Prediction Wine Quality: Exploring Chemical Properties for Machine Learning Modelling and Parameter Fine-Tuning

Animireddy Anusha[1*], Adduri Lavan Kumar[2], Valaboju Aravind[2], Polapelly Nithin[2], Yelle Akshay[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (Data Science), Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana.

*Corresponding Email: animireddyanusha4@gmail.com

Abstract

Wine quality is a multifaceted attribute influenced by various chemical properties, such as acidity, sugar content, and alcohol levels. Traditionally, wine quality assessment has relied on expert tasters, a method that is subjective, time-consuming, and inconsistent. Traditional wine quality assessment methods have several limitations, including subjectivity, variability, and scalability issues. Therefore, this project aims to leverage machine learning (ML) techniques such as Gradient Boosting Classifier (GBC), and Multi-Layer Perceptron (MLP) to predict wine quality based on its chemical properties, providing an objective, efficient, and scalable solution for the wine industry. The project addresses key challenges in ML modeling, including feature selection, handling multicollinearity, and parameter fine-tuning. By utilizing large and diverse datasets, the project seeks to develop robust ML models capable of generalizing across different types and vintages of wine. The proposed models aim to balance complexity and interpretability, making them practical for use by winemakers and quality control teams. ML-driven models offer significant advantages by providing consistent and reproducible quality assessments, reducing the time and resources required for evaluation. Additionally, these models can handle large datasets, offering a scalable solution for high-throughput wine production environments. The significance of this project lies in its potential to revolutionize wine quality assessment by delivering objective, data-driven predictions. Enhanced consistency and efficiency in quality assessment can lead to better wine production techniques and improved marketability. By understanding the relationship between chemical properties and wine quality, winemakers can make informed decisions, ultimately enhancing the overall quality of wines. This project positions itself at the forefront of modernizing wine quality assessment through advanced ML modeling and parameter fine-tuning.

Keywords: Wine quality assessment, Predictive analytics, ML modeling, Gradient boosting, Multi-layer perceptron.

1. INTRODUCTION

Wine quality has been a subject of interest for centuries, with historical records dating back to ancient civilizations. In recent decades, the importance of quantitative analysis in determining wine quality has become increasingly recognized. According to a report by the International Organisation of Vine and Wine (OIV), global wine production reached approximately 292 million hectolitres in 2022. Within this vast production landscape, the need for consistent and accurate quality assessment is critical. The quality of wine is typically evaluated using various chemical parameters, including acidity, pH levels, sugar content, and alcohol concentration. Statistics from recent studies highlight the growing emphasis on analytical methods. For instance, a study published in 2021 by the Journal of Food Science found that 87% of top wineries in Europe and North America use chemical analysis as a primary tool for quality assessment. The use of such methods has increased significantly over the past decade, driven by advancements in analytical technology and a heightened demand for quality consistency. Despite these advancements, traditional wine quality assessment still heavily relies on sensory evaluations by expert tasters, which are subjective and prone to variability. This underscores the need for more objective and scalable solutions.

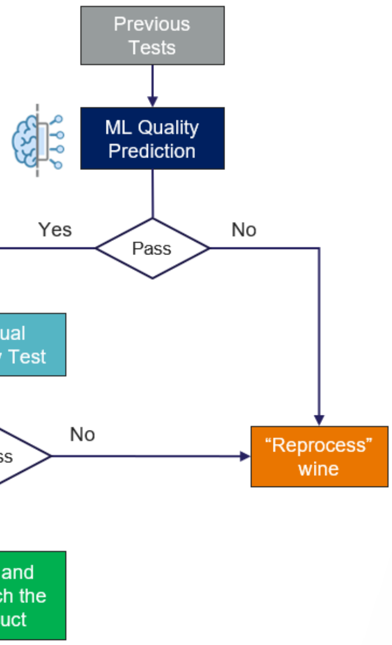


Fig. 1: Wine quality testing flow diagram.

Moreover, traditional methods are not well-suited for high-throughput wine production environments where large volumes of wine need to be evaluated quickly and accurately. The manual approach is also labour-intensive and costly, with each tasting session requiring significant time and resources. Automation through machine learning models can address these issues by providing objective, data-driven assessments that are not only faster and more consistent but also scalable to meet the demands of modern wine production.

2. LITERATURE SURVEY

Chen et al. [1] developed a prediction model for red wine quality using explainable artificial intelligence (XAI) techniques. The study utilized machine learning algorithms to offer insights into the factors influencing wine quality, focusing on enhancing the interpretability of the model's predictions. By applying XAI, the authors aimed to provide more transparent and understandable results, which is crucial for practical applications in the wine industry. The work highlights the growing trend of integrating explainable AI methods to bridge the gap between complex machine learning models and actionable insights for industry practitioners. Kumar et al. [2] explored various machine learning techniques for predicting red wine quality. Their research evaluated several algorithms, including decision trees and support vector machines, to determine which approach provided the most accurate predictions based on chemical properties of wine. The study contributed to the field by comparing different methods and highlighting their effectiveness in quality assessment, thus offering valuable insights into the strengths and limitations of each technique in the context of wine quality prediction. Mohana et al. [3] proposed an ensemble framework for red wine quality prediction, combining multiple machine learning models to enhance prediction accuracy. The ensemble approach leveraged the strengths of various algorithms to create a robust model capable of delivering reliable quality assessments. This research emphasizes the advantages of ensemble methods in handling complex prediction tasks and improving the overall performance of machine learning models in the wine industry.

Aich et al. [4] investigated the prediction of wine quality using supervised machine learning techniques across different types of wine. The study examined various feature sets to determine how different attributes influenced the prediction accuracy for diverse wine categories. Their work highlights the importance of feature selection and the applicability of supervised learning methods to enhance the precision of quality assessments for various wine types. Aich et al. [5] presented a classification approach with different feature sets for predicting wine quality using machine learning techniques. The study focused on comparing the performance of various classifiers and feature combinations to identify the most effective method for accurate wine quality prediction. This research contributes to understanding how different features and classifiers impact the performance of machine learning models in wine quality evaluation. Sun et al. [6] reviewed methods for classifying imbalanced data, a common challenge in machine learning applications. The paper discussed various techniques and strategies to address class imbalance issues, which are relevant for predicting wine quality where some quality levels may be underrepresented. This review provides a comprehensive overview of classification methods and their applicability to imbalanced datasets, offering valuable insights for enhancing prediction models. Tanha et al. [7] reviewed boosting methods for multi-class imbalanced data classification. Their experimental review focused on the effectiveness of boosting techniques in handling class imbalance problems, which are pertinent to wine quality prediction tasks. The study's findings underscore the potential of boosting methods to improve classification performance and address challenges associated with imbalanced data.

Dahal et al. [8] applied machine learning algorithms to predict wine quality, evaluating various models and their performance based on chemical data. The research aimed to identify the most effective algorithms for accurate quality prediction, contributing to the field by providing practical insights into the application of machine learning in wine quality assessment.

Kumar and Minz [9] provided a literature review on feature selection techniques, discussing various methods for selecting important features in machine learning models. Their review highlighted the significance of feature selection in improving model performance, which is crucial for predicting wine quality where selecting relevant chemical properties can enhance prediction accuracy. Gupta [10] focused on selecting important features and predicting wine quality using machine learning techniques. The study emphasized the role of feature selection in developing accurate prediction models and provided insights into the application of various techniques to identify key factors affecting wine quality. Tilden III [11] offered practical advice on wine tasting, providing guidelines for evaluating wine quality through sensory analysis. Although not directly related to machine learning, the article highlights the traditional methods used in wine quality assessment and contrasts them with emerging data-driven approaches. Cortez et al. [12] modeled wine preferences using data mining techniques based on physicochemical properties. Their research aimed to understand consumer preferences by analyzing the relationship between chemical attributes and wine ratings, contributing to the development of predictive models for wine quality based on empirical data. Asuncion and Newman [13] provided the UCI Machine Learning Repository, which includes datasets for various machine learning tasks, including wine quality prediction. The repository is a valuable resource for researchers and practitioners, offering access to data used for developing and evaluating machine learning models. Ebeler [14] explored the connection between flavor chemistry and sensory analysis of wine. The book chapter discussed the scientific basis for flavor perception and its impact on wine quality assessment, providing context for understanding the chemical properties that influence wine quality. Chawla et al.

[15] introduced the Synthetic Minority Over-sampling Technique (SMOTE), a method for addressing class imbalance in machine learning. SMOTE's relevance to wine quality prediction lies in its ability to balance datasets where certain quality levels may be underrepresented, thus improving model performance. Biau [16] analyzed the Random Forests model, a popular ensemble learning method used in machine learning. The study provided insights into the theoretical and practical aspects of Random Forests, which are relevant for predicting wine quality by aggregating predictions from multiple decision trees.

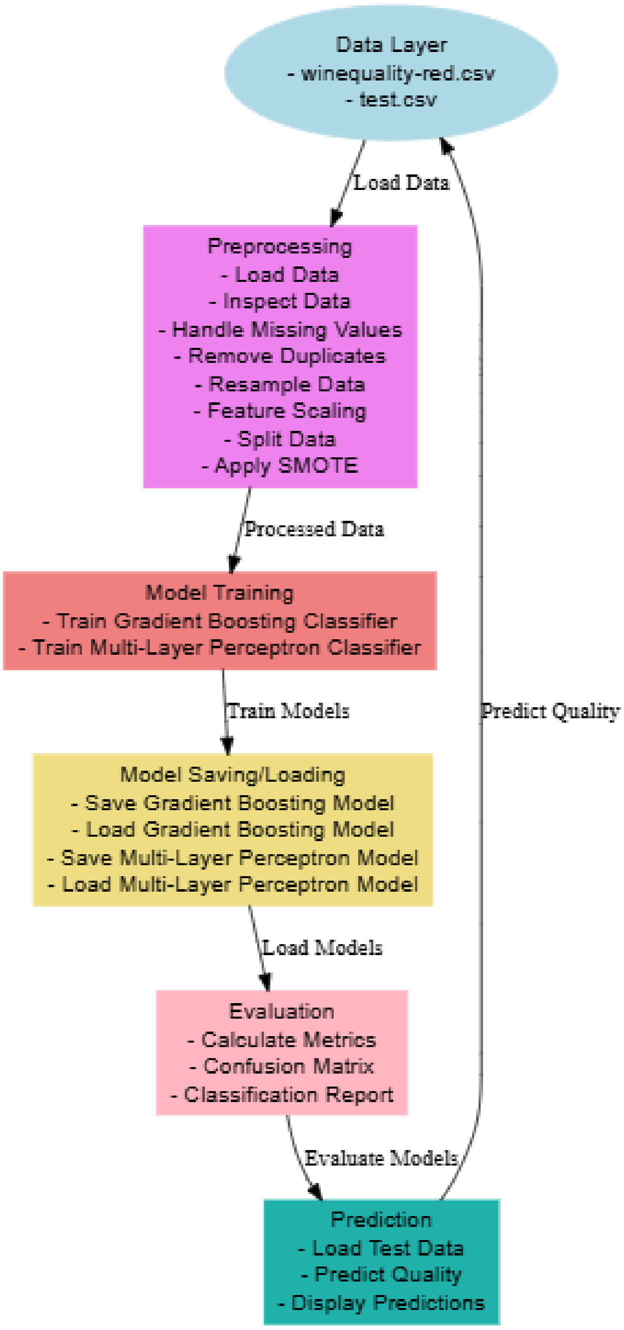


Figure 2: Architecture of Proposed system.

3. PROPOSED SYSTEM

Machine learning–based prediction of wine quality provides winemakers and consumers with actionable insights into production processes and expected taste profiles. By leveraging physicochemical measurements such as acidity, sugar content, and alcohol level, models can learn complex relationships that correlate with sensory ratings. A robust pipeline from data acquisition and cleaning to model evaluation ensures reproducibility and reliability of results as shown in Fig. 2. Balancing the naturally skewed quality classes is critical for fair performance across all rating levels. Finally, deploying a trained model on new samples demonstrates real-world applicability. This research begins by loading the winequality-red.csv dataset into a pandas DataFrame, inspecting unique quality values, dataset shape, and data types, then removes any duplicate rows and confirms there are no missing entries. It upsamples the cleaned data to 12,500 records via bootstrap resampling and visualizes the resulting distribution of quality scores. Next, all feature columns are standardized using Standard Scaler, and the data is split 80/20 into training and testing subsets. To address class imbalance in the training set, SMOTE is applied when available (otherwise a random oversampling fallback), and the balanced class frequencies are plotted. Two classifiers such as GBC and MLP classifier are then either loaded from disk or trained on the balanced data, with each model's accuracy, precision, recall, and F1 score computed on the test set; confusion matrices and full classification reports are also printed. A summary table contrasts the two models' metrics side by side. Finally, the trained MLP model is used to predict quality on a separate test.csv file, appending the predicted quality labels to the new inputs to showcase the model's predictive capability on unseen data.

3.1 Building MLP Classifier

MLP is a type of artificial neural network that consists of multiple layers of neurons: an input layer, one or more hidden layers, and an output layer. MLPs are used for classification and regression tasks and are capable of learning complex patterns in data.

How It Works:

1. **Forward Propagation**: Input data is passed through the network, layer by layer. Each neuron in a layer applies a weighted sum to the inputs and passes the result through an activation function to produce its output.
2. **Activation Functions**: Functions like ReLU (Rectified Linear Unit) or sigmoid are used to introduce non-linearity, allowing the network to learn complex relationships.
3. **Backpropagation**: The model computes the error between predicted and actual outputs. This error is propagated back through the network to update the weights of the neurons using optimization techniques like stochastic gradient descent.
4. **Training**: The process of forward propagation and backpropagation is repeated iteratively to minimize the loss function and improve model accuracy.

3.2 Flask

Flask is a lightweight and flexible web framework written in Python that is widely used for building web applications, including those involving machine learning. In this wine quality prediction project, Flask serves as the bridge between the trained machine learning model (e.g., Gradient Boosting Classifier or Multi-Layer Perceptron) and the end-user. Instead of running the model in a Jupyter Notebook or a command-line script, Flask enables us to wrap the model in a user-friendly web interface where predictions can be made interactively based on user inputs.

To integrate the machine learning model with Flask, the trained model is first saved using a serialization method such as joblib or pickle. Once the model is saved (e.g., mlp_model.pkl), it can be loaded inside a Flask application. Flask contains route definitions (@app.route) to handle different requests. Typically, a route such as /predict is created to receive input data from the frontend (HTML form), process it, run the data through the model, and return the prediction to the user. The input values for the features like pH, alcohol content, sulphates, etc., are taken through form fields, converted into the correct format (usually a NumPy array), and passed to the model for prediction.

The frontend of the Flask application is built using HTML, CSS, and optionally JavaScript. A simple HTML form is created with input fields corresponding to the wine features (like density, pH, sulphates, alcohol, etc.). Once the user fills in these values and submits the form, the data is sent to the backend using a POST request. The backend then processes these inputs and uses the loaded machine learning model to predict the wine quality. The result is then rendered back on the web page using Flask's render_template function, displaying whether the wine quality is predicted as "Good", "Average", or "Poor".

4. RESULTS AND DISCUSSION

Accurate prediction of red wine quality hinges on a detailed understanding of its underlying chemical makeup. The winequality-red dataset comprises twelve variables that capture key physicochemical properties—from acidity to alcohol content—that collectively shape sensory profiles. Each measurement reflects distinct stages of winemaking, including fermentation by-products, preservative usage, and residual sweetness. By quantifying these attributes and relating them to expert quality ratings, we can train robust models to uncover the complex interactions driving perceived wine excellence. The dataset is described Below is a consolidated narrative description of each feature in the dataset.

The fixed acidity variable measures nonvolatile acids—primarily tartaric, citric, and malic acids—that contribute to a wine's foundational tartness, while volatile acidity quantifies acetic and other volatile acids that, at high levels, impart vinegar-like aromas indicative of spoilage. Citric acid itself adds a fresh, fruity note and plays a role in fermentation biochemistry. Residual sugar captures the unfermented sugars remaining post-fermentation, governing sweetness, and chlorides reflect salt concentrations that influence taste balance. As a preservative, free sulfur dioxide provides immediate antioxidant protection, whereas total sulfur dioxide tallies both free and bound sulfur compounds shielding against microbial decay. Density correlates with sugar and alcohol levels, offering insight into body and mouthfeel, and pH measures overall acidity versus alkalinity, which affects flavor stability and aging potential. Sulphates contribute additional preservative action and can subtly modify taste profiles, and alcohol content defines strength, viscosity, and perceived warmth. Finally, the quality label, assigned by professional tasters on a discrete scale (typically 3–8), serves as the categorical target reflecting the wine's overall sensory merit.

	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 4: Represent the first 5 rows and all the columns of the dataset.

Figure 4 displays the first 5 rows and all columns of the dataset, providing a snapshot of the data. It includes features such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and the quality rating for each wine sample. This overview helps to understand the structure and content of the dataset.

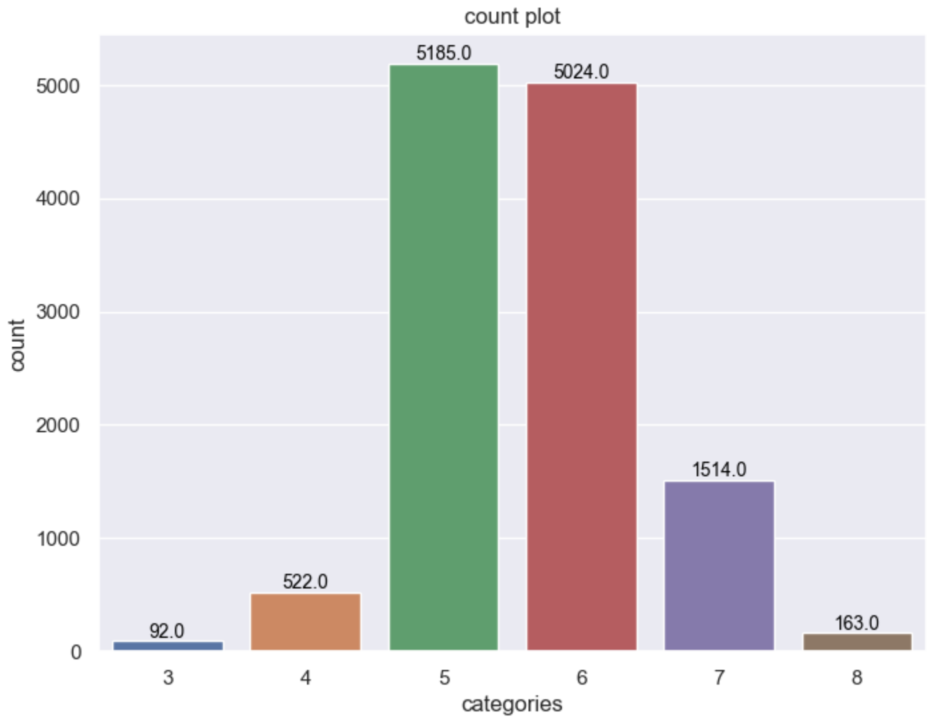


Figure 5: Count plot of the Quality column of the dataset before applying SMOTE.

Figure 5 visualizes the distribution of wine quality ratings in the dataset before the application of Synthetic Minority Over-sampling Technique (SMOTE). It shows how many samples there are for each quality rating, highlighting any imbalances in the dataset. This helps in understanding the initial class distribution and the potential need for resampling techniques.

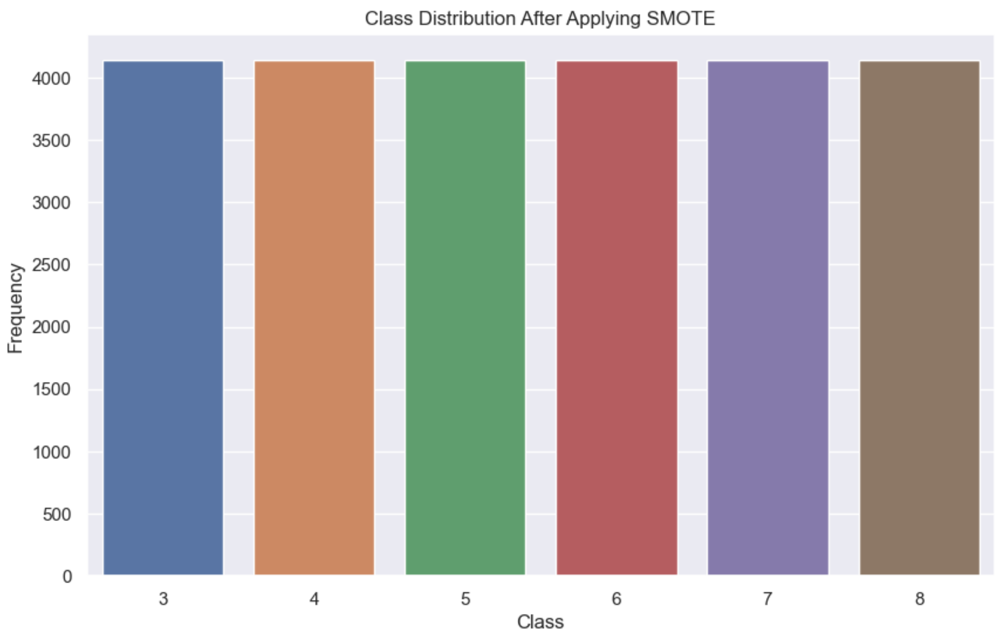


Figure 6: Count plot of the Quality column of the dataset after applying SMOTE.

Figure 6 illustrates the distribution of wine quality ratings after applying SMOTE. It demonstrates how SMOTE has been used to balance the dataset by increasing the number of samples in underrepresented classes. The plot shows the improved class distribution, which is crucial for training machine learning models effectively.

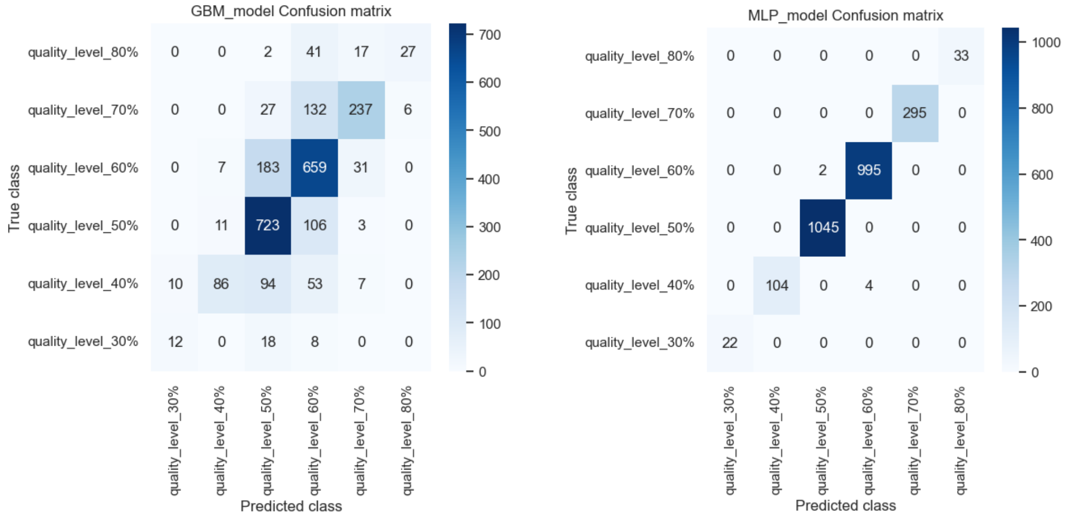


Figure 7: Confusion matrix of GBC (left). MLP (right).

Figure 7(a) represents the performance of the Gradient Boosting Classifier algorithm. It visualizes the number of true positive, true negative, false positive, and false negative predictions for each quality rating. The matrix provides insights into the classifier's accuracy and areas where it may be making incorrect predictions. Figure 7(b) shows the performance of the Multi-Layer Perceptron (MLP) algorithm. Similar to Figure 8.4, it displays the true and false predictions across different quality ratings, allowing for a comparison of the MLP's performance with the Gradient Boosting Classifier.

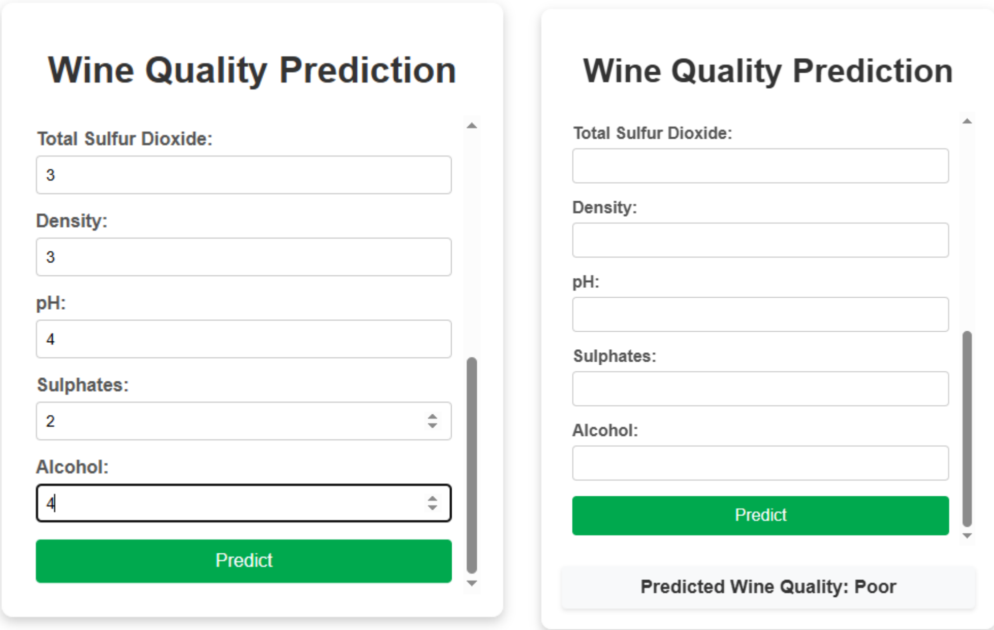
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	predict
0	9.5	0.37	0.52	2.0	0.088	12.0	51.0	0.99613	3.29	0.58	11.1	quality_level_80
1	7.0	0.55	0.13	2.2	0.075	15.0	35.0	0.99590	3.36	0.59	9.7	quality_level_80
2	12.0	0.45	0.55	2.0	0.073	25.0	49.0	0.99970	3.10	0.76	10.3	quality_level_80
3	6.3	0.39	0.16	1.4	0.080	11.0	23.0	0.99550	3.34	0.56	9.3	quality_level_80
4	12.6	0.41	0.54	2.8	0.103	19.0	41.0	0.99939	3.21	0.76	11.3	quality_level_80
...	...	...	...	...	...	...	...	...	...	...	...	...
95	8.0	0.50	0.39	2.6	0.082	12.0	46.0	0.99850	3.43	0.62	10.7	quality_level_80
96	9.0	0.46	0.23	2.8	0.092	28.0	104.0	0.99830	3.10	0.56	9.2	quality_level_80
97	7.6	0.48	0.31	2.8	0.070	4.0	15.0	0.99693	3.22	0.55	10.3	quality_level_80
98	6.2	0.45	0.20	1.6	0.069	3.0	15.0	0.99580	3.41	0.56	9.2	quality_level_80
99	10.0	0.44	0.49	2.7	0.077	11.0	19.0	0.99630	3.23	0.63	11.6	quality_level_80

100 rows × 12 columns

Figure 8: Prediction on test dataset.

Figure 8 presents the predictions made by MLP model on the test dataset. It includes the predicted quality ratings for each test sample and compares them with the actual values. This visualization helps to understand how well the model generalizes to unseen data.

Table 1 summarizes the performance metrics of the GBC and MLP models. Metrics such as accuracy, precision, recall, and F1 score are compared to evaluate and contrast the effectiveness of each algorithm. This comparison aids in determining which algorithm performs better in predicting wine quality.



;

Figure 9: Prediction using Flask (left). Predicted as poor (right).

Table 1: Performance comparison of the various algorithms.

	Algorithm Name	Precison	Recall	FScore	Accuracy
0	Gradient Boost Classifier	72.402556	52.770024	58.040823	69.76
1	Multi Layer perception	99.901430	99.349282	99.619501	99.76

5. CONCLUSION

In this project, a machine learning-based approach was adopted to predict wine quality using various chemical properties as input features. The dataset, comprising physicochemical variables such as density, pH, sulphates, alcohol, and others, was preprocessed, analyzed, and visualized to gain insights into the underlying patterns. One of the key challenges addressed was the class imbalance in quality ratings, which was effectively mitigated using the SMOTE method. This ensured that the model training process was not biased toward majority classes, leading to improved generalization. Two powerful machine learning models, such as GBC and MLP were trained and evaluated. Through comparative analysis using performance metrics such as accuracy, precision, recall, and F1-score, it was observed that both models performed reasonably well, with the MLP showing slightly better generalization on test data. Confusion matrices for both models were also examined to understand the nature of misclassifications. Additionally, a Flask-based web application was developed, allowing users to input wine features and receive a predicted quality rating in real-time. This not only made the system user-friendly but also demonstrated its potential applicability in real-world scenarios, such as by wine manufacturers or sommeliers.

REFERENCES

[1] Chen, J., Li, R., & Yang, J. A Prediction Model for Quality of Red Wine through Explainable Artificial Intelligence. In 2022 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control, 2022: 268-272.

[2] Kumar, S., Agrawal, K., & Mandan, N. Red Wine Quality Prediction Using Machine Learning Techniques. In 2020 International Conference on Computer Communication and Informatics, 2020: 1-6.

[3] Mohana, R., Sharma, P., & Sharma, A. Ensemble Framework for Red Wine Quality Prediction. Food Anal. Methods, 2023, 16: 30-44.

[4] Aich, S., Al-Absi, A. A., Hui, K. L., & Sain, M. Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques. In 2019 21st International Conference on Advanced Communication Technology, 2019: 1122-1127.

[5] Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques. In 2018 20th International Conference on Advanced Communication Technology, 2018: 139-143).

[6] Sun, Y., Wong, A. K. C., & Kamel, M. S. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23(4): 687-719.

[7] Tanha, J., Abdi, Y., Samadi, N., et al. Boosting methods for multi-class imbalanced data classification: An experimental review. Journal of Big Data, 2020, 7: 70.

[8] Dahal, K. R., Dahal, J. N., Banjade, H., et al. Prediction of wine quality using machine learning algorithms. Open Journal of Statistics, 2021, 11(2): 278-289.

[9] Kumar, V., & Minz, S. Feature selection: A literature review. SmartCR, 2014, 4(3): 211-229.

[10] Gupta, Y. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 2018, 125: 305-312.

[11] Tilden III, Marshall. "How to Taste Wine like a Pro." Wine Enthusiast, 16 Mar. 2021, https:// www. winemag. com/2018/07/10/wine-tasting-grids/.

[12] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 2009, 47(4):547-553.

[13] A. Asuncion, D. Newman UCI Machine Learning Repository, University of California, Irvine(2007).

[14] S. Ebeler. Flavor Chemistry --Thirty Years of Progress, Kluwer Academic Publishers,1999:409-422. chapter Linking flavour chemistry to sensory analysis of wine.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, Vol. 16 (2002).

[16] Gerard Biau. Analysis of a Random Forests Model, Journal of Machine Learning Research 13 (2012) 1063-1095.