Received: 3 November 2023 | Revised: 21 February 2024 | Accepted: 28 April 2024 | Published online: 21 May 2024

RESEARCH ARTICLE

Artificial Intelligence and Applications
2024, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA42021999

BON VIEW PUBLISHING

Towards Predicting the Quality of Red Wine Using Novel Machine Learning Methods for Classification, Data Visualization, and Analysis

Jovial Niyogisubizo[1,*], **Jean de Dieu Ninteretse**[2], **Eric Nziyumva**[1], **Marc Nshimiyimana**[3], **Evariste Murwanashyaka**[4] and **Erneste Habiyakare**[5]

[1]*Fujian Provincial Universities Key Laboratory of Industrial Control and Data Analysis, Fujian University of Technology, China*

[2]*Department of Construction and Real Estate, Southeast University, China*

[3]*School of Civil Engineering, Southeast University, China*

[4]*Institute of Rock and Soil Mechanics, University of Chinese Academy of Sciences, China*

[5]*School of Geosciences and Info-Physics, Central South University, China*

Abstract: There is a growing concern among consumers and the wine industry regarding the quality of wine. Traditionally, wine experts determined its quality through tasting, which was time-consuming. Therefore, there is a need to predict wine quality based on specific key features to streamline these tasks. Technological developments like machine learning approaches have replaced human assessments with computational methods. However, some of these methods have faced criticism due to their low accuracy and lack of interpretability for humans. In this paper, a stacking ensemble method is introduced and demonstrates superior predictive performance when compared to other classification techniques like logistic regression, decision trees, gradient boosting, adaptive boosting (AdaBoost), and random forest. This evaluation is based on classification metrics such as accuracy, precision, recall, and $F1$-Score, all under the same conditions. Additionally, outlier detection algorithms were employed to identify exceptional or subpar wines, though their results did not match the accuracy of classification approaches. Lastly, a feature analysis study was conducted to assess the significance of each feature in the model's performance.

Keywords: machine learning, random forest, decision trees, gradient boosting, red wine quality, stacking ensemble

1. Introduction

1.1. Background

Nowadays, emerging technologies are emphasized in all industries due to their usefulness in problem-solving. These introduced technologies are fundamental in increasing production and easing tasks. However, much effort is still required in some areas, such as product quality improvement, labor force training, and increased awareness. Though several studies on wine quality have been conducted, there is still room for improvement. Linear regression is a practical method for making predictions in various fields that is easy to implement. The correlation between the attributes was determined using linear regression. This aided in selecting the most essential quality parameters [1]. Following data analysis, it was discovered that alcohol has the most significant variation among all parameters. The higher the alcohol concentration, the better the wine quality, and the lower the density [2, 3] suggested a technique to assess wine quality by considering three distinct factors. Random forests (RF), support vector machines (SVM), and k-nearest neighbors (KNN) represent well-known classification methods highly regarded for their adaptability in various data scenarios and predictive functions. These algorithms are strongly valued for categorizing complicated patterns, adapting to complex decision boundaries, and influencing collective intelligence [3]. They employed the principal component analysis to choose relevant features and determined the RF algorithm as a classifier that performed standard benchmark techniques in generating favorable results [3]. Three different regression techniques were used, and sensitivity analysis was used to select both the model and the variables; the SVM results performed better than those obtained using neural networks and multiple regression. The proposed model is useful for testing the effects of sensory preferences.

Chen et al. [4] introduced a method to forecast wine quality by analyzing subjective taste reviews from consumers. They employed hierarchical clustering and an association rule algorithm to assess these reviews and make predictions regarding the wine's grade, achieving an accuracy of 85.25%, and [5] utilized the analytical hierarchy process (AHP) to prioritize attributes and then employed different machine learning (ML) classifiers. This approach resulted in an accuracy of 70.33% using the RF classifier and 66.54% with the SVM. These methods were used for product

[*] **Corresponding author:** Jovial Niyogisubizo, Fujian Provincial Universities Key Laboratory of Industrial Control and Data Analysis, Fujian University of Technology, China. Email: jovialniyo93@gmail.com

recommendations [6]. The study employed a user-focused clustering method, utilizing a red wine dataset for analysis. Drawing from their literature review, they assigned relative importance to various attributes and used the Gaussian distribution process to give weight to these features. To evaluate the quality, they conducted assessments within a user preference group.

Although ML techniques have demonstrated remarkable capabilities in handling outliers, missing data, and noisy information, a significant drawback is their tendency to function opaquely, akin to a "black box" approach [7]. Moreover, when decomposing the classification error, it is typical to encounter three prevalent sources of error in learning algorithms associated with a specific target function and the size of the training dataset: bias, variance, and intrinsic target noise [8]. Furthermore, classical decision trees (DT) face the issue of a lack of interpretation for humans, while artificial neural networks and SVM suffer from lower accuracy [7].

These methods benefit wine industry owners in evaluating wine quality, which is a general requirement to get a standardization certificate. The key features of good wine include "pH value, density alcohol, and other acids." To evaluate the quality of wine, one has to consider a physicochemical test, and the second is a sensory test [9]. This paper explores ML methods such as LR, DT, GB, adaptive boosting (AdaBoost), RF, and stacking ensemble to predict red wine quality, as well as data visualizations and analysis. Eventually, a feature analysis study is conducted to explain the contributing features to model performance.

1.2. Literature review

The wine industry mainly uses ML techniques in wine production. While ML models can predict the quality of wine based on physicochemical data, their application is often constrained by limited use and relies on small datasets. This section reviews a highly recommended work in the field that differentiates itself from previous research in several key characteristics. Firstly, the authors utilized a red wine dataset with eleven typical physicochemical traits, which led to a less-explored territory for wine quality forecasting. Secondly, their innovative approach to feature selection by employing ML methods, such as RF and XGBoost, provided a unique way to identify critical characteristics. It is aimed to enhance model performance through thorough standardization and hyperparameter adjustment; it is noted that their research deviates from others through the incorporation of clustering techniques for data preparation. In conclusion, their research introduced novel perspectives and applications related to wine quality prediction that have not been widely explored in the literature. These variations contributed to the structural and theoretical distinctions between their study and the existing work on the subject.

Cortez et al. [10] employed objective hypothesis testing during the certification stage to predict wine tastes. Within a significant dataset, they focused on White Vinho Verde samples from northwest Portugal. Regression analysis was applied, modeling wine preference on a continuous scale from 0 to 10. The authors used an efficient and robust process that simultaneously selects variables and models, guided by sensitivity analysis, incorporating three regression techniques. A different approach [11] evaluated a deep learning algorithm's quality forecasting using two convolution layers. This method enables winemakers to leverage deep learning for operational decision-making. Despite the experiment's limited dataset and feature set, the authors emphasized avoiding machine reliance on selecting helpful characteristics. The study by [12] contributed by developing a new technique that considers various feature selection methods, including recursive feature elimination and principal component measurement, along with nonlinear decision tree-based classifiers for analyzing performance indicators. Their investigation aims to assist wine specialists in understanding crucial elements when selecting high-quality wines.

The study [13] introduced an ML algorithm with a user interface that predicts wine quality by identifying the key factors crucial for determining it. The RF method evaluated wine quality, and KNN enhanced the model's accuracy. The resulting model categorizes wines into Good, Average, or Bad quality ratings. Kumar et al. [14] study focused on determining red wine quality using various characteristics. They used RF, SVM, and Naïve Bayes to gather data from diverse sources. The study compared outcomes between the training and testing datasets calculated several performance measures and predicted the optimal technique among the three based on the learning set outcomes. Shaw et al. [15] conducted a comparative analysis of SVM, RF, and multilayer perceptron classification algorithms for wine quality analysis. The multilayer perceptron algorithm achieved the second-highest accuracy at 78.78%, followed by the SVM algorithm at 57.29% in the comparative analysis. The RF algorithm outperforms others with the highest accuracy of 81.96%.

Bhardwaj et al. analyzed chemical and physicochemical data from New Zealand Pinot Noir wines, consisting of 18 samples with 54 characteristics [16]. Of these, 47 factors were associated with chemical data, while seven were linked to physicochemical data. The study employed four different feature selection techniques, focusing on attributes proven to be significant in at least three methods. Subsequently, seven ML algorithms were trained and tested on an original holdout sample to predict wine quality. Tiwari et al. [17] utilized a mathematical model based on industry and wine specialists' metrics for perceived quality. They validated relevant sensory and chemical concepts using ML methods. The study involved two sets of 18 New Zealand Pinot Noir wines, evaluated by experts for inherent qualities, including overall quality. The authors developed a conceptual and mathematical framework to predict wine quality, employing ML techniques to test these frameworks with a substantial dataset.

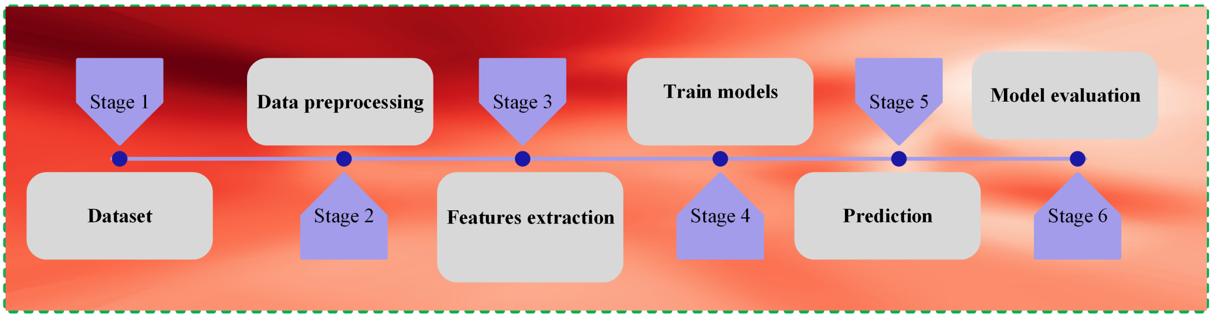
The study by [18] utilized four ML techniques, including RF, stochastic gradient descent, SVM, and LR, to forecast wine quality. Among these, RF demonstrated superior performance with an accuracy of 88%. Subsequently, [19], the red wine dataset was employed and categorized into two classes: good wine and bad wine. In a different study [20], naive Bayes, DT, SVM, and RF were employed to predict wine quality. The analysis highlighted that minimal residual sugar contributes to increased wine quality, suggesting its lesser importance than other factors like alcohol and citric acid.

2. Materials and Methods

2.1. Dataset description

This paper employs the wine dataset for all the procedures that are conducted. The dataset used includes both white and red wines. There are 4898 samples of white wine, 1599 samples of red wine and 12 physiochemical variables in each instance of both types of wine. We have "fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and quality rating." The quality rating is based on a sensory test conducted by at least three sommeliers and is scaled in 11 quality classes ranging from 0 (very bad) to 10 (very excellent). Due to flaws, using both wine collections without

Figure 1
The overall design of the study



preprocessing is impossible. The large amplitude of variable values, such as sulfates (0.3–2) vs. sulfur dioxide (1–72), is one of the significant flaws. Furthermore, some variables have values ranging from 0 to 1. This inconsistency may impact predictions because some variables are more influenced than others. A linear transformation is one method for dealing with such a problem. A linear transformation involves scaling all input values by dividing them by the highest variable value.

2.2. Design of the study

Figure 1 illustrates that the process begins with gathering and analyzing the dataset; following this initial step, a critical data preprocessing stage is executed to ensure the cleanliness and suitability of the dataset for further analysis. The dataset is divided into training and testing subsets according to a predetermined ratio for partitioning. In this study, the methodology divided the dataset into two different subsets, with 80% allocated for training and the remaining 20% reserved for testing. This division ensures that a significant portion of the data is used to train the models and techniques.

It is observed that the reserved testing data provides a robust evaluation of the performance of the models and helps to gauge their effectiveness and generalization to unseen data. This 80–20 split is a common practice in ML to balance training and testing to achieve reliable and reasonable results. The proposed methods are then trained using the training set, which is the foundation for our ML approaches. All ML models and techniques are trained with the training dataset in the prediction stage to make predictions and generate results. Finally, the performance of these methods is thoroughly evaluated based on specific performance metrics, providing a comprehensive valuation of their effectiveness and suitability for the intended task. This systematic approach ensures that the models and methods are thoroughly validated, and their performance is quantitatively assessed to inform decision-making and further modification.

2.3. Proposed method

In this paper, a novel stacking framework made up of a hybrid of LR, DT, GB, AdaBoost, and RF is proposed to predict the quality of red wine. Stacking involves the sequential use of multiple ML models to generate a new feature by aggregating the predictions made by each model. Cross-validation should always accompany model stacking to avoid overfitting models to training data. The advantage of Stacking is that it can combine the capabilities of several high-performing models to create predictions that outperform any single model in the ensemble on classification or regression challenges. As shown in Figure 2, the stacking model's architecture consists of two or more base models, also known as level-0 models, plus a meta-model. The level-1 model, often called a meta-model, combines the forecasts generated by the fundamental models. The meta-model is trained using the basic models' projections with out-of-sample data.

The following steps will help understand the procedure; in the proposed method, various classification models will handle complex and unpredictable features in the raw data by extracting valuable features. The proposed method comprises two layers; initially, the DT, GB, AdaBoost, and RF models predict temporally using a comprehensive training dataset to control each classifier's strengths. The predictions from the first layer are fed into the LR model in the second layer by predicting the quality of red wine through cross-validation, and the approach involves four key stages: feature engineering and selection supported by rationale, dataset partitioning, final prediction, and assessment, as shown in Figure 3.

The data were divided into two parts: training and test sets similar to K-fold cross-validation, and the training data was further divided into K-folds. The process involves training a base model on each of the K-1 parts and using it to make predictions for the Kth part. This process is repeated for all the folds. The base performance of the model is evaluated on the test set using the entire training dataset, which is repeated for different base models. The initial layer in stacking acts as a highly complex, nonlinear feature converter, exhibiting heterogeneous representations for various features. The base classifiers in the

Figure 2
The modeling process of stacking ensemble

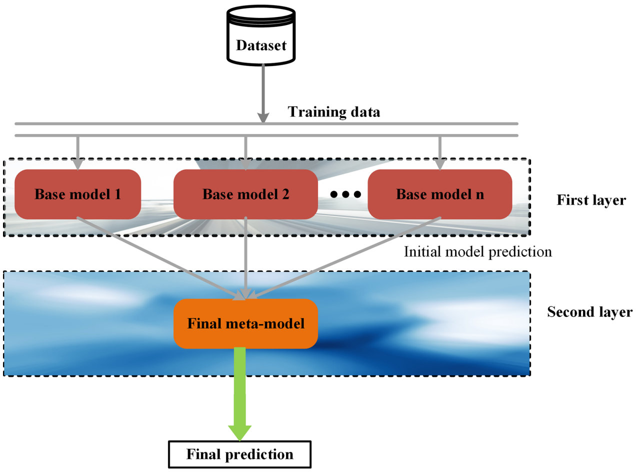
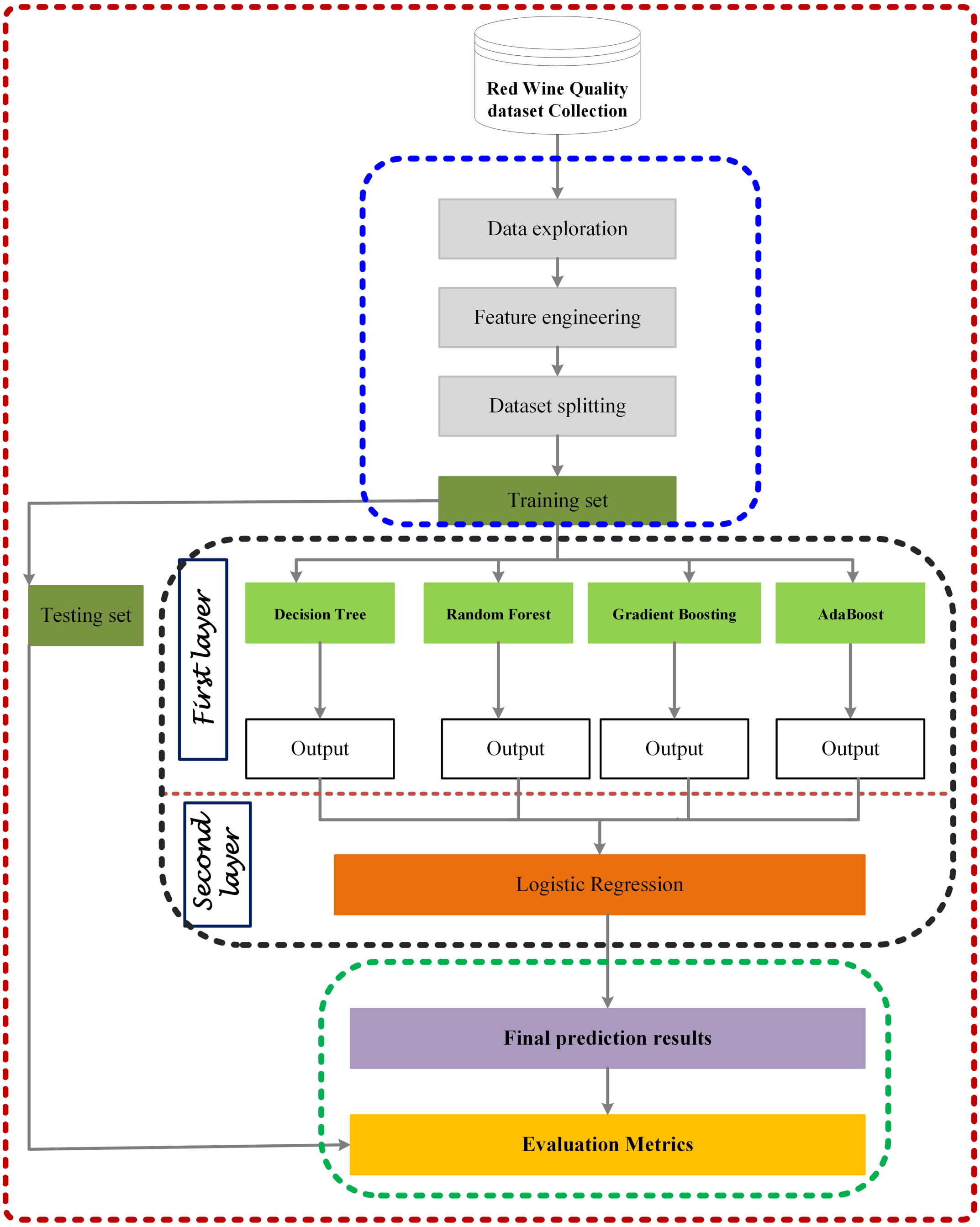


Figure 3
Proposed method



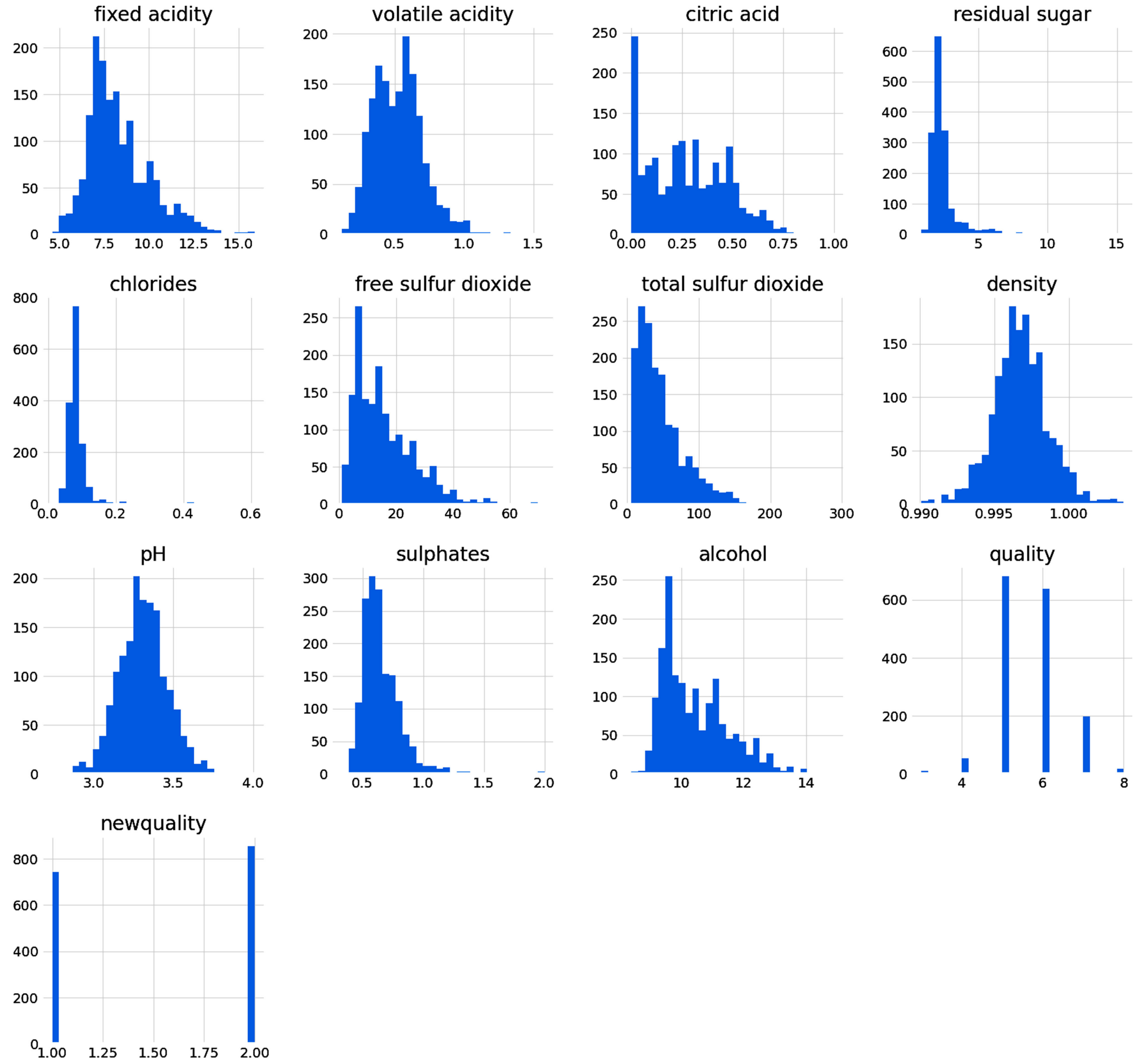
first layer must meet specific criteria to ensure effective feature extraction from raw data, including high accuracy and diversity. This study selects DT, GB, AdaBoost, and RF as the primary models for the first layer due to their varied yet proficient learning approaches. Despite their differing modeling concepts, these models were chosen for their outstanding performance in cross-validation and achieving optimal accuracy.

A second-level model uses predictions from the training set as features and is applied to the test set to generate predictions. In classification scenarios, the outputs from the base models used as inputs for the meta-model are represented as real numbers. This implies that the predictive task involves estimating continuous numerical values, making it well-suited for applications where the target variable exhibits an endless range of possible outcomes. Since the second layer extracts features through complex nonlinear transformation, overly complex classifiers in the output layer become unnecessary. Employing LR in the second layer offers a simple structure with added advantages, and integrating LR into the second layer helps prevent overfitting, contributing further to the model's robustness.

2.4. Data visualization

Based on our data, the visualization process can be used to explain fully the dataset. Visualization also shows the graphical representation of data that can be utilized to get crucial

Figure 4
Data visualization



information. Considering Figure 4, it is clear that the dataset is easily spread on features. Histograms are valuable tools for illustrating the fundamental distributional properties of variables within a dataset. They offer insights into the location of distribution peaks, the symmetry or skewness of the distribution, and the presence of outliers. The Histogram Bins plot is used to visualize all the features to check for skewness and symmetry.

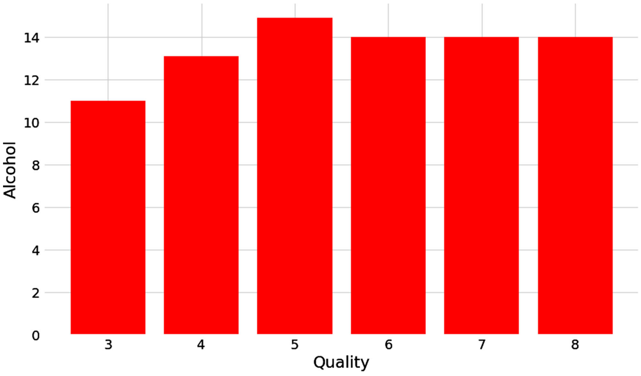
2.5. Data analysis

This section investigates the heart of our analysis by exploring and interpreting the complex data underpinning our red wine quality prediction and accurately uncovering patterns, relationships, and insights concealed within the 12 physiochemical variables with a keen eye on the quality rating. By applying various statistical, ML, and visualization techniques, we aim to differentiate the factors influencing wine quality and craft a robust predictive model to enhance the appreciation of this timeless and delightful beverage. The bivariate analysis is conducted to analyze the key features in the dataset, and based on the quality, a bar plot is constructed, as shown in Figure 5.

A graph is built to show the correlation between the dataset's features. As shown in Figure 6, it is clear that features are fully correlated to one another.

As shown in Figure 7 and the following figures, the two items do not strongly relate to the dependent variable. Therefore, as was done while analyzing the correlation heat map, we have to showcase a correlation plot to check which items are more related to the dependent variable and which are less associated with the dependent variables. From Figure 7, it is clear that the composition of citric acid increases as the quality of the wine increases. On the contrary, chloride's composition also decreases as we increase the quality of the wine. Furthermore, the sulfate level goes higher with the quality of the wine.

Figure 5
Alcohol quality vs count plot



2.6. Preliminary information on the base models utilized in the proposed two-layer ensemble ML method

To achieve the objectives of this study, the most effective approach of ML classification techniques in the first layer was proposed; a supervised ML technique designed to assign a specific class to previously unseen data points was proposed. During the prediction stage, the model with input parameters and target attributes was provided to facilitate its decision-making process. It is essential to highlight that all the practical implementations and experimental procedures were carried out using Python, which is part of its wide ecosystem of libraries and tools. The popular Sklearn, well-known for its ML and predictive analytics versatility, was proposed to facilitate the aligned development and adhere to our analyses' best practices for accuracy and reliability. The strong combination of Python and Sklearn allows us to explore the probable different algorithms and methodologies while qualifying any copyright issues associated with proprietary software. Therefore, this research benefits from the open-source nature of these tools by developing transparency and reproducibility in our findings and a concise overview of the ML classification methods utilized while building our method.

2.6.1. Logistic regression (LR)

LR stands out as a widely employed multivariate statistical technique designed to forecast the outcomes of a binary dependent variable by considering the observed values of a group of independent variables [21]. LR was also designed to deal with categorical response variables representing a binary event instead of relying on continuous parameters [22]. The whole procedure of constructing the LR model is demonstrated by [23]. The outcome projected by a LR model is expressed in a simplified form as a probability of an event, which falls within the range of 0 to 1. The result of the LR function is described as Equation (1) [24]:

$$P_x = \frac{1}{1 + e^{-C_0 + C_1 x}} \tag{1}$$

where P_x the probability of the event happening is, e is the base of the natural logarithm $C_0 + C_1$ are the parameters of the model. The coefficients of the LR model are approximated utilizing the maxi-

Figure 6
Correlation heatmap

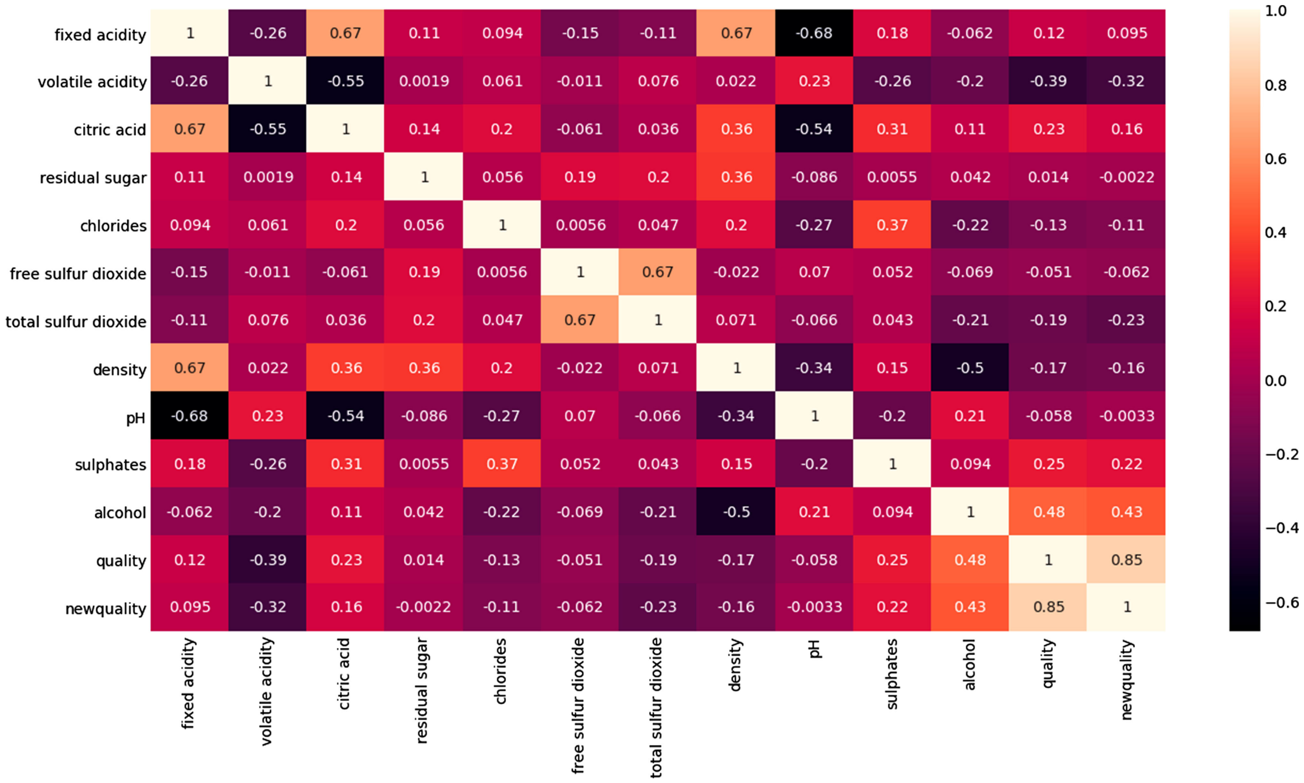
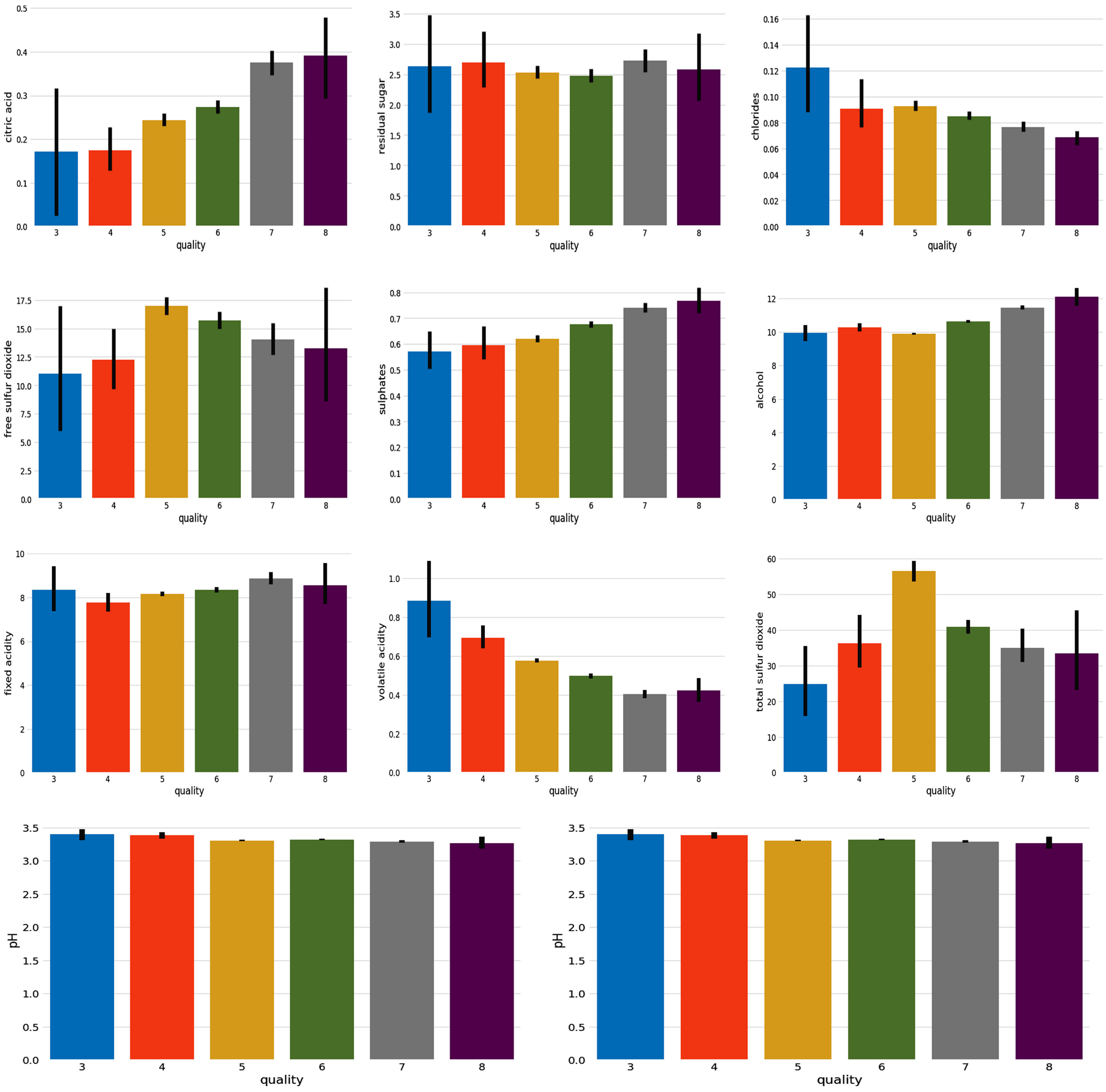


Figure 7
Bivariate analysis



mum likelihood approach, where the coefficients resulting in the most likely outcomes are chosen for the LR model [25].

2.6.2. Random forests (RF)

As outlined in the provided citations, RF is a substantial ensemble of learning techniques designed for classification and regression tasks. RF effectively connects the power of multiple classifiers, combining their outputs during the training phase [26]. This methodology turns around by creating numerous tree-structured classifiers, each dealing with independently and identically distributed random vectors. Based on the input data, each tree contributes a unitary vote for the most prevalent class [27]. RF operates as a collaborative assembly of uncorrelated DTs, effectively functioning as a set of classifiers to refine prediction outcomes. RF adopts the bootstrap aggregation principle to achieve this uncorrelation, creating subsets of training samples with replacement.

By creating the most reliable trees, cross-validation minimizes estimation and out-of-bag errors [28]. Furthermore, RF influences a randomness approach by incorporating all available features. This strategy allows RF to grow many trees, yielding trees with significant variances while mitigating bias-related issues. Subsequently, new observations are classified by aggregating the class assignments of all DTs, resulting in a robust and versatile predictive model [29].

2.6.3. Gradient boosting (GB)

The GB method is a ML algorithm considered to be multiple additive trees, introduced by [30, 31] at Stanford University to solve regression and classification problems. Friedman [30] suggested approach aims to improve the DT model by utilizing stochastic GB. This integration of GB results in a prediction model comprising a collection of distinct, simple DT algorithms. These combined models, characterized by low error rates, enhance the predictive accuracy of any given learning model, ultimately creating a high-performing ensemble learning model [32]. In GB, a method is employed where trees are organized in a sequence to improve the technique's resilience against overlapping class distributions. This approach also involves initializing a customizable loss function through gradient descent. As a result, these actions reduce the system's overall loss function, leading to a notable improvement in model accuracy [33].

2.6.4. Adaptive boosting (AdaBoost)

AdaBoost, as proposed by Freund and Schapire in 1996 [34], is an algorithm created to address regression and classification challenges. Its core concept involves sequentially developing learners and training multiple weak learners on the same dataset to enhance their overall performance. AdaBoost operates in a boosting manner but distinguishes itself by employing short DTs. Moreover, while developing the AdaBoost, every example is weighted in the training set. In the beginning, the weight distribution (vector) is initialized to:

$$weight,\ \boldsymbol{\beta}_i = \frac{1}{m} \tag{2}$$

where $\boldsymbol{\beta}_i$ represents the "*i*" training example weight while m is the number of training examples. Additionally, an initial tree is created where the performance of the trees on every training example is employed. Further, the text emphasizes the evaluation of overall errors and the calculation of subsequent iteration weights. In cases where prediction is challenging, greater significance is assigned, whereas less importance is attributed to easily predictable situations [35].

2.6.5. Decision tree (DT)

DT is a supervised learning method that can be employed to handle both regression (for continuous or discrete variables) and classification(for categorical variables) problems using a computation process that looks like a tree structure formed referring to the set of splitting rules [36]. The DT aims at predicting the target value based on several input variables. The procedure of building a DT is demonstrated in [37]. Every tree consists of many branches and nodes, the splits (set of internal nodes) and the leaves (terminal nodes). The node denotes the variable to be categorized. The branch depicts the values that a node has the potential to suppose.

Moreover, the root node, which includes all data, represents the initial point used to classify all samples in the DT. The samples are classified according to their feature values [38]. Furthermore, every distance between the root node and the leaf node implicates a decision rule demonstrating the relationship between dependent and independent variables, making the solution easily interpreted [39].

2.6.6. Performance evaluation

To evaluate the results, the methods used classification metrics such as accuracy, precision (PR), recall (Rec), and $F1$ Score ($F1$) were used. Below is a brief description of the evaluation metrics used in this study.

- **Accuracy:** denotes the number of instances correctly classified over the sum of all instances.

$$ACC = \frac{TP + T\ N}{TP + TN + FN + FP} \tag{3}$$

- **Precision:** is the proportion of predicted positive instances that are predicted as real positives.

$$\mathrm{PR} = \frac{TP}{TP + FP} \tag{4}$$

- **Recall:** computed as the proportion of actual positive instances that are correctly predicted as positive

$$\mathrm{Rec} = \frac{TP}{TP + FN} \tag{5}$$

- ***F*1-Score:** This is a critical performance metric presenting the harmonic mean of precision, recall and the metric that strikes a balance between the ability to make accurate positive predictions (precision) and the capability to capture all actual positive cases (recall) that provide a comprehensive valuation of a classification model's effectiveness.

$$\mathrm{F1\ Score} = \frac{2 * Rec * PR}{Rec + PR} \tag{6}$$

TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively.

3. Results Analysis and Discussion

To show the effectiveness of our approach, we conducted a comparative analysis, pitting our method against benchmark models and stacking ensemble. This evaluation highlights the superiority of our proposed technique and its potential to perform better than existing methodologies. The comparative study is based on the training and testing results obtained for LR, DT, RF, GB, AdaBoost, and stacking ensemble. As shown in Table 1, the stacking set performs superior for training and testing by outperforming all benchmark models with a training accuracy of 1.0 (100%) and a testing accuracy of 0.85 (85%). The RF and DT models suffered from overfitting issues.

While accuracy is a commonly used metric to assess individual model performance, exclusively relying on it can be deceptive. The model might excel in predicting the majority class but fail to

Table 1
Classification results of different models

Model	Training accuracy	Testing accuracy
LR	0.74	0.78
DT	1.0	0.74
RF	1.0	0.85
GB	0.87	0.80
AdaBoost	0.79	0.76
Stacking	1.0	0.85

Table 2
Performance measures of models for all quality levels

Model	Quality	Precision	Recall	$F1$-score
Logistic regression	High	0.79	0.75	0.77
	Low	0.77	0.81	0.79
	Overall	0.78	0.78	0.78
Decision trees	High	0.72	0.73	0.72
	Low	0.77	0.76	0.76
	Overall	0.74	0.74	0.74
Random forest	High	0.84	0.84	0.84
	Low	0.87	0.87	0.87
	Overall	0.86	0.86	0.86
Gradient boosting	High	0.80	0.78	0.79
	Low	0.80	0.83	0.81
	Overall	0.80	0.80	0.80
Adaptive boosting	High	0.74	0.75	0.74
	Low	0.78	0.78	0.78
	Overall	0.76	0.76	0.76
Stacking ensemble	High	0.83	0.86	0.85
	Low	0.88	0.86	0.87
	Overall	0.86	0.86	0.86

identify the minor one. Various performance indicators like precision, recall, and $F1$-score have addressed this concern. Table 2 shows the performance outcomes of all models for both quality levels. Notably, the stacking ensemble achieved an overall average recall of 0.86 on the testing set, implying the capability to predict nearly 86% of high-quality instances. Additionally, the average precision score of 0.86 on the testing set indicates that around 70% of predictions encompassing both high and low-quality cases are accurate.

Figure 8 shows the significance of features in assessing wine quality. The analysis reveals that chlorides appear to be the most critical factor influencing wine quality, impacting not only the taste but also the texture and structure of the wine. Total sulfur dioxide seems to be the second most critical feature and exhibits some correlation with chlorides. Notably, our findings highlight "free dioxide" as the least influential variable in our analysis. This metric measures the usage of Sulfur dioxide (SO_2) throughout the winemaking process. They primarily aim to prevent oxidation and inhibit microbial growth [40].

4. Use of Outlier Detection Algorithms to Detect the Few Excellent or Poor Wines

Considering the dataset's quality, it is evident that the classes are ordered and exhibit an imbalance. For instance, there is a notable disparity in the distribution of normal wines compared to excellent or poor ones, indicating an uneven representation of different quality levels in the dataset. Outlier detection algorithms can identify a small number of exceptional or subpar wines. The model was trained using the given 1599 instances of wine qualities, and then, the trained model was used to predict the quality of wine using SVM from scratch. SVM are well-established supervised learning models with associated algorithms. These SVM models are primarily employed to analyze data in classification and regression tasks. They excel in delineating decision boundaries and have proven effective in various applications, making them a versatile tool in ML and predictive analytics [41]. In this study, it is required to classify the given dataset as a good-quality wine and a bad-quality wine using the two attributes. The SVM model is likely to discriminate accurately between the two classes. As shown in Figure 9, the SVM built here gives a more accurate result than

Figure 8
Feature importance

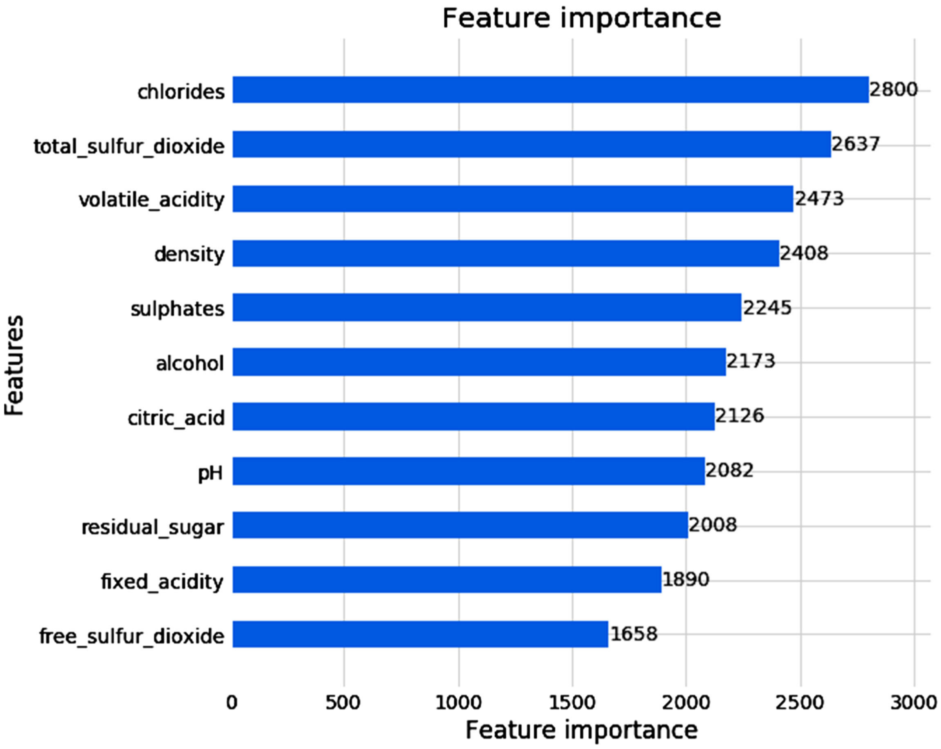
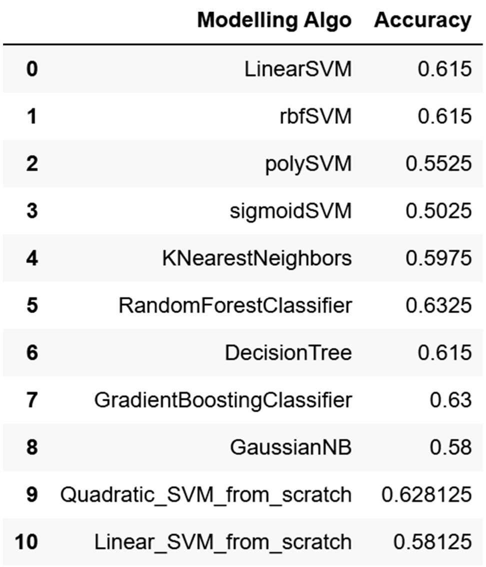


Figure 9
Performance results of outlier detection algorithms



	Modelling Algo	Accuracy
0	LinearSVM	0.615
1	rbfSVM	0.615
2	polySVM	0.5525
3	sigmoidSVM	0.5025
4	KNearestNeighbors	0.5975
5	RandomForestClassifier	0.6325
6	DecisionTree	0.615
7	GradientBoostingClassifier	0.63
8	GaussianNB	0.58
9	Quadratic_SVM_from_scratch	0.628125
10	Linear_SVM_from_scratch	0.58125

linear SVM. However, compared to the result in the section above, the stacking ensemble outperforms the Outlier detection algorithms.

5. Conclusion

This paper aimed to predict the quality of red wine using ML algorithms for classification, data visualizations, and analysis. Data analysis revealed that the features are highly correlated to each other. In most features, the composition of citric acid goes higher as we go higher in the quality of wine. On the contrary, chloride's composition also decreases as we increase the quality of the wine. Furthermore, the sulfate level goes higher with the quality of the wine. As stated, "This analysis provided a comprehensive understanding of the significance of attributes in predicting quality, highlighting the time and cost savings achieved compared to traditional methods." A comprehensive study applied diverse classification ML algorithms to predict wine quality. The results revealed that the stacking ensemble underscores the efficacy and superiority of the stacking ensemble approach in this specific context, showing its potential for more accurate and reliable wine quality predictions.

Conversely, outlier detection algorithms were employed to identify exceptional or subpar wines. However, instead of improving the performance accuracy, the results diminished over time. Furthermore, a feature analysis study was conducted to evaluate the importance of input variables on model performance. In the future, deep learning and other ML algorithms will be proposed to compare the best-performing models. This analysis will assist industries in predicting the quality of various types of wines based on specific attributes and producing good products in the future.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

The dataset utilized to support the findings of this study originates from research conducted by Cortez et al. [10] and is accessible for download from the UCI machine learning repository at: https://archive.ics.uci.edu/ml/datasets/wine+quality

The technical details, code, and data can be accessed by visiting the link: https://github.com/jovialniyo93/red_wine_quality_prediction

Author Contribution Statement

Jovial Niyogisubizo: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Project administration. **Jean de Dieu Ninteretse:** Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Eric Nziyumva:** Validation, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Supervision. **Marc Nshimiyimana:** Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Evariste Murwanashyaka:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Erneste Habiyakare:** Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision.

References

[1] Palmer, J., & Chen, B. (2018). Wineinformatics: Regression on the grade and price of wines through their sensory attributes. *Fermentation*, *4*(4), 84. https://doi.org/10.3390/fermentation4040084

[2] Mahima, Gupta, U., Patidar, Y., Agarwal, A., & Singh, K. P. (2020). Wine quality analysis using machine learning algorithms. In *Micro-Electronics and Telecommunication Engineering*, 11–18. https://doi.org/10.1007/978-981-15-2329-8_2

[3] Aich, S., Sain, M., & Yoon, J.-H. (2019). Prediction of different types of wine using nonlinear and probabilistic classifiers. In A. N. Krishna, K. C. Srikantaiah & C. Naveena (Eds.), *Integrated intelligent computing, communication and security* (pp. 11–19). Springer. https://doi.org/10.1007/978-981-10-8797-4_2

[4] Chen, B., Rhodes, C., Crawford, A., & Hambuchen, L. (2014). Wineinformatics: Applying data mining on wine sensory reviews processed by the computational wine wheel. In *2014 IEEE International Conference on Data Mining Workshop*, 142–149. https://doi.org/10.1109/ICDMW.2014.149

[5] Thakkar, K., Shah, J., Prabhakar, R., Narayan, A., & Joshi, A. (2016). AHP and machine learning techniques for wine recommendation. *International Journal of Computer Science and Information Technologies*, *7*(5), 2349–2352.

[6] Reddy, Y. S., & Govindarajulu, P. (2017). An efficient user centric clustering approach for product recommendation based on majority voting: A case study on wine data set. *International Journal of Computer Science and Network Security*, *17*(10), 103.

[7] Hashmienejad, S. H. A., & Hasheminejad, S. M. H. (2017). Traffic accident severity prediction using a novel multi-objective genetic algorithm. *International Journal of Crashworthiness*, *22*(4), 425–440. https://doi.org/10.1080/13588265.2016.1275431

[8] Kotsiantis, S., & Pintelas, P. (2004). Combining bagging and boosting. *International Journal of Computational Intelligence*, *1*(4), 324–333.

[9] Ebeler, S. E. (1999). Linking flavor chemistry to sensory analysis of wine. In R. Teranishi, E. L. Wick & I. Hornstein (Eds.), *Flavor chemistry* (pp. 409–421). Springer. https://doi.org/10.1007/978-1-4615-4693-1_35

[10] Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Using data mining for wine quality assessment. In G. Goos & J. Hartmanis (Eds.), *Lecture notes in computer science*. Springer. https://doi.org/10.1007/978-3-642-04747-3_8

[11] Agrawal, G., & Kang, D. K. (2018). Wine quality classification with multilayer perceptron. *International Journal of Internet, Broadcasting and Communication*, *10*(2), 25–30. https://doi.org/10.7236/IJIBC.2018.10.2.5

[12] Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. (2018). 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,* 139–143. https://doi.org/10.23919/ICACT.2018.8323674

[13] Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science*, *125*, 305–312. https://doi.org/10.1016/j.procs.2017.12.041

[14] Kumar, S., Agrawal, K., & Mandan, N. (2020). Red wine quality prediction using machine learning techniques. In *2020 International Conference on Computer Communication and Informatics,* 1–6. https://doi.org/10.1109/ICCCI48352.2020.9104095

[15] Shaw, B., Suman, A. K., & Chakraborty, B. (2020). Wine quality analysis using machine learning. In J. K. Mandal & D. Bhattacharya (Eds.), *Emerging technology in modelling and graphics* (pp. 239–247). Springer. https://doi.org/10.1007/978-981-13-7403-6_23

[16] Bhardwaj, P., Tiwari, P., Olejar Jr, K., Parr, W., & Kulasiri, D. (2022). A machine learning application in wine quality prediction. *Machine Learning with Applications*, *8*, 100261. https://doi.org/10.1016/j.mlwa.2022.100261

[17] Tiwari, P., Bhardwaj, P., Somin, S., Parr, W. V., Harrison, R., & Kulasiri, D. (2022). Understanding quality of Pinot Noir wine: Can modelling and machine learning pave the way? *Foods*, *11*(19), 3072. https://doi.org/10.3390/foods11193072

[18] Pawar, D., Mahajan, A., & Bhoithe, S. (2019). Wine quality prediction using machine learning algorithms. *International Journal of Computer Applications Technology and Research*, *8*(9), 385–388.

[19] Dua, D., & Graff, C. (2017). *UCI machine learning repository*. USA: University of California.

[20] Dahal, K., Dahal, J., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics*, *11*(2), 278–289. https://doi.org/10.4236/ojs.2021.112015

[21] Xu, C., Xu, X., Dai, F., Wu, Z., He, H., Shi, F., . . . , & Xu, S. (2013). Application of an incomplete landslide inventory, logistic regression model and its validation for landslide susceptibility mapping related to the May 12, 2008 Wenchuan earthquake of China. *Natural Hazards*, *68*(2), 883–900. https://doi.org/10.1007/s11069-013-0661-7

[22] Zhang, W., & Goh, A. T. C. (2013). Multivariate adaptive regression splines for analysis of geotechnical engineering systems. *Computers and Geotechnics*, *48*, 82–95. https://doi.org/10.1016/j.compgeo.2012.09.016

[23] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Linear regression. In G. James, D. Witten, T. Hastie & R. Tibshirani (Eds.), *An introduction to statistical learning* (pp. 59–126). Springer. https://doi.org/10.1007/978-1-0716-1418-1_3

[24] Bui, D. T., Lofman, O., Revhaug, I., & Dick, O. (2011). Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural Hazards*, *59*(3), 1413–1444. https://doi.org/10.1007/s11069-011-9844-2

[25] Shirzadi, A., Saro, L., Joo, O. H., & Chapi, K. (2012). A GIS-based logistic regression model in rock-fall susceptibility mapping along a mountainous road: Salavat Abad case study, Kurdistan, Iran. *Natural Hazards*, *64*(2), 1639–1656. https://doi.org/10.1007/s11069-012-0321-3

[26] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

[27] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

[28] Niyogisubizo, J., Liao, L., Zou, F., Han, G., Nziyumva, E., Li, B., & Lin, Y. (2023). Predicting traffic crash severity using hybrid of balanced bagging classification and light gradient boosting machine. *Intelligent Data Analysis*, *27*(1), 79–101. https://doi.org/10.3233/IDA-216398

[29] Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36. https://doi.org/10.1016/j.aap.2017.08.008

[30] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

[31] Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9), 1365–1381. https://doi.org/10.1002/sim.1501

[32] Zheng, Z., Lu, P., & Lantz, B. (2018). Commercial truck crash injury severity analysis using gradient boosting data mining model. *Journal of Safety Research*, *65*, 115–124. https://doi.org/10.1016/j.jsr.2018.03.002

[33] Mousa, S. R., Bakhit, P. R., Osman, O. A., & Ishak, S. (2018). A comparative analysis of tree-based ensemble methods for detecting imminent lane change maneuvers in connected vehicle environments. *Transportation Research Record*, *2672*(42), 268–279. https://doi.org/10.1177/0361198118780204

[34] Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, 148–156.

[35] Niyogisubizo, J., Liao, L., Sun, Q., Nziyumva, E., Wang, Y., Luo, L., . . . , & Murwanashyaka, E. (2023). Predicting crash injury severity in smart cities: A novel computational approach with wide and deep learning model. *International Journal of Intelligent Transportation Systems Research*, *21*(1), 240–258. https://doi.org/10.1007/s13177-023-00351-7

[36] Sabah, M., Talebkeikhah, M., Agin, F., Talebkeikhah, F., & Hasheminasab, E. (2019). Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: A case study from Marun oil field. *Journal of Petroleum Science and Engineering*, *177*, 236–249. https://doi.org/10.1016/j.petrol.2019.02.045

[37] Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, *97*(3), 322–336. https://doi.org/10.1016/j.rse.2005.05.008

[38] Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. UK: Cambridge University Press.

[39] Salimi, A., Faradonbeh, R. S., Monjezi, M., & Moormann, C. (2018). TBM performance estimation using a classification and regression tree (CART) technique. *Bulletin of Engineering Geology and the Environment*, *77*(1), 429–440.

[40] Monro, T. M., Moore, R. L., Nguyen, M. C., Ebendorff-Heidepriem, H., Skouroumounis, G. K., Elsey, G. M., & Taylor, D. K. (2012). Sensing free sulfur dioxide in wine. *Sensors*, *12*(8), 10759–10773.

[41] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

How to Cite: Niyogisubizo, J., Ninteretse, J. D., Nziyumva, E., Nshimiyimana, M., Murwanashyaka, E., & Habiyakare, E. (2024). Towards Predicting the Quality of Red Wine Using Novel Machine Learning Methods for Classification, Data Visualization, and Analysis. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA42021999