Predicting Wine Quality and Acidity using Machine Learning and CRISP-DM Methodology

César Cardoso, Rafael Pereira

December 17, 2023

Abstract

The quality of wine is a multifaceted characteristic influenced by various factors. This project employs machine learning techniques within the framework of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to predict wine quality ratings (1, 2, 3) and acidity levels. The study encompasses phases of business understanding, data exploration, preprocessing, modeling, evaluation, and deployment. Given the constraints of a small dataset, we opt for simpler models, specifically using Random Forest Classifier and Ridge Regression for classification and regression tasks, respectively. The dataset has been assumed to be clean, features have been standardized using the StandardScaler from scikit-learn, three random wines have been removed to facilitate a correct split during cross-validation with 5 folds, and models have been evaluated accordingly.

1 Introduction

The world of viticulture produces a vast array of wines, each unique in flavor, aroma, and quality. Understanding and predicting wine quality is a complex challenge that can benefit from the application of machine learning. In addition to quality, we explore the prediction of wine acidity, a crucial parameter that influences the overall taste profile. This project leverages machine learning algorithms to predict both wine quality ratings (1, 2, 3) and acidity levels.

2 CRISP-DM Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a structured framework for guiding data mining projects. Our application of CRISP-DM to the wine quality and acidity prediction project involves six key phases.

2.1 Business Understanding

In this initial phase, we aim to comprehend the business objectives and requirements surrounding the prediction of wine quality and acidity. Engaging with stakeholders and domain experts helps define the problem, establish success criteria, and set the stage for subsequent phases.

2.2 Data Understanding

Given the assumption of a clean dataset, the data understanding phase involves confirming the absence of data quality issues. We explore the dataset containing information on various wine characteristics, identify relevant features, and gain insights into potential challenges and opportunities. Additionally, three random wines were removed to facilitate a correct split during cross-validation with 5 folds.

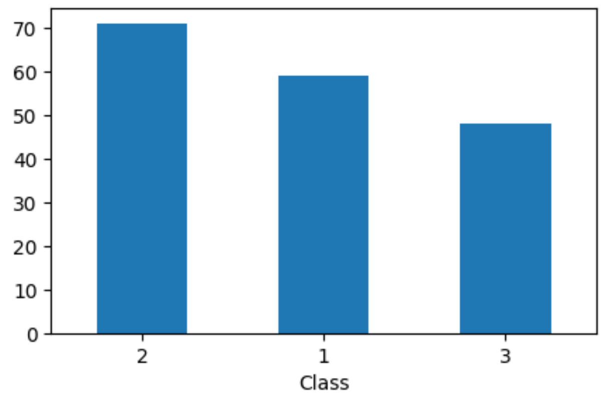

Figure 1: Wine count per class

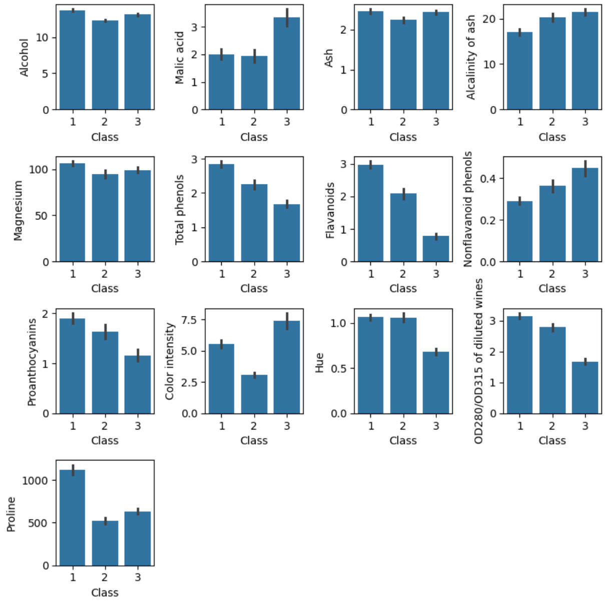

Figure 2: Features related to each class

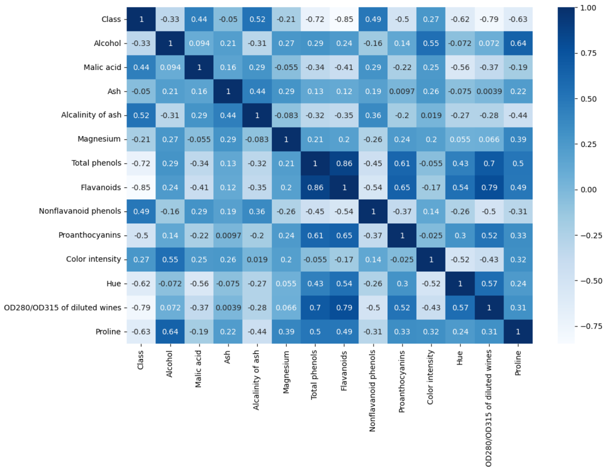

Figure 3: Correlation heatmap

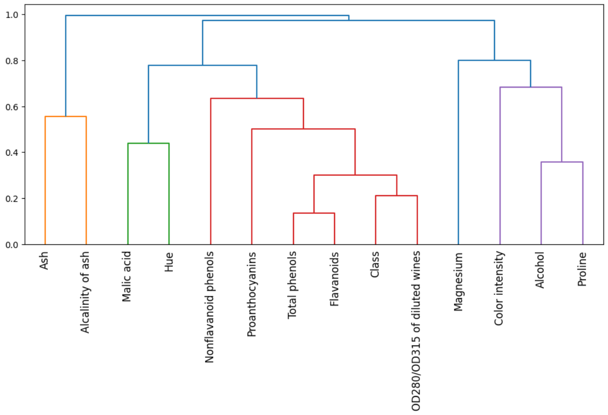

Figure 4: Clusters dendrogram (Hierarchical Clustering)

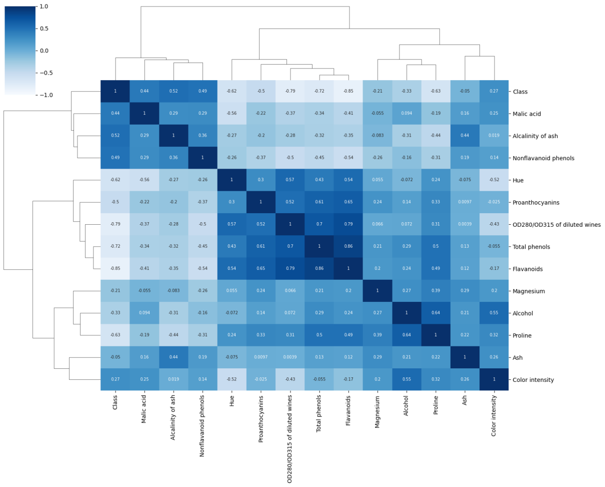

Figure 5: Correlation heatmap with Clusters dendrogram

2.3 Data Preparation

With a clean dataset, the data preparation phase focuses on feature selection, normalization, and splitting into training and testing sets. To ensure standardized features, we use the StandardScaler from scikit-learn.

2.4 Modeling

The heart of the project lies in the modeling phase, where we employ simpler machine learning algorithms to accommodate the small dataset.

2.4.1 Classification for Wine Quality

For predicting wine quality ratings (1, 2, 3), we opt for Random Forest, an ensemble learning method suitable for handling small datasets with high accuracy. We employ StratifiedKFold with 5 folds for cross-validation, considering the removal of three random wines.

2.4.2 Regression for Wine Acidity

To predict wine acidity levels, we utilize Ridge Regression, a straightforward model suitable for regression tasks on small datasets with high correlation between features. A KFold with 5 folds is used for cross-validation, accounting for the removal of three random wines.

2.5 Evaluation

To assess the performance of our models, we employ various evaluation metrics, including accuracy, precision, recall, F1 score and AUC for classification, and Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error(MAE), Normalized Mean Absolute Error (NMAE) and Root Squared (R^2) for regression. Cross-validation using StratifiedKFold(Classification) and Kfold(Regression) ensures the generalizability of our models to new and unseen data.

2.5.1 Evaluation Metrics and Formulas:

- **Accuracy:**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:**

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):**

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score:**

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **R-squared (R^2) Score:**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Normalized Mean Absolute Error (NMAE):**

$$NMAE = \frac{1}{Scale} \times \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

2.5.2 Model Performance and Results

2.5.3 Random Forest Model Performance

The Random Forest model demonstrated high accuracy in predicting wine quality. The following are the detailed performance metrics:

Overall Accuracy: 0.9886 (98.86%)

Classification Report:

Class	Precision	Recall	F1-Score	Support
1	0.98	0.95	0.96	57
2	0.96	0.97	0.97	71
3	0.98	1.00	0.99	47

Table 1: Classification report for the Random Forest model

Confusion Matrix:

	Predicted 1	Predicted 2	Predicted 3
Actual 1	54	3	0
Actual 2	1	69	1
Actual 3	0	0	47

Table 2: Confusion matrix for the Random Forest model

These results indicate the model's robustness in accurately classifying the wine quality into the three categories.

2.5.4 Logistic Regression Model Performance

The Logistic Regression model demonstrated excellent accuracy in predicting wine quality. Below are the detailed performance metrics:

Overall Accuracy: 0.9771 (97.71%)

Classification Report:

Class	Precision	Recall	F1-Score	Support
1	1.00	1.00	1.00	57
2	0.99	0.97	0.98	71
3	0.96	0.98	0.97	47

Table 3: Classification report for the Logistic Regression model

Confusion Matrix:

	Predicted 1	Predicted 2	Predicted 3
Actual 1	57	0	0
Actual 2	0	69	2
Actual 3	0	1	46

Table 4: Confusion matrix for the Logistic Regression model

These results highlight the Logistic Regression model's capability in accurately classifying the wine quality into the three categories.

2.5.5 Comparing both models using receiver operating characteristic curves (ROC) and area under the curve (AUC)

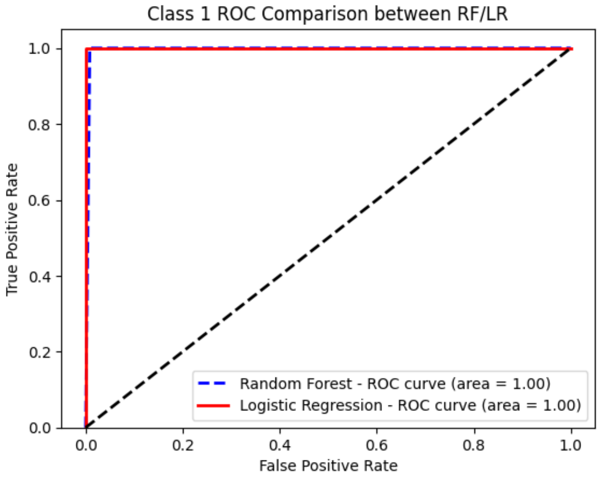

Figure 6: Class 1 ROC

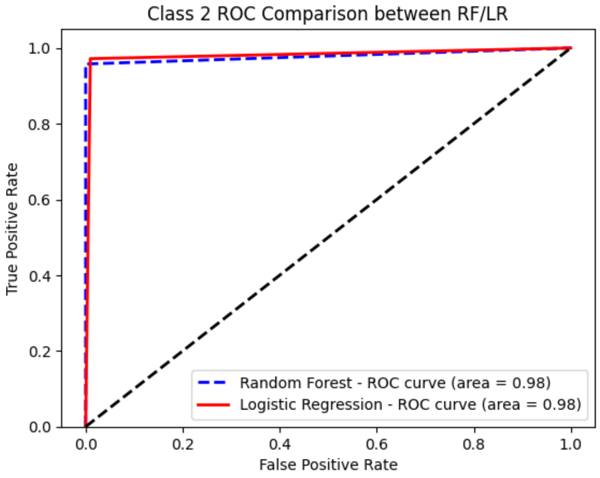

Figure 7: Class 2 ROC

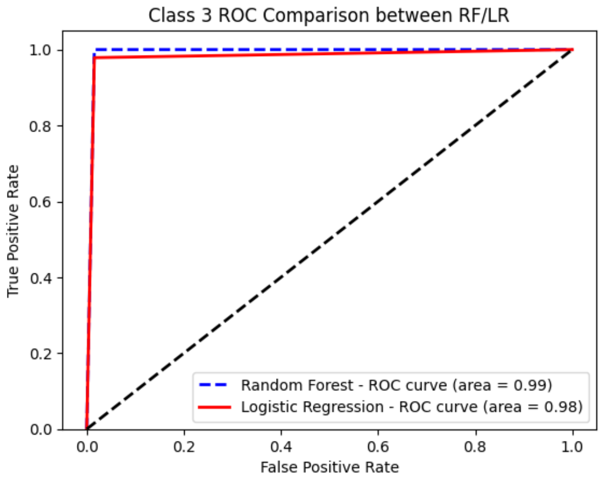

Figure 8: Class 3 ROC

2.5.6 Random Forest Regressor Model Performance

$$MSE = 0.878$$

$$RMSE = 0.934$$

$$MAE = 0.693$$

$$NMAE = 0.291$$

$$R^2 = 0.311$$

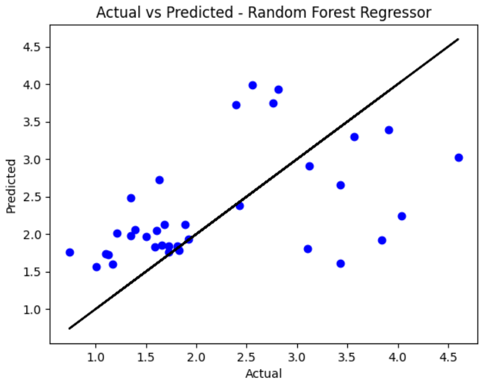

Figure 9: Actual vs Predicted Malic Acid (Random Forest Regressor)

2.5.7 Linear Regression Model Performance

$$MSE = 0.883$$

$$RMSE = 0.937$$

$$MAE = 0.721$$

$$NMAE = 0.303$$

$$R^2 = 0.289$$

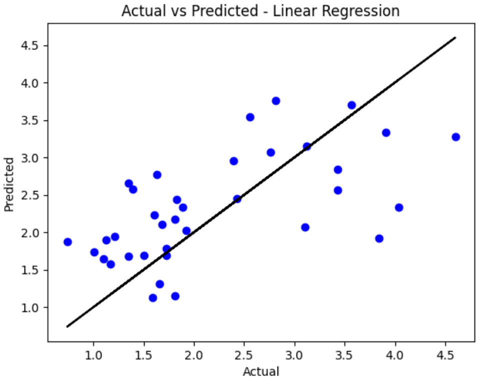

Figure 10: Actual vs Predicted Malic Acid (Linear Regressor)

2.5.8 Ridge Regression Model Performance

$$MSE = 0.880$$

$$RMSE = 0.936$$

$$MAE = 0.717$$

$$NMAE = 0.302$$

$$R^2 = 0.291$$

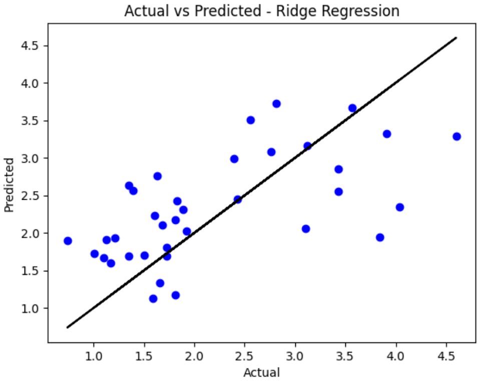

Figure 11: Actual vs Predicted Malic Acid (Ridge Regressor)

2.6 Model Selection

Random Forest Classifier ended up being the preferred model to predict wine quality since it shows the highest average AUC in all the classes and the highest accuracy.
To predict Malic Acid and given the high correlation among features in the dataset, Ridge Regression emerges as the best choice. It effectively handles multicollinearity with its L2 regularization and demonstrates a balanced performance across key metrics. While RandomForestRegressor shows strong predictive accuracy and explanatory power, its complexity and interpretability are of concern. LinearRegression, though simpler, is less effective against multicollinearity.

Ridge Regression offers a good balance between maintaining the simplicity of linear models and providing robustness against feature correlation. Its performance, with relatively low MSE, RMSE, and NMAE, along with a moderate R2 score, suggests effective predictive accuracy and a reasonable level of explanatory capability. Therefore, **Ridge Regression** is the most suitable model for this dataset, given its capability to provide reliable and interpretable results in the presence of high feature correlation.

2.7 Deployment

Upon selecting satisfactory models, we transition to the deployment phase. This involves integrating the models into a production environment, where they can make predictions on new wine data. Continuous monitoring and updates are essential to maintain the models' performance over time.

3 Conclusion

This document outlines a pragmatic approach to predicting wine quality and acidity using simpler machine learning models and the CRISP-DM methodology. Given the constraints of a small dataset, our focus on business understanding, streamlined data exploration, and the application of Random Forest Classifier and Ridge Regression in the modeling phase aims to develop interpretable and effective predictive models. The evaluation and deployment phases ensure the models' practical application in real-world scenarios.

4 References

[1] CRISP-DM. Cross-Industry Standard Process for Data Mining. https://www.crisp-dm.org/

[2] Scikit-learn: Machine Learning in Python. https://scikit-learn.org/

[3] Dealing with very small datasets. https://www.kaggle.com/code/rafjaa/dealing-with-very-small-datasets/

[4] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.

[5] Overcoming Issues with Small Data Sets when Building Machine Learning Models. https://www.youtube.com/watch?v=Ly3ogCE-GuI&t=1638s