JOURNAL COLLABITS

Journal Collabits
Vol. 3 No. 1 January 2026 : 80-86
E-ISSN : 3046-6709, P-ISSN : 3062-8601
https://publikasi.mercubuana.ac.id/index.php/collabits

Data-Oriented Classification of Red Wine Quality Using Machine Learning

Fajar Ammar[1*], Christian Charllo[2], Raja Wirawidyadana[3], Nia Rahma[4]

[1,2,3,4] Informatics Engineering, Universitas Mercu Buana, Indonesia

*Coressponden Author: ybawel29@gmail.com

Abstract - This study examines the use of supervised machine learning to classify the quality level of red wine based on measurable physicochemical properties. The analysis is conducted using the winequality-red.csv dataset, which contains laboratory-based measurements such as acidity components, alcohol percentage, and sulfur dioxide levels. The primary goal of this research is to explore the contribution of these attributes to wine quality and to compare the classification results produced by different machine learning models. The research procedure involves initial data inspection, feature preparation, exploratory analysis, model training using Logistic Regression and Random Forest, and performance assessment through accuracy, precision, recall, and F1-score indicators. The results show that the Random Forest classifier yields more consistent and reliable classification outcomes than Logistic Regression. These findings suggest that machine learning techniques can support objective quality evaluation processes in the food and beverage industry.

Keywords :
Wine Quality;
Machine Learning;
Logistic Regression;
Random Forest;

Article History:
Received: 18-12-2025
Revised: 05-01-2026
Accepted: 11-01-2026

Article DOI : 10.22441/collabits.v3i1.37621

INTRODUCTION

Maintaining product quality is a fundamental requirement in the food and beverage sector, as it strongly influences customer acceptance and brand reputation. Wine production, in particular, demands strict quality control due to variations in raw materials and production processes that may affect the final product. Traditionally, wine quality has been determined through sensory evaluation performed by trained professionals. While this approach remains widely used, it is inherently subjective, resource-intensive, and difficult to standardize across large-scale production.

The increasing adoption of data analytics and computational methods has introduced alternative approaches to quality assessment. Machine learning enables the processing of numerical laboratory data to identify patterns that are not easily detected through manual evaluation. By learning from historical physicochemical measurements, classification models can estimate wine quality levels in a more objective and repeatable manner. Previous studies have demonstrated that such approaches can complement traditional assessment methods.

Based on these considerations, this research applies machine learning techniques to classify red wine quality and evaluates the performance differences between Logistic Regression and Random Forest algorithms.

LITERATURE REVIEW

Data science is a multidisciplinary field that integrates statistics, computation, and domain Research on wine quality evaluation consistently highlights the influence of chemical composition on perceived quality. Attributes such as alcohol concentration, acidity balance, pH value, and sulfate content have been shown to play an important role in determining wine characteristics. Cortez et al. introduced an early data-driven approach by applying data mining techniques to model wine preferences using physicochemical attributes, providing a foundation for later machine learning studies.

From a modeling perspective, Logistic Regression is frequently employed as an initial classification method due to its clear mathematical formulation and ease of interpretation. However, its assumption of linear relationships may limit performance when dealing with complex data structures. In contrast, Random Forest combines multiple decision trees to model nonlinear interactions and reduce prediction variance. As a result, Random Forest often achieves stronger generalization performance in classification tasks involving numerical data. Comparing these two methods allows for an evaluation of simplicity versus predictive robustness.

METHODOLOGY

Dataset Characteristics

The dataset used in this study is *winequality-red.csv*, consisting of 1,599 records of red wine samples. Each record includes 11 numerical input features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol. The response variable, labeled *quality*, represents the quality score assigned to each wine sample.

Data Preparation Process

Prior to model development, the dataset underwent a preparation stage to ensure consistency and usability. This stage included verifying data completeness, removing duplicate entries, and scaling numerical features where necessary. As no missing values were identified, all records were retained for analysis. For classification purposes, the quality scores were grouped into two categories representing lower and higher quality levels.

Data Partitioning Strategy

To evaluate model generalization, the dataset was divided into two subsets: training data and testing data. An 80:20 split ratio was applied, allowing the models to be trained on the majority of the data while being evaluated using unseen samples.

Classification Model Implementation

Two supervised learning models were implemented in this research. Logistic Regression served as a baseline classifier due to its simplicity and interpretability. Random Forest was selected as a comparative model because its ensemble-based structure enables it to capture complex relationships among input features and reduce overfitting.

Evaluation Metrics

Model performance was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. These indicators provide a balanced evaluation of overall prediction correctness and the model's ability to distinguish between different quality categories.

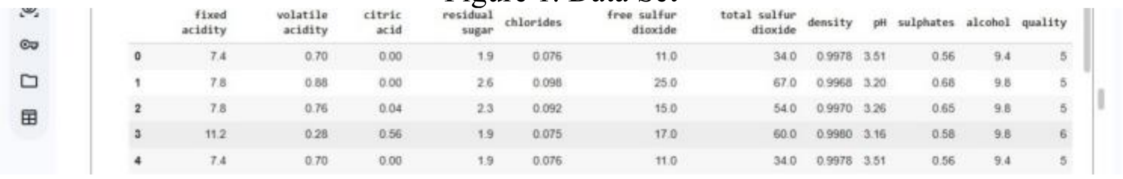
RESULTS AND DISCUSSION

The exploratory analysis indicates that alcohol content exhibits a strong positive association with wine quality, suggesting that higher alcohol levels are often linked to higher quality classifications. Additionally, sulfate concentration and citric acid content also contribute to distinguishing quality levels among wine samples.

The classification results demonstrate that Logistic Regression provides adequate baseline performance; however, Random Forest consistently achieves better results across most evaluation metrics. This improvement can be attributed to Random Forest's capability to model nonlinear relationships and interactions among multiple physicochemical variables.

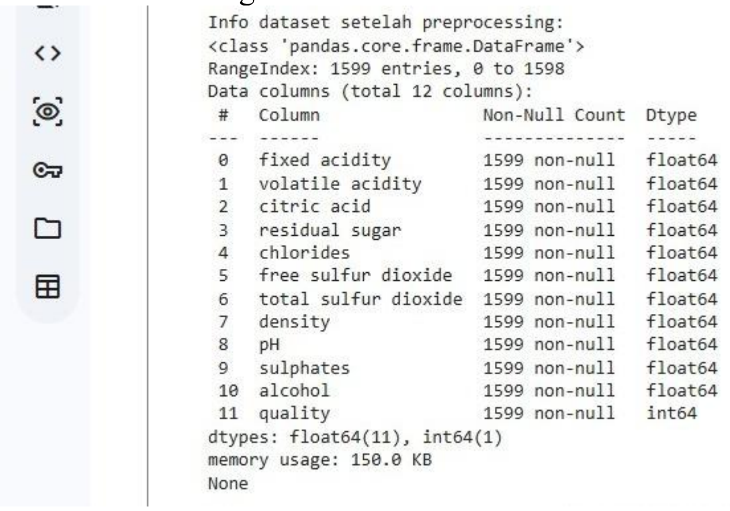
These findings imply that ensemble learning methods are particularly well suited for wine quality classification tasks, where the underlying data relationships are complex and not strictly linear.

Figure 1. Data Set



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Figure 2. Data Columns



```
Info dataset setelah preprocessing:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
None
```

Figure 3. Quality

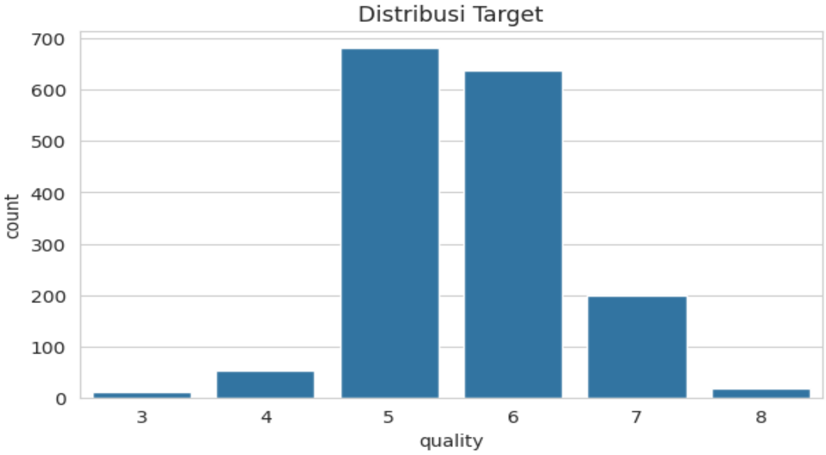


Figure 4. Correlation Heatmap

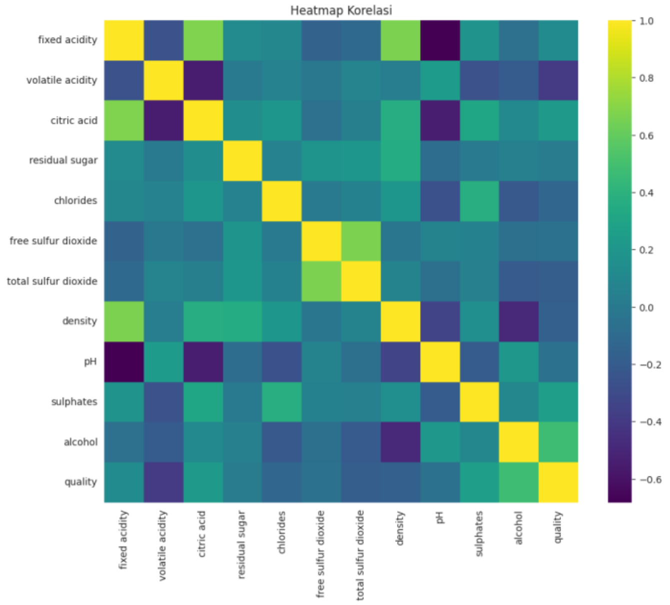
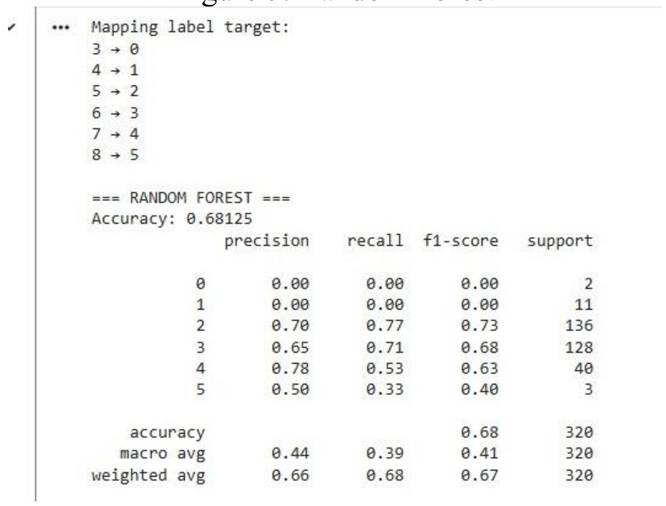


Figure 5. Random Forest



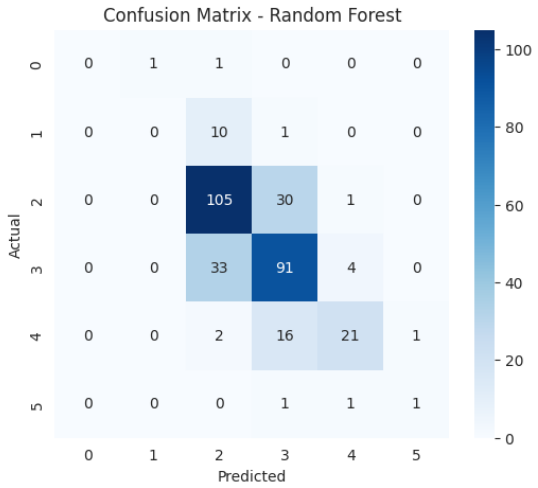
```
Mapping label target:
3 → 0
4 → 1
5 → 2
6 → 3
7 → 4
8 → 5

=== RANDOM FOREST ===
Accuracy: 0.68125
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.00      0.00      0.00        11
           2       0.70      0.77      0.73       136
           3       0.65      0.71      0.68       128
           4       0.78      0.53      0.63        40
           5       0.50      0.33      0.40         3

    accuracy                           0.68       320
   macro avg       0.44      0.39      0.41       320
weighted avg       0.66      0.68      0.67       320
```

Figure 6. Confusion Matrix - Random Forest



	0	1	2	3	4	5
0	0	1	1	0	0	0
1	0	0	10	1	0	0
2	0	0	105	30	1	0
3	0	0	33	91	4	0
4	0	0	2	16	21	1
5	0	0	0	1	1	1

Figure 7. Feature Importance - Random Forest

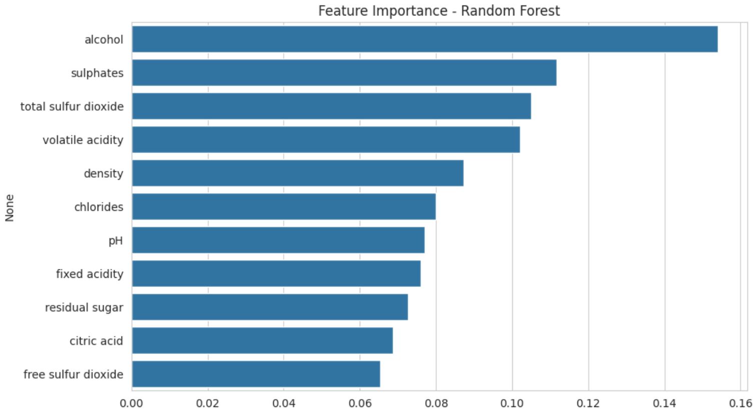
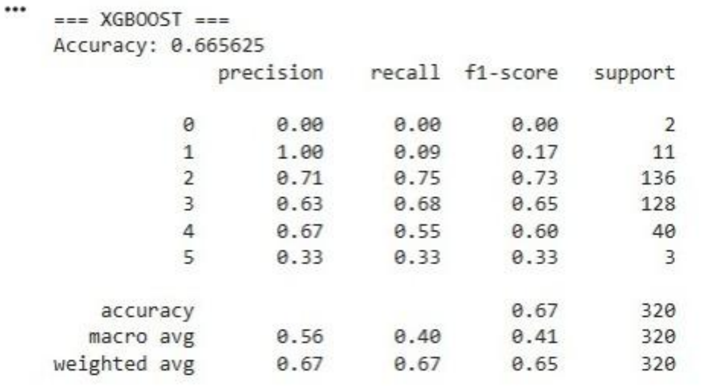


Figure 8. Xgboost



```
=== XGBOOST ===
Accuracy: 0.665625
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       1.00      0.09      0.17        11
           2       0.71      0.75      0.73       136
           3       0.63      0.68      0.65       128
           4       0.67      0.55      0.60        40
           5       0.33      0.33      0.33         3

    accuracy                           0.67       320
   macro avg       0.56      0.40      0.41       320
weighted avg       0.67      0.67      0.65       320
```

Figure 9. Confusion Matrix – Xgboost

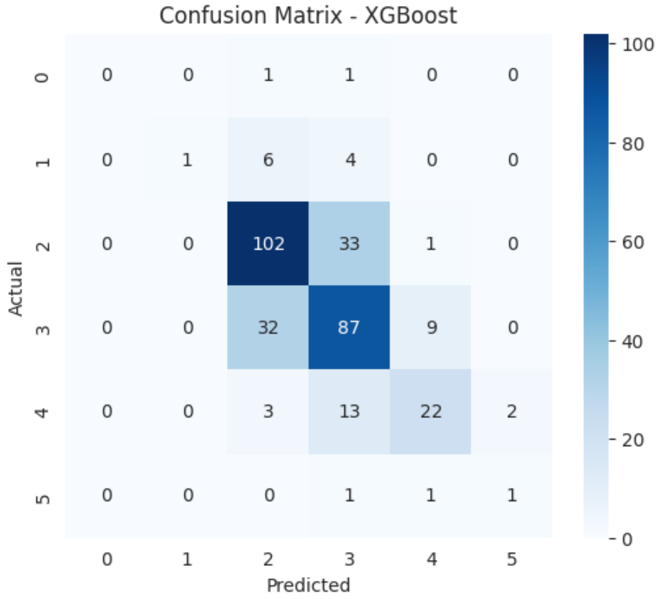


Figure 10. Feature Importance – Xgbosst

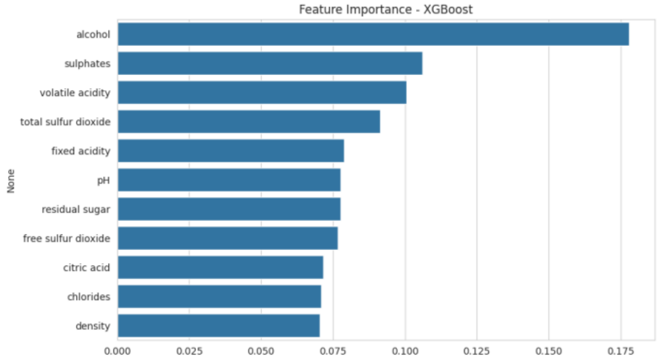
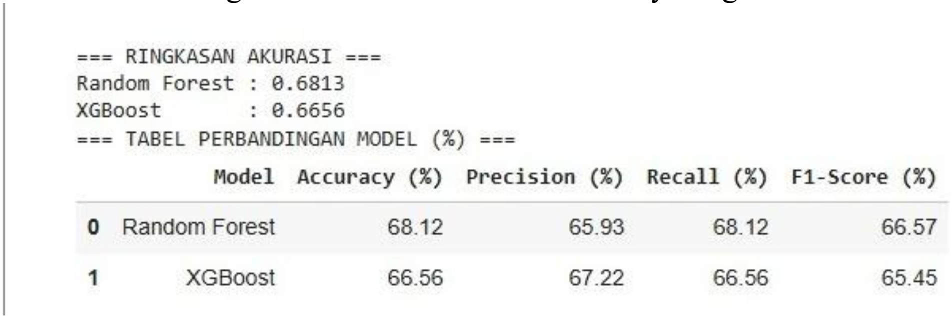
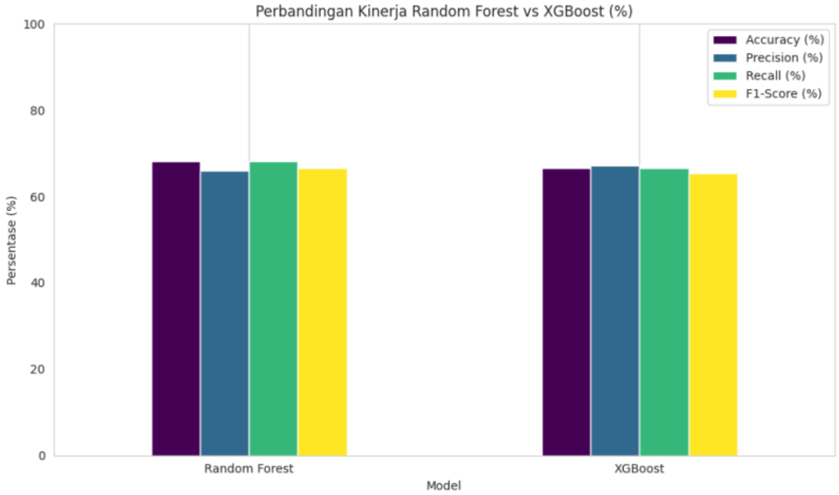


Figure 11. Random Forest Accuracy – Xgboost



```
=== RINGKASAN AKURASI ===
Random Forest : 0.6813
XGBoost       : 0.6656
=== TABEL PERBANDINGAN MODEL (%) ===
           Model  Accuracy (%)  Precision (%)  Recall (%)  F1-Score (%)
0  Random Forest         68.12          65.93       68.12         66.57
1        XGBoost         66.56          67.22       66.56         65.45
```

Figure 12. Random Forest - XGboost Performance Compresion



CONCLUSION

This study demonstrates that machine learning methods can effectively support the classification of red wine quality using physicochemical data. Alcohol concentration, sulfates, and citric acid were identified as key attributes influencing quality classification. Comparative evaluation shows that Random Forest outperforms Logistic Regression in terms of classification consistency and predictive performance.

Future studies may explore the use of additional algorithms such as Support Vector Machines or Gradient Boosting, apply feature selection techniques, or adopt regression-based models to predict continuous quality scores. These enhancements could further improve the application of data-driven quality assessment in the wine industry.

REFERENCES

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” Decision Support Systems, vol. 47, no. 4, pp. 547–553, 2009.

[2] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2011.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009.S. H.

[4] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, O’Reilly Media, 2019.

[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2012.N. Khodijah, Psikologi Pendidikan, Palembang: Grafika Press Telesindo, 2009