DOI: https://doi.org/10.52756/ijerr.2024.v45spl.003 *Int. J. Exp. Res. Rev.*, Vol. 45: 25-40 (2024)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

ISSN 2455-4855

A Novel Data Handling Technique for Wine Quality Analysis using ML Techniques

Onima Tigga*, Jaya Pal and Debjani Mustafi

Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi-834001, Jharkhand, India

E-mail/Orcid Id:

OT, otirkey@bitmesra.ac.in, https://orchid.org/0009-0006-3959-9294; *JP*, jayapal@bitmesra.ac.in, https://orchid.org/0000-0001-7843-283X; *DM*, debjani.mustafi@bitmesra.ac.in, https://orchid.org/0000-0002-7055-5031

Article History:

Received: 11th Apr., 2024
Accepted: 21st Oct., 2024
Published: 30th Nov., 2024

Keywords:

Decision tree, multi-layer perceptron, multiclass, random forest, yeo-johnson transformation

How to cite this Article:
Onima Tigga, Jaya Pal and Debjani Mustafi (2024). A Novel Data handling Technique for Wine Quality Analysis using ML Techniques. International Journal of Experimental Research and Review, 43, 25-40.
DOI:
https://doi.org/10.52756/ijerr.2024.v45spl.003

Abstract: In this era, wine is a regularly redeemed beverage, and industries are seeing increased sales due to product quality certification. This research aims to identify key wine characteristics that contribute to significant outcomes through the application of machine learning classification techniques, specifically Random Forest (RF), Decision Tree (DT) and Multi-Layer Perceptron (MLP), using white and red wine datasets sourced from the UCI Machine Learning repository. This research aims to develop a multiclass classification model using machine learning (ML) to accurately assess the quality of a balanced wine dataset comprising both white and red wines. The dataset is balanced by random oversampling to avoid biases in ML techniques for the majority class obtained by the imbalanced multiclass dataset (IMD). Furthermore, we apply a Yeo-Jhonson transformation (YJT) to the datasets to reduce skewness. We validated the ML algorithm's result using a 10-fold cross-validation approach and found that RF yielded the highest overall accuracy of 93.14%, within a range of 75% to 94%. We have observed that the proposed approach for balanced white wine dataset accuracy is 93.14% using RF, 90.83% using DT, and 75.49% using MLP. Similarly, for the balanced red wine dataset, accuracy is 89.36% using RF, 85.36% using DT, and 78.00% using MLP. The proposed approach improves accuracy by RF 23%, DT 30%, and MLP 21% for the white wine dataset. Similarly, accuracy by RF remained the same, DT 10%, and MLP 22% is improved in the red wine dataset. Additionally, the proposed approach's RF, DT, and MLP yield mean squared error (MSE) values of 0.080, 0.151, and 0.443 for the white wine dataset and 0.143, 0.221, and 0.396 for the red wine dataset. We also observed that the RF accuracy for the proposed technique is the highest among all specified classifiers for white and red wine datasets, respectively.

Introduction

In the wine industry, consumers consume the wine on its quality. Since red wine and white wine are multi-class datasets, variable selection methods are done directly instead of decomposing many two-class problems. For classification, selecting variables from a multiclass dataset is sometimes easier than a two-class dataset. There are various features for wine quality predictions, but none of them are relevant (Dahal et al., 2021). So, our goal is to find relevant features to obtain better results. For Feature selection, we used Karl Pearson Coefficient correlation and further balanced the datasets by Random Oversampling (ROS) because of its simplicity and easiness of implementation (Geethanjali et al., 2021). ROS does not necessitate complex algorithms or assumptions about the data's underlying distribution. Any dataset with a class imbalance can use it, and it doesn't necessitate prior knowledge about the dataset. Even though the random under-sampling (RUS) technique has good results, it gets

*Corresponding Author: otirkey@bitmesra.ac.in worse when less training data is available due to the potentially valuable data that may be necessary for construction purposes. The white wine and red wine data are not normally distributed, so, the Yeo-Johnson transformation must be used to normalize and make the necessary transformation. A Yeo-Johnson power transformation operates similarly to the Box-Cox transformation (BCT) (Siddiqui and Pak, 2021). The Yeo-Johnson transformation (YJT) essentially increases the values of low-variance data and decreases the values of high-variance data to create a more evenly distributed dataset. What distinguishes the Yeo-Johnson power transformation is its capability to convert data that includes negative numbers. The advantage of using the Yeo Jhonson transformation is that it preserves many of the features of the original skewed distribution, including location, range and mode, while still providing data that can be analyzed using methods that assume normality (Weisberg, 2001; Nwakuyal and Anyaogu, 2022). Data transformation ensures that our data is clean and ready for processing by the ML algorithm (Tan et al., 2022; Danrui et al., 2023). Instead, the final prediction opts for SMOTE ENN + ETC, forecasting the multiclass red wine quality dataset without converting it into a binary data set. The SMOTE ENN effectively balances the dataset, leading to enhanced accuracy in machine learning classifiers and improved performance across various metrics. They identified SMOTE ENN + ETC as the optimal model for the final red wine quality data prediction. Sampling represents a method to address the problems of imbalance class in real-world datasets. Authors have developed various sampling approaches in response to this issue. SMOTE variants represent the predominant methods for oversampling in the context of binary imbalanced datasets. Consequently, the majority of research transforms multiclass imbalanced datasets into binary balanced datasets and identifies the optimal prediction model (Uma et al., 2023). While the research work primarily focused on binary classification problems, there is a growing interest in health informatics to address the challenges associated with multi-class classification in imbalanced datasets. Multi-class imbalance problems present intricate challenges, necessitating a careful strategy to create synthetic data while preserving the relationships among the various classes. This review paper seeks to analyze oversampling methods specifically designed for medical and other datasets characterized by multi-class imbalances (Yang et al., 2024). According to Benjamin et al. (2022), recommended the implementation of new performance measurement metrics and algorithms to achieve more refined scores and enhance comparison. By implementing this approach, wineries can more accurately forecast the quality of various wine varieties, thereby improving future product outcomes.

We have used three ML techniques, RF, DT and MLP for the classification model and obtaining relevant features by using white wine and red wine datasets. The performance measures such as recall, precision, accuracy, and F1-score for comparison of the machine learning techniques have been used.

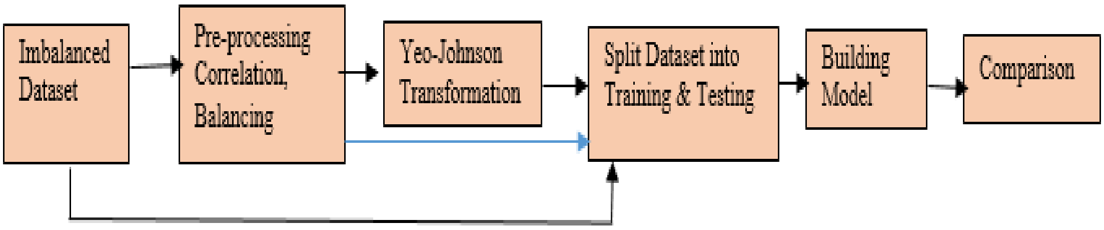

Imbalanced Dataset
Pre-processing Correlation, Balancing
Yeo-Johnson Transformation
Split Dataset into Training & Testing
Building Model
Comparison

Figure 1. Abstract workflow diagram of the research paper.

Finally, we found that the RF technique gives better results than DT and MLP techniques for wine datasets (white and red). The abstract workflow diagram of the research paper is depicted in Figure 1 as follows:

The primary aim of this research effort is to create a multiclass classification model built on ML to identify the quality of the balanced wine dataset (white & red) obtained from the UCI ML digital repository with the best accuracy. On using the pre-processed balanced wine datasets (white & red), the qualities such as 'quality 3', 'quality 4', 'quality 5', 'quality 6', 'quality 7', 'quality 8', and 'quality 9' have been identified with selected features using correlation in the datasets. The ML algorithms RF, DT, and MLP are utilized to create a classification model for determining the quality of a wine dataset. This is achieved using a 10-fold cross-validation technique (Kaliappan et al., 2023).To ensure the dependability of the model, the YJT is used to make normal distribution identical by reducing the skewness of the dataset's characteristics.

We conduct feature selection through correlation analysis, using a correlation factor of 0.6 to identify relevant features such as fixed acidity, volatile acidity, citric acid, chlorides, pH, sulphates, and alcohol from both the white and red wine datasets. Since the datasets are imbalanced, for balancing both the datasets, random oversampling and transforming the datasets white wine and red wine applied the Yeo Johnson Transformation technique. After that, the dataset is split into training and testing datasets. Hence, ML techniques were applied to training datasets, and performances were calculated.

Researchers have considered SMOTE, a technique that creates synthetic samples for minority classes to balance class distribution in imbalanced datasets. This approach enhances classifier performance metrics and mitigates bias against the majority class. SMOTE exhibits marginally lower accuracy when applied to multi-class datasets in comparison to its performance on binary datasets. In response to this issue, we have put forward a novel strategy aimed at improving the performance of a multiclass dataset.

Therefore, the contributions of this research work are defined below:

# Come up with a new way to pre-process data that includes choosing features using correlation analysis, even out multi-class datasets of white and red wine from the UCI machine repository by random over-sampling, and using Yeo-Johnson transformation techniques to normalize and turn multi-class data into a normal distribution.

# We have conducted various pre-processing comparisons: without correlation, with correlation, with correlation + ROS, with correlation + YJT, and with correlation + ROS + YJT.

# Implement machine learning methods (RF, DT, and MLP) on the filtered datasets to get improved outcomes.

# We compute performance measures such as accuracy, precision, recall, F1-score and MSE.

# Assess the newly suggested method for multi-class data pre-processing against the aforementioned machine learning methods to determine which yields superior performance while also outlining a systematic strategy for performance enhancement.

This study aims to evaluate three machine learning classifiers to see if the classification result can be enhanced by improving the data pre-processing steps.

The arrangement of this research work is as follows: Explanation of the related work is incorporated in Section 2. The exploratory data analysis and methodologies are discussed in Section 3. Section 4 illustrates the Results and Analysis. The research work concludes in Section 5. Section 6 reveals future work.

Related work

The various ML models such as ridge regression, SVM, gradient boosting (XGB) regression, and multi-layer ANN have been used to predict the quality of wine with various performance metrics. After comparison, it was observed that the gradient boosting regression model is the best with the R, MSE, and mean absolute percentage error (MAPE) of 0.6057, 0.3741, and 0.0873 respectively (Dahal et al., 2021). The technique has been proposed by Dhaliwal et al. (2022) for preprocessing the red and white wine dataset. They found that the size of the dataset has been reduced from 13 attributes to 9 attributes without any loss of performance by using ML techniques RF classifiers, Decision Trees, KNN and ANN classifiers. On the comparison of performance analysis based on accuracy and RMSE values, Random Forest performs better than the other two classifiers for predicting wine quality. According to Kumar et al. (2020), RF, SVM and NB techniques have been used to predict the quality of the wine. They have used performance metrics such as F1-score, precision, recall, accuracy, specificity, and misclassification error. They achieved that SVM gives the greatest result through an accuracy of 67.25. Geetanjali et al. (2021) used algorithms LR, DT, RF and Extra tree classifiers to detect a few excellent or poor wine qualities. To convert the categorical values into numerical values, one hot encoder is used. They obtained accuracies of RF and Extra Tree Classifier 88.19% and 88.79%, respectively. Khilari et al. (2021) focused on the analysis of the red wine dataset. LR, DT, RF, SVM, Ada Boost, and Gradient Boosting Machine Learning classifiers are used. The Random Forest algorithm outperforms other classifiers and scores an accuracy of 92%. According to Gawale (2022), ML and hybrid techniques for the prediction of wine quality. The evaluation criteria included accuracy, recall, precision, and F1-score. The comparative analysis was done among machine learning algorithms DT, RF, XGBoost, and Hybrid model implemented using DT and Random Forest. SMOTE is applied to optimize the model's performance. The outliers and null values were removed to enrich the model's performance. The accuracy of the RF algorithm is 85.57%, the DT algorithm is 79.25%, and the extreme Gradient Boost is 78.07%. Finally, a Hybrid Model of all these ML techniques is implemented which achieved an overall accuracy of 77.71%. They found that choosing the right features and balancing the data in the classification algorithms will enhance the performance of the model. One of the reasons for this variance is the fact that data for red and white wine datasets were collected in this research and implemented with ML classifiers.

As stated by Chaudhari et al. (2023), ML algorithms Logistic Regression (LR), SVC, RF, K-nearest neighbour Classifier (KNN), and DT are used and found that RF gave the best result. In addition, through the feature selection process, they used RF for feature selection and found that the quality of the wine is affected more by the alcohol content. Further, they used SMOTE for oversampling the minority class. Also, artificial intelligence for wine quality prediction is used (Patkar and Balaganesh, 2021). It implies that unpredictable acidity indicates a spoiler and can cause an upsetting scent. It is presumed that when the wine is of excellent quality. Additionally, when the liquor content in wine is higher, the wine is of generally excellent quality. Burigo et al. (2023) used imbalanced data for classifying the quality of the wine by exploring oversampling and under-sampling techniques to enhance the standard of the model in the industry. They found that the performance of RF for multiclass problems was not improved using the oversampling method for imbalanced data, whereas when using SMOTE, the performance of RF improved. Benjamin et al. (2022) recommends implementing new performance measurement metrics and algorithms to achieve more refined scores and facilitate better comparisons. By implementing this approach, wineries can more accurately forecast the quality of various wine varieties, subsequently improving future products. Uma et al. (2023) conducted a study that aims to predict the multiclass red wine quality dataset without converting it into a binary format, ultimately selecting SMOTE ENN + ETC for the final forecast. The SMOTE ENN effectively balances the dataset, leading to enhanced accuracy in machine learning classifiers and improved performance across various metrics. The optimal model identified for the final prediction of red wine quality data is SMOTE ENN combined with ETC. Sampling represents a method to report the problem of class imbalance in real-world datasets. The authors have developed various sampling approaches in response to this issue. SMOTE variants represent the predominant oversampling methods used for addressing binary imbalanced datasets. Consequently, the majority of research transforms multiclass imbalanced datasets into binary balanced datasets and identifies the optimal prediction model. Yang et al. (2024) examined binary classification issues, highlighting a growing interest in health informatics to tackle the challenges associated with multi-class classification in imbalanced datasets. Multi-class imbalance problems present intricate challenges, necessitating a careful strategy to produce synthetic data while preserving the interrelationships among the various classes. This review paper seeks to analyze oversampling methods specifically designed for medical and various datasets characterized by multi-class imbalances. Zhang et al. (2024) indicate that the C4.5 algorithm demonstrates superior performance across various medical datasets in their initial investigation into algorithmic applications within the field of medicine. The researchers identified eight established machine learning algorithms characterized by low user engagement and robust family representation to serve as foundational algorithms. The team assessed the algorithm's prediction accuracy, execution speed, and memory usage. Creating a decision tree and stepwise regression model enhances comprehension of the algorithm's relevance to medical datasets. All of the cross-verification results show that the algorithmic applicability prediction models are more than 75% accurate, which proves that the knowledge is valid and useful.

Sindayigaya and Dey (2022) report the use of a variety of machine-learning algorithms. The algorithms perform various functions, such as data mining, image processing, and predictive analytics, among others. The primary advantage of using machine learning lies in an algorithm's capacity to execute tasks following its training on data autonomously. This research introduced a multi-class classification method for technology assessment (TE) using patent information (Lee et al., 2020). The description of TE illustrates it as a conversion of technology quality into its present value. It also enables effective research and development via the application of intellectual property rights.

In this research work, the proposed multiclass classification models use the balanced white and red wine datasets, which have seven classes and seven extracted features. To avoid the biases of ML techniques due to an imbalanced multiclass dataset in the majority group, each white wine type has 2197 instances (2197 * 7) with seven features and each red wine type has 613 instances (613 * 7) with seven features, respectively.

Exploratory data analysis & methodologies

The workflow diagram is represented in Figure 2 for the proposed multiclass classification model.

In this research work, we have used four approaches which are shown in Figure 2. The data preprocessing stage employed correlation for feature selection and utilized random oversampling (ROS) to achieve dataset balance. We have completed the optimization of the model's evaluation, which has revealed its more efficient performance. We employ oversampling to achieve a balance among the classes within the datasets, as the performance analysis results show no improvement. We also subject the datasets to a Yeo-Johnson transformation (YJT) to mitigate skewness.

DOI: https://doi.org/10.52756/ijerr.2024.v45spl.003

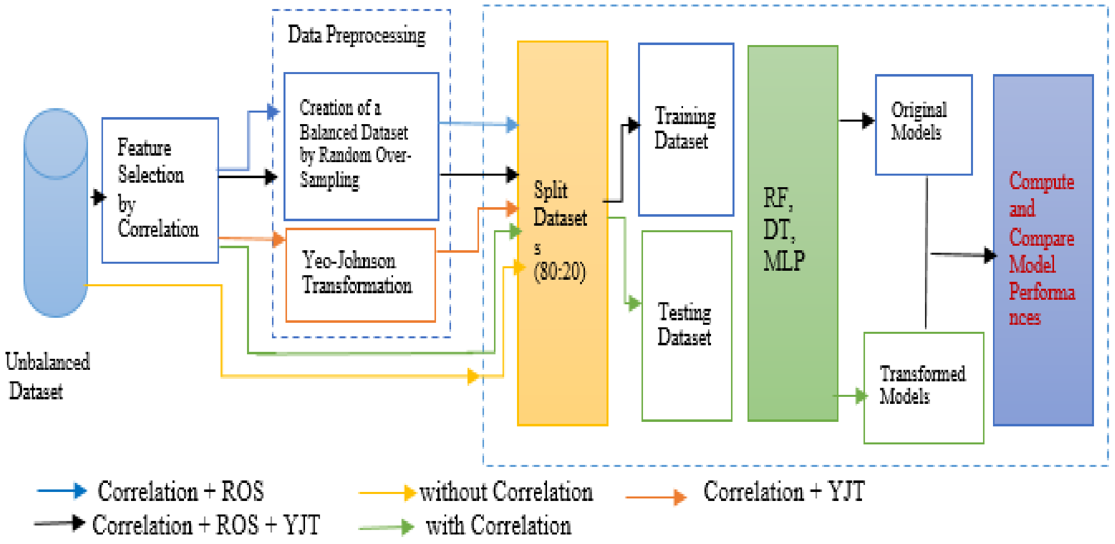


Figure 2. Workflow diagram of the research work.

In this research work, we have used four approaches which are shown in Figure 2. The data preprocessing stage employed correlation for feature selection and utilized random oversampling (ROS) to achieve dataset balance. We have completed the optimization of the model's evaluation, which has revealed its more efficient performance. We employ oversampling to achieve a balance among the classes within the datasets, as the performance analysis results show no improvement. We also subject the datasets to a Yeo-Johnson transformation (YJT) to mitigate skewness.

Description of Approaches

# Imbalanced dataset with Correlation (IDC): This approach is represented by the green solid line(→). Here, features are selected from imbalanced wine datasets using correlation. And then the model is designed using ML.

# Imbalanced dataset with Correlation and YJT (IDCY): In this approach, after feature selection from the imbalanced wine datasets Yeo-Jhonson transformation is used to reduce the skewness of the features. This approach is represented by the orange dashed line (→).

# Balanced dataset with Correlation and Random Oversampling (BDCR): Here, after feature selection from a given dataset, random oversampling is used for balancing the dataset to the design model. This approach is represented by the blue dashed line (→).

# Balanced dataset with Correlation, ROS, and YJT (BDCRY): This is our proposed approach. Here, we have used Yeo Jhonson transformation in the balanced dataset to reduce the skewness of the features. In this approach accuracy of RF is 93.14%, whereas in COY accuracy of RF is 67% in the white wine dataset. Similarly, the accuracy of RF is 89%, whereas in COY accuracy of RF is 67% in the red wine dataset. This approach is represented by the black solid line (→).

The performance measurements such as accuracy, precision, recall, F1-score, and MSE are calculated for the above four approaches. This aims of this research work is to enhance the data preprocessing steps for multi-class datasets using Yeo-Jhonson transformation.

Data

This research work used two datasets sourced from the UCI online repository (https://archive.ics.uci.edu/ml/datasets/Wine+Quality). This research employs a dataset of white wines that includes 4,898 instances, each characterized by 11 attributes. Additionally, a single quality attribute, which ranges from 3 to 9, delineates the wine's quality levels. The defining characteristics include fixed acidity, volatile acidity, citric acid, chlorides, pH, sulphates and alcohol content. We represent the quality as the output class and apply the same process to it. This study uses a Red Wine dataset that includes 1599 data instances, each characterized by 11 attributes and a single quality attribute. Table 1 provides a complete outline of the dataset's detailed descriptions.

Table 1. Datasets Description.

Name of Datasets	Instances	Attributes	Classes
White Wine	4898	11	7
Red wine	1599	11	6

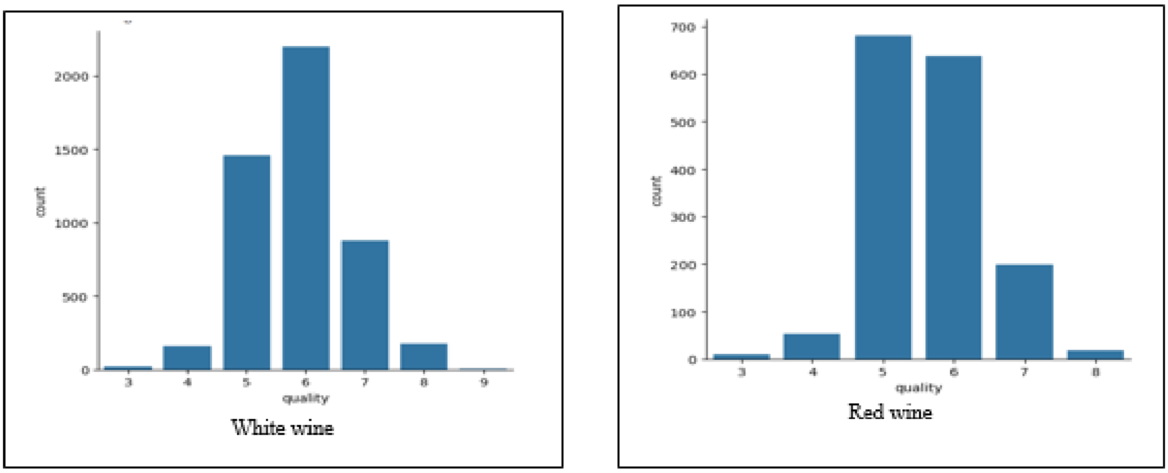
Data Visualization & Preprocessing

Visualization of imbalanced dataset

The performance of classifiers is improved by

DOI: https://doi.org/10.52756/ijerr.2024.v45spl.003

preprocessing strategies. The data distribution of the imbalanced dataset is shown in Figure 3.



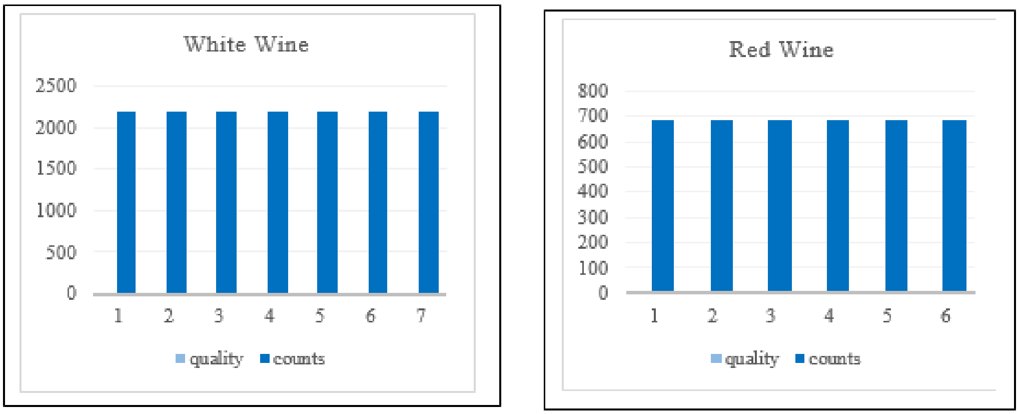
(a) (b)

Figure 3(a) & (b). Number of classes Before Random Over-Sampling.

The class-wise count of the white wine dataset is 20 for quality 3, 163 for quality 4, 1457 for quality 5, 2197 for quality 6, 880 for quality 7, 175 for quality 8, and 5 for quality 9 respectively. Similarly, the class-wise count of the red wine dataset is 10 for quality 3, 53 for quality 4, 681 for quality 5, 638 for quality 6, 199 for quality 7 and 18 for quality 8, respectively. No null values are found in both datasets.

Visualization of balanced dataset

The classification of the imbalanced dataset for predicting the model using ML techniques is an interesting task to make each class of equal size. Since an imbalanced dataset gives poor predictive results during training a model, to overcome this problem we have oversampled the minority class. Thus the data distribution of the balanced dataset is shown in Figure 4 for white and red wine datasets, respectively.



(a) (b)

Figure 4(a &b). Number of classes after Random Over-Sampling.

Correlation

In this research work, Pearson correlation coefficient r is used as shown in equation (1) :

$$r = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{k=1}^{m}(y_i - \bar{y})^2}} \qquad (1)$$

Where $\bar{x}$ = mean of independent data, $\bar{y}$ = mean of dependent data, m = number of data points, x_i = individual independent variable, y_i = individual dependent variable. Correlation (r) represents the linear relation between the independent variables and dependent variables. To overcome the risk of overfitting simple correlation coefficient (r) is used for nonlinear preprocessing (Siddiqui and Pak, 2021).

Yeo-Johnson Power Transformations

Several attempts to define transformation family variables that include negative values have been suggested. One possibility is to deliberate transformations of the form $(y+\gamma)^{\lambda}$, γ as suitably large to ensure that

DOI: https://doi.org/10.52756/ijerr.2024.v45spl.003

$y+\gamma$ is strictly positive. In principle, (γ,λ) could be estimated simultaneously, although in practice evaluations are highly variable. Transformations play a central role in regression analysis. A new family of distribution without restrictions on y and all good characteristics of the Box-Cox power family is used by Yeo and Johnson.

The transformations obtained by Yeo-Jhonson are represented by the Equation 2 as follows:

$$\psi(\lambda,y)=\begin{cases} \left((y+1)^{\lambda}-1\right)/\lambda & \text{if } \lambda\neq 0,\, y\geq 0 \\ \log(y+1) & \text{if } \lambda=0,\, y=0 \\ -\left[\left((-y+1)^{2-\lambda}-1\right)\right]/(2-y) & \text{if } \lambda\neq 2,\, y<0 \\ -\log(-y+1) & \text{If } \lambda=2,\, y<0 \end{cases} \qquad (2)$$

An exponential and monotonic Power transformation for each feature looks like more normal in distribution during certain features represent large skewness. It helps ML models to handle large skewed data. After getting the value of λ automatically, the transformed results are computed by using equation (2). It is done to get the transformed feature into a unit-variance normal distribution and zero-mean (Weisberg, 2001; Nwakuyal and Anyaogu, 2022).

Table 2. Lambda values calculated by YJT for white wine.

Sl. No.	Columns	Yeo-Johnson-lambdas(white wine)	Yeo-Johnson-lambdas(red wine)
0.	fixed acidity	-0.334514	-0.874636
1.	volatile acidity	-4.102262	-0.780009
2.	Citric acid	-1.240379	-0.325105
3.	Chlorides	-29.673929	-18.380616
4.	pH	-2.925490	-0.357025
5.	sulphates	-3.045223	-4.031033
6.	alcohol	-1.699687	-3.717320

In Table 2, it has been observed that Lambda values calculated for the given y input variable for the given datasets white wine as the selected attributes are such that fixed acidity is -0.334514, volatile acidity is -4.102262, citric acid is -1.240379, chlorides is -29.673929, pH is -2.925490, sulphates is -3.045223, and alcohol is -1.699687 respectively. Similarly, Lambda values calculated for the given y input variable for the given red wine dataset as the selected attributes are such that fixed acidity is -0.874636, volatile acidity is -0.780009, citric acid is -0.325105, chlorides is -18.380616, pH is -0.357025, sulphates is -4.031033, and alcohol is -3.717320 respectively. The distribution plot of YJT for the balanced and preprocessed white wine dataset for the selected attributes such as fixed acidity, volatile acidity, citric acid, pH, sulphates, and alcohol are shown in Figure 5.

ML Models

Training and Testing Dataset

The white wine dataset contains 4897 records and, after splitting into training data and testing data, in (80:20) ratio. The training data has 3917 records and 980 records in testing data. Similarly, the red wine dataset contains 1599 records and after splitting it into training data and testing data, in (80:20) ratio. The training data has 1279 records and 320 records in testing data.

Machine Learning Algorithms

Decision Tree

Constructing the DT, a top-to-down and greedy algorithm is used through the given sets to test every attribute at each node in the tree. Which is the basic idea of Iterative Dichotomiser ID3. The entropy is used in a set of data to calculate the amount of uncertainty. The Entropy's value always lies between 0 and 1 (Zhao et al., 2023; Tigga et al., 2022; Kaliappan et al., 2023).

Let P_1, P_2, P_m are probabilities where summation of probabilities is $\sum_{k=1}^{m} P_k = 1$, The Entropy is shown in equation 3.

$$H(P_1,P_2,\ldots\ldots\ldots,P_m)=\sum_{k=1}^{m} P_k \log(1/P_k) \qquad (3)$$

And the Gain is defined as in equation 4. For k=1 to m,

$$Gain(D,S)=H(D)-\sum_{k=1}^{m}(D_k)H(D_k) \qquad (4)$$

Random Forest (RF)

A RF classifier, which is one type of ensemble

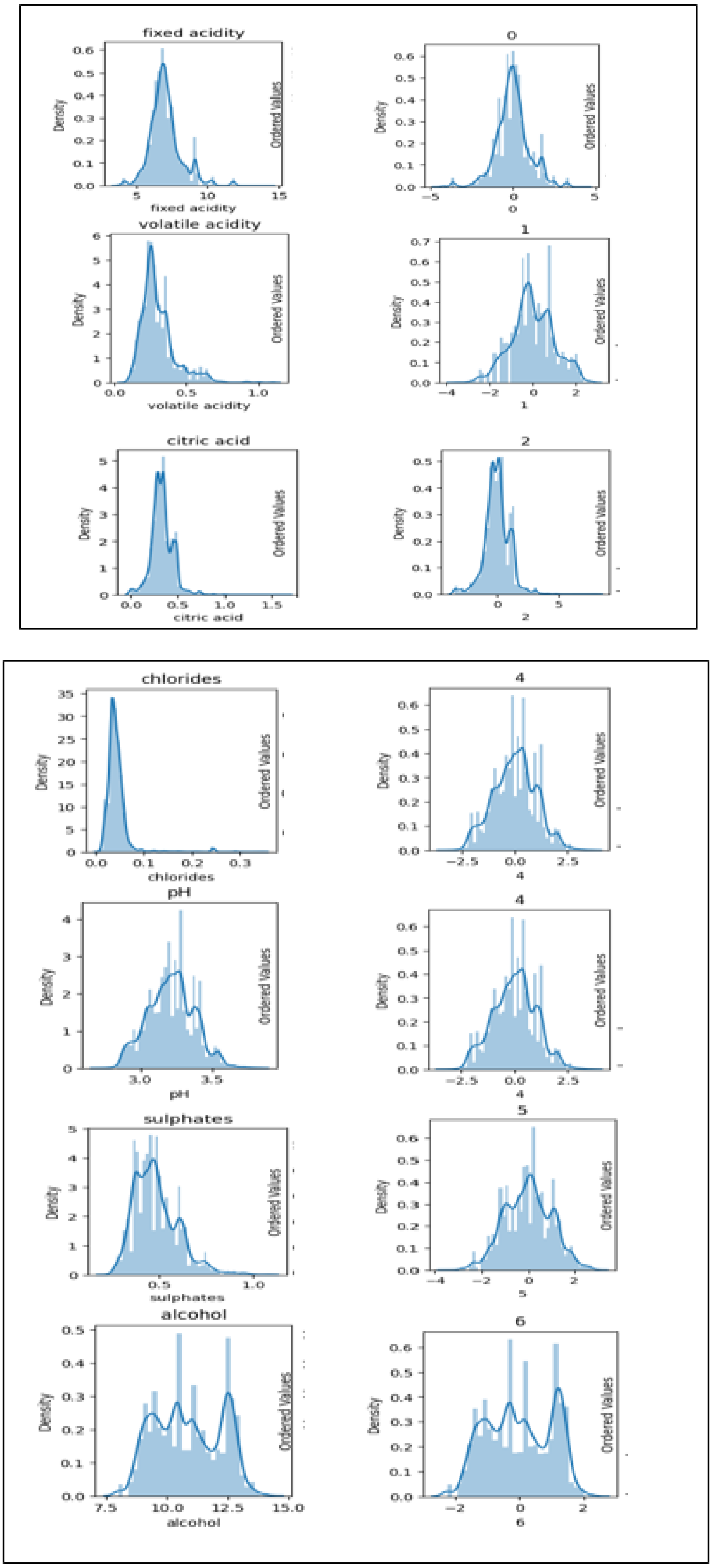


Figure 5. Distribution plot of YJT for the balanced and preprocessed white wine dataset. (a) fixed acidity (b) volatile acidity. (c) citric acid. (d) pH. (e) sulphates. (f) alcohol.

classifier (Zhao et al., 2023), in which the three important components, node size, cardinality of trees, and data points of RF, are to be set before training. After that, a divider is used to solve regression and classification problems as follows:

1. Each DT is constructed with the data points taken from the training datasets known as bootstrap sample and then such DTs are combined with RF. In the training sample, one-third data is used as test data, called as sample from the bag and then it will be returned after that.

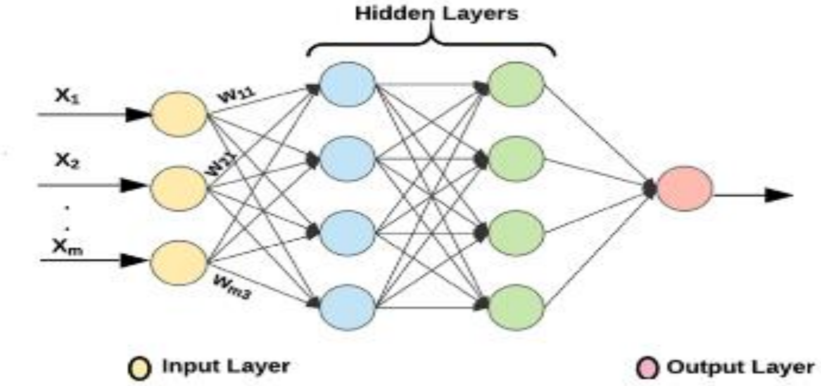


Figure 6. The schematic of an MLP, with three hidden layers and one output layer with a sigmoidal function at each node w1, w2 and w3 are the weights (Tigga et al., 2023).

2. The trees for each decision tree will be weighted and the majority voted on the most common categorical features will give the predicted category (Cao et al., 2022).

The RF technique gives the result based on the prediction of decision trees by using the output of various trees. The accuracy of the result is increased by increasing the number of trees (Kaliappan et al., 2023).

Multilayer Perceptron (MLP)

In the Multilayer Perceptron Classifier, the input layer introduces input values for the network. Classification features are performed by a hidden layer but do not represent the result, whereas the output layer works like a hidden layer and displays the results. A MLP consists input layer, 1 or more hidden layers, and 1 output layer, which is depicted in Figure 6. Each layer contains several neurons. The input layer represents the attributes of each training tuple. These inputs are passed from the input layer with weights to hidden layers. Since there are more hidden layers, the output commencing from one hidden layer will be the input of another hidden layer. Finally, the outputs of the hidden layer are nourished as inputs to the neurons of the output layer, which gives predicted results for given tuples (Zhao et al., 2023; Tigga et al., 2023; Handball et al., 2020).

In this paper, we have used the sigmoidal function as represented by Equation 5, for making a decision.

$$f(x) = \frac{1}{1+e^{-\alpha I}} \qquad (5)$$

Where α is the learning rate and I am the net activation of the neuron.

Results and analysis

Evaluation Metrics

To validate how well the machine learning algorithms perform for classification methods and compare with RF, DT and MLP techniques for (100%) of data for the white wine datasets, We have employed mean squared error, accuracy, recall, precision, and F1-score as our evaluative metrics. A confusion matrix serves as a tool for estimating the efficacy of a developed model when applied to test data. We ground the execution in the authentic values of the test data (Khilari et al., 2021).

- **Accuracy:** $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$

  (6)

- **Precision:** Precision is the ratio of true positives and predicted total positives (Cao et al., 2022).

$$\Pr ecision = \frac{TP}{TP+FP} \qquad (7)$$

- **Recall:** A Recall is defined as follows in equation (8),

$$\operatorname{Re} call = \frac{TP}{TP+FN} \qquad (8)$$

- **F1-score:** The F1-score is the harmonic mean of the precision and recall (Carpita and Goli, 2023; Saito et al., 2021).

$$F1 = \frac{2 * \Pr ecision * \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \qquad (9)$$

- **Mean Squared Error (MSE):** MSE is calculated by using the equation given in (10), where n is the number of tuples (Tigga et al., 2023; Kaliappan et al., 2023).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2 \qquad (10)$$

DOI: https://doi.org/10.52756/ijerr.2024.v45spl.003

Performance Analysis of ML Algorithms

For imbalanced wine dataset

In this research work, we have used three ML Techniques: RF, DT, and MLP for the white wine dataset. After feature selection using correlation for imbalanced datasets, the performance analysis of the above ML techniques is depicted in Table 4. The overall accuracy is 0.68, precision is 0.85 for quality 8, recall is 0.77, F1-score is 0.71 for quality 6 for the white wine dataset, whereas overall accuracy is 0.64, precision, recall & F1-score are 0.71, 0.72, & 0.72 respectively for quality 5 in red wine dataset by using RF. Similarly, for the imbalanced white wine dataset, overall accuracy is 0.61, and precision, recall, and F1-score are 0.67, 0.66, and 0.66, respectively, for quality 6 using DT. Whereas for the red wine dataset, overall accuracy is 0.57, precision, recall, and F1-score are 0.70, 0.62, and 0.66, respectively for quality 5 using DT. Similarly, for the imbalanced white wine dataset overall accuracy is 0.52, precision for quality 6 is 0.53, recall is 0.65, and F1-score is 0.58, respectively, for quality 5 using MLP. Whereas for the red wine dataset, overall accuracy is 0.55, precision, recall, and F1-score are 0.61, 0.74 and 0.67 for quality 5 using MLP.

Here, in Table 5, on using YJT for imbalanced datasets (white wine & red wine), the overall accuracy is 0.67, precision is 0.79 for quality 8, and recall & F1-score are 0.77 and 0.71 for quality 6, respectively concerning RF. Similarly, precision, recall, and F1-score are 0.71, 0.76 and 0.74, respectively, concerning RF on the red wine dataset. Similarly, for the imbalanced white wine dataset, overall accuracy is 0.60, precision is 0.67 for quality 6, recall is 0.65 for quality 5, and F1-score is 0.65 for quality 6 using DT. Whereas for the red wine dataset, overall accuracy is 0.56, precision, recall, and F1-score are 0.69, 0.62, and 0.66 for quality 5 using DT. Similarly, for the imbalanced white wine dataset overall accuracy is 0.54, precision for quality 8 is 0.67, and recall and F1-score are 0.73 & 0.62, respectively, for quality 6 using MLP. Whereas for the red wine dataset, overall accuracy is 0.56, precision is 0.67 for quality 4 and recall and F1-score are 0.71 & 0.67, respectively, for quality 5 using MLP.

Table 4. Performance Analysis of ML Algorithms for the qualities of white wine & red wine using IDC approach.

		White Wine			Red Wine		
Performance Metric ML Techniques	Quality	Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	3	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.60	0.20	0.30	0.00	0.00	0.00
	5	0.68	0.68	0.68	**0.71**	**0.72**	**0.72**
	6	0.66	**0.77**	**0.71**	0.61	0.67	0.64
	7	0.70	0.60	0.65	0.53	0.50	0.51
	8	**0.85**	0.33	0.48	0.00	0.00	0.00
	9						
DT	3	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.20	0.20	0.20	0.00	0.00	0.00
	5	0.63	0.65	0.64	**0.70**	**0.62**	**0.66**
	6	**0.67**	**0.66**	**0.66**	0.57	0.61	0.59
	7	0.57	0.55	0.56	0.43	0.48	0.45
	8	0.34	0.39	0.37	0.00	0.00	0.00
	9						
MLP	3	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00	0.00	0.00
	5	0.52	**0.65**	**0.58**	**0.61**	**0.74**	**0.67**
	6	**0.53**	0.63	0.57	0.52	0.55	0.53
	7	0.49	0.24	0.32	0.38	0.21	0.27
	8	0.00	0.00	0.00	0.00	0.00	0.00
	9						

DOI: https://doi.org/10.52756/ijerr.2024.v45spl.003

Table 5. Performance Analysis for the qualities of unbalanced wine datasets using the IDCY approach.

		White Wine			Red Wine		
Performance Metric → ML Techniques ↓	Quality	Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	3	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.64	0.23	0.34	0.00	0.00	0.00
	5	0.67	0.67	0.67	**0.71**	**0.76**	**0.74**
	6	0.66	**0.77**	**0.71**	0.63	0.70	0.66
	7	0.73	0.59	0.66	0.63	0.52	0.57
	8	**0.79**	0.33	0.47	0.00	0.00	0.00
	9						
DT	3	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.16	0.20	0.18	0.00	0.00	0.00
	5	0.63	**0.65**	0.64	**0.69**	**0.62**	**0.66**
	6	**0.67**	0.64	**0.65**	0.55	0.58	0.57
	7	0.56	0.56	0.56	0.43	0.48	0.45
	8	0.35	0.39	0.37	0.20	0.20	0.20
	9						
MLP	3	0.00	0.00	0.00	0.00	0.00	0.00
	4	0.40	0.07	0.11	**0.67**	0.00	0.31
	5	0.59	0.49	0.53	0.63	**0.71**	**0.67**
	6	0.54	**0.73**	**0.62**	0.52	0.53	0.52
	7	0.48	0.34	0.40	0.41	0.36	0.38
	8	**0.67**	0.06	0.11	0.00	0.00	0.00
	9						

For balanced wine dataset

The performance analysis of ML techniques for a balanced dataset using random oversampling (ROS) with selected features is shown in Table 6. Here, for the RF technique, overall accuracy is 0.70, precision 0.86 for quality 8, recall 0.80 & F1-score 0.74 for quality 6, respectively, on the white wine dataset. Whereas accuracy is 0.89, precision, recall & F1-score are 0.99, 1.00, & 0.99, respectively, for the red wine dataset. Similarly, for the balanced white wine dataset, overall accuracy is 0.60, precision is 0.66 for quality 6, recall is 0.64 for quality 5 & quality 6, and F1-score is 0.65 for quality 6 using DT. Whereas for the red wine dataset, overall accuracy is 0.87, precision is 0.99 for quality 8, recall is 1.00 for quality 3, 4, & 8, and F1-score is 0.99 for quality 3 & quality 8, respectively using DT. Similarly, for the balanced white wine dataset, overall accuracy is 0.52, precision for quality 6 is 0.53 and recall and F1-score are 0.65 & 0.58, respectively, for quality 5 using MLP. Whereas for the red wine dataset, overall accuracy is 0.40, precision, recall, and F1-score are 0.73, 1.00 & 0.85, respectively for quality 3 using MLP.

The performance analysis for the balanced dataset (white wine & red wine) with correlation & YJT is shown in Table 7.

1.00 for quality 3, respectively in the red wine dataset using RF.

Similarly, for the balanced white wine dataset, overall accuracy is 0.91, precision is 1.00 for quality 3 & 9, recall is 1.00 for quality 3, 4 & quality 8, 9, and F1-score is 1.00 for quality 3 & 9 using DT. Whereas for the red wine dataset, overall accuracy is 0.86, precision is 0.99 for quality 3, recall is 1.00 for quality 3, 4 & 9 and F1-score is 1.00 for quality 3, respectively, using DT. Similarly, for the balanced white wine dataset overall accuracy is 0.76, precision for quality 9 is 1.00, recall is 1.00 for quality 3 & 9 and F1-score is 1.00 for quality 9 using MLP. Whereas for the red wine dataset, overall accuracy is 0.78, precision is 0.97 for quality 3, recall is 1.00 for quality 3 & 8, and f1-score is 0.99 for quality 3 & 8 using MLP.

The graphical representation of performance analysis of ML techniques for balanced white wine and red wine datasets is depicted in figure 8.

Here, the overall accuracy is 0.93, precision 1.00, recall 1.00 & F1-score 1.00 for quality 3, quality 8, and quality 9, respectively, for the white wine dataset. Whereas the overall accuracy is 0.89, precision is 1.00 for quality 3, recall is 1.00 for quality 3, 4 & 8 and F1-score is

Table 6. Performance Analysis of ML Algorithms for the qualities of white wine & red wine using the BDCR approach.

		White Wine			Red Wine		
Performance Metric → ↓ ML Techniques	**Quality**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
RF	3	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**
	4	0.55	0.20	0.29	**1.00**	**1.00**	**1.00**
	5	0.68	0.68	0.68	0.72	0.79	0.75
	6	0.68	**0.80**	**0.74**	0.77	0.65	0.70
	7	0.79	0.62	0.69	0.92	0.98	0.95
	8	**0.86**	0.36	0.51	0.99	**1.00**	0.99
	9						
DT	3	0.00	0.00	0.00	0.99	**1.00**	**0.99**
	4	0.19	0.20	0.20	0.93	**1.00**	0.96
	5	0.64	**0.64**	0.64	0.77	0.67	0.71
	6	**0.66**	**0.64**	**0.65**	0.71	0.63	0.67
	7	0.56	0.57	0.57	0.82	0.97	0.89
	8	0.35	0.39	0.37	**0.99**	**1.00**	**0.99**
	9						
MLP	3	0.00	0.00	0.00	**0.73**	**1.00**	**0.85**
	4	0.00	0.00	0.00	0.51	0.47	0.49
	5	0.52	**0.65**	**0.58**	0.44	0.53	0.48
	6	**0.53**	0.63	0.57	0.35	0.15	0.21
	7	0.48	0.24	0.32	0.28	0.34	0.31
	8	0.00	0.00	0.00	0.53	0.49	0.51
	9						

Table 7. Performance Analysis of ML Algorithms for the qualities of white wine & red wine using the Proposed BDCRY approach.

		White Wine			Red Wine		
Performance Metrics → ↓ ML Techniques	**Quality**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
RF	3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	4	0.98	**1.00**	0.99	0.96	**1.00**	0.98
	5	0.86	0.83	0.85	0.74	0.79	0.77
	6	0.81	0.73	0.77	0.79	0.65	0.71
	7	0.86	0.96	0.91	0.90	0.98	0.94
	8	**1.00**	**1.00**	**1.00**	0.99	**1.00**	**1.00**
	9	**1.00**	**1.00**	**1.00**	-	-	-
DT	3	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**
	4	0.94	**1.00**	0.97	0.91	**1.00**	0.95
	5	0.85	0.82	0.83	0.75	0.65	0.69
	6	0.82	0.66	0.73	0.69	0.62	0.65
	7	0.84	0.94	0.89	0.84	0.97	0.90
	8	0.95	**1.00**	0.97	0.97	**1.00**	0.99
	9	**1.00**	**1.00**	**1.00**	-	-	-
MLP	3	0.98	**1.00**	0.99	**0.97**	**1.00**	**0.99**
	4	0.81	0.94	0.87	0.80	0.97	0.88
	5	0.66	0.51	0.57	0.62	0.65	0.63
	6	0.38	0.36	0.37	0.55	0.31	0.40
	7	0.58	0.56	0.57	0.70	0.86	0.77
	8	0.82	0.94	0.88	0.80	**1.00**	**0.99**
	9	**1.00**	**1.00**	**1.00**	-	-	-

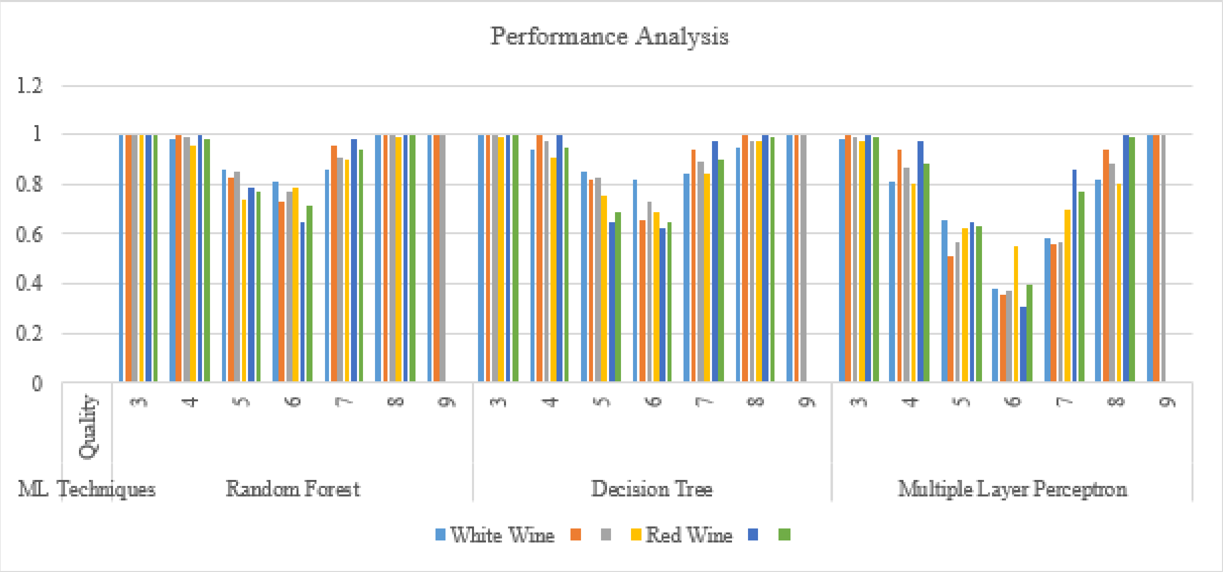


Figure 7. Performance Analysis of proposed ML Algorithms for the qualities of white wine & red wine.

We also found that the accuracy of the RF for the proposed approach is the highest among all specified classifiers for the dataset white wine. Similarly, the accuracy of the RF for the proposed technique is the highest among all specified classifiers for the dataset red wine.

Comparison Analysis

The comparison analysis of four approaches IDC, IDCY, BDCR, and **BDCRY** (proposed) is depicted in Table 9.

Table 9. Accuracy Improvement.

	White Wine			Red Wine		
ML Techniques → **Approaches ↓**	RF	DT	MLP	RF	DT	MLP
IDC	0.68	0.61	0.52	0.64	0.57	0.55
IDCY	0.67	0.60	0.54	0.67	0.56	0.56
BDCR	0.70	0.60	0.52	0.89	**0.87**	0.49
BDCRY(Proposed)	**0.93**	**0.91**	**0.75**	**0.89**	0.86	**0.78**

In Table 9, using the proposed approach (BDCRY), accuracy by RF 23%, DT 30%, and MLP 21% is improved for the white wine dataset. Similarly, in the red wine dataset, accuracy by RF remained the same, DT -0.01% and MLP 22% improved.

The performance comparison with the existing result given by the different researchers and our proposed approach is shown in Table 10 as depicted in the following table:

As shown in table 10, by comparing the results of the previous researchers with those achieved in the current research study, it can be said that the current research study has achieved far better values than the previous ones. Khilari et al. (2021) have reported 90% accuracy for LR, 90% accuracy for DT, 92% for RF, and 90% for SVM for red wine applied for binary class, which are somewhat lesser than the accuracy of the current research. M. S. Choudhari et al. (2023) have reported 79.15% accuracy for LR, 87% accuracy for SVC, 87% for KNN, 86% for DT, 91% for RF, 87% for GBC and 81% for MLP for red wine applied for binary class, which are somewhat lesser that accuracy of the current research. Gawale et al. (2022) have reported 79% accuracy for DT, 86% accuracy for RF, 78% for XGB and 78% for the Hybrid Model for red wine & white wine applied for the binary class, which is lesser than the accuracy of the current research.

The demonstration of MSE for RF, DT, and MLP for the IDC approach is larger than the Proposed technique, with 0.432, 0.694 and 0.641, respectively on the white wine dataset and 0.446, 0.690, and 0.503, respectively, on the red wine dataset is shown in Table 11.

Similarly, for the white wine dataset, the MSE of RF, DT, and MLP for the IDCY approach is larger than the proposed approach with 0.450, 0.720 and 0.643, respectively on the white wine dataset and 0.391, 0.681 and 0.534 respectively on red wine dataset. Similarly, for the white wine dataset, the MSE of RF, DT and MLP for

DOI: https://doi.org/10.52756/ijerr.2024.v45spl.003

Table 10. Performance Comparison with the existing methods with 80 : 20 split.

Authors/year	Dataset	ML Algorithms	Precision	Recall	F1-score	Accuracy in Percentage
Korade et al., 2021	Red Wine	LR	0.77	0.80	0.77	0.80
		DT	0.82	0.82	0.82	0.82
		RF	0.86	0.86	0.86	0.86
Khilari et al., 2021	Red Wine	LR	0.88	0.90	0.89	0.90
		DT	0.92	0.90	0.90	0.90
		RF	0.92	0.92	0.92	0.92
		SVM	0.88	0.90	0.88	0.90
Gawale, 2022	Red Wine & white wine	DT	0.79	0.79	0.79	0.79
		RF	0.86	0.86	0.86	0.86
		XGB	0.78	0.78	0.78	0.78
		Hybrid Model	0.78	0.78	0.78	0.78
Chaudhari et al., 2023	Red wine	LR	0.76	0.82	0.79	79.15
		SVC	0.82	0.88	0.79	87
		KNN	0.80	0.86	0.96	0.87
		DT	0.83	0.89	0.86	0.86
		RF	0.87	0.93	0.91	0.91
		GBC	0.83	0.93	0.87	0.87
		ANN	-	-	-	81.00
Proposed Method	White wine & Red wine (multiclass)	RF	0.93	0.93	0.93	**0.93**
		DT	0.90	0.91	0.90	**0.90**
		MLP	0.74	0.75	0.75	**0.75**
		RF	0.89	0.89	0.89	**0.89**
		DT	0.85	0.86	0.85	**0.86**
		MLP	0.76	0.78	0.76	**0.78**

Table 11. Mean Squared Error for RF, DT and MLP.

ML Techniques → Approaches ↓	White Wine			Red Wine		
	RF	DT	MLP	RF	DT	MLP
IDC	0.432	0.694	0.641	0.446	0.690	0.503
IDCY	0.450	0.720	0.643	0.391	0.681	0.534
BDCR	0.417	0.717	0.641	**0.122**	**0.211**	0.503
BDCRY(Proposed)	**0.080**	**0.151**	**0.443**	0.143	0.221	**0.369**

the BDCR approach is larger than the proposed technique BDCRY with 0.417, 0.717, and 0.641, respectively, on the white wine dataset. The same approach is less than the proposed technique with 0.122, 0.211, and greater value of 0.503 respectively, on the red wine dataset. MSE is obtained of RF, DT and MLP for the proposed approach with 0.080, 0.151 and 0.443, respectively, on the white wine dataset and 0.143, 0.221 and 0.396, respectively, on the red wine dataset.

Conclusion

The quality of wine directly impacts how much of it is consumed. In this research work, we used the white wine dataset, which had 7 reduced attributes and seven classes with 4898 data, as well as the red wine (1599) dataset. The performance metrics analysis such as accuracy, precision, recall, f1-score, and MSE are calculated for ML techniques RF, DT, and MLP on datasets as shown in table 1. The results conceal that the RF classifier gives the best accuracy up to 93%, among all the used ML- techniques for white and red wine multiclass datasets. Also, the results shown in table 3 show that the performance of RF and DT is almost similar. The data pre-processing stage uses random oversampling (ROS) to balance the dataset. It was done to optimize the model's evaluation, and it was found that the performance of the model was more efficient. Since the performance analysis result doesn't come better, oversampling is used to balance the classes of the datasets. Moreover, the datasets undergo a Yeo-Jhonson transformation (YJT) to reduce the skewness. We implemented the Yeo-Johnson transform because it accommodates a broader spectrum of values, including negative ones. We employed random oversampling to enhance the size of the training dataset by duplicating original examples. Random oversampling entails the replication of instances from minority classes, which enhances their representation within the dataset and fosters a more equitable distribution among all classes. In Table 8, using the proposed approach BDCRY, accuracy by RF 23%, DT 30% and MLP 21% is improved for the white wine dataset. Similarly, the accuracy of RF remained the same. DT 10% and MLP 22% are improved in the red wine dataset. Again, MSEs are obtained for (RF, DT and MLP) of the proposed approach with 0.080, 0.151, and 0.443, respectively, on the white wine dataset and 0.143, 0.221 and 0.396, respectively, on the red wine dataset. Furthermore, we have found that in using the proposed approach for a balanced white wine dataset, accuracies are 93.14% for RF, 90.83% for DT and 75.49% for MLP. Also, we found that by comparing with the proposed approach BDCRY for a balanced red wine dataset, accuracy is 89.36% by using RF, 85.94% by using DT, and 78.00% by using MLP. We also found that the accuracy of the RF for the proposed approach is the highest among all specified classifiers for the dataset white wine. Similarly, the accuracy of the RF for the proposed approach is the highest among all specified classifiers for the dataset red wine.

Future Work

We would have liked to test some concepts during BDCRY with a single threshold value. The following ideas might be tested: SMOTE has the potential to increase accuracy, but it may also generate synthetic samples that are unrealistic. We recommend under-sampling, cost-sensitive learning, and ensemble learning approaches, alternative methods for specific datasets.

Acknowledgements

We thank our Director Dr. Vandana Bhattacherjee, for fruitful suggestions on an early draft.

Conflict of Interest

The authors declared that there is no conflict of interest for this publication.

Data Availability Statement

The datasets red wine and white wine which have been used in this research work are sourced from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Wine+Quality).

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.

Burigo., R., Scott, F., Eli, K., & Nibhrat, L. (2023). Comparison of sampling Methods for Predicting Wine Quality Based on Physicochemical Properties. *SMU Data Science Review, 7*(1), 8. https://scholar.smu.edu/datasciencereview/vol7/iss1/

Cao, Y., Chen, H., & Lin, B. (2022). Wine Type Classification Using Random Forest Model. *Highlights in Science, Engineering and Technology SDPIT, 4*. https://doi.org/10.54097/hset.v4i.1032

Carpita, M., & Goli, S. (2023). Categorical Classifiers in multiclass classification with imbalanced datasets. *WILEY*. https://doi.org/10.1002/sam.11624

Chaudhari, M.S., Kiran A. A., Shahare H., Helwatkar V., Shinde, S., Janbandhu, D., & Rangari, S. (2023). VinQCheck: An Intelligent Wine Quality Assessment. I*nternational Journal of Innovative Science and Research Technology, 8*(12). https://doi.org/10.5281/zenodo.10393843

Dahal, K. R., Dahal J. N., Banjade, H., & Gaire, S. (2021). *Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11*, 278-289. https://doi.org/10.4236/ojs.2021.112015

Danrui Q., Peng, J., Yongjun H., & Wang, J. (2023). Auto-FP: An Experiment Study of automated Feature Preprocessing for tabular Data. *Open Proceedings*. https://doi.org/10.48550/arXiv.2310.02540

Dhaliwal, P., Sharma, S., & Chauhan, L. (2022). Detailed Study of Wine Dataset and its Optimization. I. *J. Intelligent Systems and Applications, 5*, 35-46. https://doi.org/10.5815/ijisa.2022.05.04

Gawale, A.S. (2022). Wine Quality Prediction using Machine Learning and Hybrid Modeling. School of Computing, National College of Ireland.

Geethanjali, T. M., Sowjanya, M.Y., Rohith, S.N., & Shubashree, B.E. (2021). Prediction of Wine Quality using Machine Learning. *Journal of Emerging Technologies and Innovative Research (JETIR), 8*(11). http://www.jetir.org/papers/JETIR2111328.pdf

Handball, I. F., Ingosan, J. S., Oyam, N. A., & Hu, Y. (2020). Classifying Wastes Using Random Forests, Gaussian Naïve Bayes, Support Vector Machine and Multilayer Perceptron. *IOP Conf. Series: Materials Science and Engineering.* https://doi.org/10.1088/1757-899X/803/1/012017

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Kaliappan, J., Bagepalli, A. R., Almal, S., Mishra, R., Hu, Y. C., & Srinivasan, K. (2023). Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise. *Diagnostics, 13*, 1692. https://doi.org/10.3390/diagnostics13101692

Khilari, N., Hadawale, P., Shaikh, H., & Kolase, S. (2021). Analysis of Machine Learning Algorithm to Predict Wine Quality. *International Research Journal of Engineering and Technology (IRJET), 08*(12). https://doi.org/10.32628/IJSRSET229235

Kumar, S., Agarwal, K., & Mandan, N. (2020). Red Wine Quality Prediction Using Machine Learning Techniques. I*nternational Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India. pp. 1-6 https://doi.org/10.1109/ICCCI48352.2020.9104095

Lee, J., Kang, J., Park, S., Jang, D., & Lee, J. (2020). A Multi-Class Classification Model for Technology Evaluation. *Sustainability, 12*, 6153. https://doi.org/10.3390/su12156153.

Nwakuyal, M. T., & Anyaogu, I. V. (2022). Implementation of Yeo-Johnson in Quantile Regression. *Benin Journal of Statistics, 5*, 123-136.

Patkar, G.S., Balaganesh D., (2021). Smart Agri Wine: An Artificial Intelligence Approach to Predict Wine Quality. *Journal of Computer Science.* https://doi.org/10.3844/jcssp.2021.1099.1103

Saito, M., Ohsato, T., & Yamanaka, S. (2021). An empirical evaluation of machine learning performance in corporate sales growth prediction. *JSIAM Letters, 13*, 25-28. https://doi.org/10.14495/jsiaml.13.25

Siddiqi, M. A., & Pak, W. (2021). An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection. *IEEE Access.* https://doi.org/10.1109/ACCESS.2021.3118361

Sindayigaya, L., & Dey, A. (2022). Machine Learning Algorithms: A Review. I*nternational Journal of Science and Research (IJSR), 11*(8).

Tan, P.N., Steinbach, M., Karpatne, A., & Kumar, V. (2022). Introduction to Data Mining. 2nd ed., Pearson Publications.

Tigga, O., Pal, J., & Mustafi, D. (2023). Performance Analysis of Machine Learning Algorithms for Data Classification. *International Conference on Machine Intelligence with Applications* (ICMIA 2023). https://doi.org/10.1063/5.0214183

Tigga, O., Pal, J., & Mustafi, D. (2023). A Comparative Study of Rule-Based Classifier and Decision Tree in Machine Learning, I*n Proceedings of 4th International Journal of Advances in Soft Computing and Intelligent Systems (IJASCIS), 02*(01), 40-47. https://sciencetransactions.com/ijascis/uploads/2023/02/j23-40-47.pdf

Uma, R., Kaladevi, R., Jebamalar, T. J., Sarasu, P., & Charles, P. V. (2023). Analysis of multiclass imbalance handling in red wine quality dataset using oversampling and machine learning techniques. *Journal of Theoretical and Applied Information Technology, 101*(19).

Weisberg, S. (2001). Yeo-Johnson Power Transformations. *National Science Foundation Grant* DUE 97-52887.

Yang, Y., Khorshidi H.A. & Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. *Front. Digit. Health, 6*, 1430245. https://doi.org/10.3389/fdgth.2024.1430245

Zhang, Y., Li, Q., & Xin, Y. (2024). Research on eight machine learning algorithms applicability on different characteristics data sets in medical classification tasks. *Front. Comput. Neurosci., 18*, 1345575. https://doi.org/10.3389/fncom.2024.1345575.

Zhao, Y., Huang, Z., Gong L., Zhu, L., Yu, O., & Gao, Y. (2023). Evaluating the Impact of Data Transformation Techniques on the Performance and Interpretability of Software Defect Prediction Models. Hindawi, I*ET Software, 2023*, 6293074. https://doi.org/10.1049/2023/6293074

How to cite this Article:

Onima Tigga, Jaya Pal and Debjani Mustafi (2024). A Novel Data handling Technique for Wine Quality Analysis using ML Techniques. *International Journal of Experimental Research and Review, 45*, 25-40.

DOI : https://doi.org/10.52756/ijerr.2024.v45spl.003