Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemdiktisaintek, No.10/C/C3/DT.05.00/2025

Evaluating the Impact of SMOTE-Based Data Balancing on Decision Bias and Algorithmic Fairness in XGBoost-Based Student Dropout Prediction

Les Endahti[1,*], Taqwa Hariguna[2], Dhanar Intan Surya Saputra[3]

[1.2,3] Magister of Computer Science, Amikom Purwokerto University, Indonesia

ARTICLE INFO

History of the article:
Received June 1, 2026
Revised July 3, 2026
Accepted August 7, 2026

Keywords:
Student Dropout Prediction
Educational Data Mining
XGBoost
SMOTE
Decision Bias
Algorithmic Fairness

Correspondence:

E-mail: endahtijhd@gmail.com

ABSTRACT

Student dropout prediction is a key application of Educational Data Mining for supporting early intervention in higher education. However, previous studies have primarily focused on improving predictive accuracy, while the effects of data balancing on decision bias and algorithmic fairness remain underexplored. This study proposes a comprehensive evaluation framework that integrates predictive performance, decision bias, and algorithmic fairness to assess the impact of the Synthetic Minority Over-sampling Technique (SMOTE) on Extreme Gradient Boosting (XGBoost) for student dropout prediction. Experiments were conducted using the publicly available *Predict Students Dropout and Academic Success* dataset containing 4,424 student records. After excluding the *Enrolled* class, the dataset was transformed into a binary classification problem consisting of 2,209 Graduate (60.9%) and 1,421 Dropout (39.1%) instances. Two models were compared: a baseline XGBoost classifier and an XGBoost classifier trained with SMOTE. Predictive performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC, while decision bias and algorithmic fairness were assessed using the False Negative Rate (FNR), Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD). The baseline model achieved higher Accuracy (93.11% vs. 92.29%), Precision (91.49% vs. 89.86%), F1-score (91.17% vs. 90.18%), and a lower FNR (0.0915 vs. 0.0951), whereas both models produced comparable ROC-AUC values (0.972). McNemar's test indicated that the difference in predictive performance was not statistically significant (p = 0.264). Although SMOTE did not improve predictive performance, it produced modest reductions in Statistical Parity Difference (0.2310–0.2218), Equal Opportunity Difference (0.0298–0.0233), and Average Odds Difference (0.0295–0.0235), indicating a slight improvement in fairness metrics while maintaining comparable discrimination capability. These findings highlight the trade-off between predictive performance and algorithmic fairness and demonstrate that evaluating predictive performance together with decision bias and fairness provides a more comprehensive assessment of educational machine learning models, supporting the development of responsible AI-based educational decision-support systems.

1. INTRODUCTION

Student dropout remains one of the most significant challenges in higher education because it negatively affects students, universities, and society. Students who discontinue their studies often experience lower employability, reduced lifetime earnings, and limited social mobility, while universities face declining graduation rates, financial losses, and reduced institutional performance (Niyogisubizo et al., 2022; Putra et al., 2025; Carballo-Mendívil et al., 2025). Consequently, improving student retention has

become a strategic priority for higher education institutions, making the early identification of students at risk of dropping out essential for timely intervention (Shafiq et al., 2022; Kim et al., 2023). Recent advances in Educational Data Mining (EDM) and machine learning have enabled universities to analyze demographic, socioeconomic, and academic characteristics to predict student dropout, thereby supporting academic advising, mentoring, counseling, and other intervention programs (Shakeel & Butt, 2015; Pal, 2012; Shafiq et al., 2022). Among various machine learning algorithms, Extreme Gradient Boosting (XGBoost) has demonstrated strong predictive capability because of its effectiveness in modeling complex nonlinear relationships while maintaining computational efficiency (Romsaiyud et al., 2024; Ridwan et al., 2024; Carballo-Mendívil et al., 2025).

Despite these advances, student dropout prediction remains challenging because educational datasets frequently exhibit class imbalance, where graduate students substantially outnumber dropout students. This issue is also reflected in the Predict Students Dropout and Academic Success dataset used in this study. The original dataset consists of 2,209 Graduate, 1,421 Dropout, and 794 Enrolled instances. After excluding the *Enrolled* class, the prediction task comprises 2,209 Graduate (60.9%) and 1,421 Dropout (39.1%) students, resulting in a moderately imbalanced class distribution that motivates the evaluation of data balancing techniques. Such imbalance may bias machine learning models toward the majority class, resulting in high overall accuracy while reducing the capability to identify minority-class instances accurately (Amin et al., 2016; Cheng et al., 2019; Kim, 2021). In educational applications, this limitation is particularly important because false negative predictions may prevent universities from identifying students who genuinely require academic support, thereby reducing the effectiveness of early intervention programs (Shafiq et al., 2022; Niyogisubizo et al., 2022). To address this issue, previous studies have widely adopted the Synthetic Minority Over-sampling Technique (SMOTE) and its variants to generate synthetic minority-class samples and improve class representation during model training (Yan et al., 2019; Sharma et al., 2022; Wang & Awang, 2024; Huang, 2025). However, the effectiveness of SMOTE varies across datasets and is typically evaluated only using conventional predictive performance metrics, such as Accuracy, Precision, Recall, F1-score, and ROC-AUC, with limited attention given to its potential impact on decision bias and algorithmic fairness (Miftahushudur et al., 2024; Kumar et al., 2025).

Beyond predictive performance, recent studies have emphasized that machine learning models should also be evaluated from the perspective of algorithmic fairness as part of Responsible Artificial Intelligence (Mitchell et al., 2021; Pessach & Shmueli, 2022; Schenk & Kern, 2024). A highly accurate classifier may still produce systematically different prediction outcomes across demographic groups, potentially resulting in unequal educational opportunities. Consequently, fairness metrics such as Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD) have been introduced to quantify demographic disparities in prediction outcomes (Mitchell et al., 2021; Pessach & Shmueli, 2022). Nevertheless, fairness evaluation remains relatively underexplored in Educational Data Mining, where most studies continue to prioritize predictive accuracy over equitable decision-making. Consequently, fairness evaluation should complement conventional predictive performance metrics to ensure that educational decision-support systems produce equitable outcomes across demographic groups.

Based on the existing literature, three methodological research gaps can be identified. First, previous studies on student dropout prediction have primarily emphasized improving predictive performance through advanced machine learning algorithms, particularly XGBoost, while giving limited attention to decision bias and algorithmic fairness (Ridwan et al., 2024; Carballo-Mendívil et al., 2025). Second, although SMOTE and its variants have been widely employed to address class imbalance, their effectiveness has generally been evaluated using conventional performance metrics, with limited investigation of their impact on fairness outcomes in educational prediction tasks (Yan et al., 2019; Sharma et al., 2022; Wang & Awang, 2024; Huang, 2025). Third, studies on algorithmic fairness have predominantly focused on defining and measuring fairness in machine learning applications (Mitchell et al., 2021; Pessach & Shmueli, 2022), but few have integrated predictive performance, decision bias, and algorithmic fairness into a unified evaluation framework for XGBoost-based student dropout prediction. Therefore, empirical evidence regarding the trade-off between predictive performance and fairness in educational machine learning models remains limited.

Therefore, this study investigates the impact of SMOTE-based data balancing on predictive performance, decision bias, and algorithmic fairness in XGBoost-based student dropout prediction. Two experimental scenarios are compared: a baseline XGBoost model trained on the original imbalanced dataset and an XGBoost model trained using a SMOTE-balanced training dataset. Unlike previous studies that primarily evaluated predictive performance or class balancing independently, this study proposes a comprehensive evaluation framework that jointly assesses predictive performance, decision bias, and algorithmic fairness within a single experimental setting. The proposed framework also enables the analysis of the trade-off between predictive performance and fairness, thereby providing broader empirical evidence to support the development of more responsible, fair, and trustworthy AI-based educational decision-support systems.

2. LITERATURE REVIEW

Student dropout prediction has become one of the primary research topics in Educational Data Mining (EDM) because of its ability to support early intervention and improve student retention. Early studies mainly employed conventional machine learning algorithms, such as Decision Trees, Naïve Bayes, Logistic Regression, and Support Vector Machines, to predict student dropout using demographic, socioeconomic, and academic attributes (Pal, 2012; Shakeel & Butt, 2015). A comprehensive review by Shafiq et al. (2022) concluded that machine learning techniques consistently outperform conventional statistical approaches in identifying students at risk of dropping out. More recently, ensemble learning methods have attracted increasing attention due to their superior predictive capability. Niyogisubizo et al. (2022) proposed a stacked ensemble model for student dropout prediction, while Romsaiyud et al. (2024), Ridwan et al. (2024), and Carballo-Mendívil et al. (2025) demonstrated that XGBoost provides competitive predictive performance because of its capability to model nonlinear relationships and heterogeneous educational data. Likewise, Kim et al. (2023) reported that optimizing machine learning models can achieve both high precision and high recall for early dropout prediction. These findings consistently indicate that XGBoost has become one of the most promising algorithms for educational prediction.

Despite the encouraging predictive performance achieved by recent machine learning models, educational datasets commonly exhibit class imbalance, where the number of graduate students substantially exceeds the number of dropout students. Such imbalance causes learning algorithms to become biased toward the majority class and often leads to poor recognition of minority-class observations (Amin et al., 2016). To address this limitation, numerous oversampling techniques have been proposed, among which the Synthetic Minority Over-sampling Technique (SMOTE) remains the most widely adopted because it generates synthetic minority samples rather than simply replicating existing observations (Cheng et al., 2019). Building upon the original SMOTE algorithm, several improvements have been introduced. Cheng et al. (2019) incorporated a grouped oversampling strategy with noise filtering, Yan et al. (2019) proposed a parameter-free cleaning mechanism, Sharma et al. (2022) integrated SMOTE with Generative Adversarial Networks, Wang and Awang (2024) developed MKC-SMOTE for multiclass classification, and Huang (2025) introduced FS-SMOTE by considering feature-space characteristics during sample generation. These studies consistently demonstrate that improving the quality of synthetic samples can enhance minority-class representation and classification performance.

However, the reported effectiveness of SMOTE is not entirely consistent across different application domains. Kim (2021) observed that oversampling may introduce noisy synthetic observations if neighborhood selection is not properly controlled, whereas Kumar et al. (2025) showed that hybrid oversampling strategies are more effective for datasets exhibiting extreme class imbalance. Similarly, recent studies combining diffusion models, semi-supervised learning, and adaptive oversampling have suggested that no single balancing technique consistently outperforms others under all data distributions (Kim et al., 2025; Jeong et al., 2026; Fu et al., 2024). These contrasting findings imply that the effectiveness of SMOTE depends strongly on dataset characteristics rather than on the oversampling algorithm alone. Therefore, its effectiveness should be evaluated empirically for each educational dataset instead of being assumed universally beneficial.

Besides predictive performance, recent advances in Responsible Artificial Intelligence have shifted research attention toward algorithmic fairness. Mitchell et al. (2021) argued that fairness cannot be

represented by a single definition because different fairness metrics evaluate different forms of discrimination. Likewise, Pessach and Shmueli (2022) emphasized that fairness should be assessed together with predictive performance since optimizing one objective does not necessarily improve the other. Schenk and Kern (2024) further highlighted that fairness constitutes an important quality dimension for machine learning systems, particularly when prediction outcomes influence human decision-making. Consequently, metrics such as Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD) have become widely adopted for evaluating demographic disparities. Nevertheless, fairness-oriented research has predominantly focused on healthcare, finance, recruitment, and public services, whereas its application in Educational Data Mining remains relatively limited.

The existing literature demonstrates that previous studies have evolved along three major methodological directions. The first direction focuses on improving predictive performance through increasingly sophisticated machine learning algorithms, particularly XGBoost, to enhance the accuracy of student dropout prediction (Niyogisubizo et al., 2022; Romsaiyud et al., 2024; Carballo-Mendívil et al., 2025). However, these studies primarily evaluate conventional performance metrics and provide limited analysis of decision bias or algorithmic fairness. The second direction addresses class imbalance using SMOTE and its variants to improve minority-class representation (Cheng et al., 2019; Sharma et al., 2022; Wang & Awang, 2024; Huang, 2025), but their effectiveness is generally assessed only in terms of predictive performance. The third direction investigates algorithmic fairness as an essential component of Responsible Artificial Intelligence (Mitchell et al., 2021; Pessach & Shmueli, 2022; Schenk & Kern, 2024), although most fairness-oriented studies have been conducted in domains such as healthcare, finance, and recruitment rather than Educational Data Mining. Consequently, limited research has integrated predictive performance, decision bias, and algorithmic fairness into a unified evaluation framework for XGBoost-based student dropout prediction. This methodological gap motivates the present study, which simultaneously evaluates these three dimensions to provide a more comprehensive assessment of responsible machine learning for educational decision-support systems.

3. RESEARCH METHODS

This study employed a quantitative experimental approach to investigate the impact of the Synthetic Minority Over-sampling Technique (SMOTE) on predictive performance, decision bias, and algorithmic fairness in student dropout prediction using the Extreme Gradient Boosting (XGBoost) algorithm. As illustrated in Figure 1, the research workflow began with the Predict Students Dropout and Academic Success dataset, followed by data preparation, including the removal of the *Enrolled* class and the transformation of the target variable into a binary classification problem consisting of *Graduate* and *Dropout* classes. The dataset subsequently underwent preprocessing, including missing-value verification, duplicate checking, categorical feature encoding, feature selection, and stratified train-test splitting using an 80:20 ratio. Two experimental scenarios were then developed: the first employed the original imbalanced training dataset to construct a baseline XGBoost classifier, while the second applied SMOTE to generate a balanced training dataset prior to XGBoost model training. Both models were evaluated using the same testing dataset to ensure a fair comparison under identical experimental conditions. Finally, the prediction results were assessed using predictive performance metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC), decision bias through the False Negative Rate (FNR), algorithmic fairness using Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD), followed by McNemar's test to determine whether the observed performance differences between the two models were statistically significant.

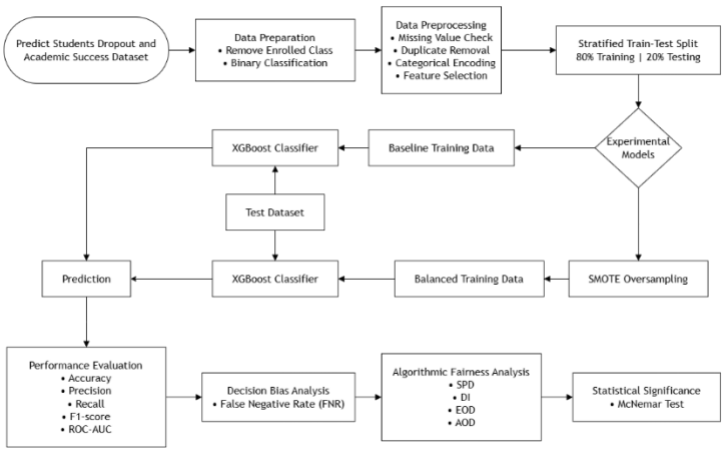

Figure 1. Research Flow

The experiments were conducted using the publicly available Predict Students Dropout and Academic Success dataset obtained from Kaggle. The dataset contains 4,424 student records, 36 predictor variables, and one target variable representing student academic status. The original dataset consists of 2,209 Graduate, 1,421 Dropout, and 794 Enrolled instances. Since the *Enrolled* class does not represent a final academic outcome, it was excluded from the analysis. Consequently, the prediction problem was transformed into a binary classification task comprising 2,209 Graduate (60.9%) and 1,421 Dropout (39.1%) instances, resulting in a moderately imbalanced class distribution.

Prior to model construction, the dataset underwent several preprocessing procedures, including dataset inspection, missing-value verification, duplicate checking, categorical feature encoding, feature-target separation, and stratified train-test splitting using an 80:20 ratio. No additional feature selection technique was applied because all predictor variables provided in the dataset were retained to preserve the original information and to ensure comparability with previous studies using the same benchmark dataset.

To overcome class imbalance, this study employed the Synthetic Minority Over-sampling Technique (SMOTE). Unlike random oversampling, SMOTE generates synthetic minority-class observations by interpolating between neighboring minority samples. The synthetic sample generation process is defined as

$$x_{new} = x_i + \lambda(x_{nn} - x_i), \lambda \in [0,1] \quad (1)$$

x_{new} denotes the synthetic sample, x_i represents the original minority observation, x_{nn} denotes one of its k-nearest minority neighbors, and λ is a random interpolation coefficient. In this study, SMOTE was implemented using sampling_strategy = auto, k_neighbors = 5, and random_state = 42 to ensure stable synthetic sample generation and experimental reproducibility.

Student dropout prediction was performed using the Extreme Gradient Boosting (XGBoost) algorithm. XGBoost constructs an ensemble of decision trees sequentially by minimizing a regularized objective function given by

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

$l(y_i, \hat{y}_i)$ denotes the loss function between the actual label y_i and the predicted label $\hat{y}_i$, while $\Omega(f_k)$ represents the regularization term that controls model complexity. The classifier was configured using n_estimators = 100, learning_rate = 0.10, max_depth = 6, subsample = 0.80, colsample_bytree = 0.80, objective = binary:logistic, and random_state = 42. The same hyperparameter configuration was maintained across both experimental scenarios to ensure that any observed differences were attributable only to the application of SMOTE. The XGBoost hyperparameters were selected based on commonly adopted configurations reported in previous studies on student dropout prediction and XGBoost classification (Ridwan et al., 2024; Carballo-Mendívil et al., 2025). No additional hyperparameter optimization, such as Grid Search or Bayesian Optimization, was performed because the primary objective of this study was to

isolate the effect of SMOTE while maintaining identical model configurations across both experimental scenarios.

The predictive performance of the proposed models was evaluated using Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). These metrics were derived from the confusion matrix. In addition, ROC-AUC was employed to evaluate the discrimination capability of the classifier across all possible classification thresholds. Besides predictive performance, this study evaluated decision bias using the False Negative Rate (FNR), which measures the proportion of actual dropout students incorrectly classified as graduates. The FNR is calculated as

$$FNR = \frac{FN}{TP + FN} \tag{3}$$

lower FNR values indicate better capability in identifying students who require early academic intervention. Furthermore, Gender was selected as the protected attribute because it is one of the most widely adopted sensitive attributes in the algorithmic fairness literature and is consistently available for all observations in the dataset. Other potentially sensitive variables, such as scholarship status and age, may also influence fairness in educational decision-making; however, these variables were not included in the present analysis and are recommended for future investigation. Four fairness metrics were employed. Statistical Parity Difference (SPD) measures the difference in positive prediction rates between demographic groups and is expressed as

$$SPD = P(\hat{Y} = 1 \mid A = u) - P(\hat{Y} = 1 \mid A = p) \tag{4}$$

u and $p$denote the unprotected and protected groups, respectively. Disparate Impact (DI) evaluates the ratio of positive prediction rates between demographic groups,

$$DI = \frac{P(\hat{Y} = 1 \mid A = p)}{P(\hat{Y} = 1 \mid A = u)} \tag{5}$$

Equal Opportunity Difference (EOD) measures the difference in True Positive Rates between demographic groups,

$$EOD = P(\hat{Y} = 1 \mid Y = 1, A = p) - P(\hat{Y} = 1 \mid Y = 1, A = u)] \tag{6}$$

Average Odds Difference (AOD) simultaneously evaluates disparities in both True Positive Rates and False Positive Rates,

$$AOD = \frac{(TPR_p - TPR_u) + (FPR_p - FPR_u)}{2} \tag{7}$$

For SPD, EOD, and AOD, values approaching zero indicate fairer prediction outcomes, whereas DI values approaching one indicate equitable treatment across demographic groups. Finally, the predictive performance of the baseline XGBoost model and the SMOTE-enhanced XGBoost model was statistically compared using McNemar's test. McNemar's test was employed to compare prediction disagreement between both models. Other statistical approaches, such as bootstrap confidence intervals or repeated cross-validation, were beyond the scope of the present study and are suggested for future research. The McNemar test statistic is calculated as

$$\chi^2 = \frac{(\mid b - c \mid -1)^2}{b + c}] \tag{8}$$

b represents the number of instances correctly classified by the baseline model but incorrectly classified by the SMOTE-enhanced model, whereas $c$represents the opposite condition. A significance level of $\alpha = 0.05$was adopted throughout the analysis. The overall experimental workflow consisted of data preprocessing, SMOTE-based data balancing, XGBoost model construction, predictive performance evaluation, decision bias assessment, algorithmic fairness analysis, and statistical significance testing.

4. RESULTS AND DISCUSSION

4.1. Predictive Performance Analysis

The predictive performance of the baseline XGBoost classifier and the SMOTE-enhanced XGBoost classifier was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. The experimental results are presented in Table 1. The experimental results indicate that the baseline XGBoost classifier achieved slightly better predictive performance than the SMOTE-enhanced model for most threshold-dependent evaluation metrics. The baseline model obtained an Accuracy of 93.11%, Precision of 91.49%, Recall of 90.85%, and F1-score of 91.17%, whereas the SMOTE-based model achieved 92.29%, 89.86%, 90.49%, and 90.18%, respectively. The decrease in Accuracy (0.82%), Precision (1.63%), Recall (0.36%), and F1-score (0.99%) indicates that applying SMOTE did not improve the classification performance of XGBoost for the investigated dataset.

Table 1. Performance Comparison between Baseline XGBoost and XGBoost + SMOTE

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Baseline XGBoost	0.9311	0.9149	0.9085	0.9117	0.9720
XGBoost + SMOTE	0.9229	0.8986	0.9049	0.9018	0.9721

In contrast, both models achieved nearly identical ROC-AUC values, with 0.9720 for the baseline model and 0.9721 for the SMOTE-enhanced model. Since ROC-AUC evaluates the discrimination capability across all possible classification thresholds, this result indicates that SMOTE preserved the ranking ability of XGBoost despite slightly reducing threshold-dependent performance. The almost identical ROC-AUC values also suggest that the predictive capability of XGBoost was already highly robust under the original class distribution.

These findings differ from several previous studies reporting that SMOTE generally improves Recall and F1-score by increasing minority-class representation (Cheng et al., 2019; Sharma et al., 2022; Wang & Awang, 2024). One possible explanation is that the student dropout dataset used in this study exhibits only a moderate class imbalance after excluding the Enrolled class. Consequently, the baseline XGBoost classifier was already capable of learning representative decision boundaries without requiring synthetic oversampling. Introducing synthetic samples slightly modified the feature distribution, resulting in a marginal reduction in classification performance.

Overall, the results indicate that applying SMOTE does not necessarily improve predictive performance when the original dataset exhibits only moderate class imbalance. Although threshold-dependent metrics (Accuracy, Precision, Recall, and F1-score) slightly decreased after oversampling, both models achieved nearly identical ROC-AUC values, indicating comparable discrimination capability across classification thresholds. This observation suggests a trade-off in which SMOTE slightly reduced predictive performance while providing modest improvements in fairness metrics, highlighting the importance of evaluating both objectives simultaneously rather than relying solely on conventional performance measures.

4.2. Decision Bias Analysis

Besides predictive performance, this study evaluated decision bias using the False Negative Rate (FNR). In educational decision-support systems, false negative predictions are particularly critical because students who are actually at risk of dropping out are incorrectly predicted as graduates, preventing universities from providing timely interventions. Table 2 presents the comparison of decision bias between the baseline XGBoost model and the SMOTE-enhanced XGBoost model using the False Negative Rate (FNR). The baseline model achieved a lower FNR of 0.0915, whereas the XGBoost + SMOTE model produced a slightly higher FNR of 0.0951. Since FNR measures the proportion of actual dropout students incorrectly classified as graduates, lower values indicate a better ability to identify students who require early academic intervention. The increase in FNR after applying SMOTE indicates that the oversampling process did not reduce decision bias and instead resulted in a marginal increase in false negative predictions. Although the difference between the two models is relatively small (0.36 percentage points), this finding suggests that generating synthetic minority samples did not improve the classifier's capability to recognize at-risk students. One possible explanation is that the original class imbalance was not sufficiently severe to

hinder the learning process of XGBoost, allowing the baseline model to learn representative decision boundaries without additional oversampling. Consequently, the synthetic samples generated by SMOTE may have slightly altered the feature distribution without providing meaningful information for improving minority-class identification. Therefore, in the context of this dataset, SMOTE was unable to reduce decision bias despite maintaining comparable overall predictive performance.

Table 2. Decision Bias Comparison Using False Negative Rate

Model	False Negative Rate
Baseline XGBoost	0.0915
XGBoost + SMOTE	0.0951

The confusion matrices of both experimental scenarios are shown in Figure 2 and Figure 3. The baseline model correctly classified 418 graduate students and 258 dropout students, while producing 24 false positive and 26 false negative predictions. After applying SMOTE, the classifier correctly identified 413 graduate students and 257 dropout students, whereas the numbers of false positives and false negatives increased to 29 and 27, respectively.

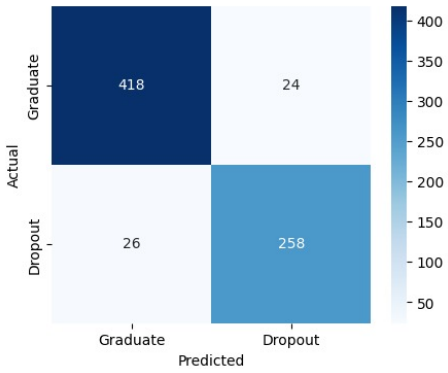

Figure 2. Confusion Matrix of Baseline XGBoost

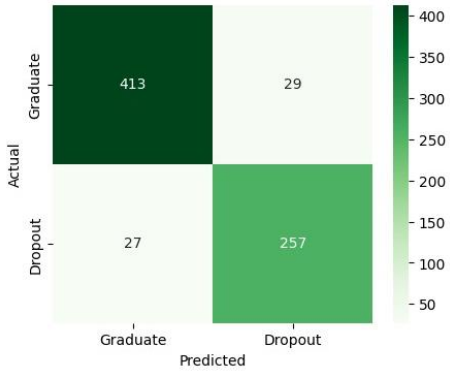

Figure 3. Confusion Matrix of XGBoost + SMOTE

Consequently, the False Negative Rate increased from 0.0915 to 0.0951, indicating that SMOTE did not reduce decision bias. Instead, one additional dropout student was incorrectly classified as a graduate. Although this increase appears numerically small, it is important from an educational perspective because every false negative prediction represents a student who may fail to receive early academic intervention.

These findings indicate that improving class balance does not automatically reduce harmful prediction errors. Similar observations have been reported in studies showing that oversampling may alter the decision boundary without substantially improving minority-class recognition, particularly when the original dataset already provides sufficient discriminatory information. Although FNR was selected because correctly identifying at-risk students is particularly important in educational intervention, future studies should complement this analysis using additional error-based measures, such as False Positive Rate (FPR), Balanced Accuracy, Matthews Correlation Coefficient (MCC), and class-specific recall, to provide a more comprehensive assessment of prediction bias.

4.3. Algorithmic Fairness Analysis

Beyond predictive performance and decision bias, this study investigated the effect of SMOTE on algorithmic fairness using Gender as the protected attribute. The fairness comparison between the baseline XGBoost model and the SMOTE-enhanced XGBoost model is presented in Table 3.

Table 3. Algorithmic Fairness Comparison

Model	SPD	DI	EOD	AOD
Baseline XGBoost	0.2310	0.5624	0.0298	0.0295
XGBoost + SMOTE	0.2218	0.5797	0.0233	0.0235

The experimental results indicate that applying SMOTE produced consistent, although relatively modest, improvements across all fairness metrics. The Statistical Parity Difference (SPD) decreased from 0.2310 to 0.2218, representing a reduction of approximately 4.0% in the disparity of positive prediction rates between male and female students. Similarly, the Equal Opportunity Difference (EOD) decreased from 0.0298 to 0.0233 (approximately 21.8%), suggesting a smaller difference in the probability of correctly identifying dropout students across demographic groups.

Comparable improvements were observed for Average Odds Difference (AOD), which decreased from 0.0295 to 0.0235, corresponding to a reduction of approximately 20.3% in the combined disparities of true positive and false positive rates. Furthermore, Disparate Impact (DI) increased from 0.5624 to 0.5797, moving closer to the ideal fairness value of 1.0, although it still remained below the commonly accepted fairness threshold of 0.80. Collectively, these results indicate that balancing the training data using SMOTE slightly reduced demographic disparities without substantially altering the predictive capability of the XGBoost classifier. Although DI increased from 0.5624 to 0.5797, the value remained substantially below the commonly referenced fairness threshold of 0.80. Therefore, while SMOTE slightly reduced demographic disparities, the resulting model cannot yet be considered free from demographic imbalance.

The observed fairness improvement can be attributed to the ability of SMOTE to generate additional minority-class samples, thereby providing a more balanced representation of dropout students during model training. A more balanced class distribution enables the classifier to learn decision boundaries that are less influenced by the majority class, which may indirectly reduce disparities in prediction outcomes between demographic groups. Nevertheless, the magnitude of improvement remained relatively small, indicating that class imbalance was not the sole source of unfairness in the prediction model.

These findings are consistent with previous studies emphasizing that fairness should be evaluated alongside predictive performance rather than treated as an independent objective (Mitchell et al., 2021; Pessach & Shmueli, 2022). Schenk and Kern (2024) further argued that algorithmic fairness constitutes an essential quality dimension of machine learning systems, particularly when prediction outcomes influence institutional decision-making. Therefore, although SMOTE alone cannot eliminate demographic disparities, its application contributes to slightly fairer prediction outcomes while maintaining comparable classification performance. This result highlights the importance of integrating predictive performance, decision bias, and algorithmic fairness into a unified evaluation framework for developing more responsible educational decision-support systems.

4.4. Statistical Significance Analysis

To determine whether the performance differences between the baseline XGBoost model and the SMOTE-enhanced XGBoost model were statistically significant, McNemar's test was conducted using the prediction results obtained from the same testing dataset. The results of the statistical test are presented in Table 4. The McNemar test produced a test statistic of 1.25 with a p-value of 0.2636, which exceeds the predefined significance level of 0.05. Consequently, the null hypothesis cannot be rejected, indicating that there is no statistically significant difference between the predictive performance of the baseline XGBoost model and the XGBoost model trained with SMOTE. In other words, the observed differences in classification outcomes are insufficient to conclude that the application of SMOTE resulted in a meaningful improvement or degradation in predictive performance.

Table 4. McNemar Test Results

Statistic	p-value	Decision
1.25	0.2636	No statistically significant difference

This statistical result is consistent with the evaluation presented in Table 1, where the differences in Accuracy, Precision, Recall, and F1-score between the two models were relatively small, while both models achieved almost identical ROC-AUC values. Although the baseline model exhibited slightly higher Accuracy (93.11% vs. 92.29%), Precision (91.49% vs. 89.86%), Recall (90.85% vs. 90.49%), and F1-score (91.17% vs. 90.18%), these differences were not statistically significant. Therefore, the minor variations observed in predictive performance are more likely attributable to normal sampling variability than to the application of the oversampling technique.

Interestingly, the statistical findings should be interpreted together with the fairness evaluation. While SMOTE did not produce statistically significant improvements in predictive performance, it consistently reduced demographic disparities across all evaluated fairness metrics, including SPD, EOD, and AOD, while slightly increasing DI toward the ideal value. This indicates that the primary benefit of SMOTE in this study was not enhancing classification accuracy, but rather improving the equity of prediction outcomes without significantly compromising model performance. These findings reinforce the importance of complementing conventional performance metrics with statistical significance testing and algorithmic fairness evaluation to obtain a more comprehensive assessment of machine learning models for educational decision-support systems. It should be noted that McNemar's test evaluates prediction disagreement between two classifiers but does not quantify uncertainty in the reported performance or fairness metrics. Consequently, future studies should complement McNemar's test with bootstrap confidence intervals, repeated cross-validation, or statistical tests specifically designed for fairness metrics to provide a more comprehensive assessment of model robustness.

4.5. Discussion

The experimental findings demonstrate that the impact of SMOTE extends beyond conventional predictive performance. While oversampling has frequently been reported to improve minority-class detection in highly imbalanced datasets (Amin et al., 2016; Cheng et al., 2019; Sharma et al., 2022; Wang & Awang, 2024), the present study shows that such improvements are not guaranteed when class imbalance is relatively moderate. The experimental results indicate that the baseline XGBoost model slightly outperformed the SMOTE-enhanced model in Accuracy, Precision, Recall, and F1-score, while both models achieved nearly identical ROC-AUC values. This finding is consistent with previous studies reporting that the effectiveness of SMOTE depends strongly on dataset characteristics, class distribution, and the quality of synthetic samples rather than the oversampling technique alone (Kim, 2021; Kumar et al., 2025; Huang, 2025). Therefore, SMOTE preserved the discrimination capability of XGBoost while producing only marginal changes in predictive performance.

More importantly, the fairness analysis reveals that SMOTE consistently reduced demographic disparities across all evaluated fairness metrics, although the observed improvements were relatively modest. Specifically, Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD) decreased, while Disparate Impact (DI) moved closer to the ideal value of one. These findings suggest that evaluating machine learning models solely using Accuracy, Precision, Recall, or F1-score may provide an incomplete assessment of model quality. Recent studies have emphasized that predictive performance should be evaluated together with fairness because highly accurate models may still produce unequal outcomes across demographic groups (Mitchell et al., 2021; Pessach & Shmueli, 2022). Furthermore, Schenk and Kern (2024) argued that algorithmic fairness should be regarded as an essential quality dimension of machine learning systems, particularly when prediction results influence institutional decision-making. Consequently, incorporating decision bias and fairness metrics provides additional insights into the practical implications of predictive models, especially for educational decision-support systems where prediction outcomes directly affect student interventions.

The results demonstrate a clear trade-off between predictive performance and algorithmic fairness. While SMOTE slightly reduced Accuracy (−0.82%), Precision (−1.63%), Recall (−0.36%), and F1-score (−0.99%), it consistently reduced SPD, EOD, and AOD and increased DI toward the ideal value. This finding suggests that evaluating educational machine learning models requires balancing predictive effectiveness with equitable decision-making rather than optimizing a single objective.

Overall, the results indicate that SMOTE should not be regarded solely as a technique for improving classification performance. Instead, its primary contribution in this study lies in producing slightly fairer prediction outcomes while preserving the strong predictive capability of XGBoost. This observation supports the growing perspective of Responsible Artificial Intelligence, which advocates balancing predictive performance with fairness and transparency to ensure equitable and trustworthy AI systems (Mitchell et al., 2021; Pessach & Shmueli, 2022; Berengueres, 2024; Amirian et al., 2025). Therefore, the findings reinforce the importance of integrating predictive performance, decision bias, and algorithmic

fairness within a unified evaluation framework, enabling educational institutions to develop more responsible, transparent, and trustworthy AI-based decision-support systems for student dropout prediction.

From a practical perspective, the findings suggest that educational institutions should not rely exclusively on predictive accuracy when deploying AI-based early warning systems. Instead, predictive performance should be monitored together with fairness metrics to ensure that model updates do not unintentionally increase demographic disparities. Periodic fairness auditing, continuous monitoring of prediction outcomes across demographic groups, and model retraining when substantial disparities emerge represent important components of responsible AI deployment in higher education. Furthermore, because the present study was conducted using a single public dataset, caution should be exercised when generalizing the findings to institutions with different demographic compositions or educational systems. Future studies should validate the proposed evaluation framework using multiple datasets and incorporate explainable AI techniques, such as SHAP, to improve transparency and support institutional decision-making.

5. CONCLUSIONS AND RECOMMENDATIONS

This study investigated the impact of the Synthetic Minority Over-sampling Technique (SMOTE) on predictive performance, decision bias, and algorithmic fairness in student dropout prediction using the Extreme Gradient Boosting (XGBoost) algorithm. The experimental results demonstrated that the baseline XGBoost model achieved slightly higher predictive performance than the SMOTE-enhanced model, with higher Accuracy, Precision, Recall, F1-score, and a lower False Negative Rate, while both models produced comparable ROC-AUC values. McNemar's test further confirmed that the observed differences in predictive performance were not statistically significant ($p = 0.2636$). Although SMOTE did not improve predictive performance, it produced modest numerical improvements in algorithmic fairness by reducing Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD), while increasing Disparate Impact (DI) toward the ideal value. However, these fairness improvements should be interpreted descriptively because their statistical significance was not evaluated. Overall, the findings indicate a trade-off between predictive performance and algorithmic fairness, suggesting that evaluating educational machine learning models requires balancing predictive effectiveness with equitable decision-making rather than optimizing predictive accuracy alone.

The primary contribution of this study is the development of a comprehensive evaluation framework that jointly assesses predictive performance, decision bias, and algorithmic fairness in student dropout prediction, thereby providing a broader perspective for responsible AI-based educational decision-support systems. Nevertheless, this study has several limitations, including the use of a single public dataset, one oversampling technique (SMOTE), one classification algorithm (XGBoost), and Gender as the only protected attribute. In addition, statistical significance was evaluated only for predictive performance using McNemar's test, whereas fairness metrics were interpreted descriptively. Future research should compare alternative class-balancing techniques, such as ADASYN, Borderline-SMOTE, SMOTEENN, and cost-sensitive learning, evaluate additional protected attributes (e.g., socioeconomic status, scholarship status, and age), validate the proposed framework using datasets from multiple educational institutions, incorporate explainable AI techniques such as SHAP or LIME, and employ more comprehensive statistical analyses, including bootstrap confidence intervals or repeated cross-validation, to strengthen the robustness of both predictive performance and algorithmic fairness evaluation.

REFERENCES

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., & Hussain, A. (2016). *Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study*. IEEE Access, 4, 7940–7957. https://doi.org/10.1109/ACCESS.2016.2619719

Amirian, S., Gao, F., Littlefield, N., Hill, J. H., Jr., Plate, J. F., Pantanowitz, L., Rashidi, H., & Tafti, A. P. (2025). *State-of-the-Art in Responsible, Explainable, and Fair AI for Medical Image Analysis*. IEEE Access, 13, 58229–58263. https://doi.org/10.1109/ACCESS.2025.3555543

Arfaoui, N. (2026). *Enhancing Machine Learning Algorithms for Imbalanced Data—A Case Study: Spam Detection*. IEEE Access, 14. https://doi.org/10.1109/ACCESS.2026.3658624

Berengueres, J. (2024). *How to Regulate Large Language Models for Responsible AI*. IEEE Transactions on Technology and Society, 5(2), 191–197. https://doi.org/10.1109/TTS.2024.3403681

Biau, G. (2012). *Analysis of a Random Forests Model*. Journal of Machine Learning Research, 13, 1063–1095.

Breiman, L. (1999). *Random Forests—Random Features*. Technical Report 567, Department of Statistics, University of California, Berkeley.

Carballo-Mendívil, B., Arellano-González, A., Ríos-Vázquez, N. J., & Lizardi-Duarte, M. del P. (2025). *Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data*. Applied Sciences, 15(16), 9202. https://doi.org/10.3390/app15169202

Cheng, K., Zhang, C., Yu, H., Yang, X., Zou, H., & Gao, S. (2019). *Grouped SMOTE With Noise Filtering Mechanism for Classifying Imbalanced Data*. IEEE Access, 7, 170668–170681. https://doi.org/10.1109/ACCESS.2019.2955086

Fu, Z., Ma, H., Wang, F., Dou, J., Zhang, B., & Fang, Z. (2024). *An Integrated Framework of Positive-Unlabeled and Imbalanced Learning for Landslide Susceptibility Mapping*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17. https://doi.org/10.1109/JSTARS.2024.3452182

Goswami, R., Garai, A., Sadhukhan, P., Ghosh, P., & Chakraborty, T. (2025). *Shape Penalized Decision Forests for Imbalanced Data Classification*. IEEE Access, 13. https://doi.org/10.1109/ACCESS.2025.3569523

Hu, J., & Szymczak, S. (2023). *A Review on Longitudinal Data Analysis with Random Forest*. Briefings in Bioinformatics, 24(2). https://doi.org/10.1093/bib/bbad002

Huang, Y. (2025). *FS-SMOTE: An Improved SMOTE Method Based on Feature Space Scoring Mechanism for Solving Class-Imbalanced Problems*. IEEE Access, 13, 148074–148082. https://doi.org/10.1109/ACCESS.2025.3597794

Jeong, J., Kahng, H., & Kim, S. B. (2026). *Semi-Supervised Learning Under Extreme Class Imbalance in Wafer Bin Map Defect Pattern Classification*. IEEE Access, 14. https://doi.org/10.1109/ACCESS.2026.3659740

Kim, D., Lee, W.-S., Ko, Y.-W., & Lee, J.-G. (2025). *Separated and Independent Contrastive Semi-Supervised Learning for Imbalanced Datasets*. IEEE Access, 13, 105712–105723. https://doi.org/10.1109/ACCESS.2025.3580738

Kim, K. (2021). *Noise Avoidance SMOTE in Ensemble Learning for Imbalanced Data*. IEEE Access, 9, 143250–143265. https://doi.org/10.1109/ACCESS.2021.3120738

Kim, S., Choi, E., Jun, Y.-K., & Lee, S. (2023). *Student Dropout Prediction for University with High Precision and Recall*. Applied Sciences, 13(10), 6275. https://doi.org/10.3390/app13106275

Kumar, R., Kim, Y.-W., & Byun, Y.-C. (2025). *Hybrid Framework Combining Diffusion-Based Image Augmentation and Feature Level SMOTE for Addressing Extreme Class Imbalance*. IEEE Access, 13. https://doi.org/10.1109/ACCESS.2025.3600622

Lee, C.-C., Comes, T., Finn, M., Pak, H., Hsu, C.-W., & Mostafavi, A. (2026). *Roadmap Toward Responsible AI in Crisis Resilience and Management*. IEEE Access, 14. https://doi.org/10.1109/ACCESS.2026.3651368

Matey-Sanz, M., Granell, C., & Mollineda, R. A. (2026). *Responsible Integration of Generative AI in Software Engineering Education*. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 21.

Miftahushudur, T., Grieve, B., & Yin, H. (2024). *Permuted KPCA and SMOTE to Guide GAN-Based Oversampling for Imbalanced HSI Classification*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 489–505. https://doi.org/10.1109/JSTARS.2023.3326963

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). *Algorithmic Fairness: Choices, Assumptions, and Definitions*. Annual Review of Statistics and Its Application, 8, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Nazri, A., Agbolade, O., & Aziz, F. (2025). *Hybrid Reinforcement Learning-Active Learning Framework for Real-Data Augmentation in Imbalanced Credit Scoring*. IEEE Access, 13. https://doi.org/10.1109/ACCESS.2025.3608032

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). *Predicting Student's Dropout in University Classes Using Two-Layer Ensemble Machine Learning Approach: A Novel Stacked Generalization*. Computers and Education: Artificial Intelligence, 3, 100066. https://doi.org/10.1016/j.caeai.2022.100066

Pal, S. (2012). *Mining Educational Data to Reduce Dropout Rates of Engineering Students*. I.J. Information Engineering and Electronic Business, 4(2), 1–7. https://doi.org/10.5815/ijieeb.2012.02.01

Pessach, D., & Shmueli, E. (2022). *A Review on Fairness in Machine Learning*. ACM Computing Surveys, 55(3), Article 51. https://doi.org/10.1145/3494672

Putra, L. G. R., Prasetya, D. D., & Mayadi. (2025). *Student Dropout Prediction Using Random Forest and XGBoost Method*. INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi, 9(1), 147–157. https://doi.org/10.29407/intensif.v9i1.21191

Qin, H. (2025). *Maximal Information Coefficient-Based Undersampling Method for Highly-Imbalanced Learning*. IEEE Access, 13, 4126–4135. https://doi.org/10.1109/ACCESS.2025.3525475

Resende, P. A. A., & Drummond, A. C. (2018). *A Survey of Random Forest Based Methods for Intrusion Detection Systems*. ACM Computing Surveys, 51(3), Article 48. https://doi.org/10.1145/3178582

Ridwan, A., Priyatno, A. M., & Ningsih, L. (2024). *Predict Students' Dropout and Academic Success with XGBoost*. Journal of Education and Computer Applications, 1(2), 1–8.

Romsaiyud, W., Nurarak, P., Phiasai, T., Chadakaew, M., Chuenarom, N., Aksorn, P., & Thammakij, A. (2024). *Predictive Modeling of Student Dropout Using Intuitionistic Fuzzy Sets and XGBoost in Open University*. Proceedings of the 7th International Conference on Machine Learning and Machine Intelligence (MLMI 2024). https://doi.org/10.1145/3696271.3696288

Schenk, P. O., & Kern, C. (2024). *Connecting Algorithmic Fairness to Quality Dimensions in Machine Learning in Official Statistics and Survey Production*. AStA Wirtschafts- und Sozialstatistisches Archiv, 18, 131–184. https://doi.org/10.1007/s11943-024-00344-2

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). *Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review*. IEEE Access, 10, 72480–72515. https://doi.org/10.1109/ACCESS.2022.3188767

Sharma, A., Singh, P. K., & Chandra, R. (2022). *SMOTified-GAN for Class Imbalanced Pattern Classification Problems*. IEEE Access, 10, 30655–30672. https://doi.org/10.1109/ACCESS.2022.3158977

Shakeel, K., & Butt, N. A. (2015). *Educational Data Mining to Reduce Student Dropout Rate by Using Classification*. Proceedings of the International Conference on Information and Communication Technologies.

Sowmya, P., & Vasudeva. (2026). *Trust and Ethics in Chatbots: A Framework for Responsible AI Interactions*. IEEE Access, 14. https://doi.org/10.1109/ACCESS.2026.3653763

Tsai, H.-C., Lee, M.-C., & Hsu, C.-H. (2025). *Fault and Severity Diagnosis Using Deep Learning for Self-Organizing Networks With Imbalanced and Small Datasets*. IEEE Access, 13. https://doi.org/10.1109/ACCESS.2025.3537659

Wang, J., & Awang, N. (2024). *MKC-SMOTE: A Novel Synthetic Oversampling Method for Multi-Class Imbalanced Data Classification*. IEEE Access, 12, 196929–196938. https://doi.org/10.1109/ACCESS.2024.3521120

Wang, Z., Xie, J., & Zhang, J. (2024). *A Robust Enhanced Ensemble Learning Method for Breast Cancer Data Diagnosis on Imbalanced Data*. IEEE Access, 12, 189776–189788. https://doi.org/10.1109/ACCESS.2024.3516376

Xing, Y. (2024). *The Influence of Responsible Innovation on Ideological Education in Universities Under Generative Artificial Intelligence*. IEEE Access, 12, 133008–133017. https://doi.org/10.1109/ACCESS.2024.3459469

Yan, Y., Liu, R., Ding, Z., Du, X., Chen, J., & Zhang, Y. (2019). *A Parameter-Free Cleaning Method for SMOTE in Imbalanced Classification*. IEEE Access, 7, 23537–23548. https://doi.org/10.1109/ACCESS.2019.2899467