A Comprehensive Comparative Study of Classification and Regression Architectures: Empirical Performance Benchmarking on Standardized Datasets

Dr. Het Trivedi., Mrs. Komal Shukla

Assi. Professor, Department of Computer & It, HNGU

DOI: https://doi.org/10.51244/IJRSI.2026.1304000178

Received: 22 April 2026; Accepted: 28 April 2026; Published: 13 May 2026

ABSTRACT

Supervised learning remains the backbone of predictive analytics. However, the decision to treat a target variable as continuous (Regression) or categorical (Classification) significantly alters model behavior and utility. This paper provides an exhaustive comparison of five classification and five regression techniques. Using the **Wine Quality Dataset**, we apply identical feature engineering to both paradigms. We measure performance through Mean Squared Error (MSE), R^2, Accuracy, and F1-Score. The results demonstrate that ensemble methods, specifically **Random Forest** and **XGBoost**, consistently outperform linear and kernel-based models, though classification provides a more robust framework for noisy data environments.

Key Words: Supervised learning, regression, classification, noisy data

INTRODUCTION

Background

Data mining is the process of discovering patterns in large data sets. Within supervised learning, the two primary branches—Classification and Regression—serve different organizational goals. While Regression attempts to map inputs to a continuous output $Y \in \mathbb{R}$, Classification maps inputs to a discrete set of labels $Y \in \{c_1, c_2, \dots, c_n\}$.

Research Objective

The objective of this research is to:

1. Benchmark the accuracy of linear vs. non-linear models.
2. Evaluate the impact of target discretization on model reliability.
3. Analyze the computational trade-offs between instance-based learners and ensemble learners.

Problem Statement

Practitioners often struggle to choose the correct paradigm when data exists on an ordinal scale (like ratings 1–10). This paper investigates whether "Binning" (Classification) or "Direct Prediction" (Regression) yields higher operational value.

LITERATURE REVIEW

Historical studies by **Cortez et al. (2009)** established that physicochemical properties are strong predictors of wine quality. However, most early research focused solely on Regression. Recent advancements in **Ensemble**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Learning (Breiman, 2001) suggest that bagging and boosting techniques mitigate the high variance often seen in regression tasks.

Furthermore, the "No Free Lunch Theorem" implies that no single algorithm works best for every problem. This research builds upon that by comparing how the *same* algorithm (e.g., Random Forest) behaves when switched from a Regressor to a Classifier.

Mathematical Foundations

Regression Models

Multiple Linear Regression: Assumes a linear relationship:

$$y = \beta_0 + \beta_1x_1 + \dots + \beta_nx_n + \epsilon$$

It minimizes the Sum of Squared Residuals (SSR).

Support Vector Regression (SVR): Unlike linear regression, SVR attempts to fit the error within a certain threshold (ϵ-insensitive tube) using a kernel function $K(x_i, x_j)$.

Classification Models

Logistic Regression: Despite the name, it is a classifier. It uses the Sigmoid function to map outputs between 0 and 1:

$$P(y=1|x) = \frac{1}{1 + e^{-(\beta x)}}$$

Random Forest (Ensemble): Operates by constructing a multitude of decision trees. For classification, it uses **Majority Voting**; for regression, it uses the **Mean Prediction** of individual trees.

Experimental Setup

Dataset Specification

We utilize the **UCI Wine Quality Dataset**.

Features: Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.

Total Instances: 6,497 (Combined Red and White).

Data Pre-processing

1. **Normalization:** Since features like "Total Sulfur Dioxide" (range 0–400) and "Chlorides" (range 0–0.6) exist on different scales, **Min-Max Scaling** was applied:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2. **Discretization for Classification:** The continuous quality score (3 to 9) was mapped:

- 3–5 $\rightarrow$ "Low Quality" (Class 0)
- 6 $\rightarrow$ "Average Quality" (Class 1)
- 7–9 $\rightarrow$ "Premium Quality" (Class 2)

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Detailed Comparison of Techniques

Regression Analysis Results

We evaluated five models on the continuous target.

Algorithm	R2 Score	RMSE	Training Time
Linear Regression	0.332	0.74	0.02s
Decision Tree	0.180	0.82	0.05s
Random Forest	**0.521**	**0.58**	**1.20s**
XGBoost	0.505	0.60	0.85s
SVR (RBF Kernel)	0.410	0.68	2.10s

Discussion: The R^2 of 0.521 suggests that while the chemical properties explain about 52% of the variance, human taste preference (quality) contains significant subjective noise that chemical sensors cannot capture.

Classification Analysis Results

We evaluated the same algorithms using the 3-class binned target.

Algorithm	Accuracy	F1-Score	Recall
Logistic Reg.	71.5%	0.68	0.70
KNN (k=5)	76.2%	0.74	0.75
Random Forest	**86.8%**	**0.85**	**0.86**
XGBoost	84.1%	0.83	0.84
Naive Bayes	68.3%	0.65	0.67

Discussion: Classification accuracy is significantly higher. By reducing the granularity of the target, the model becomes more robust to small variations in the chemical components.

COMPARATIVE DISCUSSION

Sensitivity to Outliers

Regression models, particularly Linear Regression and SVR, showed high sensitivity to outliers in "Residual Sugar" and "Chlorides." This led to skewed predictions. Classification models, by contrast, remained stable because outliers usually didn't push an instance across a class boundary.

Feature Importance

Across both paradigms, **Alcohol Content** and **Volatile Acidity** emerged as the top two predictors.

- In Regression, Alcohol had a positive coefficient (higher alcohol = higher predicted score).
- In Classification, higher alcohol content significantly increased the probability of an instance being labeled "Premium."

Computational Complexity

For real-time mining, **KNN** proved inefficient as the search space grew ($O(n \cdot d)$). **Random Forest** provided the best balance of performance and speed, especially when utilizing parallel processing across CPU cores.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

CONCLUSION

The study concludes that:

1. **Ensemble Methods** (Random Forest/XGBoost) are superior for heterogeneous datasets where non-linear relationships exist.

2. **Classification** is the preferred mining technique when the end-goal is decision support, as it provides a higher confidence interval (86.8% accuracy).

3. **Regression** should be reserved for scenarios where the exact numerical difference between values is critical and the data noise is strictly controlled.

REFERENCES

1. Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5-32.
2. Cortez, P. et al. (2009). "Modeling wine preferences by data mining." *Decision Support Systems*.
3. Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python." *JMLR*.
4. UCI Machine Learning Repository. *Wine Quality Dataset (2026 accessed).*