CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 2136–2155
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Classification of Educational Vulnerability Status Using Boosting Methods on Imbalanced Socioeconomic Data

Andi Illa Erviani Nensi, Meavi Cintani*, Mahda Al Maida, A. Qeis Tenridapi, Bagus Sartono, Aulia Rizki Firdawanti, Budi Susetyo, and Gerry Alfa Dito

Statistics and Data Science Study Program, IPB University

Abstract

Extreme class imbalance in educational data causes classification models to favor the majority class and fail to identify vulnerable students who are not attending school, despite this group requiring the greatest intervention. This study aims to develop a classification model for identifying educational vulnerability status among school-age children. The positive class corresponds to out-of-school children (minority class), while the negative class represents children who are still attending school. The dataset exhibits an extreme imbalance ratio of approximately 1:48, making conventional classification approaches ineffective in detecting minority-class observations.To address this challenge, four imbalance-aware boosting methods, namely AdaBoost-M2, SMOTEBoost, RBBoost, and RUSBoost, were compared. Model selection was conducted using stratified five-fold cross-validation on the training set, followed by final evaluation on an independent test set. Model performance was assessed using sensitivity, specificity, balanced accuracy, and Area Under the Curve (AUC), which are more appropriate than overall accuracy for highly imbalanced data. The results show that SMOTEBoost L50 with a threshold of 0.50 achieved the highest Balanced Accuracy (0.7474), Sensitivity (0.8824), and AUC (0.7745). However, its low minority-class precision (0.0452) and F1-score (0.0860) indicate a substantial false-positive trade-off. These findings demonstrate that integrating minority-class balancing strategies with boosting mechanisms can substantially improve the detection of vulnerable students in highly imbalanced socioeconomic data. Feature importance analysis revealed that education level or class was the most dominant predictor, followed by household size and frequency of internet use. Overall, this study provides empirical evidence regarding the effectiveness of imbalance-aware boosting methods for educational vulnerability detection and highlights their potential application in supporting more targeted educational interventions.

Keywords: boosting; class imbalance; educational vulnerability; SMOTEBoost; out-of-school student detection

1. Introduction

Education is a fundamental aspect that determines the quality of human resources and serves as an essential foundation for employment opportunities and social mobility. The Indonesian government has implemented a 12-Year Compulsory Education program to ensure equal access to education for all children [1]. However, this policy has not fully addressed disparities in

*Corresponding author. E-mail: meavi2501cintani@apps.ipb.ac.id

Received: December 24, 2025 Revised: June 06, 2026 Accepted: September 24, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.39670

educational access across regions. Data from the Central Statistics Agency (BPS) indicate that some adolescents aged 13–18 years remain vulnerable to educational exclusion, either due to the risk of not attending school or discontinuing schooling altogether [2]. Previous studies suggest that educational vulnerability is strongly influenced by socio-economic factors, particularly among low-income households, families whose heads work in the informal sector, and communities residing in areas with limited educational infrastructure [3].

Educational vulnerability is not only associated with formal access to schooling but also encompasses households' capacity to bear indirect educational costs, such as transportation, learning materials, and access to technology that supports the learning process [4]. Adolescents who face barriers to education tend to have lower literacy levels and limited competencies, which reduces their opportunities to obtain employment in the formal sector and increases the risk of being trapped in low-income jobs. This situation systematically reinforces the cycle of intergenerational poverty and poses a major challenge to achieving the Sustainable Development Goals (SDGs), particularly Goal 4 on inclusive and quality education [5].

The National Socioeconomic Survey (SUSENAS) provides comprehensive and nationally representative household-level data that are highly relevant for analyzing educational vulnerability in Indonesia. Key variables include ownership of the Indonesia Smart Card (KIP), participation in the Indonesia Smart Program (PIP), housing conditions, asset ownership, literacy skills, internet access, and school characteristics. These indicators collectively offer a multidimensional representation of the factors influencing adolescents' schooling status. SUSENAS-based analysis enables systematic and objective identification of vulnerable youth groups, thereby supporting evidence-based policy formulation and targeted educational interventions. Previous educational classification research has compared ordinal logistic regression and K-nearest neighbors for analyzing school accreditation rankings based on literacy and numeracy outcomes [6]. These findings demonstrate the relevance of statistical and machine-learning classification methods for analyzing educational outcomes, although severe class imbalance requires additional methodological consideration.

Modeling educational vulnerability presents a significant methodological challenge due to severe class imbalance [7]. Empirically, the proportion of adolescents who are not attending school is substantially smaller than that of those who remain enrolled, resulting in a minority class that is underrepresented in the data. This imbalance causes conventional classification algorithms to favor the majority class, leading to poor sensitivity or low recall in detecting minority-class observations [8]. Such prediction failures are particularly critical in public policy contexts, as misclassification of vulnerable individuals may result in ineffective targeting and exacerbate existing social inequalities [9].

Previous studies on educational vulnerability and school dropout prediction have primarily employed conventional machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines. Most studies focused on improving overall classification accuracy and rarely addressed the severe class imbalance commonly found in educational datasets [10]. Moreover, socioeconomic survey data such as SUSENAS present additional challenges, including overlapping class characteristics and potential measurement noise arising from heterogeneous household conditions. These characteristics often reduce the effectiveness of conventional classification methods, particularly in detecting minority-class observations [11]. Although several studies have explored boosting techniques in other domains, limited attention has been given to the use of imbalance-aware boosting algorithms for educational vulnerability classification using nationally representative socioeconomic data. Therefore, this study evaluates several boosting methods specifically designed for imbalanced classification problems and investigates their effectiveness under the complex data characteristics commonly observed in SUSENAS.

Boosting methods have been widely applied to imbalanced classification problems because they iteratively focus on observations that are difficult to classify, thereby improving minority-class

detection performance. This characteristic makes boosting particularly suitable for educational vulnerability classification using SUSENAS data, where vulnerable and non-vulnerable adolescents may exhibit overlapping socioeconomic characteristics. Despite this potential, the comparative performance of imbalance-aware boosting methods in educational vulnerability classification remains underexplored. Therefore, this study aims to evaluate and compare several boosting algorithms specifically designed for imbalanced data, namely AdaBoost-M2, SMOTEBoost, RBBoost, and RUSBoost, in identifying educational vulnerability status among adolescents. In addition, this study examines the relative importance of socioeconomic factors associated with school participation status. The findings are expected to contribute both methodologically, through the evaluation of imbalance-aware boosting approaches under extreme class imbalance conditions, and substantively, by providing evidence to support more targeted educational policies.

2. Methods

This section describes the data source, data preparation procedures, model development, hyperparameter tuning, and evaluation strategy used in this study. The analysis begins with a description of the data and research variables, followed by the implementation of imbalance-aware boosting methods and their evaluation procedures.

2.1. Data

This study utilizes data from the 2024 National Socioeconomic Survey (SUSENAS) conducted by the Central Statistics Agency (BPS). The dataset is employed to analyze the socio-economic profiles of households and individuals in relation to educational vulnerability, referred to in this study as the *Educational Vulnerability Status*. The dataset contains information on educational status, demographic characteristics, and household socio-economic conditions that are closely associated with educational participation. The unit of analysis in this study is individuals aged 13–18 years were selected, representing the secondary education age group. After completing data cleaning, age filtering, and variable selection stages, a total of 2,781 observations were obtained. The dataset initially consisted of 19 candidate predictor variables and one binary response variable. Following feature screening and information leakage assessment, two predictors (X10 and X12) were removed, resulting in 17 retained predictors used in model development.

2.2. Research Stages

2.2.1. Data Preparation

Educational vulnerability status was encoded as a binary response, with out-of-school individuals assigned to the positive class (1) and in-school individuals assigned to the negative class (0). The final analytical dataset contained 2,781 observations, comprising 57 positive-class observations (2.05%) and 2,724 negative-class observations (97.95%). The dataset was divided using a stratified 70:30 training–testing split. The training set contained 1,946 observations, including 40 out-of-school individuals, whereas the independent test set contained 835 observations, including 17 out-of-school individuals.

Stratified five-fold cross-validation was applied exclusively to the training set for model selection and hyperparameter evaluation. Within each cross-validation iteration, all data-dependent preprocessing transformations were fitted exclusively using the training fold and then applied to the corresponding validation fold. Class-balancing procedures were also restricted to the training fold, whereas the validation fold retained its original class distribution.

After the preferred configuration was selected, the model was refitted using the complete training set and evaluated once on the independent test set. Therefore, the final confusion matrix and reported test metrics represent predictions for the 835 observations in the independent test set.

Table 1: Research variables

Variables	Information	Scale
Y	School participation status	Binary
X_1	Amount of PIP assistance	Ratio
X_2	Number of household members	Ratio
X_3	Sources of household financing	Nominal
X_4	Savings account ownership	Binary
X_5	Asset/home ownership	Binary
X_6	Ability to read Latin letters	Binary
X_7	Ability to read Arabic/Hijaiyah letters	Binary
X_8	Ability to read other letters	Binary
X_9	Internet usage at home	Binary
X_{10}	Schooling status in the previous academic year	Nominal
X_{11}	Highest level/class of education ever/currently attended	Ordinal
X_{12}	School status	Binary
X_{13}	Ownership of the Smart Indonesia Card (KIP)	Nominal
X_{14}	Experience using the internet	Binary
X_{15}	Frequency of internet use	Ratio
X_{16}	Type of residential area	Binary
X_{17}	Ownership of another house by head of household/spouse/child	Binary
X_{18}	Respondent's marital status	Nominal
X_{19}	Respondent's gender	Binary

2.2.2. Data Encoding and Leakage Prevention

Binary predictors were represented as 0–1 indicators, nominal predictors were transformed into dummy variables, and ordinal predictors retained their substantively defined ordering. Within each cross-validation iteration, the encoding transformation was fitted exclusively using the training fold. The fitted transformation was then applied to the corresponding validation fold using the category structure learned from the training data [12].

Numerical predictors were standardized using StandardScaler. The scaler was fitted only on the training fold and subsequently applied to the validation fold. Although decision-tree weak learners generally do not require feature scaling, standardization was retained to support the distance-based synthetic sample generation used by SMOTEBoost [13].

All class-balancing operations associated with SMOTEBoost, RBBoost, and RUSBoost were performed exclusively within the training fold. No synthetic, oversampled, or undersampled observations were introduced into the validation folds. Therefore, each validation fold preserved the original class distribution.

2.2.3. Data Exploration

Initial data exploration was conducted to examine the imbalanced class structure and identify variables that exhibit strong associations with the target variable. This stage also served to detect variables with the potential to introduce information leakage into the modeling process. Variables X10 (schooling status in the previous academic year) and X12 (school status) were excluded because they contained information closely related to the response variable and could artificially inflate classification performance. Consequently, the subsequent modeling process was conducted using predictors that better reflect the underlying socioeconomic characteristics of adolescents.

The boosting methods evaluated in this study differ primarily in their mechanisms for handling class imbalance. AdaBoost-M2 focuses on iterative weight adjustment to emphasize difficult observations, SMOTEBoost combines boosting with synthetic minority oversampling, RBBoost employs random balancing through a combination of oversampling and undersampling, and RUSBoost integrates boosting with random undersampling of the majority class. Comparing

these approaches enables a more comprehensive assessment of how different imbalance-handling strategies affect classification performance in educational vulnerability data.

2.2.4. SMOTEBoost

SMOTEBoost combines the AdaBoost ensemble learning mechanism with synthetic minority oversampling using the Synthetic Minority Over-sampling Technique (SMOTE) at each boosting iteration. This integration produces a more balanced class distribution during the training process. The SMOTEBoost predictor generates class labels based on weighted log-likelihood weak learners at each boosting iteration [14]. The SMOTEBoost procedure consists of the following steps:

1. Identifying minority class samples.
2. Generating synthetic samples using SMOTE.
3. Combining real and synthetic data.
4. Each weak learner uses a decision tree of maximum depth two (depth = 2), with sample weights updated at each iteration.
5. Calculate the error and update the distribution weights with the formula:

$$\beta = \frac{\epsilon}{1-\epsilon} \tag{1}$$

6. Updating weights by going through the equation

$$D_i \leftarrow D_i \cdot \beta^{(1-P(h(x_i)=y_i))} \tag{2}$$

2.3. AdaBoost-M2

AdaBoost-M2 is implemented using two main functions, namely *AdaBoostM2()* which runs the boosting process by adjusting the weights based on the pseudo-loss function, and *AdaBoost-M2_predict* [15] which produces the final prediction using the aggregation of weak learners' weights.

In the boosting iteration, the following stages are carried out:

1. Initializing the initial sample weights uniformly.
2. Training a weak learner (decision tree with depth 2) using sample weights.
3. Calculating the pseudo-loss using the following formula:

$$\varepsilon_t = \frac{1}{2}\sum_{i=1}^{N}\sum_{y\neq y_i} D_t(i,y)\left(1 - h_t(x_i,y_i) + h_t(x_i,y)\right) \tag{1}$$

4. Calculating the model weight:

$$\alpha_t = \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{2}$$

5. Updating the distribution weights:

$$D_{t+1}(i,y) = \frac{D_t(i,y)\cdot\exp\left(\alpha_t\left(1 - h_t(x_i,y_i) + h_t(x_i,y)\right)\right)}{Z_t} \tag{3}$$

The final weights of the weak learners are used to determine predictions through a weighted voting mechanism.

2.4. RBBoost (Random Balance Boosting)

RBBoost is implemented through two main functions, namely *RBBoost()* to train the boosting model with random undersampling and oversampling without fixed rules, and *RBBoost_predict()* to produce the final prediction using weighted majority voting [16].

At each iteration, the following steps are performed:

1. Calculating the imbalance class ratio in the dataset.
2. Determining a new random sampling proportion such that the class distribution lies within the range of 40–60%.
3. Forming a new dataset from a combination of random undersampling of the majority class and random oversampling of the minority class.
4. Training a weak learner (decision tree with depth 2) using the balanced sample results.
5. Calculating the model error using:

$$\varepsilon_t = \sum_{i=1}^{N} D_t(i)\, \mathbb{I}\left(h_t(x_i) \neq y_i\right) \tag{4}$$

6. Updating the weights, which are then normalized, and the process continues until the iteration is complete.

$$\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t} \tag{5}$$

$$D_{t+1}(i) = D_t(i) \cdot \beta_t^{1-\mathbb{I}(h_t(x_i)=y_i)} \tag{6}$$

2.5. RUSBoost

The implementation of RUSBoost is carried out through two main functions, namely *RUSBoost()* for the training process and *RUSBoost_predict()* to produce the final prediction [17]. In general, the boosting process is carried out with a reweighting mechanism similar to AdaBoost, but at each iteration, Random UnderSampling (RUS) is performed to balance the class distribution.

The stages of the RUSBoost algorithm are explained as follows:

1. Sample Weight Initialization.
   In the first iteration, the distribution weights D_i for all samples are given uniform initial values:

$$D_i = \frac{1}{N} \tag{7}$$

   where N is the total number of samples in the training dataset.
2. Random UnderSampling (RUS).
   Random UnderSampling (RUS) is a data balancing technique that reduces the number of observations in the majority class through random sample selection until the proportions of the majority and minority classes are balanced. Unlike synthetic oversampling approaches such as SMOTE, which balance data by adding synthetic samples to the minority class, RUS does not generate new data but achieves class balance by removing some data from the majority class. Thus, RUS has the potential to reduce information in the majority class, but does not change the distribution structure of the minority class.
3. Weak Learner Training.
   Decision tree models (usually with low depth, for example depth = 2) are trained using the undersampled data by taking into account the distribution weights D_i. A shallow decision tree (depth = 2) was selected as the weak learner because boosting algorithms are designed to combine multiple weak classifiers into a strong ensemble. Shallow trees reduce model complexity and help prevent overfitting.

4. Error Calculation.
   After the model is trained, the error value is calculated based on the sample weights:

$$\varepsilon_t = \sum_{i=1}^{N} D_i \cdot \mathbb{I}\left(h_t(x_i) \neq y_i\right) \tag{8}$$

   where the value of the indicator function $\mathbb{I}(\cdot)$ is 1 if the prediction is incorrect and 0 if it is correct.
5. Beta Value Calculation.
   The learning weight for weak learners is calculated using the boosting formula:

$$\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t} \tag{9}$$

6. Sample Weight Update.
   The sample weights are updated to place greater emphasis on misclassified data:

$$D_i \leftarrow D_i \cdot \beta_t^{\mathbb{I}(h_t(x_i)=y_i)} \tag{10}$$

   If the prediction is correct, the weight decreases, and if the prediction is incorrect, the weight increases. The weights are then normalized as follows:

$$D_i = \frac{D_i}{\sum_{j=1}^{N} D_j} \tag{11}$$

7. Iteration.
   Steps 2–6 are repeated until the number of iterations (L) is reached.
8. Final Prediction
   The final prediction uses a log-weighted voting function in *RUSBoost_predict()* by combining the predictions of each weak learner with weighting based on:

$$F(x) = \sum_{t=1}^{L} \log\left(\frac{1}{\beta_t}\right) h_t(x) \tag{12}$$

2.6. Hyperparameter Tuning

This is done to find the best combination of parameters in each model [18].

Table 2: *Hyperparameter Tuning*

Parameter	**Variations Tried**
Number of *Weak Learners* (L)	10, 50, 100
Probability Threshold	0.30, 0.50, 0.70

The selected hyperparameter values were chosen based on computational efficiency considerations and recommendations from previous studies on boosting algorithms for imbalanced classification. The number of weak learners (L = 10, 50, and 100) was selected to represent low, moderate, and high boosting complexity, respectively. Meanwhile, probability thresholds of 0.30, 0.50, and 0.70 were evaluated to investigate different trade-offs between sensitivity and specificity in minority-class detection. While broader hyperparameter optimization may further improve performance, the current study focuses on comparing the effectiveness of different imbalance-aware boosting frameworks under a common experimental setting.

2.7. Model Evaluation

Model evaluation is performed to determine the best performance under imbalanced data conditions. The evaluation process uses several metrics, namely *accuracy*, *precision*, *recall* (*True Positive Rate / TPR or sensitivity*), *specificity* (*True Negative Rate / TNR*), F1-score, and *Balanced Accuracy*.

2.8. Feature Importance Analysis

The final stage of the research was to analyze the best model based on the evaluation results from the previous stage. The two highest-performing models were compared, and then a *feature importance analysis* was performed using the aggregation weights of the *weak learners* in boosting.

3. Results and Discussion

This section presents the main findings of the study, beginning with an exploration of the data characteristics and class imbalance, followed by model evaluation and comparison. The results are then discussed in relation to the ability of each boosting method to identify the minority class and the contribution of the predictor variables.

3.1. Data Exploration

3.1.1. Class Imbalance

The analysis focused on the class imbalance of the target variable, namely schooling status. The minority class is defined as children who are not attending school, while the majority class is children who are still attending school. Of the total 2,781 observations, there are 57 children (±2.05%) who are not attending school as the minority class and 2,724 children (±97.95%) who are still attending school as the majority class. This comparison produces an imbalance ratio of approximately 1:48, indicating that the size of the minority class is much smaller than the majority class. This condition indicates that the dataset is categorized as highly class imbalance and has the potential to affect the performance of the classification model if not specifically addressed.

3.1.2. Initial Analysis of Predictor Variables

The characteristics of the predictor variables were analyzed descriptively through data distribution, differences in patterns based on schooling status, and relationships between features. The results of this exploration served as the basis for assessing the potential contribution of each feature and identifying variables that require specific attention during the data preparation stage.

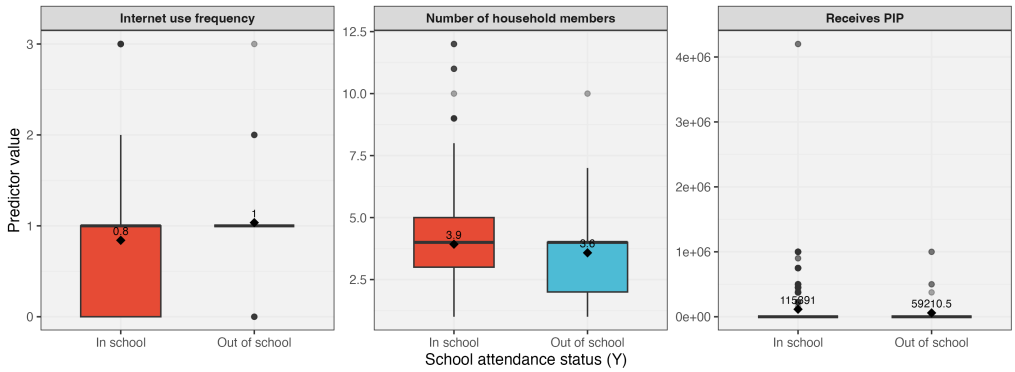


Fig. 1: Boxplot of Numeric Variables Based on Schooling Status

As shown in Figure 1, the boxplot of numeric variables based on schooling status shows a fairly wide distribution and several *outliers* in the three main numeric variables: PIP assistance, number of household members, and frequency of internet use. The PIP assistance variable is dominated by the value 0, indicating that the majority of observations did not receive this assistance. The distribution patterns between children who are still in school and those who are not in school appear to overlap, so that when these three numeric variables are observed separately, it is not possible to clearly distinguish between the two groups of schooling status.

In almost all predictor variable categories, the percentage of children out of school is relatively small. Even in categories representing vulnerable conditions, such as children who have never or

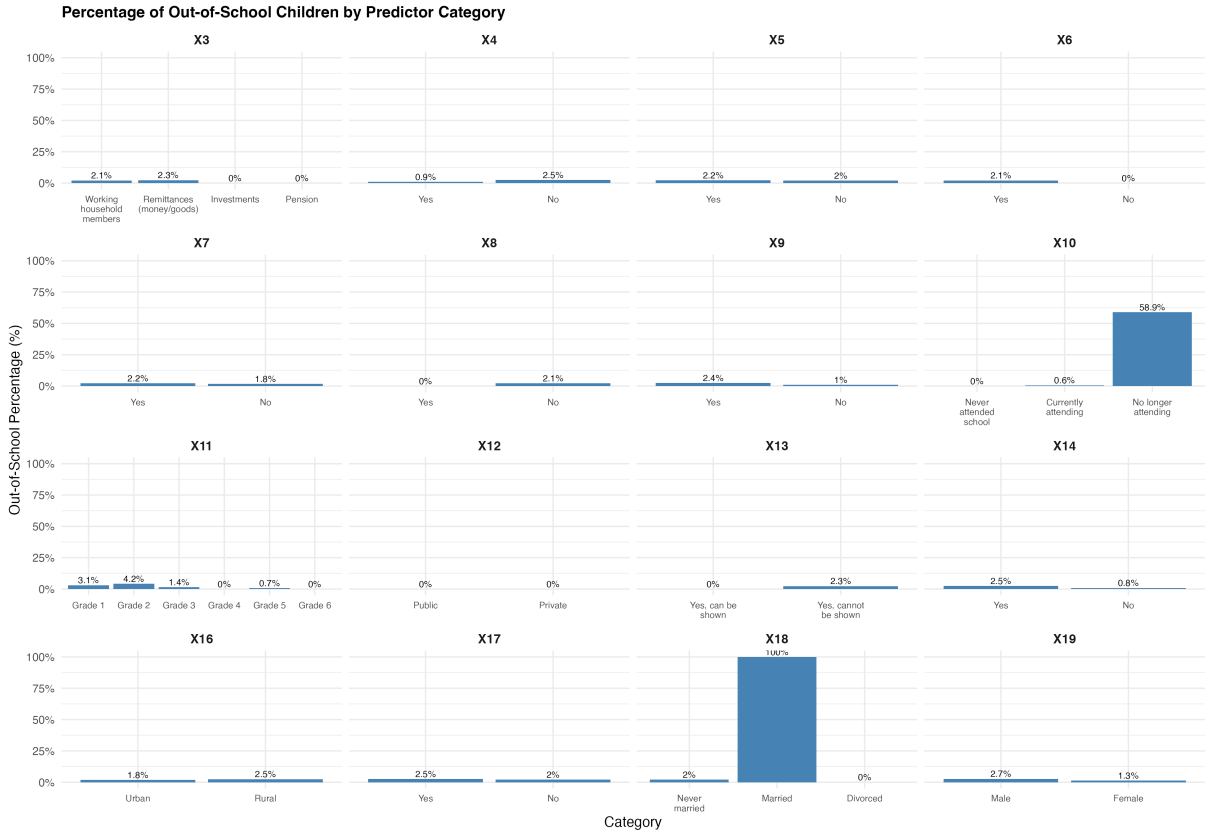


Fig. 2: Percentage of Children Not Attending School by Predictor Variable Category

are no longer attending school (X10), low-education classes (X11), and rural areas (X16), the proportion of children out of school remains low. This pattern indicates a strong class imbalance, with the number of children still in school far outnumbering those who are not.

Cramér's V values indicate that most categorical predictor variables have a weak relationship with schooling status, with values generally below 0.10. However, two variables, namely X12 and X10, show Cramér's V values that are much higher than the other variables. The high level of association in these two variables indicates the existence of overlapping information with the response variable, which has the potential to cause *information leakage* if they are still used as predictors.

3.2. Data Preparation

Based on the conceptual leakage assessment, X10 and X12 were excluded because their definitions overlapped closely with the response variable, leaving 17 predictors. The resulting dataset was divided into a training set of 1,946 observations (70%) and an independent test set of 835 observations (30%) using stratified sampling. The training set contained 40 out-of-school observations, whereas the independent test set contained 17 out-of-school observations.

Model selection and hyperparameter evaluation were conducted using stratified five-fold cross-validation within the training set. During each iteration, encoding, standardization, and class-balancing procedures were fitted or performed exclusively on the training fold. The validation fold retained its original class distribution. After model selection, the preferred configuration was refitted using the complete training set and evaluated on the independent test set. Table 3 summarizes the class distributions within the cross-validation folds of the training set.

As shown in Table 3, stratified five-fold cross-validation was performed within the training set of 1,946 observations. Each validation fold contained approximately 389–390 observations, including eight minority-class observations. The remaining 1,556–1,557 observations were used for model training within each iteration. Although stratification preserved the minority-class

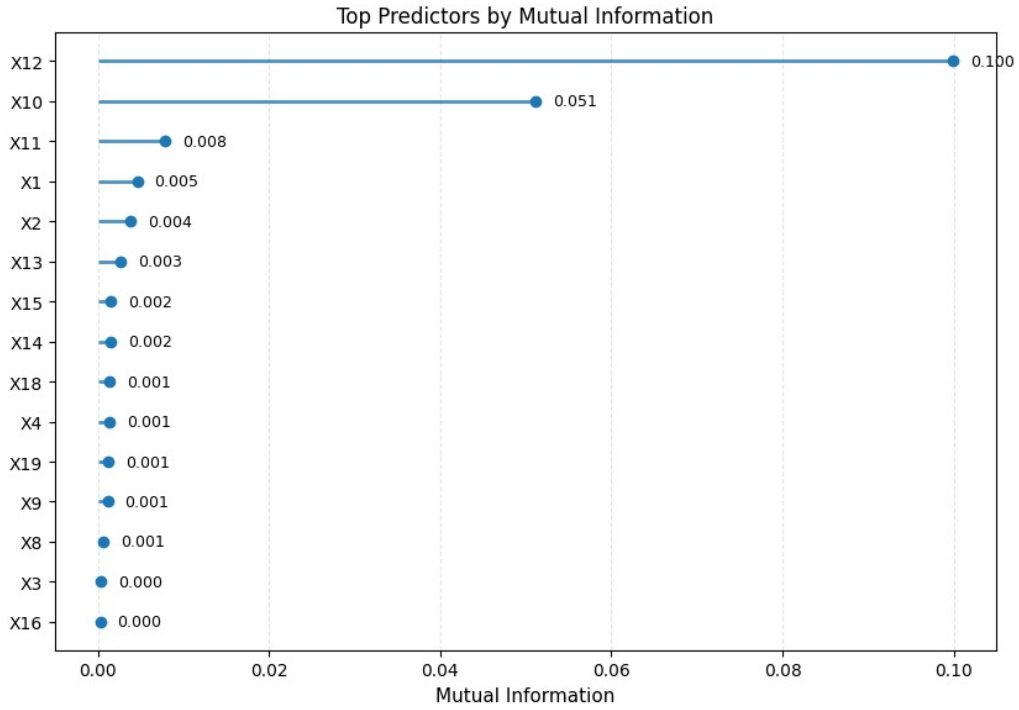


Fig. 3: Association of categorical variables with schooling status.

Table 3: Class distribution across the stratified cross-validation folds within the training set

Fold	Fold Training	Validation	Training Minority	Validation Minority
1	1556	390	32	8
2	1557	389	32	8
3	1557	389	32	8
4	1557	389	32	8
5	1557	389	32	8

proportion, the small number of positive observations in each validation fold may cause variability in minority-class performance estimates. Therefore, these estimates should be interpreted cautiously.

3.3. Random Forest

Random Forest was used as a baseline model to provide an initial overview of the performance of the classification method without explicit class imbalance handling mechanisms. This method constructs multiple decision trees using bootstrap aggregation (bagging) techniques and randomly selects a subset of variables at each node split, then combines all tree predictions using a majority rule. This approach is generally effective in reducing overfitting and capturing nonlinear relationships between predictor variables. All Random Forest configurations were identified and evaluated using metrics across thresholds and number of estimators, the Random Forest baseline results are presented in Table 4.

Table 4: Random Forest baseline performance evaluation with various thresholds

Model	Accuracy	TPR	TNR	F1-Score	Balanced Accuracy
Random Forest (Baseline)	0.9796	0.0000	1.0000	0.0000	0.5000

The Random Forest performance evaluation results presented in Table 4 show that the model produces very high accuracy (0.9796) and a True Negative Rate (TNR) of 1.00. However, the True Positive Rate (TPR) and F1-score each reach 0.00, resulting in a Balanced Accuracy of only 0.50. This pattern indicates that the model consistently predicts only the majority class,

without successfully identifying a single observation from the minority class, namely children who do not attend school.

This performance was affected by the extreme class imbalance in the research data. Under these conditions, the Random Forest learning process was dominated by the majority class, resulting in misclassifications in the minority class contributing very little to the formation of the tree structure. As a result, while the overall accuracy appeared excellent, the model's ability to detect educational vulnerability was severely limited and inconsistent with the primary research objective.

Based on these results, Random Forest is considered less effective as a primary predictive model, but remains relevant as a baseline for comparison. These results confirm that high accuracy on imbalanced data does not necessarily reflect a model's ability to recognize minority classes, and highlight the importance of using methods specifically designed to address class imbalance, such as SMOTEBoost, in the context of identifying educational vulnerability.

3.4. SMOTEBoost

SMOTEBoost combines the AdaBoost framework with SMOTE-based synthetic sample generation to improve the model's ability to recognize minority classes. At each iteration, the model not only updates the weights of misclassified observations but also adds synthetic samples from the minority class to more balance the training data distribution. Table 5 presents the performance evaluation of SMOTEBoost based on variations in the number of estimators (L) and three probability threshold scenarios.

Table 5: Performance evaluation of SMOTEBoost with various thresholds

Model Configuration	Threshold	Accuracy	TPR	TNR	F1-Score	Balanced Accuracy
SMOTEBoost L10	0.30	0.7569	0.5294	0.7616	0.0814	0.6455
	0.50	0.9796	0.0000	1.0000	0.0000	0.5000
SMOTEBoost L50	0.30	0.6168	0.8824	0.6112	0.0857	0.7468
	0.50	0.6180	0.8824	0.6125	0.0860	0.7474
	0.70	0.6180	0.8824	0.6125	0.0860	0.7474
SMOTEBoost L100	0.30	0.5581	0.8824	0.5513	0.0752	0.7168
	0.50	0.6180	0.8824	0.6125	0.0860	0.7474
	0.70	0.6180	0.8824	0.6125	0.0860	0.7474

The SMOTEBoost model shows a clear performance difference between *weak learner* configurations. At L10, a threshold of 0.50–0.70 yields very high accuracy (≈0.98), but the TPR and F1-score for the minority class are 0, so the model only predicts the majority class and does not detect out-of-school children at all (*Balanced Accuracy* 0.50). This condition is influenced by the extremely class imbalance in the data, as the model's learning process is dominated by the majority class and prediction errors in the minority class contribute very little to the loss function, especially at relatively high thresholds. As a result, despite seemingly good accuracy, the model's ability to recognize the minority class is very low.

Lowering the threshold to 0.30 increases the TPR to 0.5294 and the Balanced Accuracy to 0.6455, although with a decrease in overall accuracy. Configurations with L50 and L100 show stable TPR values around 0.88 and Balanced Accuracy values in the range of 0.72–0.75 for all threshold values, so both configurations are able to detect minority classes than L10. Based on the combined accuracy and sensitivity, SMOTEBoost L50 with a threshold of 0.50 is chosen as the main configuration because its performance is equivalent to L100 but with lower model complexity, and is considered more representative for detecting educational vulnerabilities.

To provide a more explicit evaluation of minority-class performance, Table 6 reports the precision and other evaluation measures of the selected SMOTEBoost configuration.

Table 6: Minority-class performance of the selected SMOTEBoost model

Model	Threshold	Sensitivity	Specificity	Precision	F1-Score	Bal. Acc.	ROC-AUC
SMOTEBoost L50	0.50	0.8824	0.6125	0.0452	0.0860	0.7474	0.7745

Table 7 presents the confusion matrix of the selected SMOTEBoost L50 model at the 0.50 decision threshold.

Table 7: Confusion matrix of SMOTEBoost L50 at the 0.50 threshold

Actual Class	Predicted Class	
	In School	Out of School
In School	501	317
Out of School	2	15

As shown in Table 7, the selected SMOTEBoost L50 model correctly identified 15 of the 17 out-of-school observations, while two out-of-school observations were incorrectly classified as in school. However, 317 in-school observations were incorrectly classified as out of school.

The minority-class precision was calculated as

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{15}{15 + 317} = 0.0452. \tag{13}$$

Thus, approximately 4.52% of the observations predicted as out of school actually belonged to the minority class. Although the model achieved a high sensitivity of 0.8824, its low precision of 0.0452 and F1-score of 0.0860 demonstrate that this performance was accompanied by a substantial false-positive trade-off. Therefore, the model is more appropriate for preliminary sensitivity-oriented screening, in which positively flagged individuals require further administrative or contextual verification, rather than for definitive individual classification.

The performance of SMOTEBoost in detecting minority-class observations can be attributed to its ability to simultaneously address extreme class imbalance and overlapping class characteristics in socioeconomic data. Through the Synthetic Minority Oversampling Technique (SMOTE), additional minority-class instances are generated to enrich the representation of vulnerable students, thereby reducing the dominance of the majority class. Furthermore, the boosting mechanism iteratively emphasizes observations that are difficult to classify, enabling the model to better learn complex decision boundaries. This combination is particularly relevant to the present dataset, which exhibits an imbalance ratio of approximately 1:48 and heterogeneous socioeconomic characteristics.

3.5. AdaBoost-M2

AdaBoost-M2 is used as an error-weighted *boosting* approach without explicitly manipulating the class distribution. This mechanism gives misclassified observations greater weight in subsequent iterations, allowing the model to focus more on difficult-to-predict cases, including minority classes. Table 8 shows the performance evaluation of the AdaBoost-M2 model.

At thresholds of 0.50–0.70, the model's accuracy appears very high (97–98%), but its sensitivity to minority classes is zero, resulting in the model completely failing to detect cases of educational vulnerability. This indicates an *accuracy paradox*, a situation where accuracy appears good while its actual classification ability is low due to extreme class imbalance. Balanced accuracy, which is only around 0.50, reinforces the evidence that at the default threshold, the model tends to make homogeneous predictions to the majority class and does not learn patterns that represent vulnerable students.

When the threshold is lowered to 0.30, there is a significant increase in sensitivity across all *weak learner* (L) configurations. For the L10 configuration, Balanced Accuracy reaches 0.6306,

Table 8: Performance evaluation of AdaBoost-M2 with various thresholds

Model Configuration	Threshold	Accuracy	TPR	TNR	F1-Score	Balanced Accuracy
	0.30	0.8970	0.3529	0.9083	0.1224	0.6306
AdaBoost-M2 L10	0.50	0.9796	0.0000	1.0000	0.0000	0.5000
	0.70	0.9796	0.0000	1.0000	0.0000	0.5000
	0.30	0.7150	0.6471	0.7164	0.0846	0.6817
AdaBoost-M2 L50	0.50	0.9784	0.0000	0.9988	0.0000	0.4994
	0.70	0.9796	0.0000	1.0000	0.0000	0.5000
	0.30	**0.6647**	**0.7059**	**0.6638**	**0.0789**	**0.6848**
AdaBoost-M2 L100	0.50	0.9784	0.0000	0.9988	0.0000	0.4994
	0.70	0.9796	0.0000	1.0000	0.0000	0.5000

while for the L50 and L100 configurations, it increases to 0.6817–0.6848, albeit with a decrease in specificity due to increased false positives. This performance demonstrates that *threshold tuning* is more important for classification success than simply increasing the number of estimators. AdaBoost-M2 with a low threshold is able to compensate for bias towards the majority class by prioritizing detection of the minority class without sacrificing specificity excessively.

The selection of the probability threshold directly affects the trade-off between sensitivity and specificity. Lower thresholds increase the likelihood of identifying vulnerable students but may also increase false positive classifications. Conversely, higher thresholds tend to improve specificity while reducing sensitivity. In the context of educational policy, prioritizing sensitivity is often preferable because failing to identify vulnerable students may lead to missed interventions, whereas false positive cases can still be verified through subsequent assessment processes.

3.6. RBBoost

RBBoost (*Random Balance Boosting*) is a boosting approach designed to address class imbalance without removing observations from the majority class. This model increases the weight of minority class observations while giving additional weight to hard-to-classify majority class observations. This strategy makes RBBoost adaptive in learning minority patterns without losing majority class information. Table 9 presents the evaluation of RBBoost based on variations in the number of *weak learners* (L) and three probability threshold scenarios.

Table 9: RBBoost performance evaluation with various thresholds

Model Configuration	Threshold	Accuracy	TPR	TNR	F1-Score	Balanced Accuracy
	0.30	0.7557	0.4120	0.7630	0.0642	0.5873
RBBoost L10	0.50	0.9257	0.2350	0.9400	0.1143	0.5877
	0.70	0.9257	0.2350	0.9400	0.1143	0.5877
	0.30	0.6802	0.2940	0.6880	0.0361	0.4912
RBBoost L50	0.50	0.7305	0.2350	0.7410	0.0343	0.4881
	0.70	0.7916	0.1770	0.8040	0.0333	0.4904
	0.30	0.2886	0.7060	0.2800	0.0388	0.4929
RBBoost L100	0.50	0.4551	0.6470	0.4510	0.0461	0.5491
	0.70	0.6228	0.4710	0.6260	0.0483	0.5483

Threshold of 0.50 provides the most stable performance (Accuracy = 92.57%; Balanced Accuracy = 0.5877) because it maintains sensitivity at a reasonable level without sacrificing specificity significantly.

3.7. RUSBoost

RUSBoost is a *boosting* method that combines *random undersampling* of the majority class and adaptive weighting, similar to conventional boosting algorithms. This approach is designed to reduce the dominance of the majority class in the learning process by trimming some of the majority observations, thus increasing the exposure of the model to minority class patterns. Table 10 presents an evaluation of RUSBoost's performance based on the number of *weak learners* (L) and three probability threshold scenarios.

Table 10: RUSBoost performance evaluation with various thresholds

Model Configuration	Threshold	Accuracy	TPR	TNR	F1-Score	Balanced Accuracy
RUSBoost L10	0.30	0.7557	0.4120	0.7630	0.0642	0.5873
	0.50	0.9257	0.2350	0.9400	0.1143	0.5877
	0.70	0.9257	0.2350	0.9400	0.1143	0.5877
RUSBoost L50	0.30	0.6802	0.2940	0.6880	0.0361	0.4912
	0.50	0.7305	0.2350	0.7410	0.0343	0.4881
	0.70	0.7916	0.1770	0.8040	0.0333	0.4904
RUSBoost L100	0.30	0.2886	0.7060	0.2800	0.0388	0.4929
	0.50	0.4551	0.6470	0.4510	0.0461	0.5491
	0.70	0.6228	0.4710	0.6260	0.0483	0.5483

RBBoost and RUSBoost produced identical evaluation results across the tested ensemble sizes and probability thresholds. This similarity may be attributed to the use of the same cross-validation partitions, random seed, weak learner configuration, and class-balancing ratio. Under these controlled experimental settings, both methods generated comparable classification decisions. Therefore, the identical results reflect the specific configuration used in this study and should not be generalized to other datasets or implementations.

In general, RUSBoost maintained non-zero minority-class sensitivity across the evaluated configurations, although its overall performance remained lower than that of AdaBoost-M2. The highest balanced accuracy was obtained by RUSBoost L10 at the 0.50 and 0.70 thresholds, with a value of 0.5877. These results suggest that majority-class undersampling improved minority detection to some extent, but its benefit was accompanied by reduced specificity or information loss under several configurations.

3.8. Model Comparison and Selection of the Sensitivity-Oriented Screening Model

Figure 4 compares the performance of the evaluated boosting models and supports the selection of a sensitivity-oriented screening configuration. Based on the comparison of the selected configurations from each boosting method in Figure 4, it can be seen that each model has different strengths and characteristics in handling class imbalance. AdaBoost-M2 L100 (threshold 0.30) shows high Balanced Accuracy (0.6848) with good sensitivity (TPR = 0.7059), but the overall accuracy is relatively low. RBBoost L10 and RUSBoost L10 (threshold 0.30) offer higher accuracy (0.7557) but lower sensitivity (TPR = 0.412), thus risking failure to recognize a large portion of vulnerable students. Among the evaluated configurations, SMOTEBoost L50 at the 0.50 threshold achieved a sensitivity of 0.8824 and a balanced accuracy of 0.7474. Accordingly, it was selected as the preferred sensitivity-oriented screening model among the evaluated configurations, rather than as the overall best-performing classifier across all evaluation criteria. The 0.50 threshold was retained because it provided the highest balanced accuracy while maintaining the maximum observed sensitivity of 0.8824. This selection reflects the study's primary objective of identifying as many vulnerable students as possible.

Nevertheless, the minority-class precision of 0.0452 indicates that only 4.52% of students

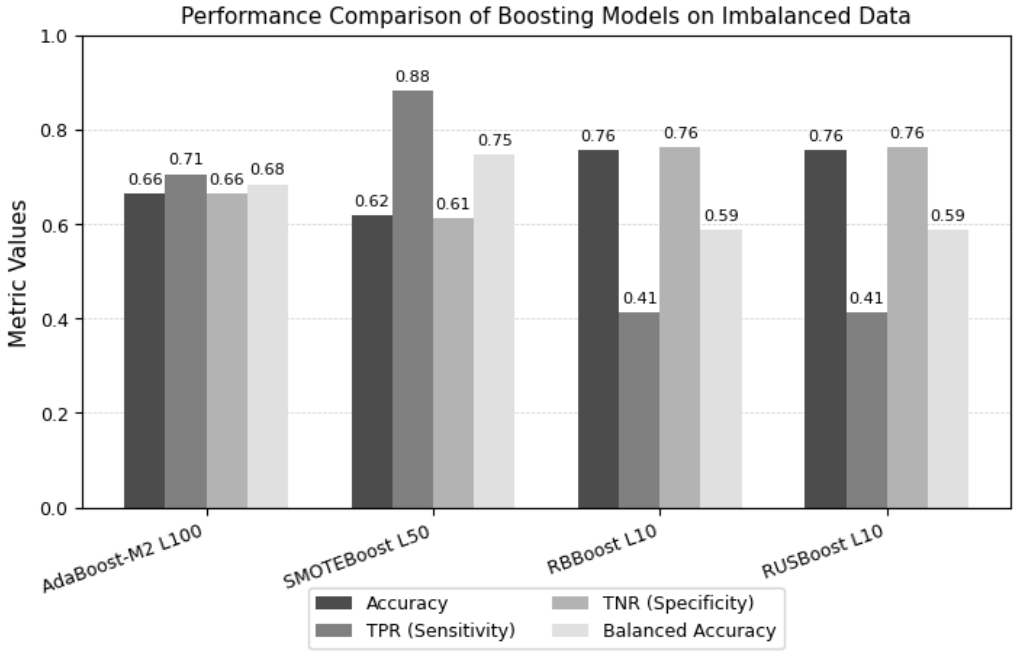


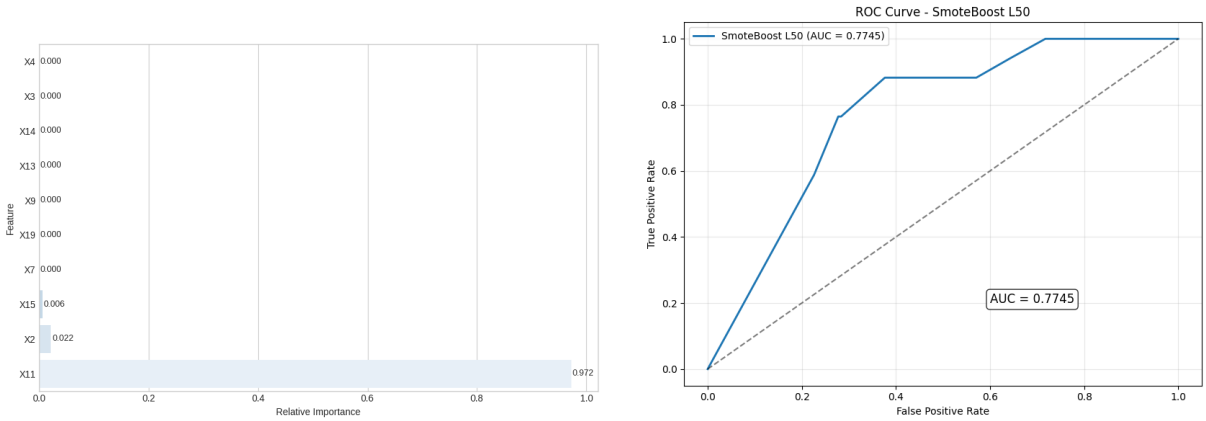
Fig. 4: Comparison of boosting models based on imbalanced datasets

flagged as vulnerable actually belonged to the minority class. Although the model correctly identified 15 of the 17 out-of-school students, it also incorrectly flagged 317 school-attending students as vulnerable. This explains the low F1-score of 0.0860 and demonstrates that the high sensitivity was achieved at the cost of a substantial number of false positives. Therefore, the model is appropriate only as a preliminary screening aid, with all positive predictions requiring further administrative or contextual verification.

The observed improvement in minority-class sensitivity is consistent with the original motivation of SMOTEBoost, which combines synthetic minority oversampling with iterative boosting to strengthen the representation of minority-class observations [14]. In addition, AdaBoost's suboptimal performance in this research is in line with findings in [17], which confirm that the effectiveness of AdaBoost is highly dependent on parameter tuning and still tends to be biased toward the majority class in skewed data distributions [19]. The low sensitivity results of RUSBoost in this study are also in line with findings in [8], which state that repeated undersampling mechanisms can eliminate majority information, causing the model to have difficulty recognizing minority patterns consistently. In addition, the trade-off pattern between increasing TPR and decreasing overall accuracy observed in this study is consistent with [16], which shows that boosting algorithms on imbalanced data tend to increase TPR at the cost of decreasing specificity. Thus, SMOTEBoost L50 at the 0.50 threshold was selected as the preferred sensitivity-oriented screening model because it prioritized minority-class detection and achieved the highest balanced accuracy among the evaluated configurations. This selection does not imply superiority across all evaluation criteria, particularly because the model's low precision and F1-score indicate a substantial number of false-positive predictions.

From a policy perspective, this trade-off may be acceptable for preliminary screening because failing to identify a vulnerable student may be more consequential than referring a non-vulnerable student for additional verification. Nevertheless, the model's positive predictions should not be used directly to determine educational vulnerability or assistance eligibility.

The feature importance analysis in Figure 5 shows that variable X11, namely the highest level/class of education ever or currently occupied, is the most dominant predictor in detecting educational vulnerability, contributes 97.2% of the total feature importance within the SMOTEBoost model. Although X11 emerged as the dominant predictor, it was retained because its



(a) Feature Importance **(b)** ROC Curve

Fig. 5: (a) Feature Importance and (b) ROC Curve SMOTEBoost L50

association with the target variable remained below the exclusion threshold used to identify potential information leakage. This means that the respondent's grade level or level is the strongest indicator in distinguishing whether a student is at risk of not attending school/stopping continuing education. This makes empirical sense because students at higher levels of education have different risks than students at the elementary level, both in terms of academic burden, level transitions, and family economic pressure. Despite its usefulness, ROC-AUC should be interpreted cautiously under severe class imbalance because it may still provide favorable results even when precision remains low. Future studies should therefore complement ROC analysis with Precision–Recall curves and PR-AUC metrics [20].

The next contributing variables, although relatively small, are X2 (number of household members/ART) and X15 (frequency of internet use). The X2 variable indicates that the larger the number of household members, the potential for economic burden and competing needs within the family increases, which can affect school sustainability. Meanwhile, X15 can describe high access to basic resources with better frequency of internet use and higher academic connectivity, thus being more protected from educational vulnerability. In contrast, other variables such as X3 (household financing source), X4 (account ownership), X7 (ability to read Latin/Arabic), X9 (home internet access), X13 (KIP ownership), and X19 (gender) do not contribute substantially to the model's predictive performance (importance value 0), indicating that these variables were not selected by the fitted model for the classification decisions under the present data and model configuration. This finding suggests that educational vulnerability in the context of this dataset may be more strongly associated with educational attainment and household structural conditions than with the other variables included in the model.

The dominance of X11 (Highest Education Level/Class) should be interpreted carefully. X12 was excluded because it directly represents current school participation, while X10 was excluded because it records school participation in the previous academic year and is therefore temporally and conceptually close to the response. In contrast, X11 describes the highest educational level or class ever or currently attended and does not directly indicate school participation at the survey reference time. Therefore, X11 was retained as a measure of educational attainment and history. However, given its conceptual proximity to the response, its dominant importance should not be interpreted as evidence of a causal relationship.

The model's ability to distinguish between the two classes was further assessed using the ROC curve, as shown in Figure 5. The use of the ROC curve and AUC values in this study is more relevant than accuracy because the data used is highly imbalanced. Accuracy tends to be biased towards the majority class and can provide high scores even if the model fails to recognize the minority class [7]. In contrast, ROC evaluates the relationship between the True Positive Rate and False Positive Rate at various thresholds, thus independent of class proportions [20]. The

AUC value represents the model's global discriminatory ability, namely the probability that the model scores observations in the positive class higher than those in the negative class. Although ROC-AUC remains useful for assessing overall discrimination ability, it should be interpreted together with sensitivity, balanced accuracy, and ideally PR-AUC under severe class imbalance. The ROC curve shows that the model line is located well above the reference diagonal line, with an Area Under Curve (AUC) of 0.7745. This value illustrates the model's strong discriminatory capacity, where the model's probability of correctly classifying individuals between vulnerable and non-vulnerable classes is relatively high.

The ROC-AUC value of 0.7745 indicates that SMOTEBoost L50 has useful discriminatory ability in ranking observations according to their likelihood of belonging to the minority class. However, ROC-AUC does not directly reflect the low precision and substantial number of false-positive predictions produced at the selected threshold. Therefore, SMOTEBoost L50 at the 0.50 threshold should be interpreted as the preferred sensitivity-oriented screening model among the evaluated configurations, rather than as an unqualified overall-best classifier.

The hyperparameter exploration in this study was limited to the number of weak learners ($L \in \{10, 50, 100\}$) and classification thresholds (0.30, 0.50, and 0.70). Other potentially influential parameters, including the SMOTE number of nearest neighbors, sampling ratio, tree depth, learning rate, and minimum number of samples per leaf, were held constant to maintain a comparable experimental setting across the boosting methods. Consequently, the reported results may not represent the optimal performance of each method. Future studies should employ a broader hyperparameter optimization strategy using nested cross-validation or other validation procedures.

4. Conclusion

This study evaluated four imbalance-aware boosting methods for detecting educational vulnerability under extreme class imbalance. Among the evaluated configurations, SMOTEBoost L50 at the 0.50 threshold provided the preferred sensitivity-oriented screening performance, achieving a sensitivity of 0.8824, balanced accuracy of 0.7474, and ROC-AUC of 0.7745. The model correctly identified 15 of the 17 out-of-school observations, demonstrating its potential for preliminary minority-class screening.

However, the minority-class precision of 0.0452 and F1-score of 0.0860 indicate that the high sensitivity was accompanied by a substantial number of false-positive predictions. Consequently, SMOTEBoost L50 should be regarded as a preliminary screening tool for identifying individuals who may require further assessment, rather than as an overall best-performing classifier or a stand-alone system for determining educational vulnerability or assistance eligibility.

Nevertheless, these findings should be interpreted with caution due to the relatively small number of minority observations, the low F1-scores observed in several configurations, and the limited hyperparameter exploration conducted in this study. In addition, the feature importance analysis revealed the strong dominance of education level/class (X11), suggesting that the predictive performance of the model may be influenced by a limited number of highly informative variables. Therefore, further investigation is needed to assess the robustness and generalizability of the findings across different datasets and contexts.

These findings suggest that the selected model may serve as a preliminary screening aid to support the identification of individuals who require further assessment. However, it should not be implemented as an autonomous early detection or decision-making system. Its positive predictions require verification using administrative records, household information, and assessments by the relevant educational authorities.

The cross-sectional design, small minority-class sample, limited hyperparameter exploration, and absence of external validation restrict the generalizability and operational readiness of the model. Future research should evaluate the model using external and longitudinal datasets, explore broader hyperparameter configurations, incorporate additional contextual predictors,

and investigate cost-sensitive learning approaches.

CRediT Authorship Contribution Statement

Andi Illa Erviani Nensi: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Visualization, Writing–Original Draft Preparation. **Meavi Cintani:** Conceptualization, Methodology, Validation, Writing–Review & Editing, Supervision, Project Administration. **Mahda Al Maida:** Software, Validation, Visualization, Writing–Review & Editing. **A. Qeis Tenridapi:** Data Curation, Investigation, Formal Analysis, Writing–Review & Editing. **Bagus Sartono:** Conceptualization, Methodology, Validation, Writing–Review & Editing, Supervision. **Aulia Rizki Firdawanti:** Investigation, Data Curation, Formal Analysis, Writing–Review & Editing. **Budi Susetyo:** Validation, Writing–Review & Editing, Supervision. **Gerry Alfa Dito:** Data Curation, Visualization, Software, Writing–Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Generative AI and AI-assisted Technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) solely to improve language clarity, grammar, and readability. The authors subsequently reviewed and edited all AI-assisted content and take full responsibility for the accuracy, integrity, and final content of the manuscript. Generative AI was not used to generate, modify, or interpret the research data or to conduct the statistical analysis.

Funding and Acknowledgments

This research was funded by personal funds and did not receive any specific grant from public, commercial, or not-for-profit funding agencies. The authors would like to express their sincere gratitude to **IPB University**, particularly the *Statistics and Data Science Study Program*, for academic support and a constructive research environment. The authors also acknowledge the Central Statistics Agency (BPS) for providing access to SUSENAS data used in this study.

Data and Code Availability

The SUSENAS microdata analyzed in this study are owned and regulated by the Central Statistics Agency (BPS) and are not publicly available. Access to the data can be requested through official BPS procedures. The code (pre-processing, modeling, and evaluation scripts) used to support the findings of this study is available from the corresponding author upon reasonable request. If required due to ethical/legal constraints, the authors may provide a reproducible workflow using synthetic/aggregated data that preserves confidentiality.

References

[1] Republic of Indonesia. *Law of the Republic of Indonesia Number 20 of 2003 concerning the National Education System*. Jakarta, 2003. https://peraturan.bpk.go.id/Home/Details/43920/uu-no-20-tahun (visited on 08/09/2026).

[2] Badan Pusat Statistik. *Statistik Pendidikan 2023*. Jakarta, 2023. https://www.bps.go.id/id/publication/2023/11/24/54557f7c1bd32f187f3cdab5/statistik-pendidikan-2023.html (visited on 08/09/2026).

[3] Suharti. "Educational inequality and household socioeconomic status in Indonesia". In: *Journal of Indonesian Economy and Business* 36.2 (2021), pp. 123–140.

[4] Nadia Fairuza Azzahra. *Addressing Distance Learning Barriers in Indonesia amid the COVID-19 Pandemic*. Jakarta, 2020. DOI: 10.35497/309162. https://hdl.handle.net/10419/249436 (visited on 08/09/2026).

[5] United Nations Development Programme. *Human Development Report 2019: Beyond Income, Beyond Averages, Beyond Today*. New York: United Nations Development Programme, 2019. https://hdr.undp.org/content/human-development-report-2019 (visited on 08/09/2026).

[6] Andi Illa Erviani Nensi, Dela Gustiara, Shalshabilla Shafa, and Budi Susetyo. "Analysis of the Relationship between Literacy, Numeracy and School Accreditation Rankings in Sulawesi Using Ordinal Logistic Regression and K-Nearest Neighbors". In: *Journal of Mathematics, Computations and Statistics* 9.2 (June 2026), pp. 280–293. DOI: 10.35580/Jmathcos11260. https://doi.org/10.35580/Jmathcos11260.

[7] Haibo He and Edwardo A. Garcia. "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[8] Bartosz Krawczyk. "Learning from imbalanced data: Open challenges and future directions". In: *Progress in Artificial Intelligence* 5.4 (2016), pp. 221–232. DOI: 10.1007/s13748-016-0094-0.

[9] Daniele Micci-Barreca. "A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems". In: *ACM SIGKDD Explorations Newsletter* 3.1 (2001), pp. 27–32. DOI: 10.1145/507533.507538. https://doi.org/10.1145/507533.507538.

[10] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. New York: Springer, 2013. DOI: 10.1007/978-1-4614-6849-3. https://doi.org/10.1007/978-1-4614-6849-3.

[11] Christophe Ambroise and Geoffrey J. McLachlan. "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data". In: *Proceedings of the National Academy of Sciences* 99.10 (2002), pp. 6562–6566. DOI: 10.1073/pnas.102102699. https://doi.org/10.1073/pnas.102102699.

[12] Gavin C. Cawley and Nicola L. C. Talbot. "On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation". In: *Journal of Machine Learning Research* 11.70 (2010), pp. 2079–2107. https://www.jmlr.org/papers/v11/cawley10a.html.

[13] Damjan Krstajic, Ljubomir J. Buturovic, David E. Leahy, and Simon Thomas. "Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models". In: *Journal of Cheminformatics* 6.1 (2014), p. 10. DOI: 10.1186/1758-2946-6-10. https://doi.org/10.1186/1758-2946-6-10.

[14] Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. "SMOTEBoost: Improving prediction of the minority class in boosting". In: *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*. Cavtat-Dubrovnik, Croatia, 2003, pp. 107–119. DOI: 10.1007/978-3-540-39804-2_12.

[15] Yoav Freund and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. DOI: 10.1006/jcss.1997.1504.

[16] Yanmin Sun, Mohamed S. Kamel, Andrew K. C. Wong, and Yang Wang. "Cost-sensitive boosting for classification of imbalanced data". In: *Pattern Recognition* 40.12 (2007), pp. 3358–3378. DOI: 10.1016/j.patcog.2007.04.009.

[17] Christopher Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. "RUSBoost: A hybrid approach to alleviating class imbalance". In: *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans* 40.1 (2010), pp. 185–197. DOI: 10.1109/TSMCA.2009.2029559.

[18] Rongfang Gao and Zhanyu Liu. "An Improved AdaBoost Algorithm for Hyperparameter Optimization". In: *Journal of Physics: Conference Series* 1631.1 (2020), p. 012048. DOI: 10.1088/1742-6596/1631/1/012048. https://doi.org/10.1088/1742-6596/1631/1/012048.

[19] T. R. Mahesh, V. Vinoth Kumar, V. Dhilip Kumar, Oana Geman, Martin Margala, and Manisha Guduri. "The Stratified K-Folds Cross-Validation and Class-Balancing Methods with High-Performance Ensemble Classifiers for Breast Cancer Classification". In: *Healthcare Analytics* 4 (2023), p. 100247. DOI: 10.1016/j.health.2023.100247. https://doi.org/10.1016/j.health.2023.100247.

[20] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.