A logo of a globe with yellow rings around it

AI-generated content may be incorrect.

GROUP ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT002-3-2-AIM

AI METHODS

INTAKE: APU2F2411CS(AI)

DATE ASSIGNED: MONDAY, 7 July 2025, 12:00 AM

DATE DUE: FRIDAY, 1 AUGUST 2025 11:59 PM

LECTURER: MR. ZAILAN ARABEE BIN ABDUL SALAM

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Student Name** | **TP Number** | **Signature** |
|  | MANREEN KAUR A/P JAGJIT SINGH | TP071290 | A black text on a white background  AI-generated content may be incorrect. |
|  | SEHBA HANIEF | TP078886 |  |
|  | Hasibul Islam | TP077662 |  |
|  | Masaki Takano | TP073279 |  |
|  | Panov Egor | TP077925 |  |

Comparative Analysis of Machine Learning Models for Wine Classification and Quality Prediction

Manreen Kaur A/P Jagjit Singh   
*School of Computing*  
*Asia Pacific University of Technology and Innovation*Kuala Lumpur, Malaysia  
manreen2004@gmail.com

Panov Egor  
*School of Computing*  
*Asia Pacific University of Technology and Innovation*  
Kuala Lumpur, Malaysia  
microsoftfinnix@gmail.com

Hasibul Islam  
*School of Computing*  
*Asia Pacific University of Technology and Innovation*  
Kuala Lumpur, Malaysia  
hirudra90@gmail.com

Masaki Takano  
*School of Computing*  
*Asia Pacific University of Technology and Innovation*  
Kuala Lumpur, Malaysia  
TP077925@gmail.comSehba Hanief  
*School of Computing*  
*Asia Pacific University of Technology and Innovation*  
Kuala Lumpur, Malaysia  
sehbahanief1@gmail.com

*Abstract*— Machine learning today requires high accuracy and versatility to address complex classification and prediction problems. In this analysis, we use the UCI Wine dataset to compare a sample of supervised learning models, specifically Logistic Regression, k-Nearest Neighbors, Random Forest, XGBoost, and Support Vector Machine (SVM). The purpose of the analysis is not only to evaluate how well the models classify multiple structured classes, but also how robust each classification model is. We are very interested in SVM because it can theoretically make nonlinear decisions. As had been previously mentioned, we examined all models with the same pre-processing methods, reducing dimensionality, and then applying noise. We pay special attention to how these models behave and intend to provide advice and guidance to discover algorithms for any future tasks.

Keywords— Support Vector Machine, Machine Learning, Wine Classification, Dimensionality Reduction, Model Comparison

# Introduction

Machine learning has become very prevalent recently, and research in specific domains such as healthcare, financial, and related Manufacturing is excited to use AI classification algorithms and prediction tools; but selecting appropriate classification models continues to be a major hurdle. We needed to identify and select models not only based on the accuracy of their prediction, but also consider factors that the model was robust, scalable, and interpretable.

The purpose of this study is to compare several supervised machine learning models using UCI Wine data. The UCI Wine dataset has data that is structured regarding the chemical and property characteristics of wine based on three different types, making it an appropriate problem to apply traditional and advanced algorithms to explore their strengths and weaknesses, in addition to comparing the model with performance of the model. The models we used are Logistic Regression, k-Nearest Neighbors, Random Forest, XGBoost, and SVM.

Instead of trying to find the single best algorithm, we are interested in seeing how much better each model can get after applying widely used and shared preprocessing procedures. We are also interested in looking at how the models behaved when noise was applied to their models. We hope this is useful information for model selection for future classification problems based on the analysis results.

# Literature Review

Machine learning techniques have increasingly become pivotal in enhancing the accuracy and efficiency of wine quality classification. One foundational study by Mor et al. (2022) explored the relevance of physicochemical features in classifying red wine types. Their analysis using Random Forest, Decision Tree, and Support Vector Classifier demonstrated that only three core features,flavonoids, proline, and color intensity could achieve classification accuracy up to 97.78 percent, outperforming models trained on the full feature set. However, the study relied on a small dataset and lacked cross-validation, raising concerns about potential overfitting and the need for stronger external evaluation methods.

Jena et al. (2023) expanded this line of research by integrating dimensionality reduction techniques such as Principal Component Analysis and Linear Discriminant Analysis with conventional classifiers like Logistic Regression, Gradient Boosting, Random Forest, and SVM. The study reported perfect classification scores after LDA transformation, although this raised concerns of overfitting due to the absence of an independent test set. Nonetheless, the findings emphasized the benefits of incorporating supervised dimensionality reduction techniques in machine learning pipelines for wine quality classification tasks.

Chen (2024) conducted a comprehensive evaluation of five ensemble learning algorithms, including Gradient Boosting, XGBoost, LightGBM, CatBoost, and Random Forest, applied to both red and white wine datasets. The study featured a rigorous experimental setup involving StratifiedGroupKFold validation, SMOTE-Tomek resampling, and Optuna-based hyperparameter tuning. Among the models tested, Gradient Boosting achieved the highest weighted F1 scores. The study also found that reducing the feature set to the top five most informative variables, such as alcohol and volatile acidity, did not significantly compromise model performance, thereby supporting the use of leaner input spaces.

Zaza et al. (2023) addressed the challenge of imbalanced datasets by using SMOTE and comparing several models, including SVM, Random Forest, Gradient Boosting, KNN, and Decision Tree. The research highlighted alcohol as the most influential feature for wine quality prediction and demonstrated that with appropriate preprocessing and feature selection, even simple models can achieve competitive accuracy. Their conclusions align with our project’s use of SMOTE and confirm the importance of key features like alcohol in enhancing model robustness.

Wu and Zhang (2023) compared multiple classification models including SVM, KNN, Random Forest, Decision Trees, and Logistic Regression on the UCI wine dataset. The authors emphasized the role of feature normalization and dimensionality reduction using PCA. They found SVM and Random Forest to be the most accurate classifiers. However, the study lacked robustness checks under noisy or imbalanced conditions, which our project addressed through simulated noise injection, AUC-ROC evaluation, and confusion matrix analysis for a more comprehensive assessment.

Collectively, these studies demonstrate the value of ensemble models, dimensionality reduction, feature selection, and proper validation techniques in the domain of wine quality classification. Each paper contributes uniquely to the methodology and evaluation strategies adopted in our group project and supports the use of models such as SVM and Random Forest along with advanced validation and preprocessing techniques.

By integrating insights from these studies and building upon their limitations, our group project provides a more comprehensive and validated analysis of ML models for wine quality prediction.

# materials and methods

### Discussion on Implementation

The UCI machine learning repository’s wine dataset was used in this paper, which can be downloaded with sklearn. datasets. load\_wine API. The data set contains 178 samples and the following 13 physicochemical characteristics of wine alcohol, flavanoids, color intensity as well one target variable made up by three types of wines expressed in terms numeric labels: Class 0,1, and2.

### Data Exploration and Visualization

Preliminary analysis was analyse-d using pandas, numpy and matplotlib / seaborn / missingno. The dataset was checked for:

* Shape and structure,
* Unique value counts,
* Statistical summary (mean, sd, min-value, and max-value),
* Nonexistence of values and duplicates,
* Correlation and heat mapping.

Histograms and boxplots were used to visualize the distributions of features; pair plots were utilized for visualizing relationships between features. No missing values were.

### Data Preprocessing

**Such pre-processing steps involved data:**

* separation of features and targets - feature (X) was separated by eliminating the column of the target.
* Train-test split: 70 percent of the data was split and used as training data, whereas 30 percent of the data was as reserved data to be tested using the train\_test\_split () method that uses stratification.
* Standardization: The numerical features were standardized by StandardScaler to make the variance and mean to zero and one respectively.
* Duplications were also evaluated and noted.

### Dimensionality Reduction

To see how many components explain variance in total, Principal Component Analysis (PCA) was done with sklearn.decomposition.PCA. Even though dimensionality reduction could be plotted, all original features were kept during classification because they had a substantial proportion of variations between the components.

### Model Development

The classification task was conducted with the help of a Random Forest Classifier (sklearn.ensemble. RandomForestClassifier). I also used hyperparameters of n\_estimators=100 and random\_state=42.

**Cross-Validation:**

* StratifiedKFold was relied on to carry out 5-fold stratified cross-validation of the training set.
* Performance measure The metric thus employed was weighted F1-score through the cross\_val\_score.

### Model Evaluation

The model obtained was used to test the test set with the following metrics:

* Accuracy
* F1-Weighted
* Classification report
* HardCoder matrix
* AUC-ROC curves (per class, label\_binarize to be used in multiclass handling).

The confusion matrix, as well as ROC curves, were plotted by using seaborn and matplotlib. The AUC was calculated on each of the classes to assess the discriminatory performance of the model.

### Feature Importance

Features of importance values from the trained Random Forest model were plotted in declining order. This study has helped to find out the most significant constituents in the categorisation task.

# Results and discussion

### Discussion on Implementation

The UCI Wine dataset was used for the wine classification role, which includes 178 samples and 13 numerical characteristics that capture chemical aspects of various wine categories. The data was normalized using StandardScaler before being reduced in dimension using Principal Component Analysis (PCA). PCA kept more than 90% of the dataset's variance, significantly lowering feature dimensionality while keeping important information.

The Support Vector Machine (SVM) technique was chosen as the fundamental model for this challenge because of its shown ability to handle nonlinear classification issues. An SVM with a radial basis function (RBF) kernel was trained and verified using a stratified 5-fold cross-validation method. The same SVM model was tested again on a separate test set to determine its generalization performance.

To imitate real-world variability, the model was tested in noisy conditions, which involved injecting Gaussian noise into the test features. This enabled robustness testing of the SVM classifier's reliability in the face of data distortion. Standard performance criteria such as accuracy, weighted F1-score, and ROC-AUC were used in the evaluation process.

### Results

|  |  |  |
| --- | --- | --- |
| Model | Accuracy (%) | F1 Score (%) |
| SVM | 0.962963 | 0.962586 |
| KNN | 0.944444 | 0.944663 |
| Random Forest | 0.925926 | 0.925431 |
| Logistics Regression | 0.925926 | 0.925660 |
| XGBoost | 0.907407 | 0.908835 |
| MLP | 0.888889 | 0.889651 |

Table 1: Model Performance Summary for SVM

Note: Other models were tested for benchmarking purposes only and are not discussed in detail.

The Support Vector Machine model outperformed all of the other models examined in terms of total classification accuracy. It achieved an accuracy of 96.30% and a weighted F1-score of 96.26% on the test set, as shown in Table 1. This demonstrates the SVM's ability to properly generalize across multi-class wine categories.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 1: SVM Accuracy and F1 Score.

As shown in Figure 1, SVM clearly exceed the other classifiers in terms of accuracy and F1-score. The RBF kernel allowed SVM to create flexible non-linear decision boundaries, which is particularly useful in datasets with overlapping class distributions. Its consistency in forecasting across all three wine classes is supported by high per-class precision and recall scores.

<p align="center"> <b>Figure 1:</b> Accuracy Comparison of SVM and Other ML Models </p>

In addition to its superior performance on clean data, the SVM model maintained good classification performance when evaluated on noisy input features. After adding controlled Gaussian noise to the test set, the SVM classifier maintained an accuracy of 94.44%, indicating its durability and stability. This increases the reliability of SVM in real-world contexts where measurement errors or data inconsistencies are widespread.

A diagram of a graph

AI-generated content may be incorrect.

Figure 2: SVM Confusion Matrix

A graph with a line and a line

AI-generated content may be incorrect.

Figure 3: SVM AUC-ROC Curve

The AUC-ROC study for SVM (Figure 3) demonstrates good separability across all classes, with area under the curve (AUC) values more than 0.95 for each class. This verifies the model's high discriminative power and balanced classification abilities across all categories.

<p align="center"> <b>Figure 2:</b> Multiclass ROC Curve – SVM </p>

Overall, the findings demonstrate that SVM is not only accurate but also generalizable and robust. It provides excellent performance without requiring lengthy tuning or sophisticated ensemble techniques. Given its efficiency on a medium-sized, multi-class dataset, SVM is recommended as the principal classification model for predicting wine quality in comparable settings.

A screen shot of a computer

AI-generated content may be incorrect.

Figure 4: All Model Algorithms Accuracy and F1-Score

# Recommendation

Based on the findings from our study and supported by the research article “Wine Quality Classification with Machine Learning Approaches,” we recommend using advanced supervised learning algorithms like **Support Vector Machine (SVM)** and **Random Forest (RF)** for wine quality classification problems. These models consistently demonstrated high accuracy and generalizability across multiple evaluations. Specifically, SVM outperformed other models in our implementation, confirming its effectiveness in handling non-linear classification tasks. For future enhancements, techniques such as **dimensionality reduction (PCA)** and **hyperparameter tuning** should be prioritized, as they significantly contribute to improving classification outcomes, as also highlighted in the literature. For future work, we recommend expanding the scope of analysis by incorporating **larger and more diverse datasets**, ideally sourced from **multiple wine-producing regions or farms across different climates and soil types**. This would not only improve the robustness and accuracy of predictive models but also make them more generalizable to real-world applications. Additionally, integrating **advanced feature engineering**, **ensemble methods**, and **deep learning architectures** could further enhance model performance. Lastly, real-time quality prediction systems may be developed by combining sensor data with trained machine learning models, creating intelligent tools for the wine industry.

# Conclusion

This study explored various machine learning algorithms, namely K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, XGBoost, MLP, and SVM—for classifying wine quality using the UCI Wine Dataset. Among them, **SVM emerged as the best performer**, achieving the highest accuracy and F1-score. These findings align with the article reviewed, where ensemble and kernel-based models were noted for their superior performance. Through detailed evaluation using confusion matrices, cross-validation, and AUC-ROC curves, our project emphasized the strengths and limitations of each model. The results validate that model selection, preprocessing, and evaluation strategies play a crucial role in achieving robust and reliable classification systems.

##### Acknowledgment

We would like to express our sincere gratitude to our module lecturer and research guide for their continuous support and constructive feedback throughout this project. Their guidance played an essential role in helping us explore, evaluate, and implement multiple machine learning approaches effectively. We also appreciate the collaborative efforts of every group member, whose contributions and teamwork ensured the successful completion of this research.

##### References

Benjamin, A. C. (2022). Wine Quality Classification Using Machine Learning Algorithms. International Journal of Computer Applications Technology and Research, 11(06), 241–246.

<https://doi.org/10.7753/ijcatr1106.1010>

**Wu, M., & Zhang, R. (2023, December).** *Wine quality classification with machine learning approaches*. ResearchGate. <https://www.researchgate.net/publication/378745496_Wine_Quality_Classification_with_Machine_Learning_Approaches>

**Mor, N. S.** (2022). *Physicochemical properties importance for type classification of wines using machine learning techniques* [Preprint]. agriRxiv. <https://doi.org/10.31220/agriRxiv.2022.00126>

**Jena, S. P., Paikaray, B. K., Pramanik, J., Thapa, R., & Samal, A. K.** (2023). *Classifications on wine informatics using PCA, LDA, and supervised machine learning techniques*. International Journal of Work Innovation, 4(1), 58–73. <https://www.researchgate.net/publication/370191000_Classifications_on_wine_informatics_using_PCA_LDA_and_supervised_machine_learning_techniques>

Chen, Z. (2024). *Wine quality prediction with ensemble trees: A unified, leak-free comparative study*. arXiv. <https://arxiv.org/abs/2506.06327>

Jena, W. P., Paikaray, B., Pramanik, J., Thapa, R., & Samal, A. K. (2023). *Classifications on wine informatics using PCA, LDA, and supervised machine learning techniques*. ResearchGate. <https://www.researchgate.net/publication/370191000_Classifications_on_wine_informatics_using_PCA_LDA_and_supervised_machine_learning_techniques>

Mor, N. S., Asras, T., Gal, E., Demasia, T., & Tarab, E. (2022). *Physicochemical properties importance for type classification of wines using machine-learning techniques* [Preprint]. agriRxiv. <https://doi.org/10.31220/agriRxiv.2022.00126>

Wu, Y., & Zhang, H. (2023). *Wine quality classification with machine learning approaches*. ResearchGate. <https://www.researchgate.net/publication/378745496_Wine_Quality_Classification_with_Machine_Learning_Approaches>

Zaza, S., Atemkeng, M., & Hamlomo, S. (2023). *Wine feature importance and quality prediction: A comparative study of machine learning algorithms with unbalanced data*. arXiv. <https://arxiv.org/abs/2310.01584>