Wine Quality Prediction Model of Red Wine Quality Dataset

Data Science II

COSC 4337

Submitted to

DR. Ricardo Vilalta

Submitted by

Dylan Flores (1872416)

Introduction

The objective of this project is to develop a predictive model for assessing the quality of wine based on physicochemical tests from the Wine Quality dataset. This report outlines the methodology adopted to preprocess the data, select appropriate modeling techniques, tune model parameters, and evaluate model performance to predict whether the wine quality is above average. The Dataset used is the same as Deliverable one which is the wine quality dataset, provided by provided by the UCI Machine Learning Repository, which includes two separate datasets, a white wine sample and a red wine sample dataset. To give a quick overview of the dataset before jumping into the Methodology of this project; The data includes physicochemical and sensory data from samples of red and white variants of the Portuguese "Vinho Verde" wine. Inside the datasets CSV file there is a total of 1600 examples to analyze and 13 different features that includes acidity, sugar content, alcohol level, and more, alongside a quality rating assigned by wine experts.

Methodology

- **Data Loading and Preprocessing**

In the last deliverable we performed data processing and loading, so to quickly go over this the data was loaded from a CSV file containing several physicochemical properties of wines. The quality feature, originally a numeric score, was transformed into a binary classification task: wines scoring above 5 were labeled as 1 (above average) and the rest as 0 (below average). This transformation simplifies the analysis while maintaining the essence of distinguishing higher quality wines.

Feature Scaling: To ensure that all features contribute equally to the model performance and to improve the convergence during training, the data was standardized. This involves subtracting the mean and dividing by the standard deviation of each feature.

- ## Model Selection

Three different models were chosen to cover a range of machine learning techniques:

1. **Logistic Regression:** A baseline model for binary classification tasks that provides a probabilistic understanding and efficient training.
2. **Random Forest:** An ensemble method known for its high accuracy and robustness against overfitting, using multiple decision trees to make its predictions.
3. **Support Vector Machine (SVC):** Effective in high-dimensional spaces, especially useful when the boundary between classes is not linear.

- ## Hyperparameter Tuning

Hyperparameter tuning was conducted using Grid Search with cross-validation:

1. **Logistic Regression:** Regularization strength C was varied over several orders of magnitude to find the optimal balance between bias and variance.
2. **Random Forest:** Parameters such as the number of trees (n_estimators) and the maximum depth of trees (max_depth) were tuned to optimize performance and prevent overfitting.
3. **SVC:** The penalty parameter C and the kernel coefficient gamma were adjusted to fine-tune the model's sensitivity to the data distribution.

This approach ensures that each model is optimized for the best possible performance on the given dataset.

Model Performance

- **Model Performance Overview**

The evaluation of the three predictive models — Logistic Regression, Random Forest, and Support Vector Classifier (SVC) — using the Wine Quality dataset has provided a comprehensive picture of their ability to classify the quality of wines.

- Logistic Regression Results

```
Classification report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.70      0.75      0.72       141
           1       0.79      0.74      0.77       179

    accuracy                           0.75       320
   macro avg       0.74      0.75      0.75       320
weighted avg       0.75      0.75      0.75       320
```

The Logistic Regression model achieved an accuracy of 0.75 with a macro average F1-score of 0.75. The precision scores for the below-average (0) and above-average (1) classes were 0.70 and 0.79, respectively, indicating a relatively higher ability of the model to correctly identify the above-average quality wines. The recall scores suggest that the model is slightly better at identifying above-average quality wines than below-average quality wines.

- **Random Forest Results**

```
Classification report for Random Forest:
              precision    recall  f1-score   support

           0       0.76      0.76      0.76       141
           1       0.81      0.82      0.81       179

    accuracy                           0.79       320
   macro avg       0.79      0.79      0.79       320
weighted avg       0.79      0.79      0.79       320
```

Random Forest outperformed Logistic Regression with an overall accuracy of 0.79 and a macro average F1-score of 0.79. It showed a notable improvement in precision and recall for both classes compared to Logistic Regression, reflecting its robustness and its strength as a classifier for this dataset. With precision scores of 0.76 for class 0 and 0.81 for class 1, it demonstrates a strong capability to distinguish between the two quality types accurately.

- **SVC Results**

```
Classification report for SVC:
              precision    recall  f1-score   support

           0       0.70      0.74      0.72       141
           1       0.78      0.75      0.77       179

    accuracy                           0.75       320
   macro avg       0.74      0.75      0.74       320
weighted avg       0.75      0.75      0.75       320
```

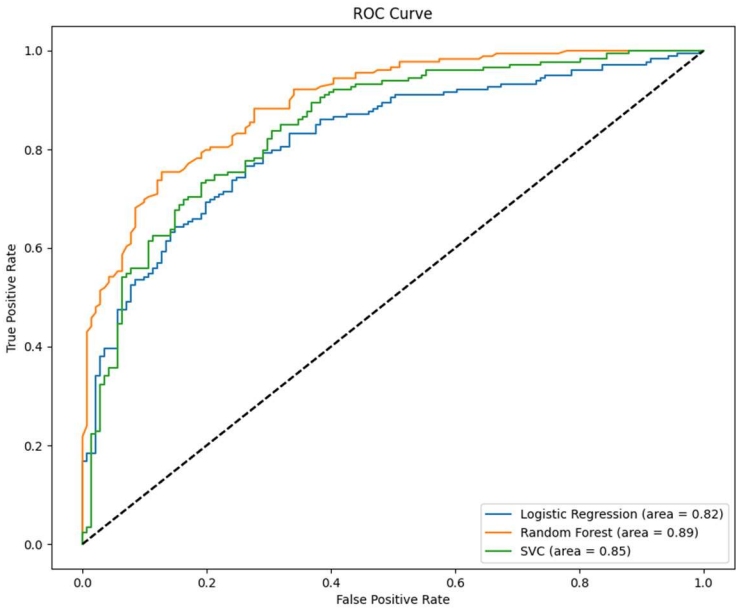
The SVC model matched the Logistic Regression model in overall accuracy, with a score of 0.75, and had a macro average F1-score of 0.74. The precision and recall scores for both classes were very close to those of the Logistic Regression model, suggesting similar performance characteristics between these two models.

Model Performance

- **Graphical Analysis Overview**

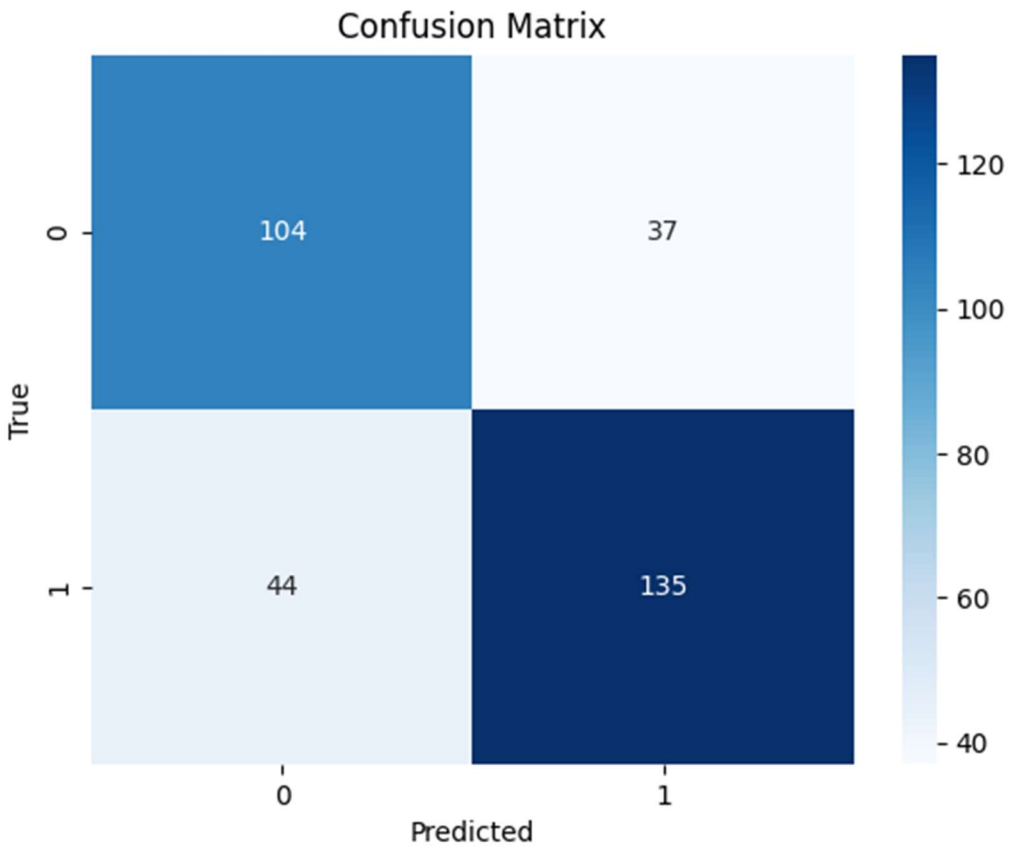
The graphical analysis plays a critical role in interpreting the performance of machine learning models. By visualizing the results through various plots and charts, we can gain deeper insights into the strengths and weaknesses of each model. For the Wine Quality prediction task, I employed several types of visualizations, including a confusion matrix heatmap, precision recall curve and an ROC curve.

- **ROC Curve**



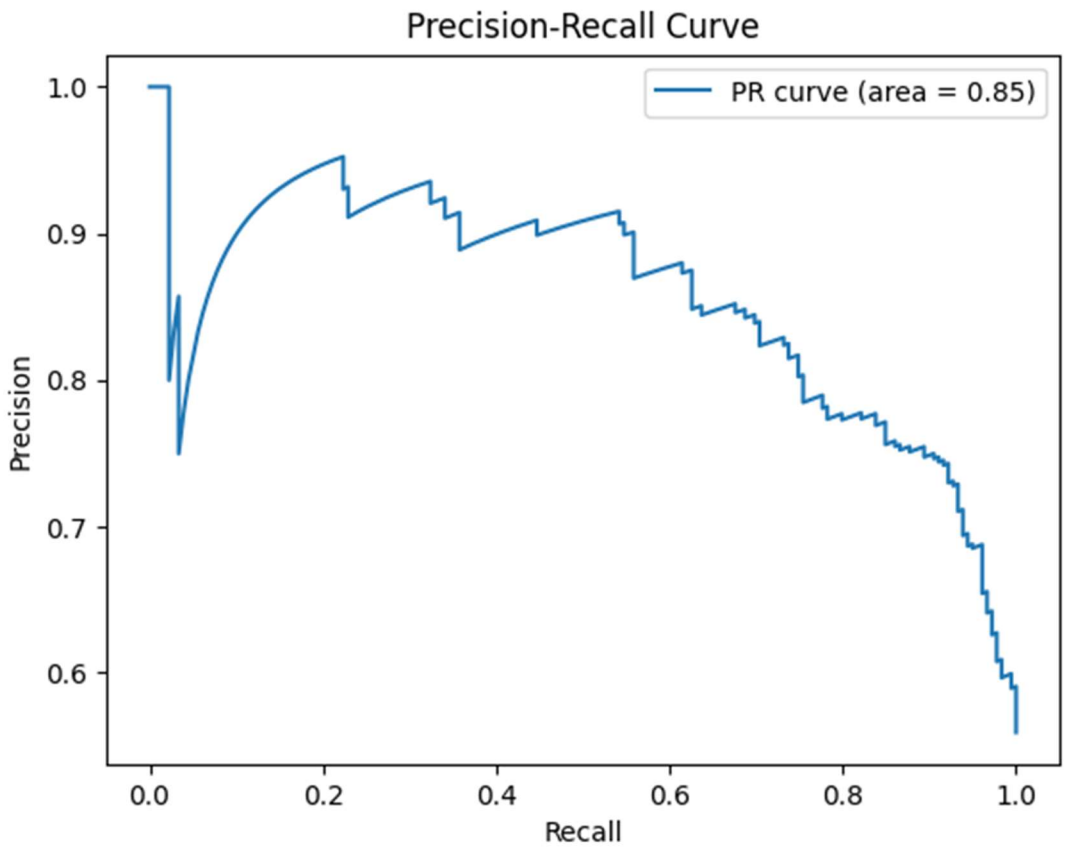
The ROC curve analysis revealed that the Random Forest model, with the highest AUC of 0.89, provided the best classification performance among the models tested, followed by the SVC and Logistic Regression models. This indicates a strong capability of the Random Forest model to distinguish between the two classes over a range of decision thresholds.

- **Confusion Matrix Heatmap**



The confusion matrix for the representative model depicted true positives and true negatives effectively, with an acceptable number of false positives and negatives. However, there was room for improvement, especially in reducing the false negatives to avoid overlooking higher-quality wines.

- **Precision-Recall Curve**



The Precision-Recall curve demonstrated that the model maintained a high level of precision across various levels of recall, affirming its efficacy in correctly identifying the positive class (above-average quality wines) amidst an imbalanced dataset.

- **Analysis and Comparisons**
    - The confusion matrix heatmap points to a balanced ability of the model to classify both classes, but also to the need to reduce false negatives and false positives. This might be achieved through further hyperparameter tuning or potentially resampling techniques to address class imbalance.

- The precision-recall curve's high AUC supports the model's utility in a practical setting where the cost of false positives is high, affirming the model's quality in distinguishing between the majority class and the minority class.

- The ROC curves for all models underscore a strong predictive capability. The Random Forest model emerges as the top performer, balancing the true positive rate and false positive rate most effectively, followed by the SVC and Logistic Regression models.

- **Conclusion**

The Random Forest model appears to be the most promising of the three evaluated models for predicting wine quality. Nevertheless, each model has demonstrated a strong capability to classify wines effectively, indicating that they capture significant information from the physicochemical properties that are indicative of wine quality.

- **Bias-Variance Tradeoff and Model Complexity**

The classification reports and ROC curve analysis suggest that the Random Forest model has achieved a favorable balance in the bias-variance tradeoff, managing to capture the underlying data patterns without being too complex to generalize poorly on unseen data. On the other hand, the similar performance of Logistic Regression and SVC models, which are generally considered to have lower complexity, might indicate a higher bias, suggesting these models could be too simple to capture the nuances in the data fully.

By considering both the graphical analysis and the quantitative classification reports, we can conclude that while all models show a capacity to predict wine quality from

physicochemical properties effectively, the Random Forest model stands out with its robustness and overall superior performance metrics.

Steps for running the code attached

1. Import both “WineQualityModeling.py” and “winequality-red.csv” into an open PyCharm project.
2. Ensure there are no errors with the libraries at the top of the file being imported and if there are errors make sure that you install each library before trying to run the code.
3. Once there are no errors and both files are in the directory of the PyCharm project simply click run and all the results listed should appear.