Red Wine Prediction Model

Saksham Saini

December 8, 2024

1 Project Objective

In this project, we explore Exploratory Data Analysis (EDA) and classification tasks to deliver balanced and reliable results. We evaluate various machine learning algorithms, categorized as follows:

- **Linear Models**: Logistic Regression
- **Tree-Based Models**: Decision Trees and Random Forests
- Support Vector Classifier: SVC
- Boosting Models: Gradient Boosting, XGBoost, LightGBM, AdaBoost
- **Distance-Based Models**: K-Nearest Neighbors

We will explore key hyperparameters for each algorithm, optimizing them to improve performance and accuracy.

2 Data Cleaning

- **Removing Outliers**: Rows with extreme values are removed, accounting for about 3.2% of the dataset.
- **Capping Data**: Values outside the interquartile range are capped.

3 Different Models

3.1 Linear Model: Logistic Regression

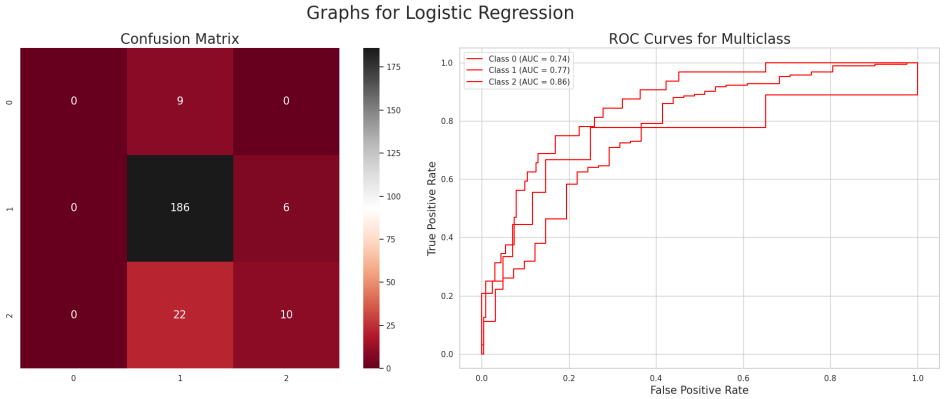


Figure 1: Logistic Regression Results

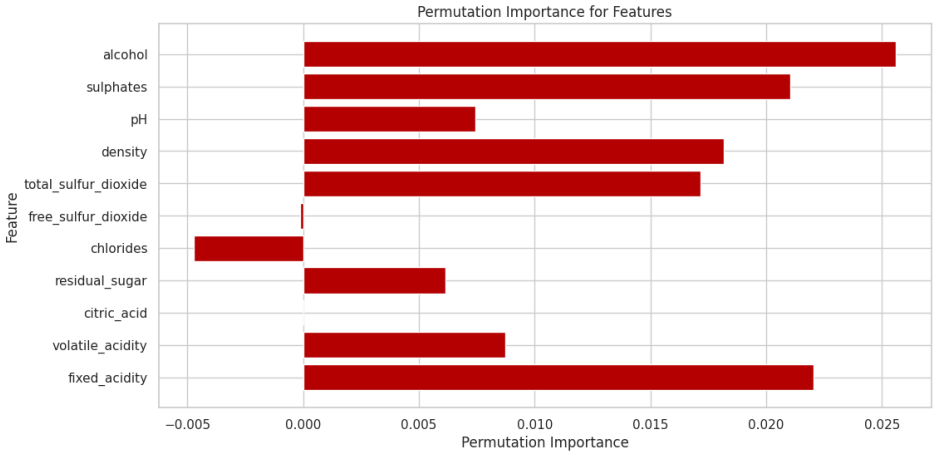


Figure 2: Model Overview

3.2 Tree-Based Model: Decision Tree Classifier

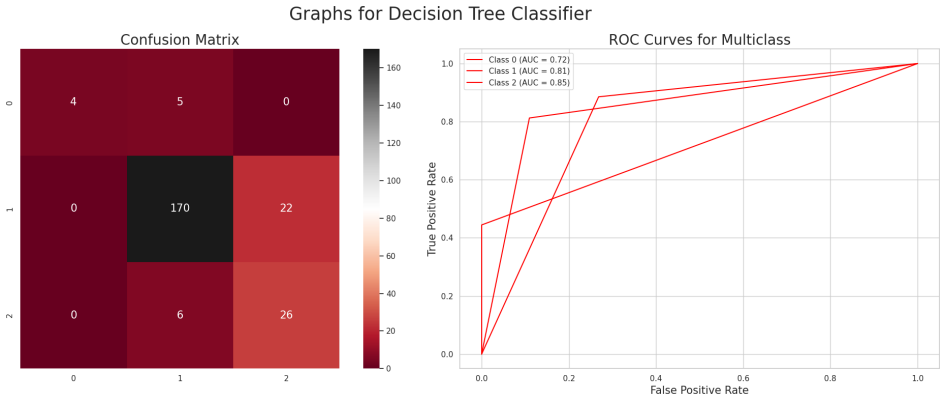


Figure 3: Decision Tree Decision Boundaries

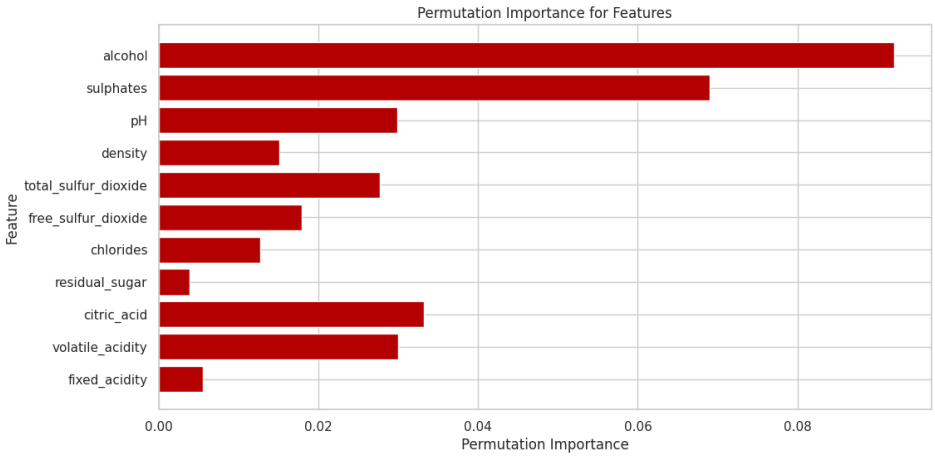


Figure 4: Decision Tree Classifier Visualization

3.3 Tree-Based Model: Random Forest Classifer

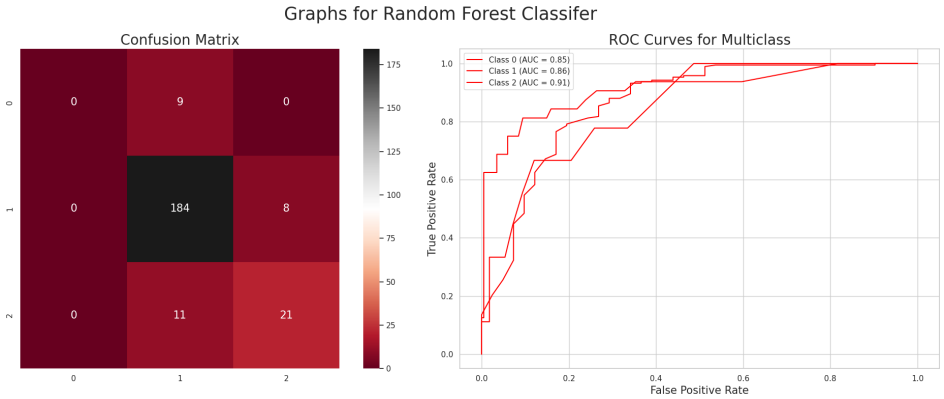


Figure 5: Random Forest

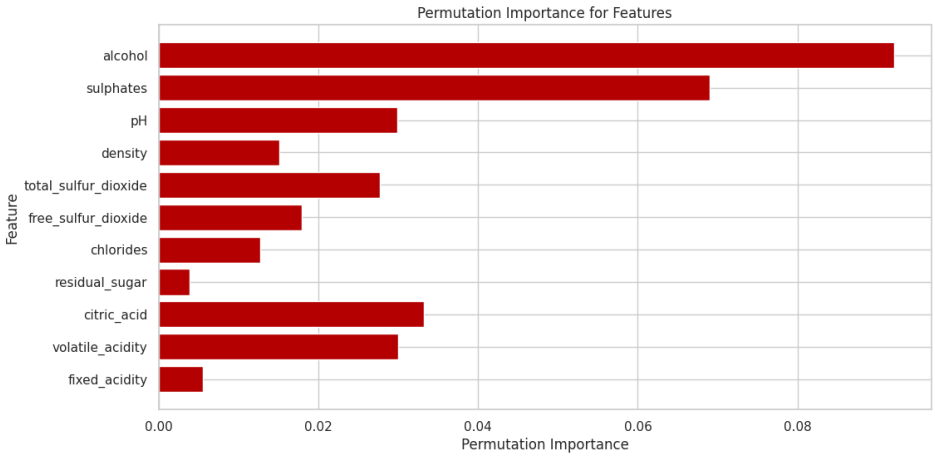


Figure 6: Random Forest

3.4 Support Vector Classifier

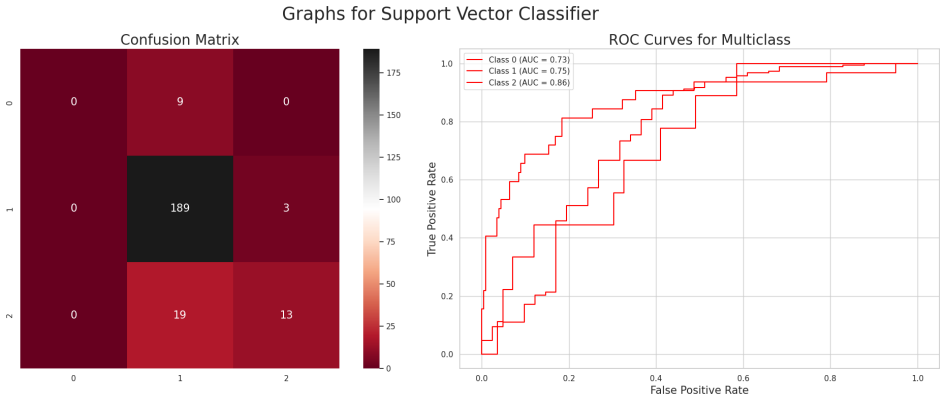


Figure 7: Support Vector Classifier

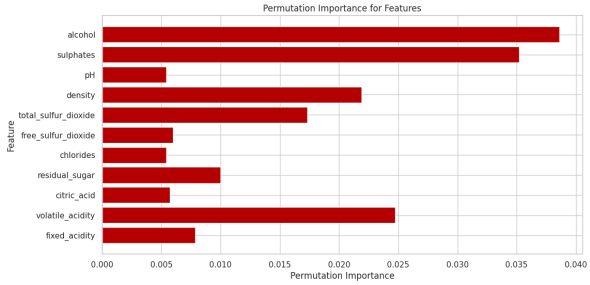
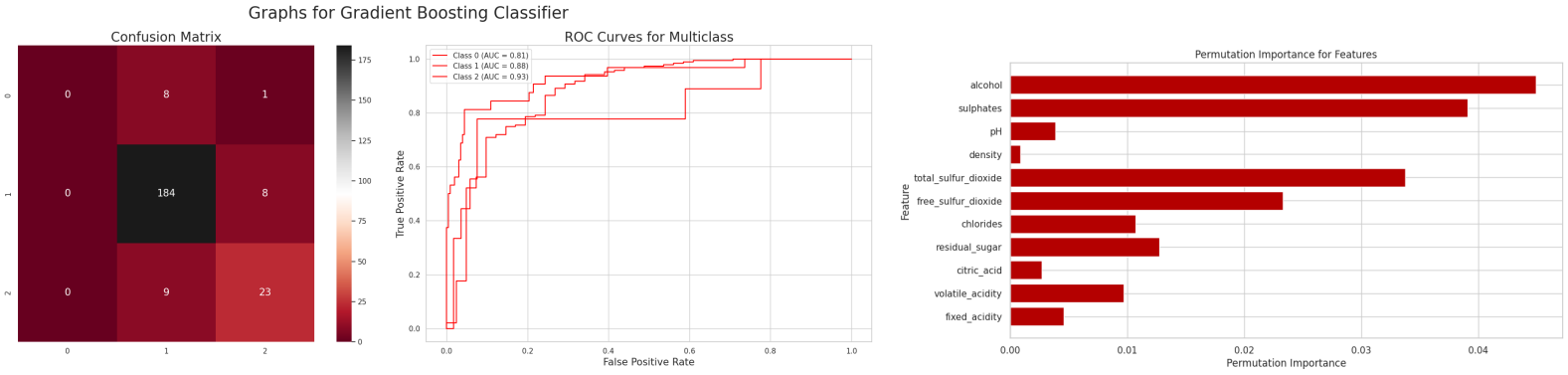
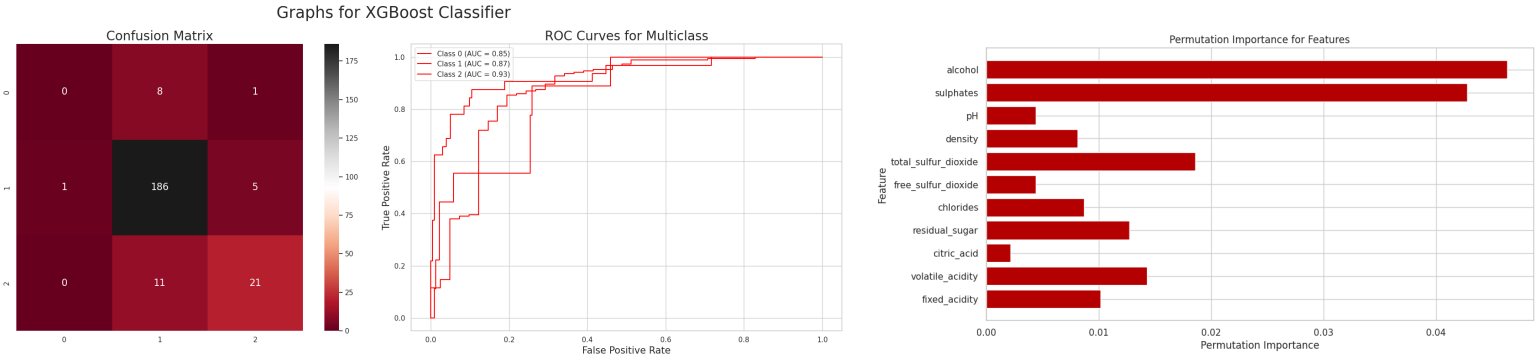


Figure 8: Support Vector Classifier

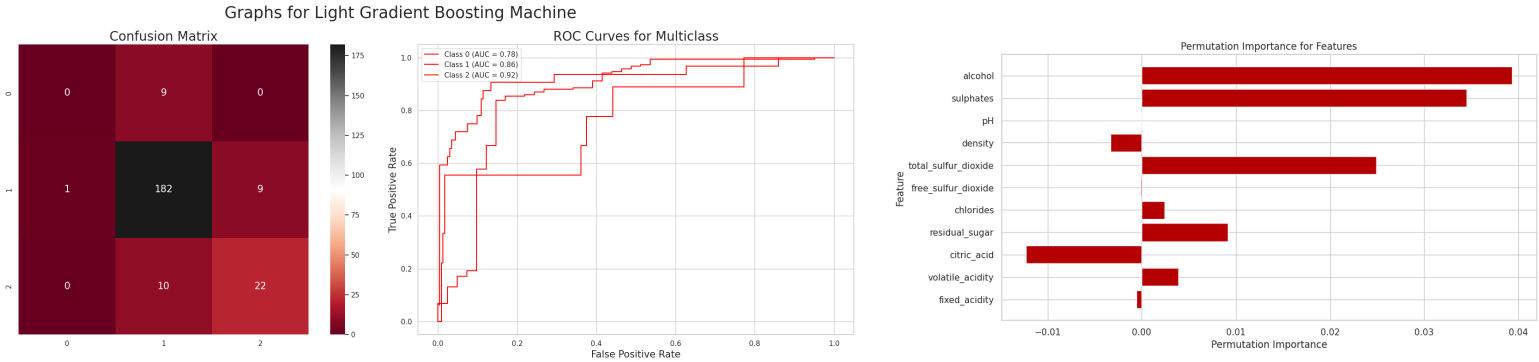
3.5 Boosting Model: Gradient Boosting Classifier



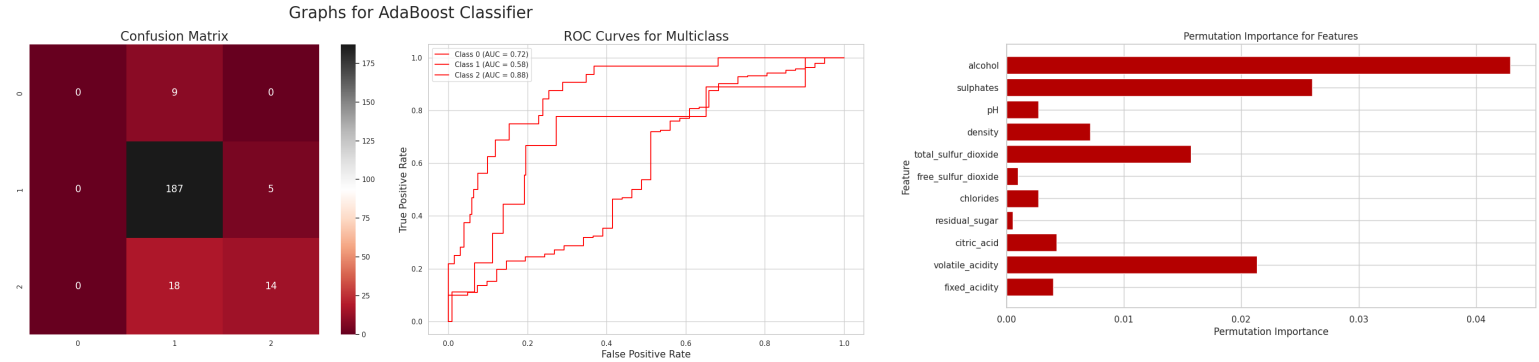
3.6 Boosting Model: XGBoost Classifier



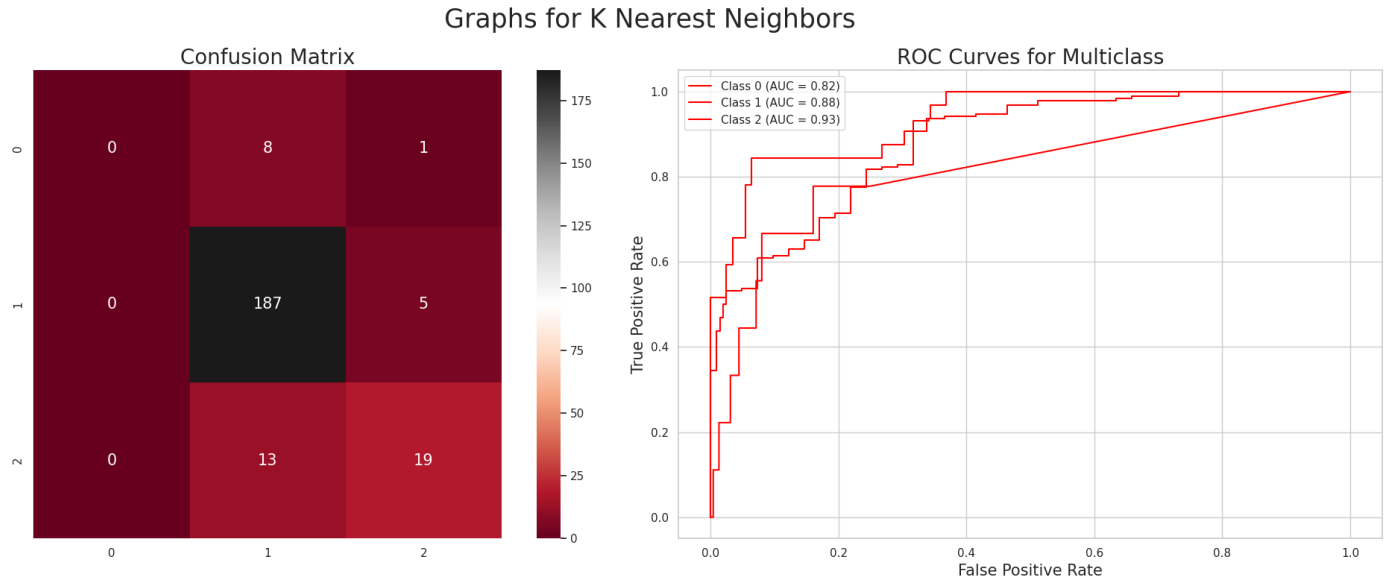
3.7 Boosting Model: Light Gradient Boosting Machine



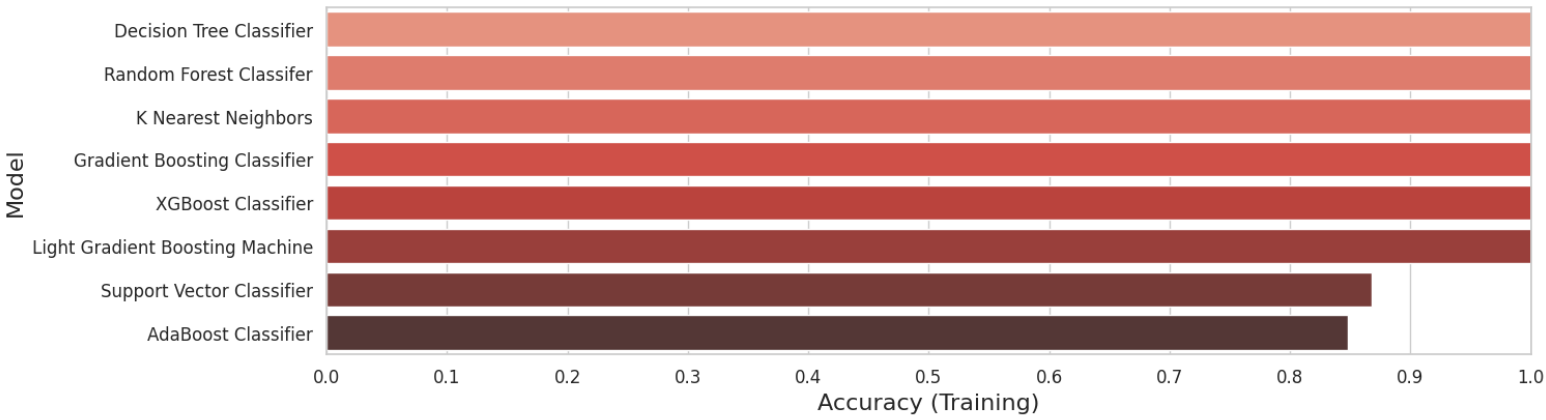
3.8 Boosting Model: AdaBoost Classifier

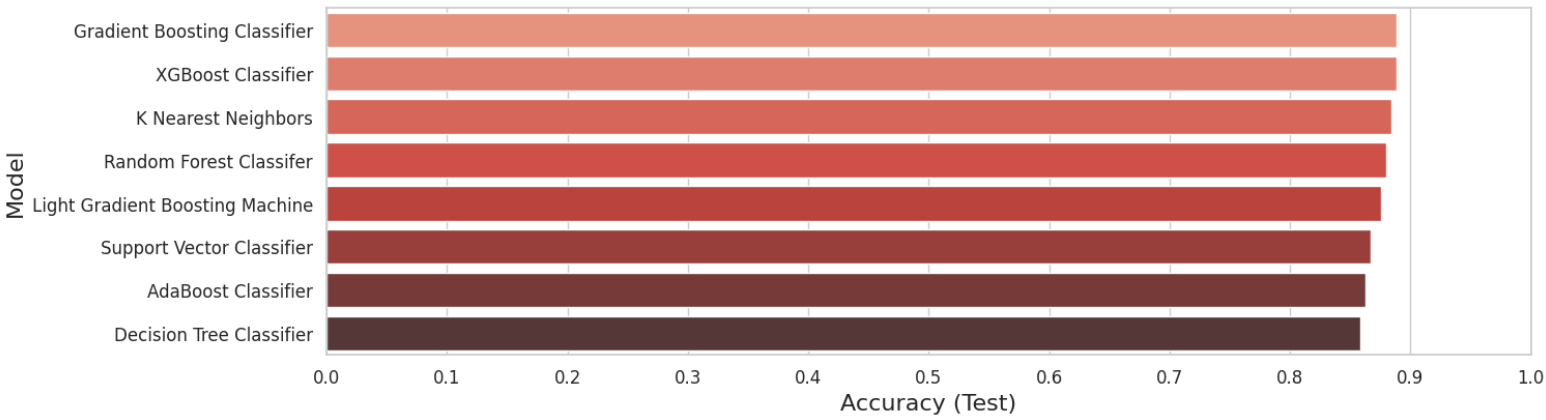


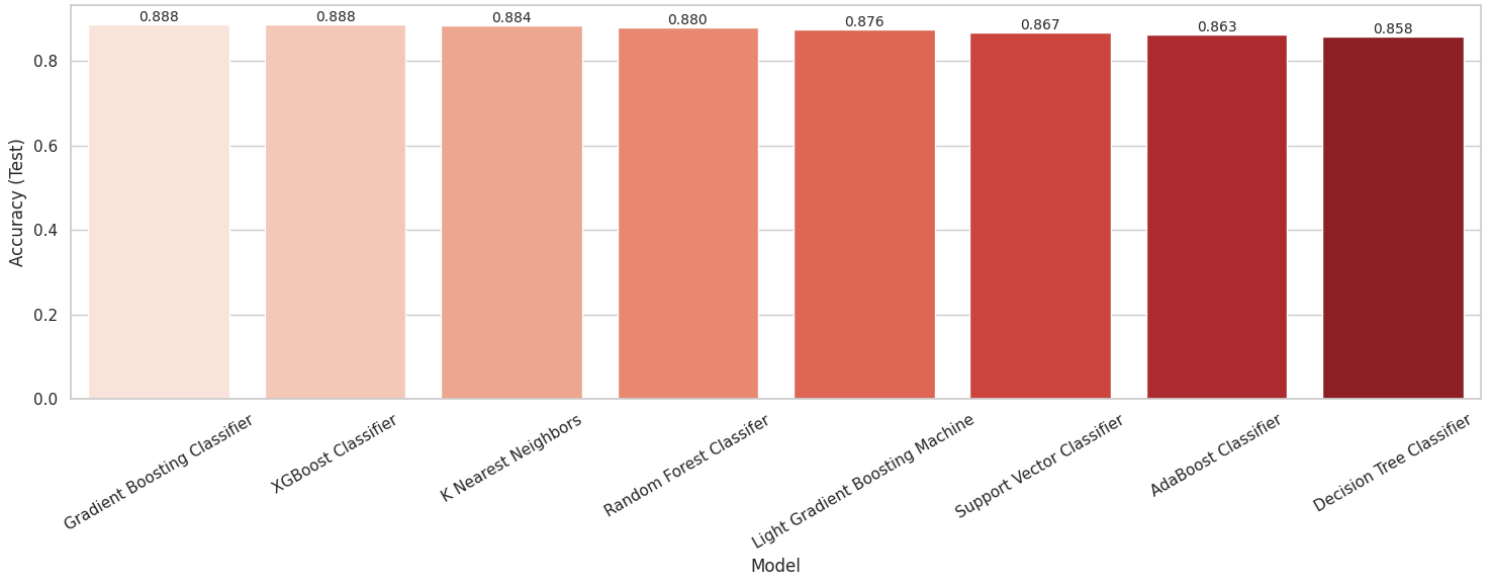
3.9 Distance-based Model: K Nearest Neighbors

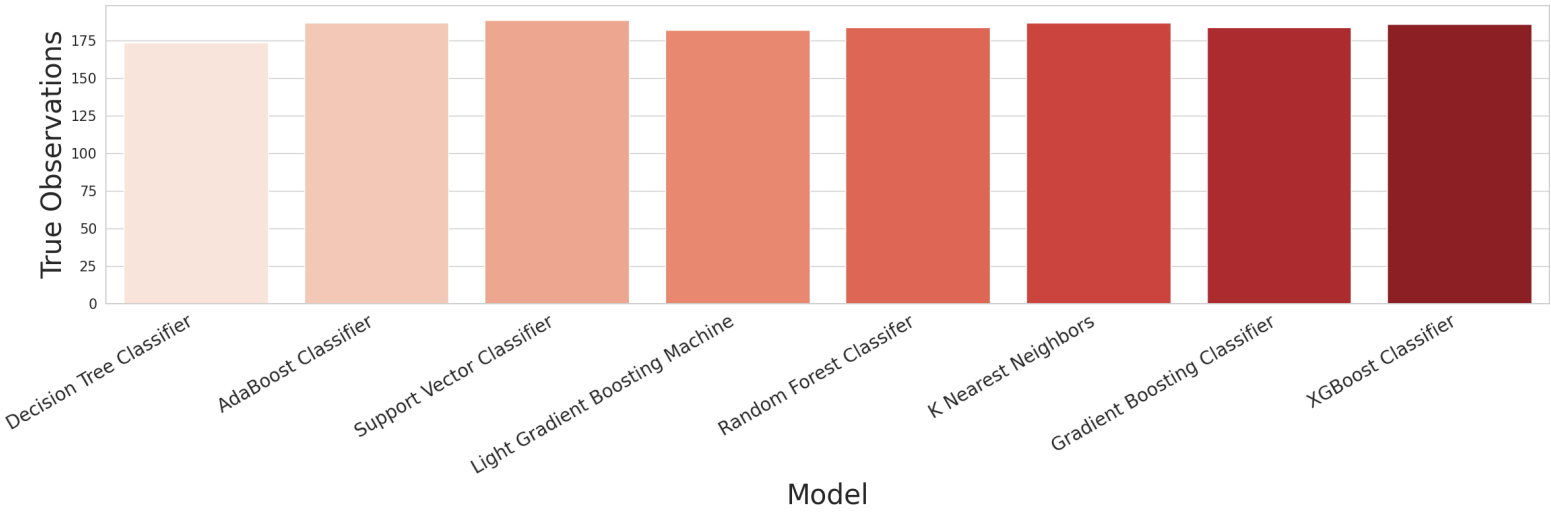


4. Compare Different Model Accuracy






0.888
0.888
0.884
0.880
0.876
0.867
0.863
0.858
Accuracy (Test)
0.0
0.2
0.4
0.6
0.8
Gradient Boosting Classifier
XGBoost Classifier
K Nearest Neighbors
Random Forest Classifer
Light Gradient Boosting Machine
Support Vector Classifier
AdaBoost Classifier
Decision Tree Classifier
Model


True Observations
0
25
50
75
100
125
150
175
Decision Tree Classifier
AdaBoost Classifier
Support Vector Classifier
Light Gradient Boosting Machine
Random Forest Classifer
K Nearest Neighbors
Gradient Boosting Classifier
XGBoost Classifier
Model

5. Conclusion

Based on the accuracy plots and the overall evaluation metrics, the XGBoost Classifier emerges as the best-performing method for this wine quality classification task.

1. **High Accuracy:** XGBoost consistently achieves high accuracy on both the training and testing datasets, indicating its ability to learn the patterns in the data effectively and generalize well to unseen data.
2. **True Observations:** XGBoost demonstrates a high number of correctly classified instances (TP + TN), further reinforcing its strong performance in accurately identifying both positive and negative cases.
3. **Robustness:** XGBoost is known for its robustness to outliers and noisy data, which is crucial for real-world datasets.
4. **Feature Importance:** The permutation importance analysis for XGBoost highlights the key features contributing to the model's predictions, providing valuable insights into the dataset.
5. **Efficiency:** Although not directly evident from the code, XGBoost is generally regarded as a computationally efficient algorithm, especially when compared to some other ensemble methods.

While other models like LightGBM and Random Forest also perform well, XGBoost's overall performance, considering accuracy, robustness, and interpretability, makes it the most suitable choice for this particular wine quality classification problem. Based on the analysis and comparison, I recommend using the XGBoost Classifier for this task. However, further fine-tuning of hyperparameters might lead to even better performance.