BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2311-2326
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2311-2326

PERFORMANCE ANALYSIS OF MODIFIED-ODBOT AND SMOTE FOR TREE-BASED CLASSIFICATION OF IMBALANCED HUMAN DEVELOPMENT INDEX DATA

Yunna Mentari Indah[1*], **Anwar Fitrianto**[2], **Indahwati**[3]

[1,2,3]*Department of Statistics and Data Science, IPB University*
Jln. Raya Dramaga, Bogor, Jawa Barat, 16680, Indonesia

*Corresponding author's e-mail: * **yunnamentari90@gmail.com***

Article Info

Article History:

Received: 15th August 2025
Revised: 27th November 2025
Accepted: 10th March 2026
Available online: 8th April 2026

Keywords:

Class imbalance;
Euclidean distance;
Mahalanobis distance;
Multiclass data;
Oversampling technique.

ABSTRACT

Classification of Human Development Index (HDI) data presents significant challenges due to severe class imbalance, where low-development regions are substantially underrepresented. This imbalance reduces classification performance because machine learning models tend to be biased toward the majority classes, making it challenging to accurately identify minority classes. This study proposes a modified ODBOT that replaces Euclidean distance with Mahalanobis distance within the oversampling mechanism (Mahalanobis-based ODBOT) and compares its performance with Euclidean-based ODBOT with and without Principal Component Analysis (PCA), as well as the conventional SMOTE technique. Four tree-based classifications were used, namely Random Forest, Double Random Forest, XGBoost, and LightGBM. The Human Development Index (HDI) data set from the Central Statistics Agency, consisting of 514 observations and four features, with an imbalance ratio (IR) of 19.0, was divided into training and testing sets (ratio 80:20) with 30 repetitions and evaluated using F1-Measure (F1-M), Geometric Mean (G-M), Area Under the Curve (AUC), and computation time. The results show that Mahalanobis-based ODBOT achieved the highest performance on the AUC evaluation metric across all classification models and the highest on the G-M evaluation metric in three of the four classification models, but required significantly longer computation time (2545.66 seconds). In contrast, the Euclidean-based ODBOT with PCA improved F1-M while reducing computation time (7.21 seconds) compared to the original ODBOT (68.23 seconds), while SMOTE consistently improved G-M and AUC across all experiments. These findings suggest that oversampling techniques should be selected based on practical application needs. Specifically, the Mahalanobis-based ODBOT can be recommended when improving prediction performance is a priority, while the Euclidean-based ODBOT with PCA or SMOTE is preferable for real-world implementations that require faster execution and lower computational cost.

How to cite this article:

Y. M. Indah, A. Fitrianto, and Indahwati, "PERFORMANCE ANALYSIS OF MODIFIED-ODBOT AND SMOTE FOR TREE-BASED CLASSIFICATION OF IMBALANCED HUMAN DEVELOPMENT INDEX DATA", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2311-2326, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Class imbalance is a well-known problem in socio-demographic data analysis, like the Human Development Index (HDI), where the count of low HDI regions (minority class) is much less than the number of high HDI regions (majority class). In such conditions, the model often provides biased prediction results towards the majority class and causes a decrease in overall performance [1]. Furthermore, the highly imbalanced class distribution typically causes classification models, particularly decision tree-based algorithms, to fail in classifying minority regions with adequate sensitivity, resulting in reduced accuracy in detecting areas needing development [2]. In the context of Indonesia, this challenge becomes especially relevant, as the Human Development Index (HDI) reached 0.728 in 2023, placing the country in the high human development category and ranking 113th out of 193 countries according to United Nations Development Programme (UNDP) yet substantial disparities persist across provinces, where low-development regions constitute only a small minority compared to high-development regions. This imbalance highlights the need for an enhanced HDI classification method that can effectively handle skewed class distributions, ensuring the fair and accurate identification of regions requiring development attention [3]. While this number indicates some level of progress at the national level, Indonesia still faces considerable regional inequality, posing significant challenges that must be addressed to promote equitable development. Multi-class classification is relevant in regional development because it can help group regions according to indicators like HDI to make more precise and ultimately better government policy interventions [4]. Badan Pusat Statistik (BPS) classifies the HDI into very high, high, intermediate, and low. However, regions based on districts/cities are unevenly distributed across HDI categories, resulting in an imbalanced dataset where the high category accounts for 62.84%. In comparison, the intermediate category (22.37%) and the low category have fewer cases at 3.31% [5].

This imbalanced data is a common challenge in developing effective classification models, as it increases the risk of bias toward majority classes [6]. Accurately detecting intermediate and low categories for development prioritization is particularly difficult, since models tend to favor majority classes while underdetecting minority classes [7]. Addressing class imbalance in training data is crucial to developing reliable classification models. Multiple data balancing methods are available to address this challenge, including oversampling and undersampling. SMOTE (Synthetic Minority Over-sampling Technique) is one of the most well-known and widely used oversampling methods, as it generates synthetic minority samples through interpolation [8]. Several extensions of the SMOTE algorithm have been developed, including Outlier-SMOTE, which specifically focuses on generating additional synthetic instances for minority class outliers. By giving greater weight to these infrequent observations, the approach enhances the model's ability to capture underrepresented structures within the data, thereby mitigating the misclassification of minority class samples [9]. Additionally, SOA-SMOTE uses a One-Class Support Vector Machine (OCSVM) to identify representative samples from the minority class and subsequently carry out synthetic oversampling [10].

Furthermore, Lusito et al. [11] employed FROID (Features Reduction and Outlier Detection), a data preprocessing method that utilizes unsupervised representation learning to address data imbalance by optimizing the dataset through feature reduction using PCA (Principal Component Analysis) and outlier detection via the Local Outlier Factor. The abbreviation FROID highlights its dual function: simultaneously reducing data dimensionality and identifying extreme observations, thereby enhancing the quality and reliability of subsequent modeling. Several approaches to addressing imbalanced data problems using clustering were also developed. Mirzaei et al. [12] conducted a clustering-based approach using DBSCAN to select the most representative samples of the majority class, while FCM-KSMOTE [13] used fuzzy clustering to group data and address noise and unclear class boundaries before the data synthesis process. In addition, the ODBOT (Outlier Detection-Based Oversampling Technique) method was introduced by Ibrahim [14]. ODBOT's performance utilizes outlier detection with K-Means clustering, employing Euclidean Distance and the Weighted Bat Algorithm (WBBA) for cluster optimization in both minority and majority classes within data clusters. The Euclidean distance used in K-Means clustering in the original ODBOT framework (Euclidean-based ODBOT) has shown reasonable effectiveness in generating synthetic samples for imbalanced classification tasks. However, research conducted by Kumari and Gupta [15] explains that Euclidean distance has several limitations, particularly its sensitivity to feature scale differences and correlations between variables. The Mahalanobis distance is a better measure because it takes into account the correlation between variables (i.e., their covariance), and is more effective when used on datasets with heterogeneous feature distributions that vary in range. Based on this, this study modifies ODBOT by

replacing Euclidean distance with Mahalanobis distance (Mahalanobis-based ODBOT) as an oversampling method for handling imbalanced data. This modified approach then performs classification using decision tree-based algorithms, such as Random Forest, due to their flexibility, stability, and interpretability. When combined with the proper oversampling technique, decision tree-based models can improve classification performance on imbalanced data [16]. Therefore, this study aims to analyze the performance of the Mahalanobis-based ODBOT method applied to multi-class HDI data. The effectiveness of the proposed method is evaluated by comparing it with the SMOTE technique and the Euclidean-based ODBOT approach combined with Principal Component Analysis (PCA). PCA is only applied in conjunction with Euclidean-based ODBOT and is not used in conjunction with SMOTE. The evaluation utilizes the 2023 HDI data, which is imbalanced, comprising 514 observations that represent all districts and cities in Indonesia. It employs four decision tree classifications: Random Forest, Double Random Forest, XGBoost, and LightGBM. It is expected that Mahalanobis-based ODBOT will improve classification performance for underrepresented HDI categories compared to other methods, especially in terms of F1-Measure, Geometric Mean, and Area Under the Curve (AUC). In addition, this study is also expected to provide practical recommendations for selecting data imbalance handling methods that are appropriate for HDI data characteristics.

2. RESEARCH METHODS

RStudio version 4.3.2 was used to implement the analysis in this study and comprised the following significant steps:

1. Data exploration and visualization include descriptive analysis to view data summaries, distribution proportions of variable Y (HDI) to determine the distribution of categories in the HDI variable, and identifying class imbalances. Additionally, the correlation matrix is used to detect multicollinearity in independent variables.
2. The dataset was divided into 80% training data and 20% testing data to train and evaluate the classification model. The division is performed in a stratified manner to maintain the class distribution in both subsets, repeating the process 30 times to obtain optimal and stable results.
3. Imbalanced classes in the training data were treated in five scenarios:
   a. Without handling imbalanced,
   b. Oversampling by the SMOTE,
   c. Oversampling by the ODBOT algorithm with Euclidean distance,
   d. PCA is used for dimensionality reduction before using Euclidean-based ODBOT,
   e. Modified ODBOT with Mahalanobis distance.
4. Random Forest, Double Random Forest, XGBoost, and LightGBM models are used for classification modeling on both the imbalanced and balanced datasets.
5. Model performance was evaluated using 5-fold cross-validation with standard parameter values, without hyperparameter tuning, to demonstrate the relative competitiveness of the model performance generated by various types of decision tree algorithms. Model performance was also evaluated based on Area Under the Curve (AUC), F1-Measure (F1-M), and Geometric Mean (G-M), while computation time was recorded for each method. Each experiment was repeated 30 times to ensure the reliability of the results. Stratified training-testing splits were applied to maintain class distribution across subsets. Evaluation metrics were then averaged across repetitions to obtain stable estimates of performance. The step-by-step procedure is illustrated in Fig. 1.

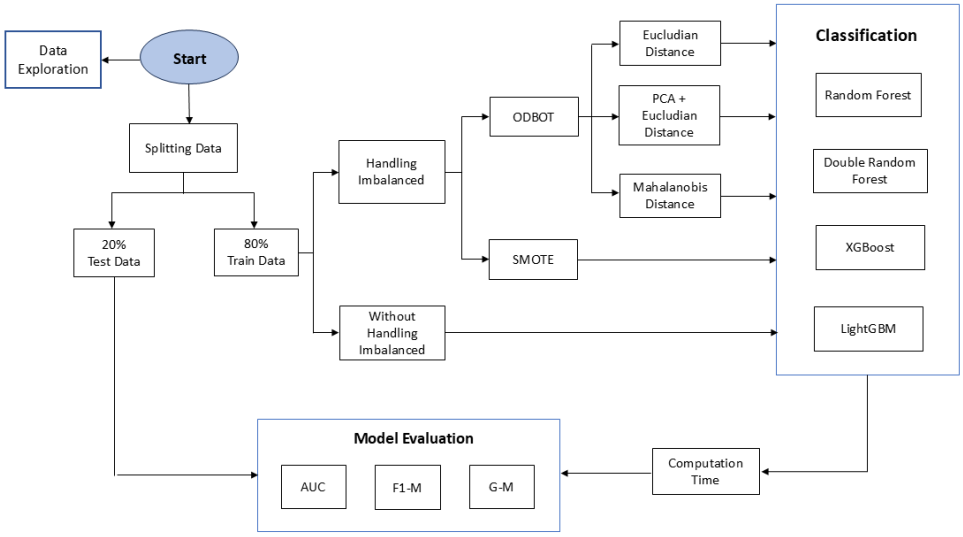

Figure 1. Flowchart of Procedure Analysis

2.1 Dataset

This study used empirical data for the year 2023, obtained from the official website of Badan Pusat Statistik (BPS). The dataset consists of 514 observations representing all regencies and municipalities across Indonesia, with the HDI serving as the target variable (Y), which is classified into four categories. There are four predictor variables used in this study. The rationale for selecting these specific variables is based on their relevance to the HDI and their ability to capture key aspects of regional development. A detailed description of the variables and their data types is presented in Table 1.

Table 1. Description of the Target Variable and Predictor Variables

Variables	Description	Type of Data
Y	HDI	Ordinal
X_1	Percentage of Poor Population	Continuous
X_2	Open Unemployment Rate (TPT)	Continuous
X_3	Labor Force Participation Rate (TPAK)	Continuous
X_4	Percentage of Households with Access to Improved Drinking Water Sources	Continuous

According to BPS [5], the level of human development in a given region over a specific period can be classified into four categories as follows:

1 = Very High: HDI ≥ 80;

2 = High: 70 ≤ HDI < 80;

3 = Intermediate: 60 ≤ HDI < 70;

4 = Low: HDI < 60.

2.2 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE is an oversampling technique that generates synthetic samples for the minority class by interpolating between each minority instance and its k-nearest neighbors (k-NN) within the feature space. By considering the local neighborhood of minority instances, SMOTE effectively increases the representation of underrepresented classes while preserving the underlying data distribution. The samples generated are not exactly similar to the original minority class distribution, but the method remains widely accepted because it is incredibly efficient in dealing with class imbalance in datasets [17].

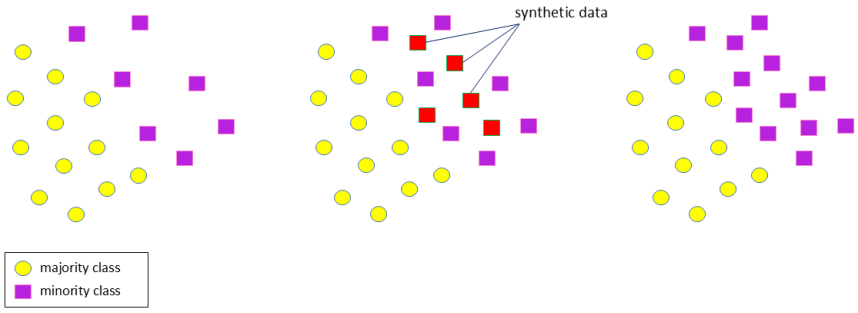

Figure 2. Illustration of SMOTE

The SMOTE algorithm integrates an oversampling rate parameter that controls the count of synthetic samples produced for each minority class observation. This parameter defines the proportion of additional synthetic data produced during oversampling [18].

2.3 Outlier Detection-Based Oversampling Technique (ODBOT)

ODBOT was initially proposed by Ibrahim in 2021 [14] as an oversampling method specifically designed to address two critical challenges in imbalanced learning: class overlap and the generation of effective synthetic samples for the minority class. Compared to binary classification, the problem becomes more complex in multiclass imbalanced datasets, where multiple minority classes may be involved. Nevertheless, most existing oversampling techniques have not been optimally adapted to handle such multiclass scenarios. ODBOT was explicitly developed to overcome this limitation and provide a more effective solution for multiclass imbalanced data. The main steps of the ODBOT method are as follows:

1. Identification of class-wise instance counts: The number of cases in each class is identified to quantify class imbalance.
2. Calculation of the number of synthetic samples (NSS): The number of synthetic instances required for each minority class is computed based on the imbalance ratio.
3. Clustering of minority and majority classes: Both minority and majority class instances are clustered with two clusters using the Weighted Bat Algorithm K-Means (WBBA-KM) method, which utilizes Euclidean distance as the distance metric [19]. Cluster centers for each class are then determined.
4. Selection of the optimal minority cluster: *The Sum of the Minority Cluster Distance*s (SMCD) is calculated to evaluate the separation between each minority cluster ($MinCen$) and the centers of the majority clusters ($MajCen$). The SMCD is defined as follows:

$$SMCD_{i,j} = \sum_{l=1}^{k} \left| MinCen_{i,j} - MajCen_{l} \right|, j = 1,2, \dots, k. \quad (1)$$

5. The cluster with the highest SMCD value is selected as the optimal minority cluster: $Max\left\{SMCD_{i,1}, SMCD_{i,2}, \dots, SMCD_{i,k}\right\}$. Conversely, a low SMCD value indicates that the corresponding minority cluster may be an outlier susceptible to class overlap issues [14].
6. Generation of synthetic samples: Based on the calculated NSS, synthetic data points are generated within the limits of the selected optimal minority cluster using the following interpolation formula:

$$C = a * (max_{minor} - min_{minor}) + min_{minor}, \quad (2)$$

where $a \in [0,1]$: a randomly generated scalar, max_{minor} : maximum values of the optimal minority cluster and min_{minor} : minimum values of the optimal minority cluster.

2.4 Mahalanobis Distance

Mahalanobis distance is a statistical measure introduced by Prasanta Chandra Mahalanobis in 1936, which is used to determine the distance between a point and a multivariate data distribution. In the original ODBOT framework, Euclidean distance is used during the clustering and outlier detection steps to measure the distance between instances. In the modified Mahalanobis-based ODBOT, Mahalanobis distance replaces

Euclidean distance, specifically in these steps, allowing the algorithm to account for correlations among features and differences in scale when identifying clusters and detecting outliers [20]. The Mahalanobis distance is mathematically defined as follows [21]:

$$D_M(x) = \sqrt{(x-\mu)^T\, \theta^{-1}(x-\mu)}, \quad (3)$$

where $\boldsymbol{x}$: an observation vector, $\boldsymbol{\mu}$: the mean vector, dan $\boldsymbol{\theta^{-1}}$: an inverse covariance matrix of the observation.

2.5 Random Forest (RF)

Random Forest is an ensemble learning method composed of multiple base learners in the form of decision trees, constructed using the bagging technique and the random subspace method. These techniques introduce diversity among the individual models, where each tree selects the optimal partition from a randomly chosen subset of features at each non-leaf node [22].

2.6 Double Random Forest (DRF)

DRF constructs each decision tree using the entire training dataset, in contrast to the traditional Random Forest, which relies on randomly drawn bootstrap samples. As a result, trees in Double Random Forest (DRF) tend to be larger, exhibiting greater depth and a higher number of nodes, due to the inclusion of more unique instances in the training process. DRF also applies bootstrap sampling and random feature subset selection at each node to determine the optimal splitting rule, after which the original data are passed down to the child nodes. This approach helps to reduce bias and enhance tree diversity [23].

2.7 Extreme Gradient Boosting (XGBoost)

XGBoost was developed as a scalable machine learning system for tree boosting. The primary advantage of XGBoost is the incorporation of a regularization component into the loss function, which considers both the complexity of the resulting ensemble and the predictability of each split. Additionally, XGBoost enables users to mitigate model overfitting by adjusting various hyperparameters, including single tree complexity, forest complexity, learning rate, regularization parameters, column subspace, dropout, and others [24].

2.8 Light Gradient Boosting Machine (LightGBM)

LightGBM is a type of gradient boosting that refers to a light version. This algorithm is designed for faster and more efficient tree-based learning methods. In the gradient boosting framework, trees are built sequentially, unlike random forests, which build trees independently for each example. LightGBM uses a leaf-wise tree development algorithm, where splitting is performed at the leaf level if the tree is unbalanced. Specifically, information gain is used to determine the split at each node. Additionally, the GOSS (Gradient-based One-Side Sampling) function in LightGBM determines the split point based on variance gain, which helps improve efficiency in the tree splitting process [25].

2.9 Confusion Matrix of Multiclass

The confusion matrix in binary classification has a dimension of 2×2, while in multiclass classification, it has a dimension of $N \times N$, where N is the number of different class labels, such as $K_1, K_2, \ldots, K_N$ [26]. The True Positive (TP) and True Negative (TN) values indicate the number of positive and negative test examples that the model correctly classified. Meanwhile, the False Negative (FN) and False Positive (FP) values indicate the number of positive and negative test examples that the model incorrectly classified. The main objective in a classification model is to minimize the FN and FP values. In addition, classification errors occur when the actual class does not match the predicted class. Table 2 shows the confusion matrix for multiclass classification.

Table 2. Confusion Matrix of Multiclass Classification

		Prediction			
		C_1	C_2	…	C_N
Actual	C_1	$C_{1,1}$	FP	…	$C_{1,N}$
	C_2	FN	TP	…	FN
	…	…	…	…	…
	C_N	$C_{N,1}$	FP	…	$C_{N,N}$

3. RESULTS AND DISCUSSION

3.1 Descriptive Data Analysis

Descriptive data analysis is needed prior to modeling, giving an understanding of the characteristics of the data. During this stage, the preliminary exploration is done to analyze the structure of data, the distribution of each independent variable, and class proportions of the response variable. The analysis results show that there are no missing values in the dataset, ascertaining its completeness. Additionally, Table 3 summarizes the distribution of values for each predictor variable.

Table 3. Summary of Descriptive Statistics for Predictor Variables

	Minimum	Maximum	Median	Mean
X_1	2.27	40.01	9.62	11.50
X_2	0.05	11.65	4.02	4.34
X_3	42.81	96.60	69.63	70.20
X_4	2.00	100.00	91.17	86.65

Based on Table 3, variables X_1, X_2, X_3, and X_4 have different value ranges, with the most noticeable differences observed in X_2 and X_4. It is essential to consider that distance-based methods, such as the Euclidean distance used in the ODBOT algorithm, are sensitive to the scale of the variables. Tightly ranged variables have the potential to dominate distance calculation, thereby causing biased synthetic minority oversamples in oversampling [27]. In addition, X_1, X_2, and X_3 variables have greater mean values than their medians, indicating positively skewed distributions. Such a pattern typically occurs when most values are found in a relatively low range, lower than the mean, with high values being rare but tending to pull the mean towards a higher value. However, the differences between the mean and median for X_2 and X_3 are relatively small, suggesting mild skewness. In contrast, the X_4 variable shows a changing pattern of distribution where the mean is smaller than the median and indicates a negatively skewed distribution. This shows that while most areas are well-served with high levels of improved drinking water access, some areas remain significantly underserved.

The asymmetric distribution patterns in the dataset should be explicitly addressed when developing social classification models. Strategies such as stratified sampling, oversampling techniques (e.g., SMOTE or ODBOT), or applying appropriate class weights in the loss function can mitigate class imbalance, thereby maintaining both interpretability and predictive accuracy. In addition to the characteristics of each predictor variable, the distribution of class proportions in the target variable (Y) also needs to be examined to understand the underlying class imbalance that presents a challenge in classification tasks. The detailed class proportions are presented in Fig. 3.

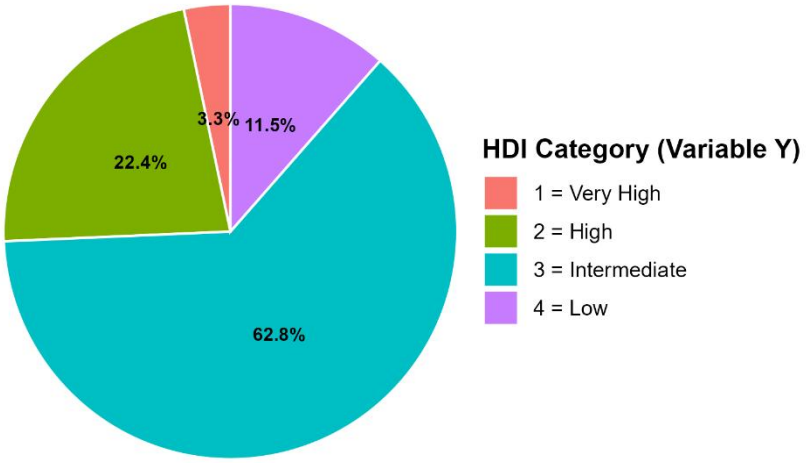

Figure 3. Proportion of Target Variable (Y)

Fig. 3 illustrates the distribution of classes in the target variable spread over four classes. Category 3 is the most common with 323 observations (62.8%), followed by category 2 with 115 observations (22.4%), category 4 with 59 observations (11.5%), and category 1 is the least represented class with only 17 observations (3.3%) out of the total 514 observations. Based on the proportion of the majority class (323) divided by the smallest minority class (17), an imbalance ratio (IR) of 19.0 was obtained, indicating a class

imbalance because IR > 1. This significant class imbalance risks reducing the model's responsiveness to minority classes, as the model tends to focus predictions on the majority class [28]. The study showed that learning algorithms tend to overgeneralize toward the majority class without balancing intervention, thus ignoring crucial minority classes. Adaptive oversampling and cost-sensitive techniques have also been shown to improve the performance of minority classes by improving metrics such as recall and G-mean [29].

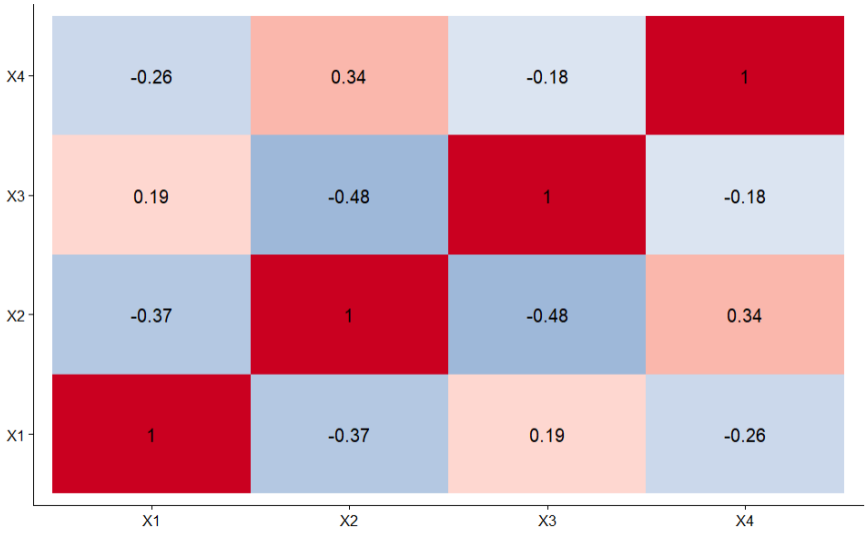

Figure 4. Correlation Matrix Value of Predictor Variables

Four predictor variables in this study are continuous; therefore, correlation analysis is required to identify multicollinearity. As seen in Fig. 4, the correlation among the independent variables ranges from -0.48 to 0.34, indicating a weak to moderate positive or negative linear relationship. Therefore, there is no multicollinearity among the variables at large. According to Shrestha [30], intercorrelations between independent variables below the ±0.70 level do not generally pose a multicollinearity problem in prediction analysis.

3.2 Performance Evaluation of the ODBOT Method

After descriptive analysis, the next step involved splitting the dataset into two subsets: 80% for training and 20% for testing. An imbalanced data handling procedure was applied to tree-based classifiers, including Random Forest, Double Random Forest, XGBoost, and LightGBM. The evaluation employed 5-fold cross-validation repeated 30 times independently, resulting in 150 train-test evaluations for each classifier to ensure stable and reliable performance estimates. All models in this study were developed using default parameter settings, without performing hyperparameter tuning. This constitutes a significant limitation, as it may prevent the models from achieving optimal predictive performance and therefore weakens the generalizability and strength of the study's conclusions. Although the models built from each algorithm do not necessarily reflect their optimal performance, standard parameters are intended to illustrate the relative competitiveness of the classification performance across different decision tree-based algorithms. One of the main challenges in distance-based ODBOT is its sensitivity to data scale and correlation between variables. To address this, this study modified ODBOT by using Mahalanobis distance in the WBBA-KM clustering algorithm, as outlined in Eq. (3), which previously employed Euclidean distance. Unlike Euclidean distance, Mahalanobis distance considers the correlation between features. It scales the distance according to the variance of each feature, as well as the creation of relevant synthetic samples in unbalanced datasets. PCA was also applied before Euclidean-based ODBOT as an alternative to reduce data dimensionality and correlation between variables without losing essential information. Therefore, three ODBOT approaches, namely Euclidean-based ODBOT, Mahalanobis-based ODBOT, and PCA + Euclidean-based ODBOT, are compared to evaluate the effectiveness of each in handling data imbalance.

Table 4. Eigen Values and Variation in PCA Components

Data Dimension	Eigen Value	Percentage of Variance	Cumulative Percentage of Variance
Dim 1	2.0635	51.59	51.59

Dim 2	0.7767	19.42	71.00
Dim 3	0.6401	16.00	87.01
Dim 4	0.5198	12.99	100.00

Table 4 presents the PCA results, showing the cumulative variance explained by each principal component. The first four components (Dim 1–Dim 4) account for 51.59%, 71.00%, 87.01%, and 100.00% of the total variance, respectively, and were used as input for subsequent analyses, including modeling and oversampling procedures.

Table 5. Average Performance Measures of Euclidean-based ODBOT, Mahalanobis-based ODBOT, and Euclidean-based ODBOT combined with PCA

Models	Matrix Evaluation	ODBOT		
		Euclidean	Mahalanobis	PCA + Euclidean
RF	F1-M	0.5698	0.5988	**0.6353**
	G-M	0.6843	**0.7139**	0.6842
	AUC	0.8626	**0.8720**	0.8702
DRF	F1-M	**0.5368**	0.5319	0.4743
	G-M	**0.6872**	0.6808	0.5869
	AUC	0.8351	**0.8405**	0.7789
XGBoost	F1-M	0.5802	0.6002	**0.6344**
	G-M	0.6972	**0.7186**	0.7020
	AUC	0.8605	**0.8718**	0.8531
LightGBM	F1-M	0.5793	0.6053	**0.6163**
	G-M	0.6948	**0.7224**	0.6653
	AUC	0.8618	**0.8763**	0.8598
Computation Time (Seconds)		68.23	2545.66	**7.21**

Based on the performance presented in Table 5, the values that were previously highlighted indicate the best performance for each evaluation metric (F1-M, G-M, AUC) among the three ODBOT methods compared: Euclidean-based ODBOT, Mahalanobis-based ODBOT, and PCA + Euclidean ODBOT. In other words, these highlighted values represent which ODBOT method achieved the optimal result for a specific metric and model. Mahalanobis-based ODBOT performance consistently produced the highest AUC values across all the tree-based classification models used, demonstrating its strong ability to separate classes effectively. In addition, the G-M values achieved with the Mahalanobis-based ODBOT were higher than those obtained by the other two approaches for most models, except for DRF. This indicates that the Mahalanobis distance-based ODBOT technique is better at maintaining classification balance between minority and majority classes.

However, the PCA + Euclidean-based ODBOT technique produced the highest F1-M values for the RF, XGBoost, and LightGBM models. This means that applying PCA as a preprocessing step before the Euclidean-based ODBOT can improve the model's ability to identify minority classes while still retaining essential information from the original data. Regarding efficiency, the PCA + Euclidean-based ODBOT approach demonstrated significant superiority with a computation time of only 7.21 seconds. This value is much lower than the Mahalanobis-based ODBOT computation time (2545.66 seconds) and even the standard Euclidean-based ODBOT (68.23 seconds). The Mahalanobis distance is computationally intensive due to the calculation of covariance matrices and the need for structure-dependent matrix inversions based on inter-feature relationships. This computational complexity increases processing time and resource requirements, which should be considered when applying Mahalanobis-based ODBOT in practical classification tasks.

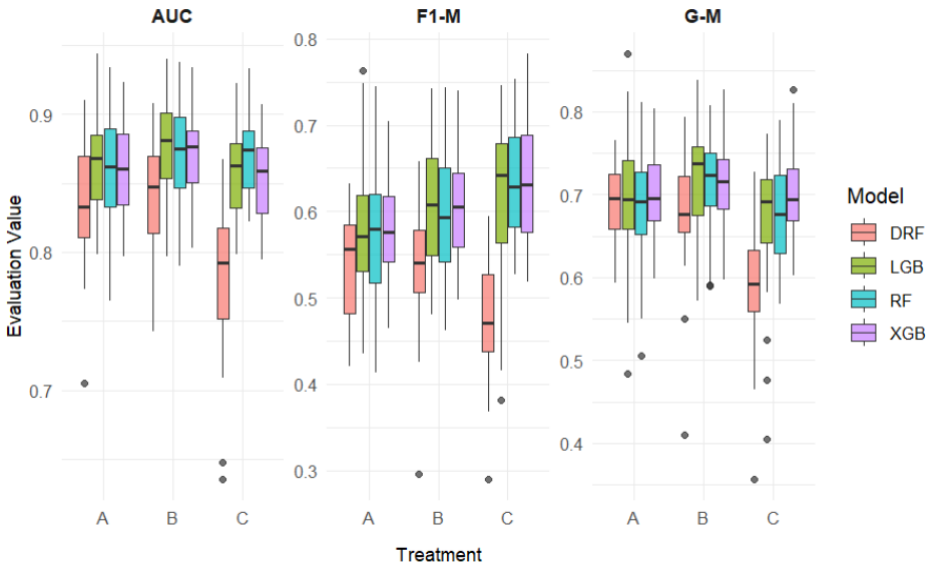

Figure 5. Boxplot of Metric Evaluation Distribution Based on Model and Handling of Imbalanced Class (A: Euclidean-based ODBOT, B: Mahalanobis-based ODBOT)

Fig. 5 shows how the Mahalanobis-based ODBOT consistently produces the highest median and a more uniform distribution across all models for the AUC measure, which indicates its best ability to distinguish classes. Conversely, according to the F1-M value, the PCA + Euclidean-based ODBOT approach performs better than the RF, XGB, and LGB models. It indicates the use of PCA in augmenting minority class identification. However, its performance is unstable and poor in the DRF model due to a profound fall in the median and outliers. In the G-M metric, the Mahalanobis-based ODBOT method performs better with the maximum median value in almost all models. Whereas PCA + Euclidean-based ODBOT achieves competitive performance in some models, its performance becomes less stable for DRF.

3.3 Performance Evaluation of NONE, SMOTE, and ODBOT Method

The following presents an analysis of the performance results of the classification model with various treatments for handling imbalanced data, namely without treatment (NONE), SMOTE, and ODBOT with the Euclidean, Mahalanobis, and PCA + Euclidean approaches.

Table 6. Average Performance Metrics of All Treatments

Models	Matrix Evaluation	NONE	SMOTE	ODBOT		
				Euclidean	Mahalanobis	PCA + Euclidean
Random Forest	F1-M	0.6392	**0.6421**	0.5698	0.5988	0.6353
	G-M	0.6849	**0.7773**	0.6843	0.7139	0.6842
	AUC	0.8840	**0.9012**	0.8626	0.8720	0.8702
Double Random Forest	F1-M	**0.5888**	0.5555	0.5368	0.5319	0.4743
	G-M	0.6551	**0.7321**	0.6872	0.6808	0.5869
	AUC	0.8493	**0.8707**	0.8351	0.8405	0.7789
XGBoost	F1-M	0.6278	0.5840	0.5802	0.6002	**0.6344**
	G-M	0.6907	**0.7420**	0.6972	0.7186	0.7020
	AUC	0.8683	**0.8831**	0.8605	0.8718	0.8531
LightGBM	F1-M	**0.6250**	0.5947	0.5793	0.6053	0.6163
	G-M	0.6703	**0.7626**	0.6948	0.7224	0.6653
	AUC	0.8767	**0.8980**	0.8618	0.8763	0.8598

The bold values in Table 6 indicate that SMOTE generally outperforms other methods, particularly in terms of the Geometric Mean (G-M) and Area Under the Curve (AUC) metrics across most tree-based classifiers, highlighting its effectiveness in enhancing classification balance and inter-class discrimination. However, for the F1-Measure (F1-M), the XGBoost and LightGBM models achieve higher scores without any imbalanced data handling, suggesting that oversampling does not always improve both precision and recall simultaneously. This outcome highlights the importance of carefully selecting oversampling techniques based on the specific performance metrics prioritized in the study, rather than assuming uniform improvement

across all evaluation criteria. Meanwhile, the Mahalanobis-based ODBOT provides competitive results, especially on G-M, although it does not consistently outperform SMOTE. This suggests that the selection of handling methods should be tailored to the characteristics of the data and the desired evaluation priorities. Next, an analysis of variance (ANOVA) was performed on the three evaluation metrics (F1-M, G-M, and AUC) using a factorial randomized design to determine the differences between treatments. The first factor was the classification model, and the second factor was the data balancing technique, with an α value of 0.10. The null hypothesis (H_0) is that there are no significant differences in the classification model, data balancing technique, and interaction between the two factors.

Table 7. Analysis of Variance for The HDI Data on The F1-M Metric

Source of Variation	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Classification Models	3	0.2444	0.0815	17.698	6.82e-11***
Balancing Techniques	3	0.1811	0.0604	13.113	3.12e-08***
Interaction	9	0.0620	0.0069	1.495	0.147
Error	464	2.1362	0.0046		

significant code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Based on Table 7, the results of the variance analysis based on the F1-M evaluation metric obtained p-values in the classification model (6.82e-11) and balancing technique (3.12e-08) < α (0.10), meaning that H_0 is rejected, so there is a significant difference in the F1-M average value in the classification model and the balancing technique. Therefore, the Duncan post-hoc test was conducted to determine the specific differences between the classification model and the data balancing technique.

Table 8. Duncan Test Results for The F1-M Metric of The HDI Data

Classification Models	Average	Groups
RF	0.6125	a
LightGBM	0.6011	a
XGBoost	0.5980	a
DRF	0.5532	b
Balancing Techniques		
NONE	0.6202	a
SMOTE	0.5941	b
ODBOT Mahalanobis (M-ODBOT)	0.5841	b
ODBOT Euclidean	0.5665	c

The results of Duncan's test of the classification model factors in Table 8 show that in the HDI data, the RF model is the best model with the highest average F1-M, but there is no significant difference between the LightGBM and XGBoost models, while the DRF model shows the lowest average F1-M value and has a significant difference from the three classification models. Meanwhile, further test results on the balancing technique factor show that M-ODBOT is not significantly different from SMOTE. In HDI data, without imbalance handling, it produces the highest average F1-M value and is significantly different from the three data balancing techniques.

Table 9. Analysis of Variance for The HDI Data on the G-M Metric

Source of Variation	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Classification Models	3	0.0545	0.0182	5.101	0.0018**
Balancing Techniques	3	0.4120	0.1374	38.587	<2e-16***
Interaction	9	0.0417	0.0046	1.301	0.2339
Error	464	1.6516	0.0036		

significant code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Furthermore, Table 9 shows that analysis of variance based on the G-M evaluation metric obtained a p-value in the classification model (0.0018) and balancing technique (<2e-16) < α (0.10), thus, rejecting H_0, meaning that there is a significant difference in the G-M mean values in the classification model and the balancing technique. Table 10 shows the results of Duncan's post hoc test for the classification model and balancing technique.

Table 10. Duncan Test Results for the G-M Metric of The HDI Data

Classification Models	Average	Groups
RF	0.7151	a
LightGBM	0.7125	a
XGBoost	0.7121	a
DRF	0.6888	b
Balancing Techniques		
SMOTE	0.7535	a
ODBOT Mahalanobis (M-ODBOT)	0.7089	b
ODBOT Euclidean	0.6909	c
NONE	0.6753	d

Further analysis based on the G-M value in the classification model did not reveal any significant differences from the F1-M value, indicating that there was no significant difference between the RF, LightGBM, and XGBoost models, with the RF model showing the best performance. Meanwhile, the DRF model obtained the smallest average G-M value and differed significantly from the three classification models. Based on the balancing technique, the highest average G-M value was found in the SMOTE technique, which was significantly different from the other techniques. M-ODBOT followed, showing the second-best performance. The analysis results also show that the technique without data imbalance handling produces the smallest average G-M value, indicating that data imbalance handling can improve the model's ability to recognize both majority and minority classes in low HDI areas. Overall, there are differences between balancing techniques based on the G-M metric.

Table 11. Analysis of Variance for The HDI Data on The AUC Metric

Source of Variation	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Classification Models	3	0.0732	0.0244	22.212	1.87e-13***
Balancing Techniques	3	0.0698	0.0233	21.176	7.15e-13***
Interaction	9	0.0053	0.0006	0.535	0.849
Error	464	0.5099	0.0011		

significant code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of variance was also performed on the AUC evaluation metric. Based on Table 11, the p-value obtained for the classification model (1.87e-13) and balancing technique (7.15e-13) was < α (0.10), meaning that H_0 was rejected, indicating a significant difference in the mean AUC values between the classification model and the balancing technique in the HDI data. Therefore, further analysis was conducted to refine the classification model and optimize the balancing technique.

Table 12. Duncan Test Results for The AUC Metric of The HDI Data

Classification Models	Average	Groups
RF	0.8799	a
LightGBM	0.8782	a
XGBoost	0.8709	b
DRF	0.8489	c
Balancing Techniques		
SMOTE	0.8883	a
NONE	0.8696	b
ODBOT Mahalanobis (M-ODBOT)	0.8652	b
ODBOT Euclidean	0.8550	c

Based on the AUC value, further analysis in Table 12 reveals that there is no significant difference between the RF and LightGBM models. However, both models differ significantly from the XGBoost and DRF models. Furthermore, further analysis of the balancing techniques shows that the SMOTE technique obtained the highest average AUC and was significantly different from the other techniques. On the other hand, the performance of M-ODBOT yielded a higher average AUC than that of ODBOT Euclidean.

3.4 Discussions

The present results underscore the crucial importance of selecting suitable strategies for addressing class imbalance in datasets, particularly when employing tree-based classification algorithms. The Mahalanobis-based ODBOT demonstrated superior class separability, achieving higher Area Under the

Curve (AUC) and Geometric Mean (G-M) scores compared to Euclidean-based ODBOT methods. These observations are consistent with the findings of Yao and Lin [31], which demonstrated that the Mahalanobis distance enhances sensitivity to covariance structures among variables, thereby improving classification performance in imbalanced settings. Recent studies further corroborate that Mahalanobis-based approaches effectively capture internal data structure in multivariate contexts, particularly when feature correlations are heterogeneous [32], [33].

Conversely, the PCA + Euclidean-based ODBOT achieved the highest F1-Measure scores across several classifiers, including Random Forest, XGBoost, and LightGBM, while reducing computational cost relative to the Mahalanobis-based variant. This outcome illustrates a practical trade-off between class separability and computational efficiency. The utility of PCA in reducing dimensionality and enhancing model generalization has been well documented [34], and combining PCA with oversampling has proven effective in other contexts; for example, PCA + SMOTE has been shown to improve classifier performance for high-dimensional imbalanced datasets [35]. Additional oversampling strategies also support these findings. GMOTE, which adapts the Mahalanobis distance to better account for local outliers and minority class tail distributions, improves classification metrics relative to standard SMOTE [36]. Similarly, MAHAKIL integrates k-means clustering with Mahalanobis distance within a genetic algorithm framework, generating synthetic minority samples that preserve both density and covariance structure [37].

These insights have practical implications. When the priority is to maximize class separation and sensitivity, such as correctly identifying underdeveloped regions in the Human Development Index (HDI), the Mahalanobis-based ODBOT is a strong candidate due to its sensitivity to covariance structures. Conversely, in operational settings constrained by computational resources or when F1-Measure is prioritized, PCA + Euclidean-based ODBOT offers a more efficient trade-off by reducing dimensionality and computational burden [38], [39]. (Nevertheless, the study has several limitations. First, all models were trained using default hyperparameters, which may have constrained the performance of both oversampling techniques and classifiers. Future studies should incorporate hyperparameter optimization via grid search, Bayesian optimization, or evolutionary algorithms to maximize the performance of both Mahalanobis-based oversampling and tree-based classifiers [40]. Second, the computation of Mahalanobis distance requires inversion of the covariance matrix, imposing a substantial computational burden, especially for larger or high-dimensional datasets [41]. Third, the current work focused exclusively on data-level balancing techniques. Integrating algorithm-level strategies, such as cost-sensitive learning, or hybrid approaches combining Mahalanobis-based oversampling with optimization techniques, could further enhance performance and generalizability [42]. Future research should also evaluate the scalability of Mahalanobis-based methods on larger, more complex datasets and explore hybrid strategies that combine data-level and algorithm-level balancing. Such approaches, including cost-sensitive ensemble learning or adaptive oversampling, may provide robust performance across diverse real-world applications.

4. CONCLUSION

This study demonstrates that utilizing HDI data, the Mahalanobis-based ODBOT approach is more effective in recognizing minority and majority classes than the Euclidean-based ODBOT. However, this increase in accuracy comes at the cost of greater computational load. Combining PCA and Euclidean-based ODBOT improves computational time efficiency. Meanwhile, the application of SMOTE consistently improves the G-M and AUC metrics across all decision tree-based models. In addition, the results of the variance analysis show that the performance of the RF, LightGBM, and XGBoost classification models does not show significant differences in the F1-M and G-M evaluation metrics. Meanwhile, the data balancing technique produces different performances for each evaluation metric. However, the M-ODBOT performance produces a higher average value for the three evaluation metrics than the Euclidean ODBOT.

Author Contributions

Yunna Mentari Indah: Conceptualization, methodology, Writing-Original Draft, Data Curation, Software, Visualization. Anwar Fitrianto: Conceptualization, Resources, Draft Preparation, Validation, and Writing–Review and Editing. Indahwati: Formal Analysis, Validation, and Writing–Review and Editing. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was supported by the Department of Statistics and Data Science, School of Data Science, Mathematics, and Informatics (SSMI). The author also acknowledges financial support from the Indonesia Endowment Fund for Education (LPDP), Ministry of Finance of the Republic of Indonesia.

Acknowledgement

The authors extend sincere thanks to the School of Data Science, Mathematics and Informatics, IPB University, for the kind support and assistance that led to this article. Also, thanks to the reviewers and other parties who gave some constructive suggestions to make this article more significant and contribute to academic things.

Declarations

The authors declare no competing interests.

Declaration of Generative AI and AI-assisted Technologies

ChatGPT was used exclusively to improve readability and grammatical structure. No AI tool was used to generate, modify, or influence the research data, methodology, results, or interpretations. All content was verified by the authors for accuracy and consistency with the study.

REFERENCES

[1] B. Liu and G. Tsoumakas, "DEALING WITH CLASS IMBALANCE IN CLASSIFIER CHAINS VIA RANDOM UNDERSAMPLING," *Knowledge-Based Syst.*, vol. 192, p. 105292, 2020. doi: https://doi.org/10.1016/j.knosys.2019.105292.

[2] C. O. Vázquez, S. vanden Broucke, and J. De Weerdt, *HELLINGER DISTANCE DECISION TREES FOR PU LEARNING IN IMBALANCED DATA SETS*, vol. 113, no. 7. Springer US, 2024. doi: https://doi.org/10.1007/s10994-023-06323-y.

[3] UNDP, *Human Development Report 2025: A MATTER OF CHOICE: PEOPLE AND POSSIBILITIES IN THE AGE OF AI*. New York: United Nations Development Programme, 2025. [Online]. Available: https://hdr.undp.org/content/human-development-report-2025. doi: https://doi.org/10.2139/ssrn.5353261

[4] Y. M. Indah, R. Aristawidya, A. Fitrianto, E. Erfiani, and L. M. R. D. Jumansyah, "COMPARISON OF RANDOM FOREST, XGBOOST, AND LIGHTGBM METHODS FOR THE HUMAN DEVELOPMENT INDEX CLASSIFICATION," *Jambura J. Math.*, vol. 7, no. 1, pp. 14–18, 2025. doi: https://doi.org/10.37905/jjom.v7i1.28290.

[5] Badan Pusat Statistik, "INDEKS PEMBANGUNAN MANUSIA 2023," vol. 18, pp. 1–282, 2023.

[6] J. H.-Osorio, A. A.-Meza, G. D.-Santacoloma, A. O.-Gutierrez, and G. C.-Dominguez, "RELEVANT INFORMATION UNDERSAMPLING TO SUPPORT IMBALANCED DATA CLASSIFICATION," *Neurocomputing*, vol. 436, pp. 136–146, May 2021. doi: https://doi.org/10.1016/j.neucom.2021.01.033.

[7] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "CLASSIFICATION OF IMBALANCED DATA: REVIEW OF METHODS AND APPLICATIONS," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, p. 012077, 2021. doi: https://doi.org/10.1088/1757-899X/1099/1/012077.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002. doi: https://doi.org/10.1613/jair.953.

[9] V. P. K. Turlapati and M. R. Prusty, "OUTLIER-SMOTE: A REFINED OVERSAMPLING TECHNIQUE FOR IMPROVED DETECTION OF COVID-19," *Intell. Med.*, vol. 3–4, no. November, p. 100023, 2020. doi: https://doi.org/10.1016/j.ibmed.2020.100023.

[10] P. Gnip, L. Vokorokos, and P. Drotár, "SELECTIVE OVERSAMPLING APPROACH FOR STRONGLY IMBALANCED DATA," *PeerJ Comput. Sci.*, vol. 7, pp. 1–22, 2021. doi: https://doi.org/10.7717/peerj-cs.604.

[11] S. Lusito, A. Pugnana, and R. Guidotti, *SOLVING IMBALANCED LEARNING WITH OUTLIER DETECTION AND FEATURES REDUCTION*, vol. 113, no. 8. Springer US, 2024. doi: https://doi.org/10.1007/s10994-023-06448-0.

[12] B. Mirzaei, B. Nikpour, and H. Nezamabadi-Pour, "AN UNDER-SAMPLING TECHNIQUE FOR IMBALANCED DATA CLASSIFICATION BASED ON DBSCAN ALGORITHM," in *2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS)*, IEEE, Sep. 2020, pp. 21–26. doi: https://doi.org/10.1109/CFIS49607.2020.9238718.

[13] H. Zhou, J. Tong, Y. Liu, K. Zheng, and C. Cao, "AN OVERSAMPLING FCM-KSMOTE ALGORITHM FOR IMBALANCED DATA CLASSIFICATION," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 10, p. 102248, Dec. 2024. doi: https://doi.org/10.1016/j.jksuci.2024.102248.

[14] M. H. Ibrahim, "ODBOT: OUTLIER DETECTION-BASED OVERSAMPLING TECHNIQUE FOR IMBALANCED DATASETS LEARNING," *Neural Comput. Appl.*, vol. 33, no. 22, pp. 15781–15806, 2021. doi: https://doi.org/10.1007/s00521-021-06198-x.

[15] P. Kumari and S. Gupta, "COMPARATIVE ANALYSIS BETWEEN EUCLIDEAN DISTANCE METRIC AND MAHALANOBIS DISTANCE METRIC," *Int. J. Innov. Res. Technol. Sci. www.ijirts.org*, vol. 12, no. 2, 2024, [Online]. Available: www.ijirts.org

[16] F. Wang, M. Zheng, K. Ma, and X. Hu, "RESAMPLING APPROACH FOR IMBALANCED DATA CLASSIFICATION BASED ON CLASS INSTANCE DENSITY PER FEATURE VALUE INTERVALS," *Inf. Sci. (Ny).*, vol. 692, p. 121570, Feb. 2025. doi: https://doi.org/10.1016/j.ins.2024.121570.

[17] D. Elreedy, A. F. Atiya, and F. Kamalov, "A THEORETICAL DISTRIBUTION ANALYSIS OF SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) FOR IMBALANCED LEARNING," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024. doi: https://doi.org/10.1007/s10994-022-06296-4.

[18] K. S. Raslan, A. S. Alsharkawy, and K. R. Raslan, "IHHO-SMOTE: A CLEANSED APPROACH FOR HANDLING OUTLIERS AND REDUCING NOISE TO IMPROVE IMBALANCED DATA CLASSIFICATION," *Int. J. Comput. Appl.*, vol. 186, no. 32, pp. 975–8887, 2024. doi: https://doi.org/10.5120/ijca2024923849.

[19] M. H. Ibrahim, "WBBA-KM: A HYBRID WEIGHT-BASED BAT ALGORITHM WITH K-MEANS ALGORITHM FOR CLUSTER ANALYSIS," *Politek. Derg.*, vol. 25, no. 1, pp. 65–73, 2022. doi: https://doi.org/10.2339/politeknik.689384.

[20] T. R. Etherington, "MAHALANOBIS DISTANCES FOR ECOLOGICAL NICHE MODELLING AND OUTLIER DETECTION: IMPLICATIONS OF SAMPLE SIZE, ERROR, AND BIAS FOR SELECTING AND PARAMETERISING A MULTIVARIATE LOCATION AND SCATTER METHOD," *PeerJ*, vol. 9, 2021. doi: https://doi.org/10.7717/peerj.11436.

[21] K. Dashdondov and M.-H. Kim, "MAHALANOBIS DISTANCE BASED MULTIVARIATE OUTLIER DETECTION TO IMPROVE PERFORMANCE OF HYPERTENSION PREDICTION," *Neural Process. Lett.*, vol. 55, no. 1, pp. 265–277, Feb. 2023. doi: https://doi.org/10.1007/s11063-021-10663-y.

[22] M. A. Ganaie, M. Tanveer, P. N. Suganthan, and V. Snasel, "OBLIQUE AND ROTATION DOUBLE RANDOM FOREST," *Neural Networks*, vol. 153, pp. 496–517, Sep. 2022. doi: https://doi.org/10.1016/j.neunet.2022.06.012.

[23] S. Han, H. Kim, and Y. S. Lee, "DOUBLE RANDOM FOREST," *Mach. Learn.*, vol. 109, no. 8, pp. 1569–1586, 2020. doi: https://doi.org/10.1007/s10994-020-05889-1.

[24] O. Sagi and L. Rokach, "APPROXIMATING XGBOOST WITH AN INTERPRETABLE DECISION TREE," *Inf. Sci. (Ny).*, vol. 572, pp. 522–542, 2021. doi: https://doi.org/10.1016/j.ins.2021.05.055.

[25] M. J. Sai, P. Chettri, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "AN ENSEMBLE OF LIGHT GRADIENT BOOSTING MACHINE AND ADAPTIVE BOOSTING FOR PREDICTION OF TYPE-2 DIABETES," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, 2023. doi: https://doi.org/10.1007/s44196-023-00184-y.

[26] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "MULTICLASS CONFUSION MATRIX REDUCTION METHOD AND ITS APPLICATION ON NET PROMOTER SCORE CLASSIFICATION PROBLEM," 2021. doi: https://doi.org/10.3390/technologies9040081.

[27] A. Bolívar, V. García, R. Alejo, R. F.-Juárez, and J. S. Sánchez, "DATA-CENTRIC SOLUTIONS FOR ADDRESSING BIG DATA VERACITY WITH CLASS IMBALANCE, HIGH DIMENSIONALITY, AND CLASS OVERLAPPING," *Appl. Sci.*, vol. 14, no. 13, 2024. doi: https://doi.org/10.3390/app14135845.

[28] S. S. Rawat and A. K. Mishra, "REVIEW OF METHODS FOR HANDLING CLASS IMBALANCE IN CLASSIFICATION PROBLEMS," 2024, pp. 3–14. doi: https://doi.org/10.1007/978-981-97-0037-0_1.

[29] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "BOOSTING METHODS FOR MULTI-CLASS IMBALANCED DATA CLASSIFICATION: AN EXPERIMENTAL REVIEW," *J. Big Data*, vol. 7, no. 1, 2020. doi: https://doi.org/10.1186/s40537-020-00349-y.

[30] N. Shrestha, "DETECTING MULTICOLLINEARITY IN REGRESSION ANALYSIS," *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, Jun. 2020. doi: https://doi.org/10.12691/ajams-8-2-1.

[31] L. Yao and T. Lin, "EVOLUTIONARY MAHALANOBIS DISTANCE-BASED OVERSAMPLING FOR MULTI-CLASS IMBALANCED DATA CLASSIFICATION," *Sensors*, vol. 21, no. 19, p. 6616, Oct. 2021. doi: https://doi.org/10.3390/s21196616.

[32] S. Lusito, A. Pugnana, and R. Guidotti, "SOLVING IMBALANCED LEARNING WITH OUTLIER DETECTION AND FEATURES REDUCTION," *Machine Learning*, vol. 113, no. 8, pp. 5273–5330, 2024. doi: https://doi.org/10.1007/s10994-023-06448-0.

[33] I. Naglik and M. Lango, "GMMSAMPLING: A NEW MODEL-BASED, DATA DIFFICULTY-DRIVEN RESAMPLING METHOD FOR MULTI-CLASS IMBALANCED DATA," *Machine Learning*, vol. 113, no. 8, pp. 5183–5202, 2024. doi: https://doi.org/10.1007/s10994-023-06416-8.

[34] Y. Han and I. Joe, "ENHANCING MACHINE LEARNING MODELS THROUGH PCA, SMOTE-ENN, AND STOCHASTIC WEIGHTED AVERAGING," *Appl. Sci.*, vol. 14, no. 21, p. 9772, Oct. 2024. doi: https://doi.org/10.3390/app14219772.

[35] G. A. Mulla, Y. Demir, and M. Hassan, "COMBINATION OF PCA WITH SMOTE OVERSAMPLING FOR CLASSIFICATION OF HIGH-DIMENSIONAL IMBALANCED DATA," *Bitlis Eren Univ. Fen Bilimleri Dergisi*, vol. 10, no. 3, pp. 858–869, 2021. doi: https://doi.org/10.17798/bitlisfen.939733.

[36] S. J. Yang and K. J. Cha, "GMOTE: GAUSSIAN BASED MINORITY OVERSAMPLING TECHNIQUE FOR IMBALANCED CLASSIFICATION ADAPTING TAIL PROBABILITY OF OUTLIERS," *arXiv preprint arXiv:2105.03855*, 2021.

[37] Y. Zhang, T. Zuo, L. Fang, J. Li, and Z. Xing, "AN IMPROVED MAHAKIL OVERSAMPLING METHOD FOR IMBALANCED DATASET CLASSIFICATION," *IEEE Access*, vol. 9, pp. 16030–16040, 2020. doi: https://doi.org/10.1109/ACCESS.2020.3047741.

[38] I. D. Mienye and Y. Sun, "PERFORMANCE ANALYSIS OF COST-SENSITIVE LEARNING METHODS WITH APPLICATION TO IMBALANCED MEDICAL DATA," *Informatics in Medicine Unlocked*, vol. 25, p. 100690, 2021. doi: https://doi.org/10.1016/j.imu.2021.100690.

[39] B. Zhu, X. Jing, L. Qiu, and R. Li, "AN IMBALANCED DATA CLASSIFICATION METHOD BASED ON HYBRID RESAMPLING AND FINE COST SENSITIVE SUPPORT VECTOR MACHINE," *Computers, Materials & Continua*, vol. 79, no. 3, 2024. doi: https://doi.org/10.32604/cmc.2024.048062.

[40] T. Aswani, J. M. Gummadi, and G. Sharada, "A RANDOM FOREST-BASED MACHINE LEARNING FRAMEWORK WITH PCA, SMOTE, AND SHAP FOR EFFICIENT AND INTERPRETABLE CORONARY ARTERY DISEASE PREDICTION," *Informatica*, vol. 49, no. 22, 2025. doi : https://doi.org/10.31449/inf.v49i22.7998.

[41] N. G. Siddappa and T. Kampalappa, "IMBALANCE DATA CLASSIFICATION USING LOCAL MAHALANOBIS DISTANCE LEARNING BASED ON NEAREST NEIGHBOR," *SN Computer Science*, vol. 1, no. 2, p. 76, 2020. doi: https://doi.org/10.1007/s42979-020-0085-x.

[42] M. Fachrie, A. Musdholifah, and R. Pulungan, "EFFECTIVENESS OF DATA RESAMPLING AND ENSEMBLE LEARNING IN MULTICLASS IMBALANCE LEARNING," *Artificial Intelligence Review*, vol. 58, no. 12, p. 368, 2025. doi : https://doi.org/10.1007/s10462-025-11357-w.