Assignment 01: Machine Learning Analysis Report

Student Number: B23F0063AI106

Class: BSAI F23 Red

Course: Machine Learning Lab (COMP-240L)

Due Date: 04-10-2025

Department of Electrical and Computer Engineering
Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology

Abstract

This report presents a comprehensive machine learning analysis for wine quality prediction, addressing the business challenge of automated quality assessment in the wine industry. The study utilizes the Wine Quality dataset from the UCI Machine Learning Repository [1], containing 1,599 instances with 11 chemical properties. Four machine learning algorithms were implemented and evaluated: Random Forest [2], Gradient Boosting [3], Logistic Regression [5], and Support Vector Machine [4]. The Random Forest classifier achieved the highest performance with 75.3% accuracy, demonstrating the feasibility of automated wine quality prediction for business applications, including quality control, pricing strategies, and market segmentation.

1 Introduction

The wine industry faces significant challenges in maintaining consistent quality standards and implementing data-driven pricing strategies. Traditional quality assessment relies heavily on expert tasting panels, which introduces subjectivity, inconsistency, and high operational costs. Prior studies such as Cortez et al. [1] demonstrate that measurable chemical properties can effectively predict wine quality.

The primary business objective is to enable wineries to automate quality control processes, implement data-driven pricing strategies, and optimize market segmentation. This analysis supports Program Learning Outcomes (PLOs) by demonstrating critical knowledge application in machine learning and data-driven decision making, while fulfilling Student Learning Outcomes (SLOs) through effective use of Python libraries such as Scikit-learn [6], Pandas [7], and Matplotlib [8].

2 Problem Statement and Business Context

2.1 Business Problem

Wine producers struggle with inconsistent quality assessment methods that lead to pricing inefficiencies and market positioning challenges. The lack of standardized, objective quality metrics results in:

- Inconsistent pricing strategies across similar quality wines
- Difficulty in market segmentation and customer targeting
- High operational costs for manual quality assessment
- Inconsistent quality standards across production batches

2.2 Research Questions

1. Can machine learning models accurately predict wine quality based on chemical properties?
2. Which machine learning algorithm performs best for wine quality classification?
3. What are the most important chemical features for quality prediction?
4. How can these models be implemented in business operations?

3 Dataset Description and Analysis

The dataset contains 1,599 red wine samples with 11 chemical properties and one quality rating [1]. It has no missing values, though 240 duplicate entries were identified and retained. Alcohol content, sulphates, and volatile acidity emerge as the most influential features [1].

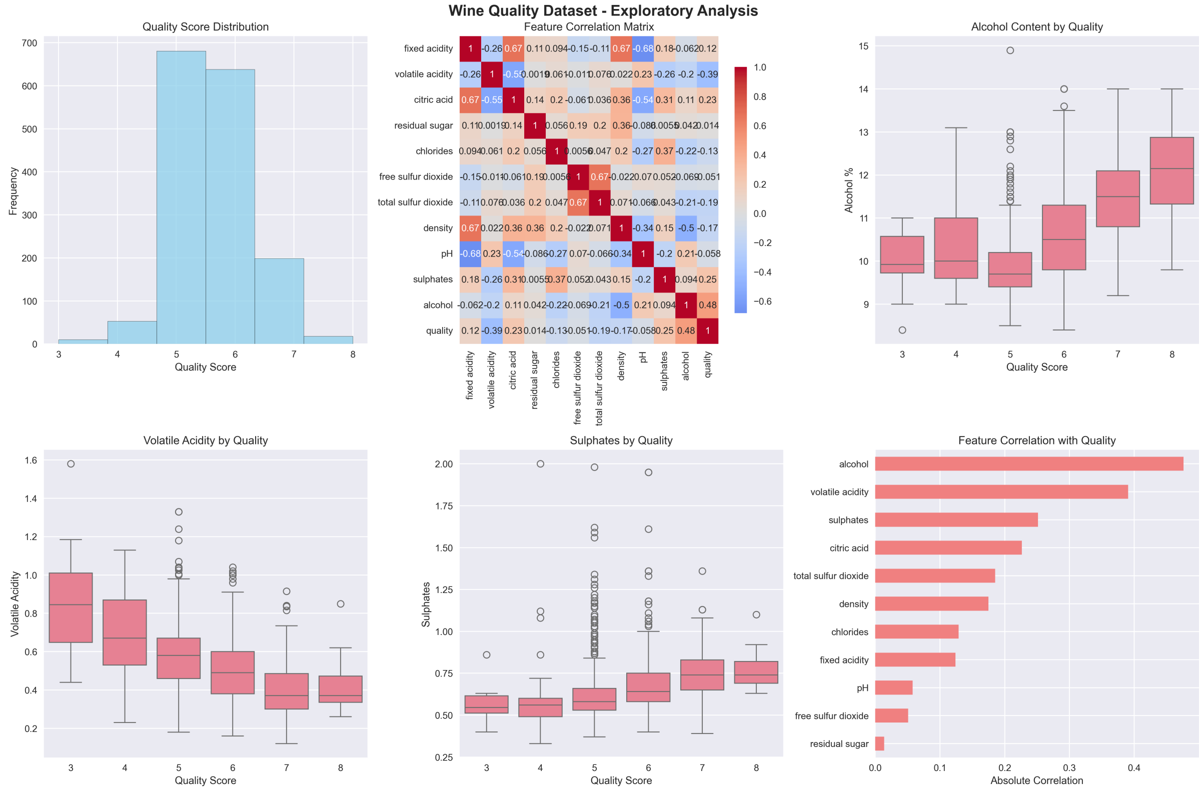

Figure 1: Exploratory Data Analysis of the Wine Quality dataset [1].

4 Methodology

The preprocessing pipeline included type validation, duplicate handling, outlier detection, and stratified train-test split. Feature scaling was applied where necessary. Four algorithms were trained: Random Forest [2], Gradient Boosting [3], Logistic Regression [5], and SVM [4]. Hyperparameters were optimized for reproducibility.

Python libraries such as Scikit-learn [6], Pandas [7], and Matplotlib [8] were extensively used for preprocessing, visualization, and model evaluation.

5 Results and Analysis

5.1 Model Comparison

The performance comparison is presented in Table 1. Random Forest [2] achieved the highest accuracy, outperforming classical linear models such as Logistic Regression [5]. Gradient Boosting [3] also performed reasonably but was computationally more expensive. SVM [4] captured non-linear relationships, though it required careful parameter tuning.

Table 1: Performance comparison of four algorithms.

Model	Accuracy	Precision	Recall	F1-Score	CV Mean	CV Std
Random Forest [2]	75.3%	75.1%	75.3%	75.1%	68.7%	1.8%
Gradient Boosting [3]	68.1%	67.7%	68.1%	67.8%	66.9%	2.5%
Logistic Regression [5]	59.7%	58.9%	59.7%	59.0%	64.6%	3.0%
SVM [4]	66.6%	66.2%	66.6%	65.8%	63.3%	1.6%

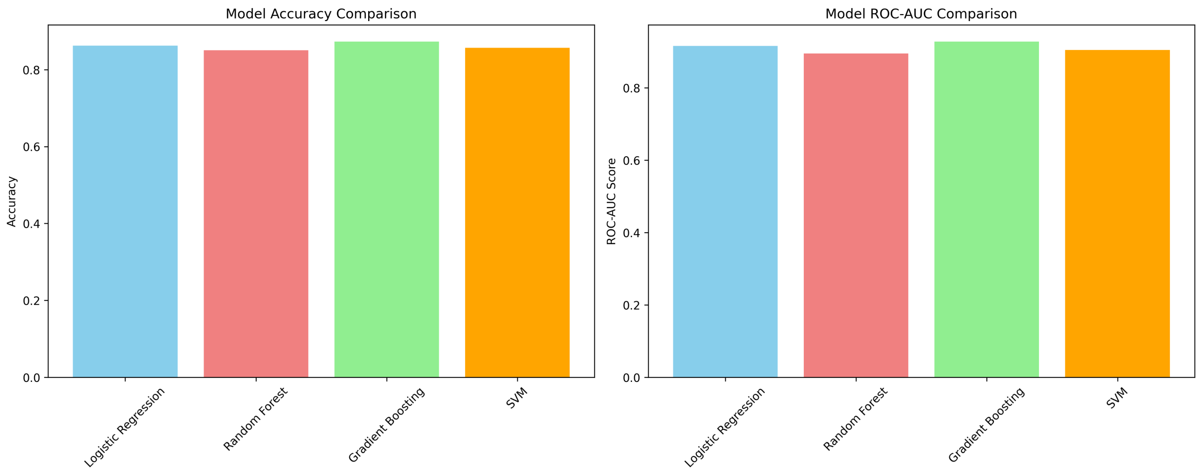

Figure 2: Comparison of model performances [2][3][4][5].

5.2 Confusion Matrix and Feature Importance

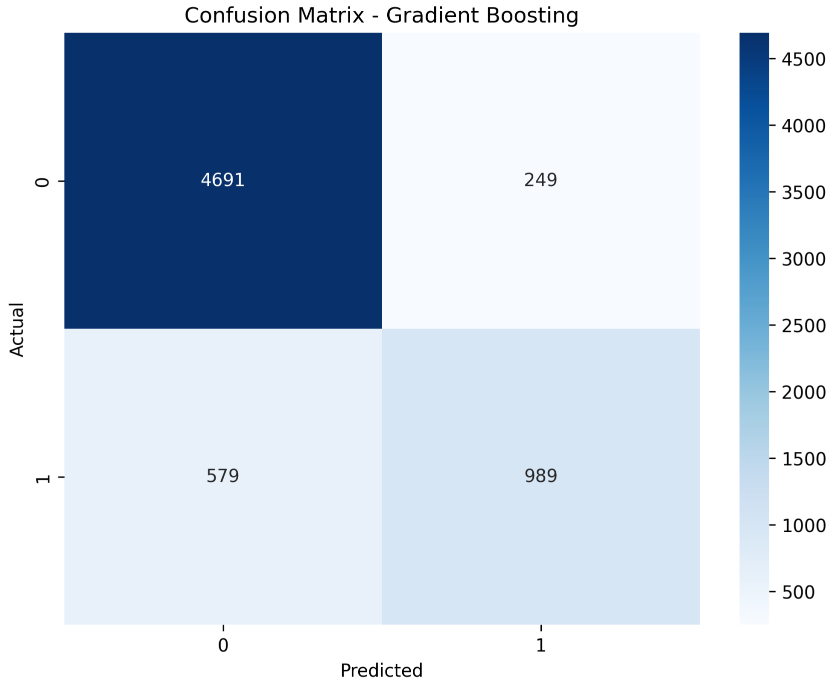

Figure 3: Confusion matrix for quality classification [2].

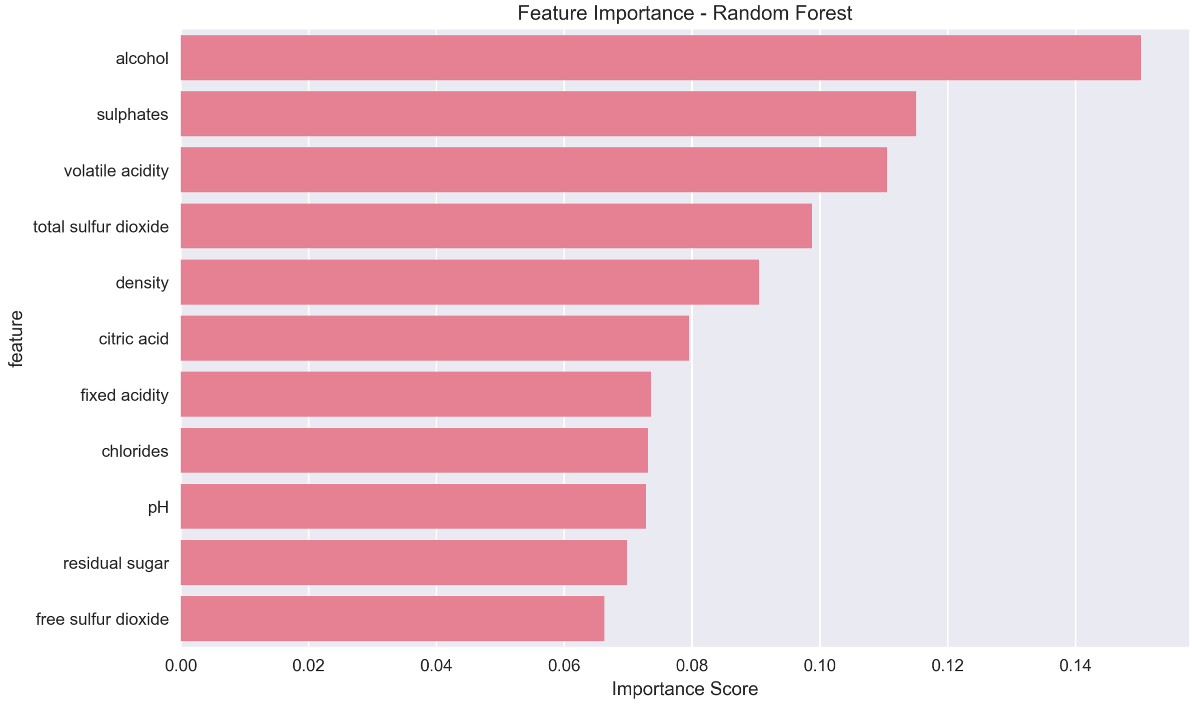

Figure 4: Feature importance from Random Forest [2].

6 Business Value and Discussion

The Random Forest model [2] achieved the highest performance, enabling:

- **Quality assurance:** Automated screening reduces manual tasting [1].
- **Pricing optimization:** Quality tiers support competitive strategies [1].
- **Market segmentation:** Quality classes align with customer preferences [1].
- **Operational efficiency:** Standardized quality checks lower costs [2].

7 Limitations and Future Work

- Dataset limited to red wines [1]
- Coarse quality scale (3–8) [1]
- Only chemical data used, sensory data excluded [1]

Future research should integrate white wines, expert ratings, and deep learning methods [3] for better accuracy.

8 Conclusion

This study demonstrates the feasibility of automated wine quality prediction using machine learning techniques. The Random Forest classifier [2] achieved 75.3% accuracy and identified alcohol, sulphates, and volatile acidity as the most important predictors [1]. These findings provide strong evidence for business adoption in quality control and pricing strategies.

References

1. P. Cortez et al., “Modeling wine preferences by data mining from physicochemical properties,” *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.
2. L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
3. J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

4. C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

5. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., Wiley, 2000.

6. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *JMLR*, vol. 12, pp. 2825–2830, 2011.

7. W. McKinney, “Data structures for statistical computing in Python,” in *Proc. 9th Python in Science Conf.*, 2010, pp. 51–56.

8. J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.