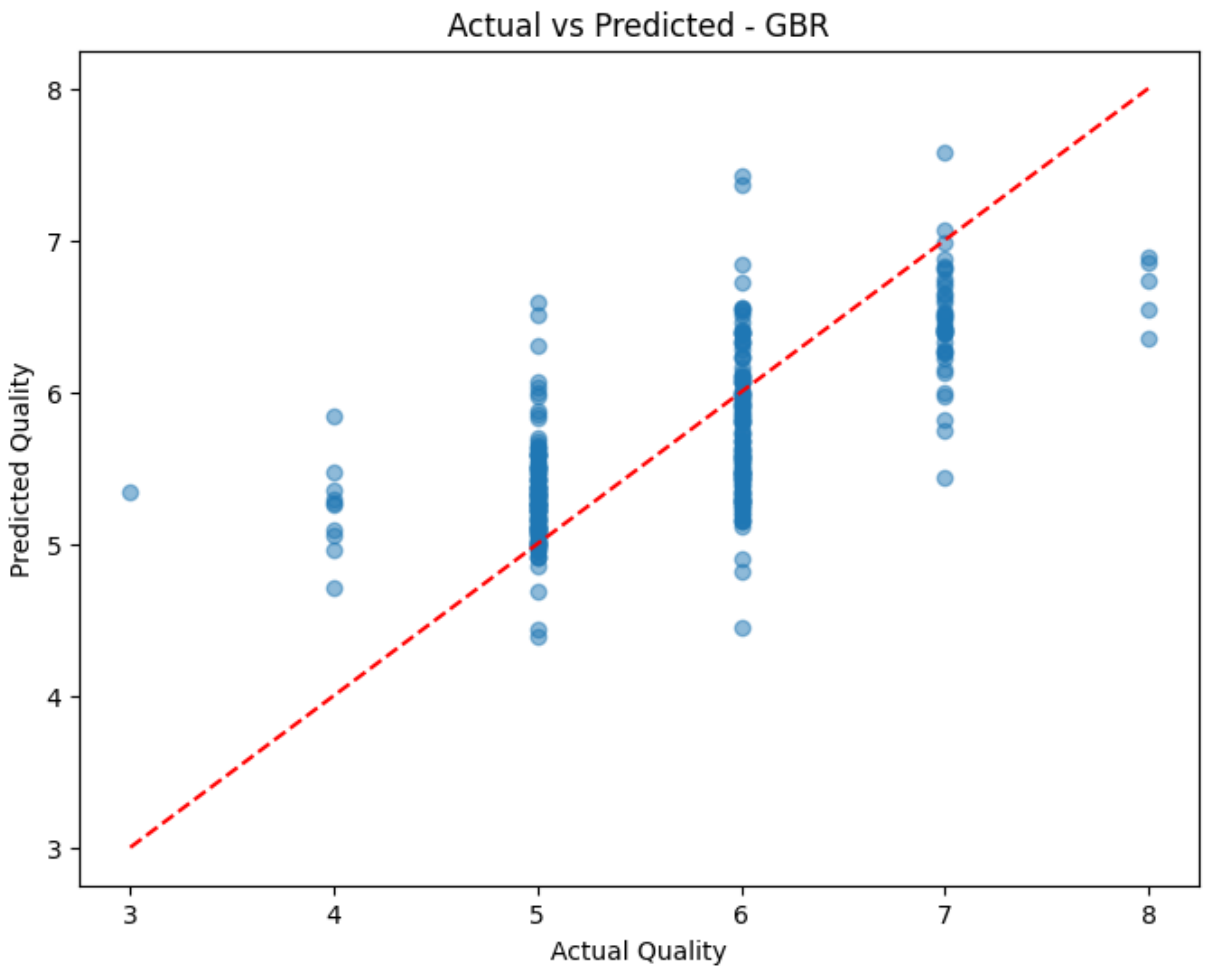
Feature Importance (Gradient Boosting)

This graph shows which dataset features the model considered most useful for predicting wine quality. According to the graph, alcohol had the highest importance, followed by sulphates and volatile acidity.



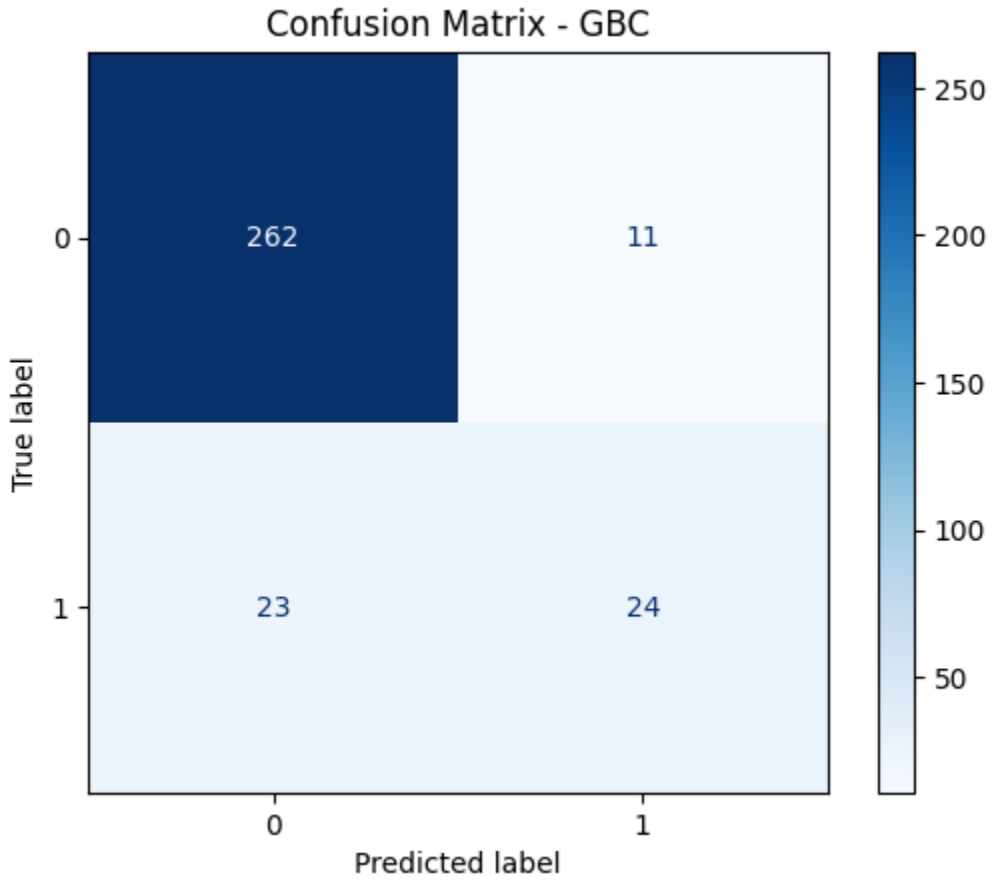
Actual vs Predicted Plot

This chart shows how close the predicted wine quality values are to the actual ones. Most points are near the red line, which means the Gradient Boosting model made accurate predictions overall.



Confusion Matrix

The confusion matrix shows the model predicted most low-quality wines correctly, with 262 out of 273. It also identified 24 out of 47 high-quality wines, but there's still room to improve its accuracy for high-quality predictions.



.

Discussion

The results show that Gradient Boosting was the most effective model for both regression and classification tasks. It handled non-linear relationships well and improved significantly with hyperparameter tuning. The MLP models also performed well after tuning, though they required more time to train and were slightly less accurate. Linear and Logistic Regression served as useful baselines but were limited by their linear assumptions, especially in a dataset with more complex patterns.

Conclusion

This project demonstrated how different machine learning models perform on the Red Wine Quality dataset using both regression and classification approaches.
Through comparison, it became clear that **Gradient Boosting** outperformed other models due to its ability to handle complex data and its responsiveness to tuning.
The project helped me develop a deeper understanding of model selection, evaluation metrics, feature importance, and the role of cross-validation in building reliable models.

For future work, I would explore:

- Handling class imbalance using resampling techniques
- Trying different tree depths and boosting parameters
- Feature selection or dimensionality reduction
- Using cross-validation (as done in GridSearchCV) for more robust evaluation

Reference

https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data