Journal of Theoretical and Applied Information Technology

15th October 2023. Vol.101. No 19
© 2023 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ANALYSIS OF MULTICLASS IMBALANCE HANDLING IN RED WINE QUALITY DATASET USING OVERSAMPLING AND MACHINE LEARNING TECHNIQUES

UMA RANI V[1], KALADEVI R [2], JEBAMALAR TAMILSELVI J[3], SARASU P[4], CHARLES PRABU V[5]

[1]Associate Professor, Saveetha Engineering College, Department of Computer Science and Engineering, Chennai, Tamil Nadu, India.
[2]Associate Professor, Panimalar Engineering College, Department of Information Technology, Chennai, Tamil Nadu, India.
[3]Associate Professor, SRM Institute of Science and Technology, Department of Computer Science, Ramapuram, Chennai, Tamil Nadu, India.
[4] Professor, Kalasalingam Academy of Research and Education, Department of Computer Science and Engineering, Tamil Nadu, India.
[5] Assistant Professor, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Department of Computer Science and Engineering, Chennai, Tamil Nadu, India.
E-mail: [1]umaranibharathy@gmail.com, [2] kalaramar26@gmail.com, [3] jebamalj@srmist.edu.in, [4] sarasujivat@gmail.com, [5] charlesprabu.vt@gmail.com

ABSTRACT

Wine quality is very important in the wine industry and is determined by its features and flavors. The primary goal of this study is to balance the wine quality data by generating synthetic data and using a machine learning model to predict. For the current research, a multiclass unbalanced red wine quality data set is obtained from UCI resources. To balance the multiclass imbalance red wine quality data set, SMOTE and its six derivatives, including SMOTENC, SMOTE EN, SMOTE Tomek, SVM-SMOTE, Borderline SMOTE, and SMOTE ENN, are used. To predict red wine quality, seven machine-learning approaches, including Logistic Regression, Decision Tree, Random Forest, Extra Tree, XG-Boost, AdaBoost, and Bagging classifier, were trained and assessed. According to the results of this experiment, a combination of SMOTE ENN+ ETC has a higher precision of 0.96, recall of 0.96, specificity of 0.99, f1 score of 0.96, geometric mean of 0.97, and indexed balance score of 0.95 than all other SMOTE variations. SMOTE ENN + ETC has a higher accuracy of 95.8% when compared to other models. As a result, this combination is utilized to forecast red wine quality.

Keywords: *Wine Quality, Multiclass imbalance, SMOTE and its Variants, Oversampling Machine Learning*

1 INTRODUCTION

The result of technological advancements and the industrial revolution, people purchase food and beverages both online and offline [1]. Quality is one of the most important factors to consider when judging food because poor quality is harmful to human health. Evaluate the quality of food purchased online based on customer feedback. This, however, is insufficient for assessing quality; we require an automated analysis based on the contents [2]. The goal of this study is to examine red wine quality assessment methods using Machine Learning (ML).

One of the major challenges in such autonomous real-time data processing is multiclass imbalance, which increases the misclassification rate of machine learning approaches [5,6,7]. When the total number of samples in multiple classes of real-world classification datasets is biased or imbalanced or skewed, this is referred to as multi class imbalance [8,9]. Skewed data leads to poor prediction and an increased error rate in classification. To deal with such imbalanced dataset classification, researchers have devised a number of strategies.

Sampling is one approach to the problem of class imbalance in real-world datasets. To address this issue, authors have developed a variety of sampling approaches [10]. SMOTE variants are the most popular oversampling approaches for dealing with binary imbalanced datasets. As a result, most studies convert multiclass imbalanced datasets to binary balanced datasets and select the best prediction model.

The main purpose of this research is to

i) Conduct research on how to deal with multiclass imbalance problems without turning them into binary class problems.

ii) To balance the red wine multiclass data set, use popular SMOTE modifications such as SMOTE EN, SMOTE ENC, Borderline SMOTE, SVM SMOTE, ADASYN, and SMOTE ENN.

iii) Choosing the best prediction model and sampling technique by comparing the performance of different sampling methods and machine learning classifiers such as RFC, DTC, XGB, ABC, and BC.

iv) Finally, it chooses the best prediction model and deploys it for testing.

The remaining portions of this work are organized as follows. The literature review in Section 2 highlights previous research as well as the main purpose of this study. Sections 3 and 4 discuss the proposed methodology for forecasting red wine quality, as well as experimental results before and after sampling and ML methodologies. Section 5 concludes with a discussion of the findings and future directions.

2 REVIEW OF LITERATURE

Wine is an alcoholic beverage that comes in a variety of colors, including red and white, due to the various grapes used. Wine is distributed in 31 million tones around the world, which is enormous. According to experts, the aroma, flavor, and color of wine distinguish it. Initially, researchers concentrated on categorizing or forecasting red wine quality based on feedback and physiochemical properties [11].

Several research projects have been undertaken to predict quality using machine learning techniques, with only a small portion of the work concentrating on dataset balancing. According to Liu et al. [12], Spanish DO wines are classified into five types based on smell, taste, and texture. They assess wine quality using Partial Least Square structural equation models (PLSEM) and data from the 2005 to 2015 vintages. They use 13 fragrance markers (vegetal, woody, leather, toasted, fresh, dried...), three flavor indicators (sweet, acidity, bitter), and 19 mouth feel markers (dry, smooth, sandy, hard, persistent...).

Jingxian et al [13] use a SMOTE-based decision tree to describe the chemical characteristics of pinot noir and related wines with the goal to forecast astringency, sweetness, sourness, bitterness, and clarity. Bhardwaj et al [14] created a SMOTE and ML model to predict wine quality in various New Zealand regions. They forecast classifier performance by using accuracy, recall, the F1-score, and the ROC curve. They use seven classifiers for classification: XGB, RF, GNB, Ada boost, SGD, SVM, DTC, and KNN. Their survey found that the hybrid SMOTE + AdaBoost model outperforms other models in terms of accuracy.

Gu et al [15] have collected Chinese red wine samples and predicted the chemical information using two machine learning (ML) techniques: orthogonal partial least squares discriminant analysis (OPLS-DA) and SVM models. Mohana et al [16] used SMOTE-based ensemble learning to predict. They use three classifiers for prediction: RFC, XGB, and DTC. In comparison, SMOTE + RFC achieves 98.7 accuracy, according to their analysis.

Qadrinin [17] has created SMOTE and Ada-boost to deal with skewed data. According to this analysis, the Ada-boost SMOTE model outperforms the original AdaBoost model by 78.4%. Dahal et al [18] compare the performance of Ridge Regression (RR), Support Vector Machine (SVM), Gradient Boosting Regressor (GBR), and multi-layer Artificial Neural Network (ANN) in predicting wine quality. Several parameters influencing wine quality are investigated. In terms of MSE, R, and MAPE, GBR outperforms all other models, with values of 0.3741, 0.6057, and 0.0873, respectively.

Devika Pawar et al [19] have proposed a decision support system that is integrated with an automatic wine quality prediction system to accelerate the quality prediction process. The system can choose variables related to wine quality using feature selection approaches. For classification, machine learning methods such as RF, Stochastic Gradient Descent, Logistic Regression, and SVC are employed.

To predict wine quality, [20,21] employs SVM, linear regression, and neural networks. They use selected characteristics and metrics to make predictions. Bhavya et al [22] use a feature selection mechanism such as a genetic algorithm, simulated annealing approaches, and machine learning to choose the best attributes. Selected features outperform linear, nonlinear, and probabilistic models.

Lee et al. [23] have suggested a rule-based quality analytics system for fine-tuning wine physiochemical features as well as hidden wine quality patterns. They employed IoT to collect real-time data and then used rule mining to examine the relationship between wine's physicochemical and sensory properties. Ye et al [24] have suggested an MF-DCCA to assess the relationship between physiochemical data and red wine quality using a dataset from the UC Irvine repository and a machine learning (ML) technique for prediction.

Burigo et al. [25] synthesize a Portuguese wine multiclass imbalance dataset into a binary data set for prediction. For prediction, they employ an under sampling and oversampling technique, as well as a Nave Bayes classifier. According to this analysis, SMOTE EN has a higher F1 score of 0.91 when compared to other models. Kong et al [26] have used an ADASYN+XGB to predict the operating mode of fused magnesium. The proposed model is compared against SVM, LGB, RF, and Ada-boost models. ADASYN+XGB has a prediction accuracy of 92.5%, according to their studies.

Based on existing analysis, a small amount of effort was done to balance the dataset before prediction. SMOTE is a well-known oversampling mechanism for balancing real-world datasets [27,28].

This study looks at the popular Synthetic Minority Oversampling Technique (SMOTE) and its variations for dealing with multi-imbalance in red wine quality prediction. Logistic Regression (LRC), Decision Tree (DTC), Random Forest (RFC), Extra tree (ETC), XGBoost (XGB), AdaBoost (ABC), and Bagging classifier (BC) are used for prediction. Metrics including as precision, recall, specificity, geometric mean, f1 score, indexed balance, and accuracy are used to assess the performance of ML techniques.

3 MATERIALS AND METHODS

This section outlines the proposed approaches for predicting wine quality. Figure 1 depicts the overall flow. In order to train the model, the dataset is first collected and preprocessed. After preprocessing, SMOTE and its variants are used to balance the dataset. The balanced dataset is partitioned into k folds for machine learning model training. The test data set is used to forecast the performance of machine learning models. Machine learning models' performance is evaluated, and the best model is utilized to create predictions.

The following are the key processes:

1.Preprocessing

2.SMOTE sampling and its variants

3.Machine learning approaches

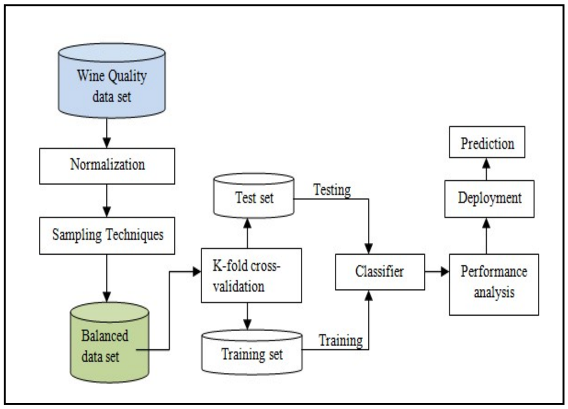


Figure 1: Multiclass imbalance handling and prediction in Redwine quality dataset

3.1 Preprocessing

The red wine quality dataset was initially gathered from UCI repositories established by Cortez et al [29]. It has been divided into ten physiochemical input features, each of which indicates an output quality level ranging from 3 to 8 (excellent to poor). The min-max scalar standardization method is used to standardize the red wine quality dataset. Using the equation (1), the min max scalar formulates the features in a specific range.

$$x' = \frac{x - \text{minimum}\ (x)}{\text{maximum}\,(x) - \text{minimum}\,(x)} \qquad (1)$$

3.2 SMOTE sampling and its variants

This subsection discusses SMOTE and its six important data balancing variations, which are SMOTE EN, SMOTE ENC, Borderline SMOTE, SVM SMOTE, and SMOTE ENN [30]. The principles of these methods are as follows.

3.2.1 SMOTE

SMOTE is a synthetic minority oversampling technique developed by Nitesh Chawla et al [31]. Let 'm' represent the number of samples required to balance the data set. SMOTE generates a new synthetic minority sample S_{new} using the equation (2).

$$S_{new} = M_j + (M_j - M_k) \times \delta \qquad (2)$$

Here M_j is the minority sample, M_k is k-nearest neighbor, and is a random number ranging from 0 to 1. SMOTE augments the current k minority class samples with new synthetic minority samples ('m'). The synthetic minority samples are developed till m samples are produced.

3.2.2 SMOTE EN

SMOTE EN [32] is a SMOTE variant that produces synthetic samples based solely on category features. It ignores the continual method of sample production. The number of neighbors is set to 5 by default.

3.2.3 SMOTE ENC

SMOTE ENC [33]is a SMOTE variation that generates synthetic samples from encoded nominal and continuous features. It ensures that the data set contains only nominal characteristics before calculating distance with chi(X). The extent of association with a specific minority class is represented by chi(X). If the data set contains continuous characteristics, the median of the continuous variable's standard deviation multiplied by X which makes the value comparable to other continuous attributes.

3.2.4 Borderline SMOTE

Han et al [34] was developed a borderline SMOTE to balance the dataset. SMOTE generates new m synthetic data samples from the border line (m) of minority class samples. It generates the 'm' new synthetic sample S_{new} using the equation (3).

$$S_{new} = B_j + R_j \times D_j \qquad (3)$$

Here B_j is the minority sample in the border line segment, R_j is a random number in the range [0,1], and D_j is the distance between the sample's selected 'm' nearest neighbors.

3.2.5 SVM SMOTE

SVM SMOTE, an oversampling technique, was developed by Ngugen et al [35]. It uses the support vector machine concept, interpolation, and extrapolation to generate new synthetic samples from minority samples (m). It selects the k nearest neighbors of the m minority samples along the extension line and uses equation (4) to create the new m synthetic samples.

$$S_{new} = M_j + (M_j - N_k) \times \delta \qquad (4)$$

Here M_j represents the minority sample, N_k represents the k-nearest neighbor in the extension line, and δ is a random number between 0 and 1.

3.2.6 ADASYN

ADASYN [36] is a synthetic adaptive algorithm that generates a random sample set depending on class distribution. It first computes the degree of imbalance and the number of synthetic samples required. It discovers K-neighbors using Euclidian distance and uses them to generate new synthesized samples.

3.2.7 SMOTE ENN

SMOTE-ENN is a mixture of oversampling (SMOTE) and under sampling (Edited Nearest Neighbor) created by G. Batista [37]. SMOTE ENN creates synthetic samples and uses ENN to clean them. SMOTE ENN begins by selecting 'm' minority samples at random. Following the selection of m samples, it finds the k-nearest neighbor of 'm' samples M_j and generates a new synthetic sample using equation (2). The nearest neighbor criteria are applied to determine whether or not the created sample S_{new} is a noisy overlapping sample. Overlapping samples are removed from the collection, and non-overlapping m samples are added.

3.3 Machine Learning Techniques

The popular six classifiers are used for training in this proposed ML selection. The red wine quality data set is separated into two parts for training the ML model: training data set and test data set. ML approaches train the model using training data sets, and the best classifier is picked for final prediction. The suggested model is trained using LRC, DTC, RFC, ETC, ABC, and XGB [38,39,40,41,42]. The following are the principles of these methods.

3.3.1 LRC

Logistic regression classifier (LRC) uses a weighted sum of input features to estimate the probability of each sample belonging to one of several classes and outputs a logit of the outcome via equation (5).

$$p' = \sigma(\theta^T.X) \qquad (5)$$

Where p' is the estimated probability and σ (.) refer sigmoid function.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3.2 DTC

Decision Tree Classifier (DTC) computes the likelihood of a sample belonging to multiple classes 'k' by traversing the tree from root to instance and returning the ratio of this node's training instance of class 'k'. It compares all features on all samples at each node.

3.3.3 RFC

Random Forest Classifier (RTC) is a decision tree ensemble trained with the bagging method. It adds unpredictability to DTC by searching for the best feature from a set of features.

3.3.4 ETC

Extra Tree Classifier (ETC) is a forest of extremely randomized tree ensembles that find the best possible threshold for each sample feature at each node.

3.3.5 ABC

Ada Boost Classifier is a boosting strategy that trains the model based on the base classifier, calculates the weighted error rate of the learnt model, and then updates or boosts the weights before training to enhance it.

3.3.6 XGBC

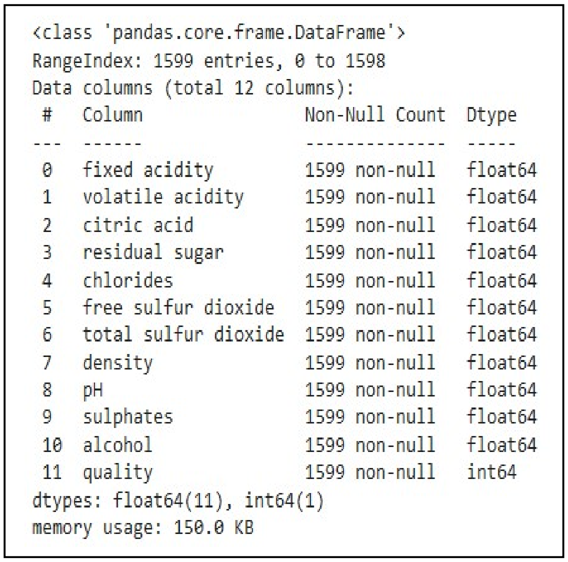
XGBoost Classifier (XGBC) is a gradient boost algorithm that works by gradually adding predictors and discovering new predictors based on the mistake rate. Finally, predictions are generated by combining all predictors.

4 RESULTS AND DISCUSSION

For experimental results, this study makes use of Google Collaboratory notebook, as well as the scikit imbalanced lean and sklearn packages. The gradio package is used for prediction testing.

4.1 Dataset Description

The imbalanced red wine quality dataset is initially obtained from UCI libraries. Redwine quality dataset has 1599 data items and ranks wines from 3 to 8 based on a variety of characteristics. Figure 2.a) depicts the parameter as well as its descriptions. Figure 2.b) depicts the histogram for each parameter. To fit data in the range 0 to 1, an imbalanced multiclass dataset is pre-processed and normalized using the min max scalar technique. Figure 3 depicts the count of each class and data distribution for each of the six different quality classification ranges ranging from 3 to 8.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Figure 2.a: Redwine quality dataset description

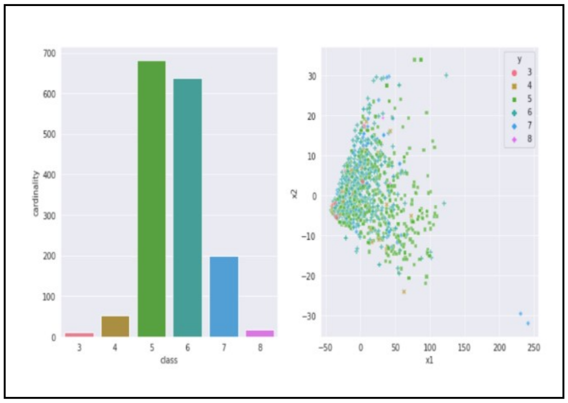


Figure 3: Count of red wine quality classes and its data distribution before sampling

Figure 3 indicates that red wine quality classes 3, 4, 7, and 8 have significantly fewer incidences than classes 5 and 6.

The multiclass dataset is balanced using oversampling SMOTE variations, and the resulting count of balanced red wine grade classes and data distribution is shown in Figure 4. The sampling helps to increase the number of samples in class 3,4,5 and 7.

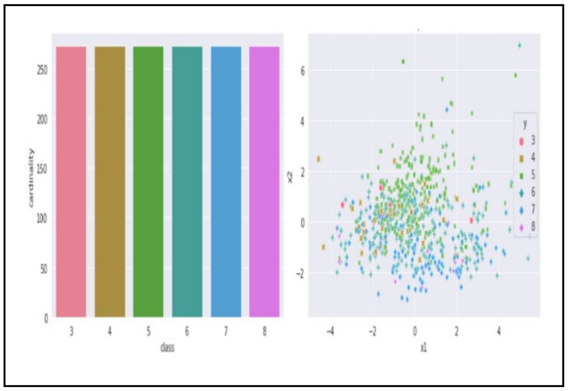


Figure 4: Count of red wine quality classes and its data distribution after sampling

4.2 Performance metrics

The following performance metrics are used to assess the ML performance.

1. Precision(pre) - specify the accuracy of positive prediction calculated by equation (6)

$$Pre = \frac{TP}{TP + FP} \qquad (6)$$

2. Recall (rec)– Recall or sensitivity or true positive rate is the ratio of positive instances that are correctly detected by the ML technique. It is calculated by equation (7).

$$Rec = \frac{TP}{TP + FN} \qquad (7)$$

3. Specificity (spe) is the ratio of true negative Instances that are correctly detected by ML technique.

$$Spe = \frac{TN}{TN + FP} \qquad (8)$$

4. f1 score – is the harmonic mean of precision and recall values. It ranges from 0 to 1 evaluated by equation (9).

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \qquad (9)$$

5. Geometric mean (geo) - is the squared root of product of sensitivity and specificity shown in equation (10).

$$geo = \sqrt{Rec \times spe} \qquad (10)$$

6. Indexed balance Accuracy (iba) - is the average of recall obtained on each class.

$$iba = \big(1 + \alpha\,(Rec - Spe)\big)(Rec \times spe) \quad (11)$$

4.3 Analysis of machine learning techniques for wine quality prediction before sampling

Initially, ML techniques are used to predict the unbalanced red wine quality dataset. Table 1 displays the performance of the ML classifier prior to sampling. Without balancing the data set, ML techniques RFC (precision= 0.69, recall= 0.72, specificity = 0.82, f1 score =0.70 and iba=0.58) and ETC (precision= 0.68, recall= 0.71, specificity = 0.81, f1 score =0.70 and iba=0.57) perform better in terms of metric precision, recall, specificity, f1 score, and indexed balance accuracy. Similarly, ETC has 72.81 accuracy, RFC has 71.87 accuracy, BC has 71.25 accuracy, and the other models, LRC, DTC, XGBC, ABC, and BC, have less than 70% accuracy. This is shown in Figure 5.

Table 1: Performance of ML techniques before sampling

ML	pre	rec	Spe	f1	Geo	iba	Acc
LRC	0.59	0.63	0.74	0.61	0.65	0.45	63.4
DTC	0.65	0.63	0.80	0.64	0.70	0.50	63.4
RFC	0.69	0.72	0.82	0.70	0.74	0.58	71.8
ETC	0.68	0.71	0.81	0.70	0.73	0.57	72.8
XGBC	0.64	0.66	0.79	0.64	0.69	0.50	65.6
ABC	0.54	0.48	0.77	0.49	0.58	0.35	40.6
BC	0.69	0.72	0.82	0.70	0.74	0.58	71.25

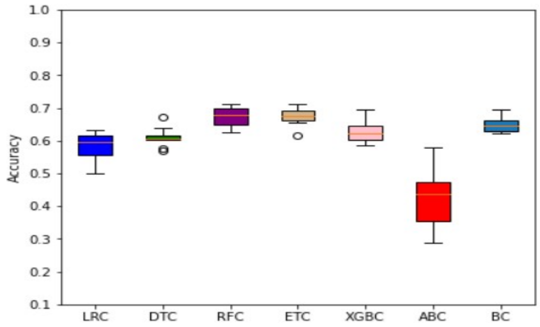


Figure 5: A boxplot shows accuracy of ML techniques before sampling

Because of the imbalanced data problem, the ML technique provides low accuracy (75%), precision (70%), recall (73), and f1 score (=70). The ML method has a low indexed balance score (60%) and a low geometric mean (75). For classes 3, 4, 7, and 8, the ML technique provides lower performance metric values.

4.4 Analysis of machine learning technique for wine quality prediction after sampling

The red wine quality data set contains unbalanced data for quality classes 3, 4, 7, 8, and 9 when compared to classes 5 and 6. The dataset is balanced using oversampling techniques SMOTE, SMOTE EN, SMOTE ENC, Borderline SMOTE, SVM SMOTE, SMOTE Tomek, and hybrid sampling SMOTE ENN. The balanced multiclass dataset is used to assess the performance of the ML classifier. The ML performance has improved across all classes, with accuracy exceeding 75% and all other metric scores improving as well.

Table 2: Performance of ML techniques after sampling

ML	pre	Rec	spe	f1	Geo	iba	Acc
After SMOTE							
LRC	0.47	0.48	0.90	0.47	0.65	0.41	48.16
DTC	0.77	0.77	0.95	0.77	0.85	0.73	78.24
RFC	0.87	0.87	0.97	0.87	0.92	0.84	87.04
ETC	**0.87**	**0.87**	**0.97**	**0.87**	**0.92**	**0.85**	**87.28**
XGBC	0.75	0.76	0.95	0.75	0.84	0.72	76.03
ABC	0.44	0.41	0.88	0.41	0.59	0.35	41.44
BC	0.84	0.84	0.97	0.84	0.90	0.81	84.47
After SMOTE ENC							
LRC	0.48	0.48	0.90	0.47	0.65	0.42	48.41
DTC	0.77	0.77	0.95	0.77	0.85	0.73	77.13
RFC	0.84	0.85	0.97	0.84	0.90	0.82	84.71
ETC	**0.87**	**0.87**	**0.97**	**0.87**	**0.92**	**0.85**	**87.28**
XGBC	0.77	0.78	0.96	0.77	0.85	0.74	77.99
ABC	0.4	0.4	0.88	0.40	0.58	0.34	40.09
BC	0.80	0.81	0.96	0.80	0.88	0.77	80.56
After SMOTE EN							
LRC	0.59	0.61	0.92	0.59	0.74	0.54	60.54
DTC	0.85	0.85	0.97	0.85	0.91	0.82	85.08
RFC	0.87	0.87	0.97	0.87	0.92	0.84	86.67
ETC	**0.88**	**0.88**	**0.97**	**0.88**	**0.92**	**0.85**	**87.5**
XGBC	0.84	0.84	0.97	0.84	0.90	0.81	83.86
ABC	0.61	0.59	0.92	0.59	0.72	0.53	58.86
BC	0.86	0.85	0.97	0.85	0.91	0.82	85.45
After Borderline SMOTE							
LRC	0.65	0.66	0.93	0.65	0.77	0.61	66.25
DTC	0.82	0.82	0.96	0.82	0.89	0.79	82.2
RFC	0.86	0.87	0.97	0.86	0.91	0.84	82.27
ETC	**0.87**	**0.87**	**0.97**	**0.87**	**0.92**	**0.85**	**87.40**
XGBC	0.81	0.82	0.96	0.81	0.88	0.78	81.9
ABC	0.54	0.53	0.9	0.53	0.68	0.47	52.3
BC	0.84	0.85	0.97	0.84	0.90	0.81	84.5
After SVM SMOTE							
LRC	0.57	0.58	0.91	0.57	0.72	0.51	57.82
DTC	0.80	0.80	0.95	0.79	0.87	0.75	79.60
RFC	0.84	0.84	0.96	0.84	0.89	0.79	83.5
ETC	**0.85**	**0.85**	**0.96**	**0.85**	**0.90**	**0.81**	**85.27**
XGBC	0.75	0.76	0.94	0.75	0.84	0.71	75.76
ABC	0.48	0.48	0.88	0.47	0.64	0.41	47.5
BC	0.82	0.82	0.95	0.82	0.88	0.77	81.74
After SMOTE Tomek							
LRC	0.52	0.53	0.91	0.51	0.68	0.46	52.61
DTC	0.81	0.81	0.96	0.81	0.88	0.77	81.38
RFC	0.87	0.88	0.97	0.87	0.92	0.85	87.61
ETC	**0.88**	**0.89**	**0.98**	**0.88**	**0.93**	**0.86**	**88.63**
XGBC	0.76	0.78	0.95	0.77	0.85	0.73	77.65
ABC	0.83	0.83	0.97	0.83	0.89	0.80	83.39
BC	0.83	0.83	0.95	0.83	0.88	0.77	82.7
After SMOTE ENN							
LRC	0.60	0.63	0.90	0.61	0.73	0.55	63.24
DTC	0.89	0.89	0.98	0.89	0.93	0.87	89.32
RFC	0.96	0.96	0.99	0.95	0.97	0.94	95.6
ETC	**0.96**	**0.96**	**0.99**	**0.96**	**0.97**	**0.95**	**95.89**
XGBC	0.92	0.92	0.98	0.92	0.95	0.89	91.78
ABC	0.73	0.72	0.93	0.72	0.82	0.66	72.07
BC	0.93	0.93	0.99	0.93	0.96	0.92	93.4

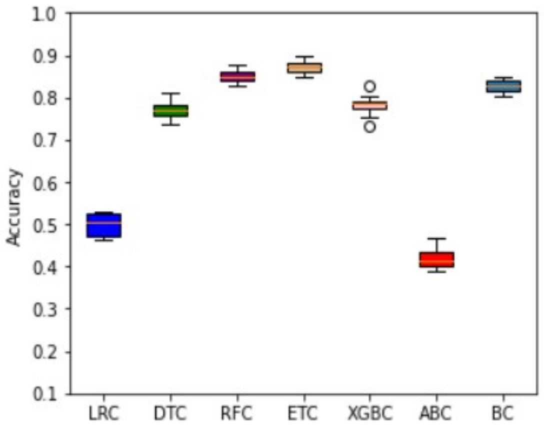


Figure 6: A boxplot shows accuracy of ML techniques after SMOTE

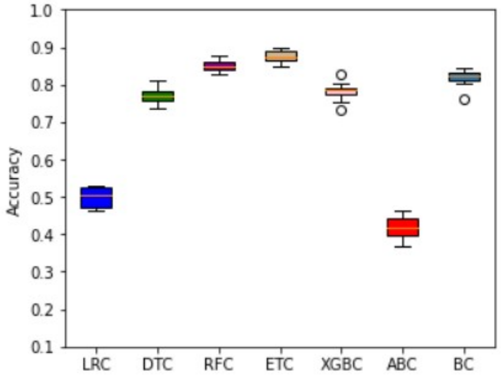


Figure 7: A boxplot shows accuracy of ML techniques after SMOTE ENC

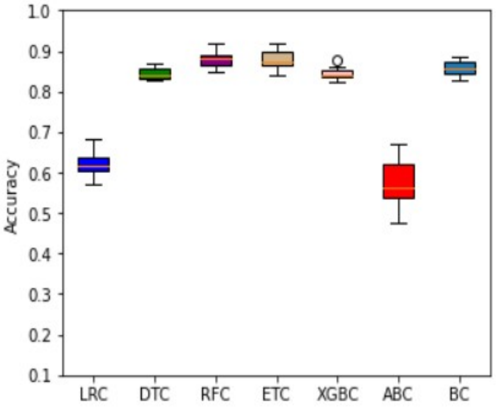


Figure 8: A boxplot shows accuracy curve of ML techniques after SMOTE EN

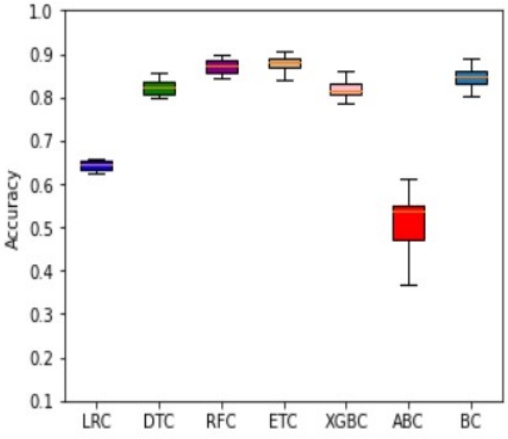


Figure 9: A boxplot shows accuracy curve of ML techniques after Borderline SMOTE

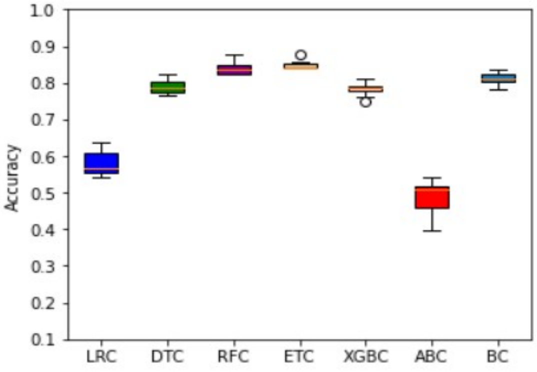


Figure 10: A boxplot shows accuracy curve of ML techniques after SVM SMOTE

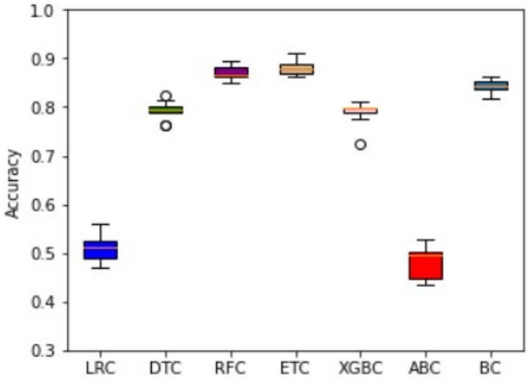


Figure 11: A boxplot shows accuracy curve of ML techniques after SMOTE Tomek

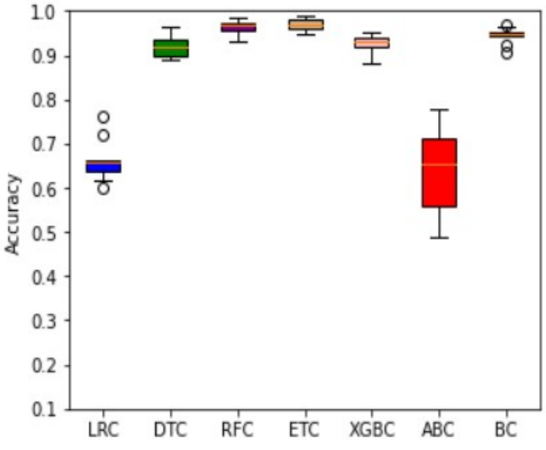


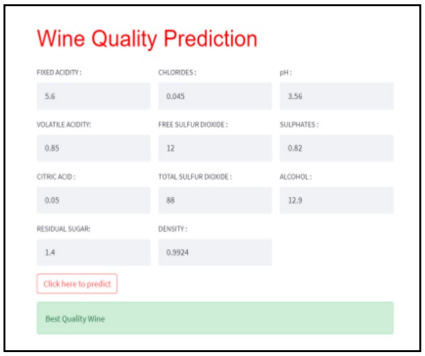
Figure 12: A boxplot shows accuracy curve of ML techniques after SMOTE ENN

Table 2 shows the precision, recall, specificity, f1 score, geomean, and indexed balance accuracy of the classifier. When compared to other classifiers, ETC provides higher precision 0.87, recall 0.87, specificity 0.97, f1 score 0.87, geometric mean 0.92, iba score 0.85, and accuracy 87.28 after SMOTE sampling. In comparison to other classifiers, ETC provides higher precision 0.87, recall 0.87, specificity 0.97, f1 score 0.87, geometric mean 0.92, iba score 0.85, and accuracy 87.28 after SMOTEENC sampling. In comparison to other classifiers, ETC provides higher precision 0.88, recall 0.88, specificity 0.97, f1 score 0.88, geometric mean 0.92, iba score 0.85, and accuracy 87.5 after SMOTE EN sampling. ETC provides higher precision 0.87, recall 0.87, specificity 0.97, f1 score 0.87, geometric mean 0.92, and iba score 0.85 after Borderline SMOTE sampling.

After SVM SMOTE sampling, ETC provide higher precision 0.85, recall 0.85, specificity 0.96, f1 score 0.87, geometric mean 0.92, iba score 0.85 and accuracy 87.28 compared to other classifiers. After SMOTE TOMEK sampling, ETC provide higher precision 0.88, recall 0.89, specificity 0.98, f1 score 0.88, geometric mean 0.93, iba score 0.86 and accuracy 88.63 compared to other classifiers. The combination of SMOTE ENN+ ETC provide a higher precision of 0.96, recall of 0.96, specificity of 0.99, f1 score of 0.96, geometric mean of 0.97, and iba score of 0.95 than all other SMOTE variants, ML techniques.

The boxplot Figure 6 to 12 shows the accuracy of classifier using multiclass balanced red wine data set. From this boxplot, SMOTE+ ETC provides 87.28% accuracy, whereas SMOTE+ RFC provides 87.04% accuracy, which is greater than other ML approaches. In a SMOTE ENC-based ML combination, SEMOTE ENC+ ETC provides superior accuracy (87.28%), whereas SMOTE+ETC does not. SMOTE EN + ETC provides 87.5% accuracy in SMOTE EN based ML combinations, and SMOTE Tomek + ETC provides 88.63% accuracy in SMOTE Tomek based ML combinations. DTC, RFC, ETC, XGBC, and BC provide accuracy above 88% in SMOTE ENN-based ML combinations, with SMOTE ENN+ ETC providing better accuracy of 95.89%.

In this research, multiclass red wine quality dataset is predicted without turning it into binary data set and SMOTE ENN + ETC is selected for final prediction. The SMOTE ENN helps to balance the dataset and it increase the accuracy of machine learning classifiers and increase the performance of other metrics. So, the best model SMOTE ENN + ETC is selected for final prediction of red wine quality data. Finally, this model is tested using python-gradio tool. The prediction is shown in following Figure 13.



Wine Quality Prediction

FIXED ACIDITY: 5.6
CHLORIDES: 0.045
pH: 3.56
VOLATILE ACIDITY: 0.85
FREE SULFUR DIOXIDE: 12
SULPHATES: 0.82
CITRIC ACID: 0.05
TOTAL SULFUR DIOXIDE: 88
ALCOHOL: 12.9
RESIDUAL SUGAR: 1.4
DENSITY: 0.9924

Best Quality Wine

Figure 13: Wine Quality Prediction using gradio tool

5 CONCLUSION

Seven different SMOTE variations and seven ML approaches are utilized to examine the chemical parameters of red wine in order to forecast its quality. According to this study, the ML approach has a 72% accuracy due to multiclass unbalanced red wine data. To balance the red wine dataset, the over sampling methods SMOTE and its variants SMOTE ENC, SMOTE EN, Borderline SMOTE, SVM SMOTE, SMOTE Tomek, and SMOTE ENN are used. This study employs LRC, DTC, RFC, ETC, XGBC, ABC, and BC for prediction. Precision, recall, specificity, f1 score, geometric mean, and indexed balance accuracy are used to evaluate the performance of the ML approach. According to the experimental results, the combined SMOTE ENN outperforms the other oversampling strategies. The hybrid SMOTE ENN improves the accuracy of ensemble machine learning techniques RFC, DTC, ETC, and boosting ML approach to greater than 90%. This hybrid technique improves the f1 score and G-mean to above 95%. As a result, SMOTE ENN + ETC outperforms other models and is selected for implementation. In the future, this work will be expanded by including other physiochemical features in various types of wines and using machine learning models to predict them. It will use meta heuristic algorithm to fine tune the performance of machine learning models. The sampling method is tested with other multiclass imbalance datasets.

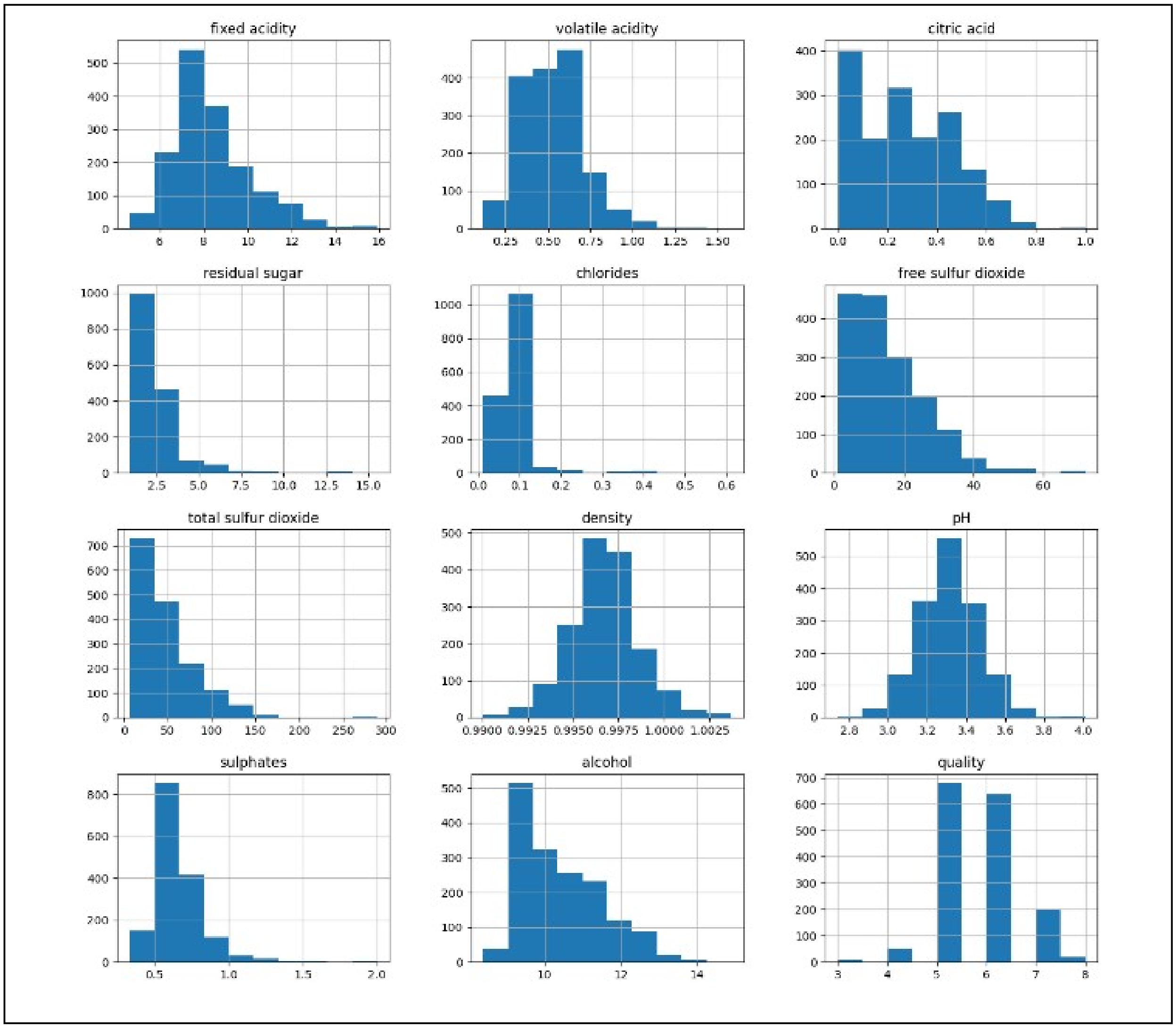


Figure 2.b: Histogram of parameters in Redwine Quality Dataset

REFERENCES:

[1] Khalafyan,Z.Temerdashev, A. Abakumov, Y.Yakuba, O. Sheludko, & Kaunova, A,"Multidimensional analysis of the interaction of volatile compounds and amino acids in the formation of sensory properties of natural wine" in Heliyon, Vol.**2,**2023.

[2] Armstrong, EJ Claire, Adam M. Gilmore, Paul K. Boss, Vinay Pagay, and David W. Jeffery, "Machine learning for classifying and predicting grape maturity indices using absorbance and fluorescence spectra",in *Food Chemistry*, Vol.**403,2023** pp.134321.

[3] Kalopesa, Eleni, Konstantinos Karyotis, Nikolaos Tziolas, Nikolaos Tsakiridis, Nikiforos Samarinas, & George Zalidis. "Estimation of Sugar Content in Wine Grapes via In Situ VNIR–SWIR Point Spectroscopy Using Explainable Artificial Intelligence Techniques", in *Sensors,* Vol.**23,** No 1,2023 pp.1065.

[4] Jiang, Liang, Yu Qiu, Morphy C. Dumlao, William A. Donald, Christopher C. Steel & Leigh M. Schmidtke. "Detection and prediction of Botrytis cinerea infection levels in wine grapes using volatile analysis", in *Food Chemistry*,Vol.21,No.**4**,2023,pp.136120.

[5] S.Mani, R.A.Krishnankutty, S.Swaminathan & P.Theerthagiri, "An investigation of wine quality testing using machine learning techniques", IAES International Journal of

Artificial intelligence, Vol.12, No.2, 2023, pp.740-747.

[6] M.Tiwari, H.Pandey, A.Mukherjee, & R.F Sutar,"Artificial Intelligence in Food Processing", in *Novel Technologies in Food Science,* No.**6** ,2023,pp.511-550

[7] V. Cardoso Schwindt, M.M Coletto, M.F Diaz & Ponzoni, I., "Could QSOR modelling and machine learning techniques be useful to predict wine aroma?", in *Food and Bioprocess Technology*, Vol.**16**,2023,pp. 24-42.

[8] C.L Udeze, I.E Eteng, & A.E Ibor, "Application of Machine Learning and Resampling Techniques to Credit Card Fraud Detection", *in Journal of the Nigerian Society of Physical Sciences*, Vol.**4,** 2022, pp.769. https://doi.org/10.46481/jnsps.2022.

[9] M. Lango & J. Stefanowski,"What makes multi-class imbalanced problems difficult? An experimental study," in *Expert Systems with Applications,* no.***199*** ,2022,116962.

[10] James Palmer, Victor S. Sheng, Travis Atkison, & Bernard Chen, "Classification on Grade, Price, and Region with Multi-Label and Multi-Target Methods in Wine informatics", in *Big Data Mining and Analytics,*no.**3**,2020,
DOI: 10.26599/BDMA.2019.9020014.

[11] Nattane Luíza da Costa, Leonardo A. Valentin, Inar Alves Castro, Rommel Melgaço Barbosa,"Predictive modeling for wine authenticity using a machine learning approach", in *Artificial Intelligence in Agriculture*, vol.**5**,2021,pp.157-162.

[12] S. Liu, A.R Vega, & M. Dizy," Assessing ultra-premium red wine quality using PLS-SEM". LWT, **5**(2023),114560.

[13] Jingxian, Paul A. Kilmartin, Brent R. Young, Rebecca C. Deed, & Wei Yu. "Decision trees as feature selection methods to characterize the novice panel's perception of Pinot noir wines",**6** (2023).

[14] P.Bhardwaj, P.Tiwari, Jr K Olejar, W.Parr & D. Kulasiri, "A machine learning application in wine quality prediction", in *Machine Learning with Applications*,Vol. **8**,2022,100261.

[15] H.W Gu, H.H Zhou, Y.Lv, Q. Wu, Y. Pan, Z.X Peng, & X.L.Yin, "Geographical origin identification of Chinese red wines using ultraviolet-visible spectroscopy coupled with machine learning techniques," in *Journal of Food Composition and Analysis*, No.**119**,2023,pp. 105265.

[16] R.Mohana, P. Sharma, & A. Sharma. "Ensemble Framework for Red Wine Quality Prediction." Food Analytical Methods **16** (2023) 30-44.

[17] L.Qadrini," Handling Unbalanced Data with Smote Adaboost,".in Jurnal Mantik, vol.**6** ,2022,pp.2332-2336.

[18] K.R Dahal, J. N. Dahal, H. Banjade & S. Gaire. "Prediction of wine quality using machine learning algorithms." In *Open Journal of Statistics, vol.* **11**,2021,pp.278-289.

[19] Devika Pawar, Aakanksha Mahajan, Sachin Bhoithe, "Wine Quality Prediction using Machine Learning Algorithms", in *International Journal of Computer Applications Technology and Research*, vol.**8**,2019, pp.385-388.

[20] Mohit Gupta & C.Vanmathi, "A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality", International *Journal of Recent Technology and Engineering* (IJRTE),**10**,(2021).

[21] Yogesh Gupta,"Selection of important features and predicting wine quality using machine learning techniques"*,Procedia Computer Science*,no.**125**,2018,pp.305-31.

[22] AG.Bhavya, "Wine Quality Prediction Using Different Machine Learning Techniques", International Journal of Science, Engineering and Technology, vol.**8**,no.4,2020.

[23] Lee, Carmen KH, Kris MY Law, and Andrew WH Ip. "A rule-based quality analytics system for the global wine industry", Journal of Global Information Management (JGIM) vol.29, no. 3,2021,pp.256-273.

[24] Ye, Chao, Ke Li, and Guo-zhu Jia. "A new red wine prediction framework using machine learning." in Journal of Physics: Conference Series, vol. 1684, no.1,2020,012067. IOP Publishing.

[25] Burigo, Robert, Scott Frazier, Eli Kravez & Nibhrat Lohia. "Comparisn of Sampling Methods for Predicting Wine Quality Based on Physicochemical Properties" SMU Data Science Review, **no.7** 2022,pp. 1- 8.

[26] W.Kong, Z.Su, & Y.Ding," Prediction of Fused Magnesium Operating Mode Based on ADASYN-XGBoost",in *International Journal of Machine Learning and Computing*, Vol.**12**,no.5,2022.

[27] Muntasir Nishat, F.Faisal, L.Jahan Ratul, Al-Monsur, Ar-Rafi, S.M.Nasrullah, & M. R. H Khan,"A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset". Scientific Programming, vol.**3** ,2022,pp.1-17.

[28] F Yang, K.Wang, L.Sun, M.Zhai, J.Song & H.Wang, "A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis",in *BMC Medical Informatics and Decision Making*, vol.**22**,2022, 344.

[29] P.Cortez, A. Cerdeira, F.Almeida, T.Matos, & J.Reis, "Modeling wine preferences by data mining from physicochemical properties", *Decision support systems,* vol.47,no.**4**, 2009,pp. 547-553.

[30] H.He, & Y.Ma, " Imbalanced learning: foundations, algorithms, and applications, (2013).

[31] Chawla, Nitesh,"SMOTE: synthetic minority over-sampling technique." In *Journal of artificial intelligence research*, vol.**16** (2002) 321-357.

[32] M.L Zhang, Y.K Li, H. Yang & X.Y Liu," Towards class-imbalance aware multi-label learning". IEEE Transactions on Cybernetics, **52** 2020, pp.4459-4471.

[33] M. Mukherjee & M. Khushi,"SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features", Applied System Innovation, vol.**4,no.1,** 2021,pp.1-18.

[34] H. Han, W. Y Wang & B. H Mao "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning". Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, A, no.**1**,2005,pp.878-887.

[35] Almajid, A. S,"Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction", Journal of Advances in Information Systems and Technology, no.**3** ,2021,pp. 67-74.

[36] UmaRani, V. Saravanan, and J. Jebamalar Tamilselvi. "A Hybrid Grey Wolf-Meta Heuristic Optimization and Random Forest Classifier for Handling Imbalanced Credit Card Fraud Data." *International Journal of Intelligent Systems and Applications in Engineering*, Vol.11, No. 9,2023, pp. 718-734.

[37] M. Muntasir Nishat, F.Faisal, I.Jahan Ratul, A.Al-Monsur, A. M. Ar-Rafi, S. M Nasrullah, & M. R. H. Khan, "A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset". Scientific Programming, 2022, pp.1-17.

[38] V.Umarani, Anitha Julian & J. Deepa," Sentiment Analysis using various Machine Learning and Deep Learning Techniques". Journal of the Nigerian Society of Physical Sciences, **No.3** ,2021, pp.385–394.

[39] David Opeoluwa Oyewola, Dada, Ndunagu , J. N Abubakar Umar, & S.A, A,"COVID-19 Risk Factors, Economic Factors, and Epidemiological Factors nexus on Economic Impact: Machine Learning and Structural Equation Modelling Approaches" , Journal of the Nigerian Society of Physical Sciences, no 3,2022,pp.395–405. https://doi.org/10.46481/jnsps.2021.173.

[40] Ibidoja,F.P Shan, Mukhtar, J.Sulaiman, & M. K. Majahar Ali, "Robust M-estimators and Machine Learning Algorithms for Improving the Predictive Accuracy of Seaweed Contaminated Big Data", Journal of the Nigerian Society of Physical Sciences, No.**5**,2023,pp.1137-1148. https://doi.org/10.46481/jnsps.2023.1137