Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Comparing Self-Paced Ensemble and RUSBoost for Imbalanced Poverty Classification in West Java*

Nur Andi Setiabudi[1‡], Bagus Sartono[2], Utami Dyah Syafitri[3], Komang Budi Aryasa[4]

[1,2,3] Study Program on Statistics and Data Science, IPB University, Indonesia
[4] Divisi Digital Business and Technology, Telkom Indonesia
[‡]corresponding author: nur.andi@apps.ipb.ac.id

Abstract

Class imbalance remains a major challenge in classification modelling that frequently leads to biased predictive models. This study aimed to compare two ensemble techniques based on an undersampling approach, namely Self-Paced Ensemble and RUSBoost, for handling imbalanced classification in poverty identification in West Java. The results suggested that RUSBoost consistently outperformed Self-Paced Ensemble across the most critical metrics. It showed better balance in classification outcomes. When the objective is to maximize the identification of poor households, the default threshold in the RUSBoost model was prefered. On the other hand, if precision is prioritized due to limited resources, the Youden Index threshold offers a better alternative. Given the overall evaluation metrics, RUSBoost with the default threshold was suggested as the most reliable and well-balanced option among the compared models for classifying poor households in West Java under imbalanced data condition.

Keywords: ensemble learning, imbalance classification, RUSBoost, Self-Paced Ensemble, undersampling

* Received: Jan 2025; Reviewed: Jun 2025; Published: Des 2025

1. Introduction

Poverty is a multidimensional problem that involves economic, social, and political aspects. The issue is because it relates to the most basic needs in life, and it is a global problem faced by many countries. Addressing this persistent challenge demands more effective and efficient solutions, particularly through the application of data analytical techniques to develop predictive and classification models. These models can utilize empirical data to identify populations at risk of poverty. In modeling context, classifying poverty status is challenging due to the imbalanced data problem.

Imbalance data is characterized by a disproportionate number of observations across different response classes. The imbalance ratio (IR), which quantifies the disparity between the majority and minority classes, can vary considerably depending on the specific application. An IR ranging from 100:1 to 1,000:1 is considered as extreme imbalance (Hasanin et al., 2019). In poverty data, non-poor households form the larger group or majority class, while poor ones are the smaller group or minority class. For instance, in West Java, Indonesia, 7,6% of household were classified as poor in 2023 (Badan Pusat Statistik, 2023). If not handled through appropriate techniques, it often leads to poor model performance. Such models typically exhibit a bias towards the majority class, as their training process is predominantly influenced by majority class, while treating minority class as noise. Consequently, the model struggles to accurately learn the distinctive characteristics of the minority class. Furthermore, global accuracy metrics can be misleading, as models may favor the majority class and score high accuracy (Rahmadini & Santoso, 2025), but they may still perform poorly on the minority class (Wang et al., 2021).

A variety of methodologies have been developed to mitigate the effects of class imbalance in classification. In general, these methodologies are divided into three groups (Altalhan et al., 2025). The first involves interventions at the data level, primarily through resampling techniques designed to rebalance class distributions. This includes oversampling, which augments the number of minority class observations, and undersampling, which reduces the number of majority class observations. While oversampling can be effective, it carries the risk of overfitting and often demands substantial computational resources (Zhang et al., 2021), hence less practical for large-scale datasets. On the other hand, while computationally lighter, undersampling risks the loss of important information from the majority class. The second category of approaches focuses on modifying classification algorithms to account for the significance of the minority class during the training process. An example is cost-sensitive learning, where different misclassification costs are assigned to each class. However, the pre-defined cost matrix, which typically needs expert domain knowledge for its accurate construction, is often not available (Liu et al., 2020). The final group consists of hybrid methods designed to combine the strengths of each approach while reducing their respective limitations. This includes combined sampling techniques, algorithmic resampling strategies, and ensembles of resampled datasets (Altalhan et al., 2025).

An increasingly popular and effective strategy for handling data imbalance is the application of resampling procedures within ensemble learning algorithms. Ensemble methods have demonstrated robust performance to handle imbalanced datac classification. When applying ensemble models to datasets with clearly defined features, undersampling proved more effective than oversampling because it delivers strong classification performance with significantly lower computational cost (Jeong et al., 2022). This finding has spurred the development and implementation of several ensemble techniques incorporating undersampling, eg. Balanced Random Forest (Agusta & Adiwijaya, 2018; Fulazzaky et al., 2024), UnderBagging which combines undersampling with bagging (Galar et al., 2012; Permatasari et al., 2016), RUSBoost that combine undersampling and boosting (Fulazzaky et al., 2024; Galar et al., 2012; Permatasari et al., 2016; Seiffert et al., 2010) and Self-Paced Ensemble (SPE) (Bano et al., 2024; Chen et al., 2023; Liu et al., 2020; Ristea & Ionescu, 2021).

The final two algorithms discussed utilize ensemble learning that perform undersampling in the learning process. RUSBoost is known for its computational efficiency, which combines random undersampling with a boosting algorithm (Seiffert et al., 2010). SPE introduces an innovative learning mechanism that iteratively adjusts the training process based on the classification hardness of the data that enhances adaptability to the complex characteristics of large datasets (Liu et al., 2020).

This study compared the performance of SPE and RUSBoost in classifying the poverty status of households in West Java. Due to class imbalance, performance evaluation focused on more relevant metrics for this scenario rather than accuracy. Evaluation was conducted using two threshold approaches: the default (0.5) and an optimal threshold based on the Youden Index (Hassanzad & Hajian-Tilaki, 2024). The better model was then chosen for further analysis, including variable importance and partial dependence plots, to investigate the effects of features on poverty status.

2. Data and Methodology

2.1 Data

The data used in this study were obtained from the 2023 National Socio-Economic Survey (Survei Sosial Ekonomi Nasional/Susenas) in West Java (Badan Pusat Statistik, 2023b). There was a total of 25,890 households, of which 1,073 (4.14%) were poor households unweighted, representing an imbalanced condition, as shown in Table 1. Households are considered poor if their monthly per-capita expenditure is less than the district/city (kabupaten/kota) poverty line.

Table 1 Proportion of Household in West Java Based on Poverty Status

Poverty Status	Number of Household	Proportion Based on Sample	Weighted Proportion (Official)
Poor	1,073	4.14%	7.62 %
Non-poor	24,817	95.86%	92.38%

The response for modeling was poverty status (unweigted) while the explanatory variables consisted of 20 categorical and six numeric variables (Table 2), including education, financial inclusion, housing condition, basic infrastructure, health and social ulnerability and household structure.

2.2 Data Analysis and Modeling

Data analysis and modeling in this study were performed as follows:

a. Prepare Susenas data and poverty line data of West Java in 2023
b. Select relevant variables based on previous research
c. Feature engineering:
   1) Define a response variable by comparing capita expenditure and poverty line
   2) Regroup class for independent variables

Table 2 List of Variable and Its Description

Variable	Description	Type
Response		
Y	Household poverty status	Categoric
Education		
X_1	Years of schooling of the head of household	Numeric
X_4	Percentage of illiterate household members	Numeric
X_{23}	Highest education level attained in the household	Categoric
X_{14}	Flag indicating ownership of KIP/PIP (Smart Indonesia Program card)	Categoric
Economic and Financial Inclusion		
X_3	Percentage of bankable household members	Numeric
X_5	Flag indicating whether the household receives cash transfers	Categoric
X_9	Flag indicating BPNT (Bantuan Pangan Non-Tunai) recipient	Categoric
X_{10}	Flag indicating PKH (Program Keluarga Harapan) recipient	Categoric
X_{11}	Flag indicating KKS (Kartu Keluarga Sejahtera) ownership	Categoric
X_{12}	Flag indicating receipt of regional/local government assistance	Categoric
Housing Conditions		
X_{15}	Roofing material of the house	
X_{16}	Wall material of the house	Categoric
X_{17}	Flooring material of the house	Categoric
X_{18}	Floor area (in square meters) of house	Numeric
X_6	Flag indicating land ownership	Categoric
Basic Infrastructure Access		
X_7	Flag indicating access to the internet	Categoric
X_{19}	Source of lighting	Categoric
X_{20}	Type of cooking fuel used	Categoric
X_{21}	Flag indicating access to proper sanitation	Categoric
X_{22}	Main source of drinking water	Categoric
Health and Social Vulnerability		
X_2	Flag indicating whether the household is considered vulnerable	Categoric
X_8	Flag indicating illness without outpatient treatment	Categoric
X_{13}	Flag indicating ownership of BPJS PBI (health insurance for the poor)	Categoric
X_{24}	Flag indicating whether the household is food insecure	Categoric
Household Structure		
X_{25}	Number of household members	Numeric

d. Splitting data into training and test set with proportion 80:20
e. Using training data set:
   1) Standardized all numeric variables
   2) Convert all categorical variables into dummy variables
   3) Train and tune hyperparameters for both SPE & RUSBoost using Grid Search and statified 5-fold cross validation and store the best model
f. Using test data set:
   1) Standardized all numeric variables using parameters from training data set
   2) Convert all categorical variables into dummy variables
   3) Using the best model obtained from training, predict response class

g. The prediction results were evaluated using the following metrics: AUC ROC, AUC PR, Geometric Mean, F1 Score, Recall (sensitivity), specificity, MCC, and Cohen's Kappa. For threshold-dependent metrics, use default threshold (0.5) as well as optimal threshold based on Youden Index.
h. Make conclusions and recommendations.

2.3 Self-Paced Ensemble (SPE)

Self-Paced Ensemble (SPE), developed byLiu et al., (2020), is an ensemble learning with undersampling mechanism to reduce trivial and noise observations while giving greater weight to more important data. Sampling in SPE is controlled by a self-paced procedure meaning that every observation contributes to the process. SPE iteratively selects the most informative sample from the majority class based on the *hardness* distribution, which is an error function, such as absolute error, squared error, and cross entropy. SPE utilizes other learning algorithms, such as decision trees, as its underlying model.

Suppose D is a training dataset containing all observations (x, y), with data sets from the minority class P and data sets from the minority class N. Because the data is imbalanced, $|P| < |N|$. Let F be the ensemble classification model composed of T classification models ft, while $F(x_i)$ is the probability that observation x_i is in the positive (minority) class. The hardness for observation (x,y) is a function of $H(x,y,F)$ (Zhang et al., 2021). Based on this hardness value, the training data is divided into k bins. The observation in the l-th bin, B_l, is defined as:

$$B_l = \left\{(x,y) | \frac{l-1}{k} \leq H(x,y,F) \leq \frac{l}{k}\right\}$$

Using the training data set D, the hardness function H, the base classifier f iterated T times, and k bins, the SPE is constructed through the following stages (Liu et al., 2020):

a. Initiate dataset S using random undersampling technique, so that $|P| \approx |N_0|$
b. Train classifier f_0 using training set S
c. For t = 1 to T:
   1) Ensemble:

$$F_t(x) = \frac{1}{t}\sum_{j=0}^{t-1} f_j(x)$$

   2) Cut majority set into k bins with respect to $H(x,y,F)$: $B_1, B_2, \ldots, B_k$
   3) Average hardness contribution in l-th bin:

$$h_l = \sum\nolimits_{s \epsilon B_l} \frac{H(x_s, y_s, F_i)}{|B_l|}$$

   4) Update *self-paced factor*

$$\alpha = tan\left(\frac{i\pi}{2t}\right)$$

   5) Unnormalized sampling weight of l-th bin: $w_l = \frac{1}{h_l + \alpha}$
   6) Under-sample from l-th bin with $\frac{w_l}{\sum_m p_m} \cdot |\mathrm{P}|$ sample
   7) Train f_t using newly under-sampled subset

d. Return final *ensemble:*

$$F(x) = \frac{1}{T}\sum_{m=1}^{T} f_m(x).$$

In steps 3.b and 3.c, the hardness value is updated to select the most suitable samples for the current ensemble model. Step 3.e uses the tangent function to control the growth of the self-paced factor α. The α value controls the weight reduction rate for undersampling, which gradually reduces the weights on large bins. In the initial iteration, $\alpha = 0$, and in the final iteration, $\alpha \rightarrow \infty$. When α is small, the training procedure focuses on informative borderline examples, so noise and outliers do not significantly affect the model. Furthermore, as α increases, the training process focuses on difficult data but still uses a few trivial samples to prevent overfitting.

2.4 RUSBoost

RUSBoost (Seiffert et al., 2010), as its name suggests, applies random undersampling to the boosting algorithm. At each iteration, the training process is performed using the random undersampled data after weighting. Samples in the majority class are initially ignored, and the boosting process is performed step by step on the remaining data. This process is repeated according to the specified number of iterations.

Given training data D with M observations and y as the binary response variable, where P is the minority observation and N is the majority observation, the RUSBoost algorithm is described as follows:

a. Define initial weight for each observation $w_1(i) = 1 / M$; $i = 1, 2, \ldots, M.$
b. For each iteration t, where t = 1, 2, …, $T,$ do:
   1) Draw dataset S using random undersampling technique*,* so that $|P| \approx |N_t|$
   2) Fit base classifier f_t using dataset $S_t,$ with respect to w_i*:*
   3) Calculate classification error
   $$\epsilon_t = \sum_{(i,y);\, y_i \neq y} w_t(i)(1 - f_t(x_i, y_i) + f_t(x_i, y))$$
   4) Calculate $\alpha_t = \frac{\epsilon_t}{1-\epsilon_t}$
   5) For each mis-classified observation, recalculate the weight
   $$w_{t+1} = w_t(i)\alpha_t^{\frac{1}{2}(1+f_t(x_i,y_i)-f_t(x_i,y:y\neq y_i)}$$
c. Final predictions is class with highest value:
$$F(x) = \underset{y\in Y}{argmax} \sum_{t=1}^{T} f_t(x, y)\log\frac{1}{\alpha_t}$$

2.5 Evaluation Metrics for Imbalance Class Classification

Accuracy often fails to adequately reflect a model's performance in scenarios with imbalanced class distributions. Therefore, there are more relevan evaluation criteria derived from the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) that are extracable from a confusion matrix (Liu et al., 2020). In addition, the area under receiver operating characteristic curve (AUC-ROC), the area under precision-recall curve (AUC-PR) and Cohen's Kappa (McHugh, 2012) can also be used to evaluate classification performance in imbalance data. Confusion matrix-based metrics for imbalance classification evaluation are shown in Table 3.

Table 3 Confusion Matrix Based Metrics for Imbalance Classification Evaluation

Metrics	Formula
Recall (sensitivity)	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Spesificity	$\frac{TN}{TN+FP}$
F1-Score	$2 \cdot \frac{Recall\ \times Precision}{Recall +\ Precision}$
Geometric Mean	$\sqrt{Recall\ \times Precision}$
Matthew's Correlation Coefficient	$\frac{TP\ \times TN - FP\ \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

3. Results and Discussion

3.1 Results

A grid search with stratified 5-fold cross-validation was used to find the best hyperparameters for both Self-Paced Ensemble (SPE) and RUSBoost models. For SPE, the best model was achieved with 20 bins, 200 estimators, and a maximum tree depth of 10. For RUSBoost, the best combination was a learning rate of 0.05, 100 estimators, and the same maximum depth of 10. Both models use decision tree as base classifier. These settings were used to train the final models for comparison.

Table 4 Comparison of SPE and RUSBoost in Poverty Status Classification

Metrics	Self-Paced Ensemble (SPE)		RUSBoost	
	Default	Optimal	Default	Optimal
AUC ROC	0.8148	0.8148	0.8185	0.8185
AUC PR	0.1594	0.1594	0.1701	0.1701
Geometric Mean	0.7286	0.6671	0.7436	0.6932
F1 Score	0.2055	0.2179	0.1900	0.2198
Recall (Sensitivity)	0.6550	0.5100	0.7250	0.5600
Precision	0.1219	0.1386	0.1094	0.1368
Specificity	0.8104	0.8726	0.7628	0.8580
Matthew's Correlation Coefficient	0.2211	0.2112	0.2153	0.2207
Cohen's Kappa	0.1501	0.1674	0.1318	0.1682

The analysis result in Table 4 shows that both models demonstrated quite similar performance in predicting poverty status in West Java, as indicated by relatively high AUC-ROC values. The SPE model produced an AUC-ROC of 0.814, whereas RUSBoost was slightly better (0.8185). The AUC-PR, that is particularly more relevant in the context of imbalanced data than AUC-ROC, exhibited RUSBoost's superiority (0.170) over SPE (0.159). Both AUC-ROC and AUC-PR values remained constant across both thresholds because these metrics were not threshold-dependent.

RUSBoost demonstrated the highest recall at 0.725 using default threshold 0.5, meaning that it was able to detect approximately 72.5% of all poor families. The SPE recorded a recall of 0.655, which, although quite good, was still lower. Specificity, that reflects the ability to identify non-poor households, was higher for SPE (0.810), than for RUSBoost (0.763). RUSBoost recorded higer G-Mean value of 0.744 compared to SPE's 0.729. However, SPE slightly outperforms RUSBoost in terms of F1 Score (0.206 vs. 0.190). SPE recorded an MCC of 0.221, which was slightly higher than RUSBoost (0.215). For Kappa, SPE obtained 0.150, whereas RUSBoost recorded 0.132.

The probability threshold adjustment using the Youden Index shifted the results. The recall decreased drastically in both models: from 0.655 to 0.510 in SPE and from 0.725 to 0.560 in RUSBoost. Specificity increased from 0.8104 to 0.8726 in SPE and from 0.7628 to 0.8580 in RUSBoost. The G-mean decreased to 0.6671 in SPE and 0.6932 in RUSBoost. In contrast, the F1 Score increased to 0.2179 in SPE and to 0.2198 in RUSBoost. The MCC value in SPE decreased slightly to 0.2112, while in RUSBoost it increased to 0.2207. Kappa increased in both models, from 0.1501 to 0.1674 in SPE, and from 0.1318 to 0.1682 in RUSBoost.

Based on previous evaluation results, the RUSBoost model with the default threshold was considered as the best model (among the compared models) for classifying household poverty status. To understand how this model makes decisions, feature importance that measures the contribution of each feature to the model performance was calculated using permutation feature importance (PFI) approach (Quay, 2022) as shown in Figure 1. The PFI of the RUSBoost model shows that the most influential factors were the number of household members, the percentage of bankable household members, and the floor area of the house, followed by indicators of access to information, education, and social vulnerabilities.

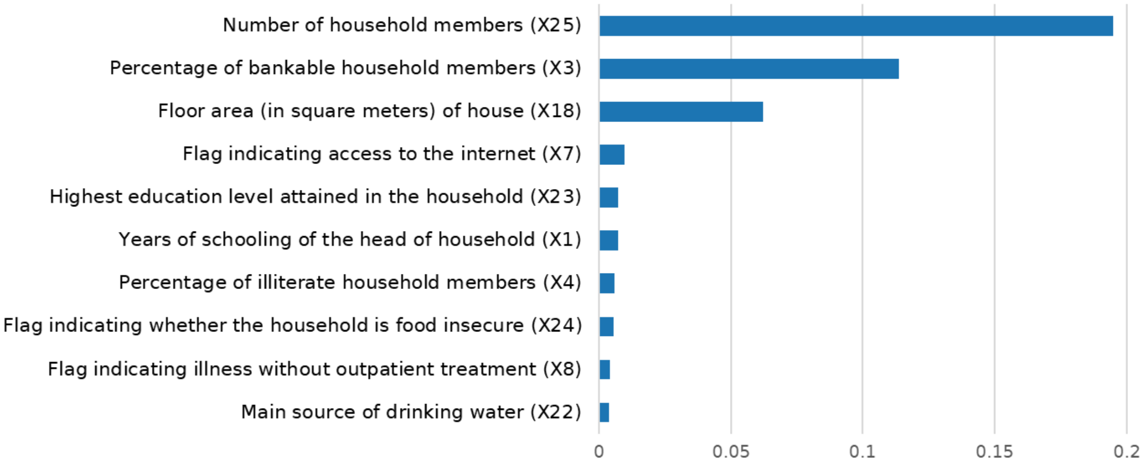

Figure 1 Top 10 Most Important Variables in the RUSBoost Model

The partial dependence plot of the RUSBoost model (Figure 2) confirms that the number of household members, percentage of bankable household members, and floor area of the house are the most influential factors in predicting poverty in West Java. The probability of poverty increases sharply up to five household members, whereas financial access and the size of the dwelling are inversely related to the risk of poverty. Educational variables, such as the head of household's years of schooling and illiteracy rate, affect poverty status, although to a smaller degree. Overall, the model considers a combination of demographic, economic, and educational aspects to determine the poverty status of a household.

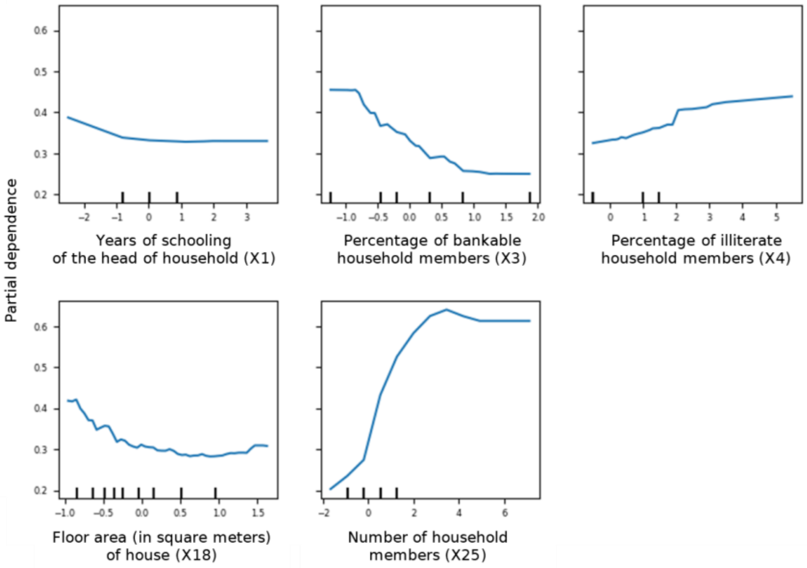

Figure 2 Partial Dependence Plot of Numeric Variables in the RUSBoost Model

3.2 Discussion

The findings indicate that RUSBoost was consistently better than SPE in most important metrics, regardless of the threshold applied. RUSBoost achieved higher AUC-ROC/PR and G-mean. RUSBoost also has better F1 Score after threshold adjustment. This suggests that RUSBoost was more effective in handling imbalanced data and provided better classification balance between poor and non-poor households.

The higher recall achieved by RUSBoost at the default threshold demonstrated its capability to identify more poor households. This is important because the poverty case, which is minority class, represents the target of interest. However, this improvement came with a trade-off. RUSBoost was more aggressive than SPE in assigning poor labels, represented by the lower specificity, resulting in more false positives. In contrast, SPE was more conservative, which explains its higher specificity but lower recall. The decrease in recall and the increase in specificity after threshold adjustment using Youden Index reflect a classic trade-off between detecting poor households and avoiding false positives. In practical terms, policymakers must consider whether missing true poverty cases (false negatives) or mislabeling non-poor households (false positives) is more costly. The increase in F1 Score for both models after threshold adjustment indicates improved precision in positive predictions, particularly in RUSBoost. Meanwhile, the improvement in MCC and Kappa values shows that adjusting the threshold increases the overall agreement between the predicted and actual labels, despite the lower recall.

The feature importance analysis using PFI suggests that poverty in West Java is strongly influenced by demographic burden (household size), economic capacity (bankable household members), and housing conditions (floor area). The partial dependence plot demonstrates how these variables affect the probability of poverty. Larger households increase poverty risk, while better financial access and larger living spaces reduce it. Meanwhile, education-related factors contribute moderately.

These results highlight that a combination of demographic, economic, and educational factors explain poverty status. The RUSBoost model with the default threshold provides the most balanced and effective classification tool for identifying poor households in imbalanced datasets.

4. Conclusions and Recommendations

RUSBoost consistently outperforms SPE across most key evaluation criteria, regardless of the threshold type used for predicting poverty classification in West Java, where class imbalance was present. It offers better classification balance, achieves the highest AUC-ROC and AUC-PR scores, and produces the strongest F1-score after threshold adjustment. When the goal is to identify as many poor households as possible, the default threshold of RUSBoost is ideal, delivering high recall along with a solid G-mean. However, if prioritizing precision is necessary, the Youden Index threshold can be used, with the trade-off of undetected poor households.

For future research, it is recommended to explore other ensemble methods such as SMOTEBoost, EasyEnsemble, or BalancedRandomForest to assess whether they offer improvements over RUSBoost and SPE. Testing the models on different regions or timeframes also can be done to evaluate their generalizability. In addition, applying spatial-based modeling algorithms could capture geographic patterns of poverty. Finally, integrating explainability techniques like SHAP or LIME to understand how the models make decisions, which is crucial for policy interpretation.

References

Agusta, Z. P., & Adiwijaya, A. (2018). Modified balanced random forest for improving imbalanced data prediction. *International Journal of Advances in Intelligent Informatics*, *5*(1), 58. https://doi.org/10.26555/ijain.v5i1.255

Altalhan, M., Algarni, A., Turki-Hadj Alouane, M., Altalhan, M., Algarni, A., & Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. *IEEE Access*, *13*, 13686–13699. https://doi.org/10.1109/ACCESS. 2025.3531662

Badan Pusat Statistik. (2023a). *Jumlah dan Persentase Penduduk Miskin Menurut Kabupaten/Kota di Provinsi Jawa Barat, 2023*.

Badan Pusat Statistik. (2023b). Survei Sosial Ekonomi Nasional (SUSENAS) Tahun 2023. In *BPS*. BPS.

Bano, S., Zhi, W., Qiu, B., Raza, M., Sehito, N., Kamal, M. M., Aldehim, G., & Alruwais, N. (2024). Self-paced ensemble and big data identification: a classification of substantial imbalance computational analysis. *Journal of Supercomputing*, *80*(7), 9848–9869. https://doi.org/10.1007/S11227-023-05828-6

Chen, Y., Du, X., & Guo, M. (2023). Self-paced ensemble for constructing an efficient robust high-performance classification model for detecting mineralization anomalies from geochemical exploration data. *Ore Geology Reviews*, *157*, 105418. https://doi.org/10.1016/j.oregeorev.2023.105418

Fulazzaky, T., Saefuddin, A., & Soleh, A. M. (2024). Evaluating Ensemble Learning Techniques for Class Imbalance in Machine Learning: A Comparative Analysis of Balanced Random Forest, SMOTE-RF, SMOTEBoost, and RUSBoost. *Scientific Journal of Informatics*, *11*(4), 969–980. https://doi.org/10.15294/SJI.V11I4.15937

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. In *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* (Vol. 42, Issue 4). https://doi.org/10.1109/TSMCC.2011.2161285

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced Big Data challenges: investigating data sampling approaches. *Journal of Big Data*, *6*(1), 107. https://doi.org/10.1186/s40537-019-0274-4

Hassanzad, M., & Hajian-Tilaki, K. (2024). Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. *BMC Medical Research Methodology 2024 24:1*, *24*(1), 84-. https://doi.org/10.1186/S12874-024-02198-2

Jeong, D. H., Kim, S. E., Choi, W. H., & Ahn, S. H. (2022). A Comparative Study on the Influence of Undersampling and Oversampling Techniques for the Classification of Physical Activities Using an Imbalanced Accelerometer Dataset. *Healthcare*, *10*(7), 1255. https://doi.org/10.3390/HEALTHCARE10071255

Liu, Z., Cao, W., Gao, Z., Bian, J., Chen, H., Chang, Y., & Liu, T.-Y. (2020). Self-paced Ensemble for Highly Imbalanced Massive Data Classification. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 841–852. https://doi.org/10.1109/ICDE48307.2020.00078

McHugh, M. L. (2012). Interrater Reliability: The Kappa Statistic. *Biochem Med (Zagreb)*, *22*(3), 276–282.

Permatasari, Y., Sartono, B., & Permatasari, Y. (2016). Penanganan Masalah Kelas Tidak Seimbang Dengan Rusboost Dan Underbagging (Studi Kasus: Mahasiswa Drop Out SPs IPB Program Magister). [Institut Pertanian Bogor]. In *Master Theses*. http://repository.ipb.ac.id/handle/123456789/80118

Rahmadini, R. (Rina), & Santoso, B. J. (Bagus). (2025). Machine Learning-Based Prediction of Divorce Verdicts Using Posita Data and Imbalanced Data Handling: A Case Study in Padang Sidempuan. *International Journal of Advances in Data and Information Systems*, *6*(2), 460–478. https://doi.org/10.59395/IJADIS.V6I2.1405

Ristea, N. C., & Ionescu, R. T. (2021). Self-paced ensemble learning for speech and audio classification. *Interspeech*, *2*, 1276–1280. https://doi.org/10.21437/ INTERSPEECH. 2021-155

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *40*(1). https://doi.org/10.1109/TSMCA.2009.2029559

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*, 64606–64628. https://doi.org/10.1109/ACCESS.2021.3074243

Zhang, Y., Chen, H. C., Du, Y., Chen, M., Liang, J., Li, J., Fan, X., & Yao, X. (2021). Power transformer fault diagnosis considering data imbalance and data set fusion. *High Voltage*, *6*(3), 543–554. https://doi.org/10.1049/hve2.12059