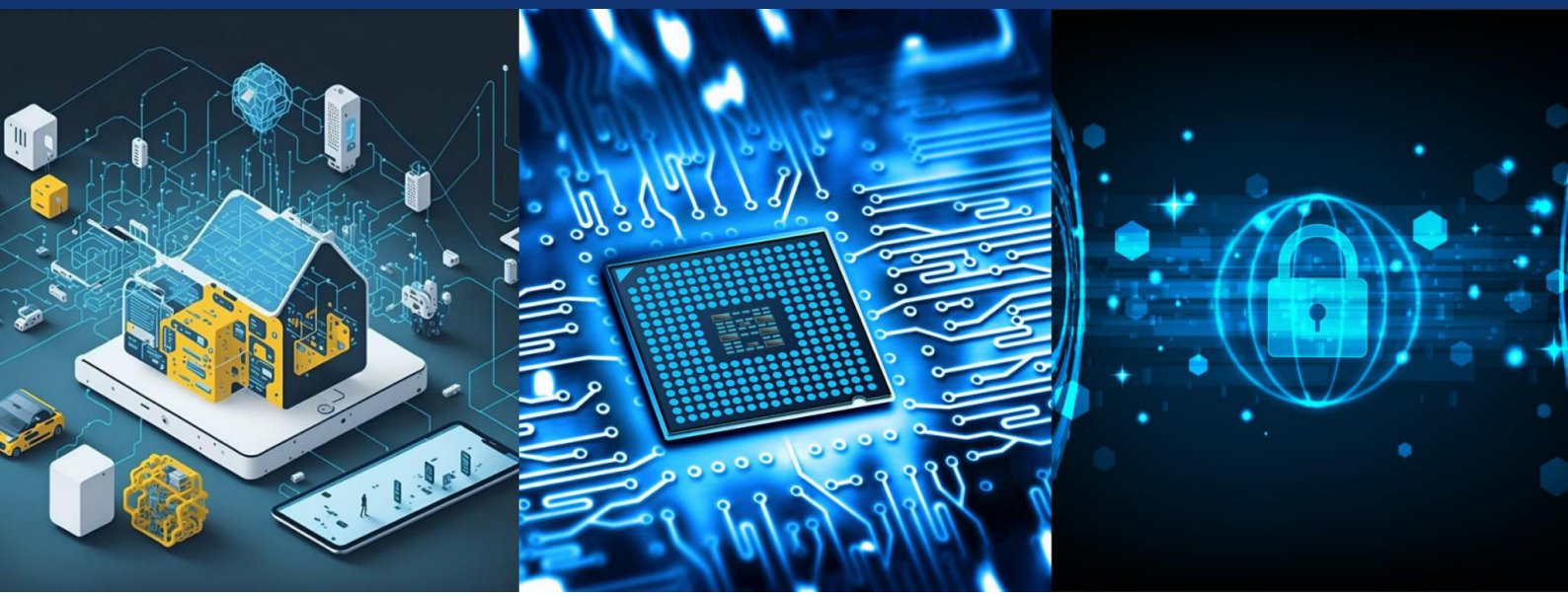
IJIRCCE

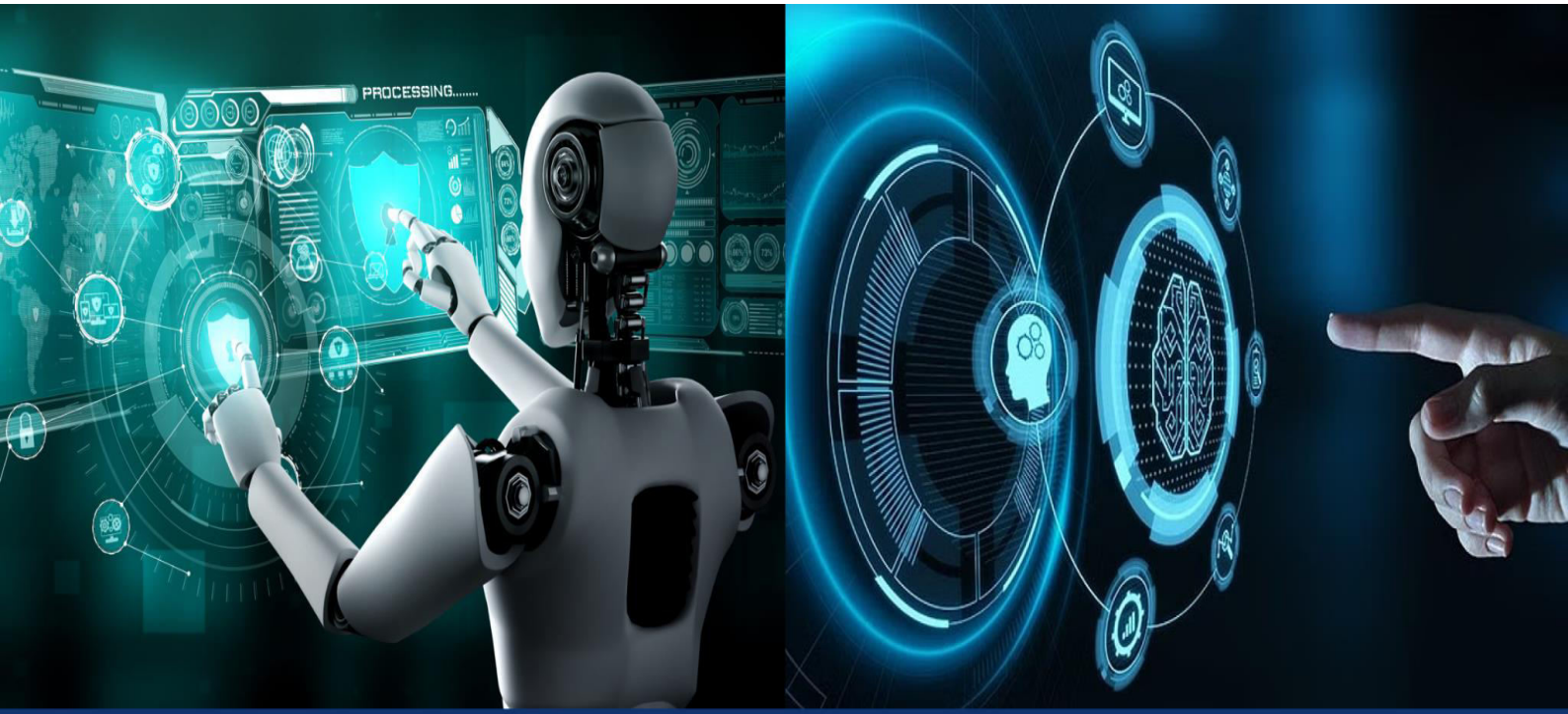
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 10, October 2026

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410019

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Alcohol Quality Prediction

Er. Sanjeev Jigali[1], Adarsh Prakash Ambore[2], Chandru Fakkirappa Lamani[3], Raghavendra Shrinivas Meharwade[4], Akash Vishnuraddi Yeresheemi[5]

Assistant Professor, Department of Computer Science and Engineering, STJIT, Ranebennur, India[1]

B.E. Students, Department of Computer Science and Engineering, STJIT, Ranebennur, India[2–5]

ABSTRACT: Wine quality estimation from physicochemical measurements is a useful machine-learning benchmark, but a prediction interface must communicate uncertainty and avoid presenting a sensory score as a health or safety judgment. This paper describes an evaluated machine-learning dashboard for estimating wine quality from laboratory-style measurements. The work uses the UCI Wine Quality data, which contains red and white Vinho Verde wine records, and compares a set of regression approaches before selecting a Random Forest model for the deployed application. On a held-out test set of 1,300 records, the selected model achieved a root mean squared error (RMSE) of 0.600, mean absolute error (MAE) of 0.432, and coefficient of determination (R^2) of 0.528; 66% of predictions were within 0.5 quality points of the recorded score. The system is implemented in Python with Streamlit and supports single-record prediction, batch CSV scoring, feature-importance visualisation, prediction history, and PDF reporting. Explanations are presented as model-based indicators rather than causal conclusions. The results suggest that physicochemical attributes contain useful but incomplete information about sensory quality. The system is a decision-support prototype and does not assess alcohol purity, toxicity, drinkability, or regulatory compliance.

KEYWORDS: Wine quality prediction, machine learning, Random Forest, regression, Streamlit, explainable AI, data visualisation.

I. INTRODUCTION

Sensory wine quality is influenced by chemical composition, production practices, storage, and human perception. Laboratory measurements can describe selected physicochemical properties, but they do not fully capture aroma, taste, balance, provenance, or individual preference. Predictive modelling therefore offers an analytical aid rather than a replacement for sensory evaluation. The Wine Quality dataset introduced by Cortez et al. is widely used to investigate relationships between physicochemical measurements and expert quality scores [1], [2].

This project develops a practical application around that prediction task. Instead of reporting a model score alone, the application provides a dashboard through which a user can enter measurements, request an estimate, inspect global feature importance, score a CSV file, review previous predictions, and export a report. The design emphasises reproducibility, a clear separation between prediction and interpretation, and explicit limitations.

The contributions of this work are: (1) a documented regression workflow using a fixed held-out evaluation set; (2) an implemented Random Forest model embedded in a Python/Streamlit dashboard; (3) user-facing prediction, batch-processing, and report-generation functions; and (4) safeguards against interpreting a sensory quality estimate as a health or safety certification.

II. RELATED WORK

Cortez et al. studied wine preference modelling using physicochemical laboratory data and data-mining methods, establishing the dataset and problem formulation commonly reused in later studies [1]. The UCI Machine Learning Repository distributes red and white wine records with physicochemical attributes and sensory quality labels [2]. Because the target is an ordered score rather than a direct chemical property, both regression and classification formulations are possible; regression preserves the numerical distance between quality scores.

Subsequent studies have evaluated multiple machine-learning algorithms and tuning strategies for wine-quality prediction. For example, a 2023 IEEE conference paper investigated machine-learning approaches with hyperparameter

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

optimisation for red-wine quality prediction [3]. Such results should be compared cautiously because dataset versions, train/test partitions, target encodings, and evaluation metrics can differ. The present work therefore reports the test metrics available for its own fixed split and does not claim a universal ranking over prior methods.

The implementation uses standard machine-learning tooling and a lightweight web interface. Scikit-learn provides estimator, preprocessing, model-selection, and evaluation utilities [4], while Streamlit supports the interactive Python application [5].

III. DATASET AND PROBLEM FORMULATION

A. Data Description

The source dataset contains Portuguese Vinho Verde wine samples. The commonly combined release contains 1,599 red-wine records and 4,898 white-wine records, for 6,497 observations in total [2]. Each observation includes 11 physicochemical measurements: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The recorded sensory quality score is the target. The combined workflow also retains wine type as a categorical indicator so that the model can distinguish red from white samples.

B. Problem Definition

Let x_i represent the input measurements for sample i and y_i its recorded quality score. The model estimates $\hat{y}_i = f(x_i)$. The task is treated as supervised regression, and errors are measured against the dataset's recorded scores. These scores are human sensory labels; they are not direct measurements of purity, safety, or health effects.

C. Data Partitioning

The data is divided into training and held-out test partitions using an 80:20 split and a fixed random seed. The resulting split contains 5,197 training samples and 1,300 test samples. Stratification is based on wine type and clipped quality categories to help preserve the distribution across the partitions. The test partition is reserved for final evaluation rather than model fitting.

IV. METHODOLOGY

A. Preprocessing and Candidate Models

The workflow checks the expected input columns and handles the combined dataset's wine-type field. The candidate regressors include Ridge Regression, Decision Tree, Random Forest, Extra Trees, and Gradient Boosting. These models provide a range of baselines, from a regularised linear estimator to nonlinear tree ensembles. Five-fold cross-validation was used during model comparison. A GridSearchCV procedure was applied to Random Forest using three-fold cross-validation and candidate values for tree depth, minimum leaf size, and the number of features considered at each split.

B. Model Selection

The selected deployed model is a Random Forest regressor configured with 300 trees, unlimited maximum depth, a minimum leaf size of one, and max_features = 0.5. Random Forest aggregates predictions from many decision trees and can represent nonlinear relationships and interactions in tabular data. It was retained for the application, including its tree-agreement display. The available experiment record does not contain a complete, verified metric table for every candidate model; accordingly, this paper does not invent missing values or claim that the selected model outperforms every alternative.

C. Evaluation Metrics

Performance is summarised using RMSE, MAE, and R^2. MAE is the average absolute difference between predicted and recorded quality scores. RMSE penalises larger errors more strongly because errors are squared before averaging. R^2 measures the proportion of test-target variance accounted for by the model relative to a constant-mean baseline. The proportion of predictions within 0.5 score points is also reported as an application-oriented tolerance measure. This tolerance is not a probability that an individual prediction is correct.

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410019

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Application and Explanations

The Streamlit application provides a single-sample form, sample loading, a batch CSV workflow with separator detection and column-name normalisation, model analytics, prediction history backed by SQLite, and downloadable PDF reports generated with ReportLab. The interface can display the estimate on multiple score scales; any transformed scale should be read as a display convention and not as a new ground-truth label.

Global feature importance is derived from the fitted tree ensemble. For a local explanation, the application replaces one input feature with a typical value for the relevant wine type and observes the change in prediction. This is a simple sensitivity comparison, not a causal explanation, and it may miss feature interactions. The displayed tree-agreement indicator represents the fraction of individual tree outputs lying within 0.5 quality points of the ensemble mean. It is not a calibrated confidence interval or a statistically validated probability of correctness.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION

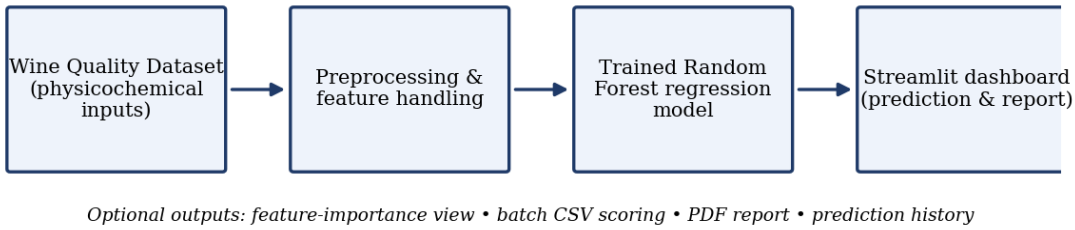


Fig. 1. High-level workflow of the prediction application.

The application is organised into data input, preprocessing, model inference, and presentation layers. Users supply the eleven physicochemical measurements and wine type; the application checks and arranges the input features before passing them to the saved estimator. Results are rendered with explanatory information and can be exported. Batch mode applies the same expected feature schema to rows in a CSV file. Prediction-history storage is local SQLite, so persistence depends on the deployment environment; ephemeral cloud filesystems may not preserve history across restarts.

The project is implemented in Python. Streamlit provides the user interface, scikit-learn supplies the estimator and evaluation utilities, pandas supports tabular data handling, SQLite stores history, and ReportLab generates PDF reports. The architecture intentionally avoids a separate API service for the Streamlit deployment path, reducing deployment complexity for a student project.

VI. RESULTS AND DISCUSSION

TABLE I. AVAILABLE HELD-OUT PERFORMANCE RESULTS

Metric	Held-out result
Test samples	1,300
RMSE	0.600
MAE	0.432
R^2	0.528
Within ±0.5 points	66%
Extra Trees RMSE / MAE	0.586 / 0.383

© 2026 IJIRCCE | Volume 14, Issue 10, October 2026 DOI: 10.15680/IJIRCCE.2026.1410019

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The deployed Random Forest achieved RMSE = 0.600, MAE = 0.432, and R^2 = 0.528 on 1,300 held-out samples. On average, predictions differed from recorded scores by approximately 0.43 points, while the RMSE indicates that larger deviations remain present. An R^2 of 0.528 indicates that the model explains a little over half of the variance in the held-out quality labels under this split; substantial variation remains unexplained. Sixty-six percent of estimates fell within 0.5 score points of the recorded label, which also means that roughly one-third did not meet that tolerance.

The experiment record additionally reports Extra Trees with RMSE = 0.586 and MAE = 0.383. These values are numerically lower than the reported Random Forest errors, but the available record does not verify the full evaluation context and remaining metrics for that model. The Random Forest was retained in the deployed dashboard for implementation reasons, including its tree-level agreement indicator. A rigorous model-selection decision should compare all candidates on the identical split and cross-validation protocol, report uncertainty or repeated-split results, and document the reason for choosing a model whose test errors may not be the lowest.

TABLE II. RECORDED FEATURE-IMPORTANCE ESTIMATES (RANDOM FOREST)

Feature	Approx. importance
Alcohol	≈ 21%
Volatile acidity	11.5%
Density	9.7%
Free sulfur dioxide	9.0%
Wine type	≈ 0.2%

The recorded feature-importance estimates identify alcohol (about 21%), volatile acidity (11.5%), density (9.7%), and free sulfur dioxide (9.0%) among the more influential variables; wine type contributes about 0.2% in that fitted model. These are model-specific importance estimates, not evidence that changing a variable will cause a particular sensory outcome. Correlated measurements can share or distort importance, and importance values may vary with data partition and model configuration.

VII. LIMITATIONS, RESPONSIBLE USE, AND FUTURE WORK

First, the data represents Vinho Verde wines from a particular region and dataset collection. Generalisation to other wine styles, countries, production methods, or laboratory procedures is not established. Second, sensory quality labels are imbalanced and inherently subjective. Third, feature importance and one-feature replacement explanations are limited and should not be interpreted causally. Fourth, the model has not been validated as a food-safety, toxicity, purity, adulteration, or regulatory-compliance detector. Any standards-screening component is a preliminary rule-based aid and requires verification against current authoritative requirements by a qualified professional.

Future work should publish the complete metric table for every candidate, compare repeated or nested cross-validation, add error analysis by wine type and quality band, test calibration and prediction intervals where appropriate, and conduct independent validation on external samples. Automated tests, schema validation, model-version tracking, drift monitoring, and persistent production storage would strengthen reliability. Deployment documentation should also state that cloud-hosted local SQLite history may be ephemeral.

VIII. CONCLUSION

This paper presented a machine-learning workflow and Streamlit dashboard for estimating wine sensory quality from physicochemical measurements. The deployed Random Forest produced an RMSE of 0.600, MAE of 0.432, and R^2 of 0.528 on the documented held-out set. The application combines single and batch predictions with feature-importance views, prediction history, and PDF reporting. Results show that measured attributes provide useful but incomplete predictive information. The system should be used for educational analysis and exploratory decision support only; it does not certify safety, purity, or health outcomes.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IX. ACKNOWLEDGMENT

The authors thank their project guide, the Department of Computer Science and Engineering, STJIT Ranebennur, and the maintainers of the UCI Wine Quality dataset.

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, doi: 10.1016/j.dss.2009.05.016.

[2]P . Cortez, "Wine Quality," UCI Machine Learning Repository, 2009, doi: 10.24432/C56S3T. [Online]. Available: https://archive.ics.uci.edu/dataset/186/wine+quality

[3] "Enhancing red wine quality prediction through Machine Learning approaches with Hyperparameters optimization technique," in Proc. 2nd Int. Conf. Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2023, doi: 10.1109/ICEEICT56924.2023.10157719. [Author and page details to be verified on IEEE Xplore before submission.]

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

[5] Streamlit, "Streamlit documentation." [Online]. Available: https://docs.streamlit.io/ (accessed Oct. 22, 2026).



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details