Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Benchmarking Machine Learning Classifiers for Wine Cultivar Identification from Chemical Analysis: A Reproducible Evaluation on the UCI Wine Dataset

Atul Gaikwad

atulgaikwad23092003@gmail.com

Sanjivani Universty, Kopargaon, Maharashtra, India https://orcid.org/0009-0000-5534-7106

Short Report

Keywords: Wine classification, machine learning, food authenticity, supervised learning, cross-validation, reproducibility

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11576250/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Benchmarking Machine Learning Classifiers for Wine Cultivar Identification from Chemical Analysis: A Reproducible Evaluation on the UCI Wine Dataset

Atul Gaikwad

ORCID ID: 0009-0000-5534-7106
Department of Artificial Intelligence and Data Science
Sanjivani University, Maharashtra, India

***Abstract*—*Identifying the grape cultivar behind a wine from its chemical composition is a classic multi-class pattern-recognition task relevant to food authenticity and quality control. This study benchmarks logistic regression, a radial-basis-function support vector machine (SVM), k-nearest neighbors, random forest, and gradient boosting on the UCI Wine dataset distributed with scikit-learn, which contains 178 samples, 13 chemical measurements, and three cultivar classes. A stratified 80/20 train–test split was fixed with random seed 42, and five-fold stratified cross-validation was run on the training set. Logistic regression, RBF SVM, and k-nearest neighbors each reached 97.22% held-out accuracy, random forest reached 100%, and gradient boosting reached 94.44%. Because the test set contains only 36 samples, one prediction changes accuracy by 2.8 percentage points, and the cross-validation results rank the models differently. These results show strong class separability on a small, well-known benchmark; they do not establish performance on new vintages, regions, or laboratories.*

Index Terms*—*Wine classification, machine learning, food authenticity, supervised learning, cross-validation, reproducibility

I. INTRODUCTION

Chemical profiling is widely used to verify the origin and variety of food and beverage products. Machine learning can turn such multivariate measurements into automated classifiers, yet reported accuracy on small benchmarks can be misleading when the evaluation protocol is not clearly described. A held-out test set and cross-validation give complementary estimates of generalization within a dataset, although neither replaces validation on independent samples.

This study asks: how do common supervised classifiers compare on a standard wine-cultivar benchmark when they are evaluated with the same split and metrics? The contributions are (1) a controlled comparison of five conventional classifiers, (2) a reproducible split and cross-validation protocol, and (3) an explicit discussion of how small test sets limit the conclusions that can be drawn.

II. PROBLEM STATEMENT AND OBJECTIVES

The task is multi-class classification: assign a wine sample to one of three cultivars using 13 numeric chemical measurements. The goal is not to propose a production authentication system, but to measure benchmark performance of commonly used models under a transparent protocol.

The objectives are to compare predictive performance using accuracy and macro-averaged precision, recall, and F1-score, to report cross-validation variability, and to identify limitations that would need to be addressed before any real-world use.

III. RELATED WORK

The wine dataset originates from chemical analysis of wines grown in the same region of Italy but derived from three different cultivars [1], and it is available through the UCI Machine Learning Repository [2]. It is a common teaching and benchmarking dataset for classification and dimensionality-reduction methods. Support vector machines [4], random forests [5], and gradient boosting [6] are widely used general-purpose learners, and cross-validation is a standard approach for estimating accuracy and selecting models [7]. This paper uses these established methods and focuses on a reproducible baseline comparison rather than proposing a new algorithm.

IV. DATASET AND METHODOLOGY

A. Dataset

The Wine dataset included with scikit-learn [3] contains 178 observations, 13 numeric predictors, and a three-class target. The predictors include alcohol, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, and proline. The classes contain 59, 71, and 48 samples respectively. The dataset is small and curated and does not represent wines from other regions or production years.

B. Experimental Protocol

Data were divided once into a stratified training set (80%, 142 samples) and held-out test set (20%, 36 samples) using random_state=42. Cross-validation was restricted to the training set and used five stratified folds with shuffle=True and random_state=42. The held-out test set was used for one final evaluation after fitting each model on the full training set.

C. Models and Preprocessing

Five classifiers were evaluated: logistic regression (maximum 3,000 iterations), RBF-kernel SVM, k-nearest neighbors (k = 5), random forest (500 trees), and gradient boosting. StandardScaler was fitted within a scikit-learn pipeline for logistic regression, SVM, and k-nearest neighbors, so scaling statistics were learned only from the training portion of each fold. Tree-based models were trained on the original features. Hyperparameters were kept at the stated or default settings; no hyperparameter search was conducted.

D. Metrics

Accuracy and macro-averaged precision, recall, and F1-score were computed on the held-out set; macro averaging gives each cultivar equal weight. Cross-validation accuracy is reported as mean ± standard deviation across five folds. Confusion matrices are reported with rows as actual class and columns as predicted class, in class order 0, 1, 2.

V. RESULTS

A. Cross-validation and held-out performance

Table I summarizes the results. Logistic regression, RBF SVM, and k-nearest neighbors each achieved 0.972 held-out accuracy. Random forest classified all 36 test samples correctly (1.000), and gradient boosting achieved 0.944. In cross-validation on the training set, the RBF SVM had the highest mean accuracy (0.986), logistic regression and random forest followed (0.979), then k-nearest neighbors (0.951) and gradient boosting (0.909, with the largest spread of 0.099). The held-out and cross-validation rankings therefore differ, and with such a small test set these differences should not be read as proof that one method is superior.

Model	CV Accuracy (mean ± SD)	Test Acc.	Macro Precision	Macro Recall	Macro F1
Logistic Regression	0.979 ± 0.028	0.972	0.978	0.967	0.971
RBF SVM	0.986 ± 0.017	0.972	0.978	0.967	0.971
k-Nearest Neighbors	0.951 ± 0.035	0.972	0.970	0.976	0.972
Random Forest	0.979 ± 0.028	1.000	1.000	1.000	1.000
Gradient Boosting	0.909 ± 0.099	0.944	0.951	0.943	0.945

TABLE I. TRAINING-SET CROSS-VALIDATION AND HELD-OUT TEST RESULTS

B. Classification errors

The held-out confusion matrices were: logistic regression [[12, 0, 0], [0, 14, 0], [0, 1, 9]], RBF SVM [[12, 0, 0], [0, 14, 0], [0, 1, 9]], k-nearest neighbors [[12, 0, 0], [0, 13, 1], [0, 0, 10]], random forest [[12, 0, 0], [0, 14, 0], [0, 0, 10]], and gradient boosting [[12, 0, 0], [1, 13, 0], [0, 1, 9]]. Logistic regression and SVM each made a single error (one class-2 sample predicted as class 1), k-nearest neighbors made one error (one class-1 sample predicted as class 2), gradient boosting made two errors, and random forest made none. These counts are benchmark-specific and are not expected error rates in practice.

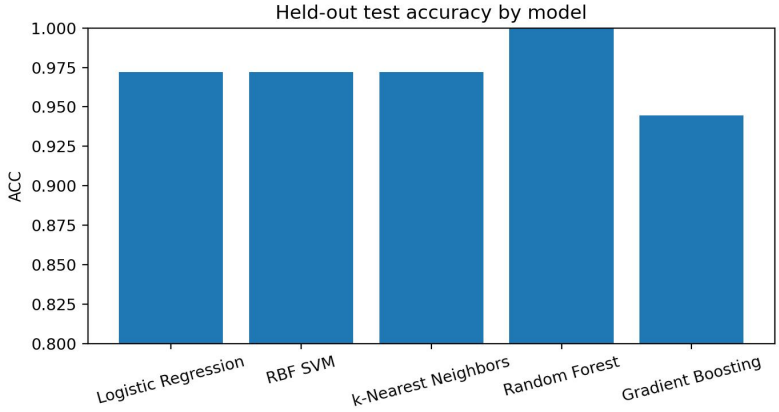


Fig. 1. Held-out test accuracy by classifier.

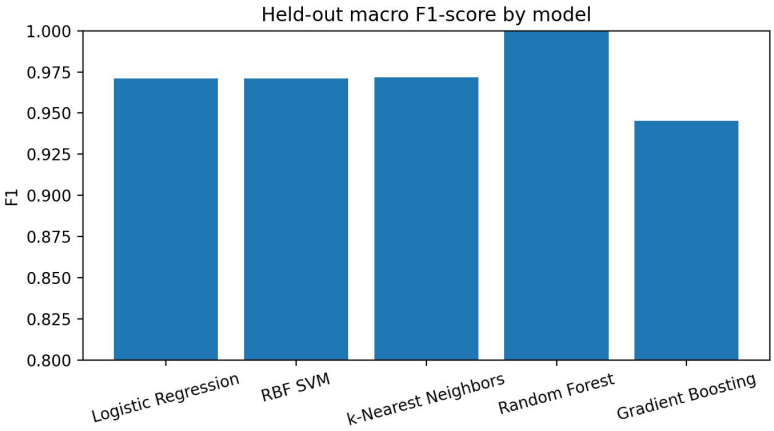


Fig. 2. Held-out macro F1-score by classifier.

VI. DISCUSSION

All five models separated the three cultivars well, which is consistent with the strong class structure of this dataset. Simple, well-scaled models such as logistic regression and the RBF SVM matched or approached more complex ensembles, and the SVM was the most accurate and stable model in cross-validation. Random forest's perfect held-out score is encouraging but rests on 36 test samples; its cross-validation accuracy was equal to that of logistic regression, so it should not be declared the best model on this evidence.

Gradient boosting showed the lowest and most variable cross-validation accuracy. With default settings and only 142 training samples, ensembles of many small trees can be sensitive to the particular folds; no tuning was attempted here to test that explanation. Repeated cross-validation or nested cross-validation with hyperparameter search would give more reliable model rankings.

VII. THREATS TO VALIDITY

Several limitations constrain the conclusions. First, the dataset has only 178 observations from a single region, so external validity is limited. Second, a single random train–test split with 36 test samples produces highly split-dependent estimates. Third, default or lightly specified hyperparameters were used, so this is not a comprehensive model search. Fourth, no independent laboratory, vintage, or region was evaluated, and no calibration or feature-importance analysis was performed. Finally, correlated chemical measurements may make results sensitive to measurement protocols.

VIII. ETHICAL AND PRACTICAL CONSIDERATIONS

The classifiers in this study are research baselines and should not be used for regulatory authentication or labeling decisions. A practical system would require data from multiple producers and years, validation against reference methods, uncertainty estimates, and documented quality control of the measurement process. The dataset contains no personal data.

IX. CONCLUSION

This study compared logistic regression, RBF SVM, k-nearest neighbors, random forest, and gradient boosting on the UCI Wine benchmark using a fixed stratified split and five-fold cross-validation. Three models reached 97.22% test accuracy, random forest reached 100%, and gradient boosting reached 94.44%, while cross-validation favored the RBF SVM. The findings provide a reproducible benchmark comparison, not evidence of readiness for real-world authentication. Future work should include repeated or nested cross-validation, hyperparameter tuning, external datasets, and calibration analysis.

DATA AND CODE AVAILABILITY

The dataset is distributed through scikit-learn's built-in loader (load_wine). The experiment used Python, scikit-learn 1.8.0, NumPy, and Matplotlib with a fixed random seed of 42. For a public preprint, the author should upload the exact code, package versions, and instructions to a version-controlled repository so that the reported results can be independently reproduced.

ACKNOWLEDGMENT

The author acknowledges the maintainers of scikit-learn and the researchers who collected and documented the Wine dataset.

REFERENCES

[1] M. Forina et al., “PARVUS: An extendible package for data exploration, classification and correlation,” Institute of Pharmaceutical and Food Analysis and Technologies, Genoa, Italy, 1988.

[2] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, School of Information and Computer Sciences, 2019. doi: 10.24432/C5PC7J.

[3] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

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

[5] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[6] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[7] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proc. IJCAI, 1995, pp. 1137–1143.

Author note: This manuscript reports a small benchmark experiment and has not been peer reviewed. Confirm all references, reproduce the experiment in your own environment, and review the reported metrics before submission. A DOI is assigned by the chosen repository or registration service; it cannot be guaranteed by the manuscript itself.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- T1.png