ITC 1/55
Information Technology and Control
Vol. 55 / No. 1/ 2026
pp. 143-165
DOI 10.5755/j01.itc.55.1.39399

Enhancing Fairness and Explainability in Student Performance Prediction Using Bias Mitigation Techniques

Received 2024/11/07 | Accepted after revision 2025/04/02

HOW TO CITE: Hou, N., Chen, H. (2026). Enhancing Fairness and Explainability in Student Performance Prediction Using Bias Mitigation Techniques. *Information Technology and Control*, 55(1), 143-165. https://doi.org/10.5755/j01.itc.55.1.39399

Enhancing Fairness and Explainability in Student Performance Prediction Using Bias Mitigation Techniques

Na Hou
Department of International Exchange, Jilin International Studies University, Changchun, Jilin 130117, China
e-mail: houn992@nenu.edu.cn;

Hongyao Chen*
International School of Chinese Studies, Northeast Normal University, Changchun, Jilin 130024, China

Corresponding author: tudoup2024@163.com

Ensuring fairness in artificial intelligence (AI)-driven student performance prediction remains a critical challenge, as biases in educational data can lead to unfair treatment of certain demographic groups. This study aims to develop fair and explainable AI models for predicting student performance in secondary education. Specifically, we investigate how bias mitigation techniques can be integrated with explainability methods to improve both fairness and interpretability without compromising predictive accuracy. We analyze a real-world dataset from Portuguese schools and apply machine learning models including Random Forests, XGBoost, and Logistic Regression. To mitigate bias, we implement fairness constraints and employ Adversarial Debiasing Representation Learning (ADRL). Post-hoc explainability is achieved using SHapley Additive Explanations (SHAP) to reveal the most influential factors in model predictions. Our findings demonstrate that bias mitigation techniques successfully reduce fairness violations while maintaining high predictive performance. The Bias Severity Index decreases from 0.35 to 0.08, and Demographic Parity improves from 15.3% to 4.2%. SHAP analysis reveals that factors such as study time, parental education, and previous grades have the most significant influence on student performance predictions. This study integrates fairness-aware learning with explainability tools, ensuring that AI models in education remain both equitable and interpretable.

KEYWORDS: Student Performance, Fairness in AI, Bias Mitigation, Explainability, Adversarial Debiasing, Model Transparency, Educational Data Mining, Predictive Modeling, AI.

1. Introduction

The use of Artificial Intelligence (AI) in educational settings has grown significantly in recent years, driven by the desire to improve student outcomes and optimize institutional decision-making [15]. Predictive models, powered by AI and machine learning techniques, have been increasingly adopted to forecast student performance, identify at-risk students, and enhance personalized learning [16]. However, there are growing concerns regarding the fairness, transparency, and ethical implications of these technologies, especially in light of their potential to perpetuate bias and inequity [5].

AI models can identify subtle relationships between student attributes, such as socioeconomic background, prior academic performance, and engagement levels, and educational outcomes [9]. As a result, AI is now being used to aid institutions in policy development, resource allocation, and curriculum design [10, 17]. The ability to generate predictions in real-time further enhances the value of AI, providing educators with actionable insights that can impact student support interventions [12]. The impact of these predictions is not solely determined by their accuracy; the fairness of the algorithms and the transparency of the decision-making processes are equally important for ensuring that these technologies do not reinforce existing disparities [20].

The use of AI in predicting student performance presents significant challenges, particularly concerning bias. Bias in educational machine learning systems can originate from two distinct sources: bias in the data and bias in the models, with the latter often being a consequence of the former. Data bias arises from historical inequalities, sampling imbalances, or underrepresentation of certain demographic groups, leading to skewed patterns in training datasets. If left unaddressed, these biases propagate into model bias, where the algorithm learns and reinforces disparities, potentially leading to unfair predictions.

Bias in AI models arises when algorithms disproportionately favor or disadvantage certain groups, often reflecting the underlying biases present in the training data. In educational contexts, these biases may lead to skewed predictions that harm historically marginalized student populations, such as those from lower socioeconomic backgrounds, minority ethnic groups, or students with disabilities. If left unaddressed, such biases can result in inequitable outcomes, exacerbating educational inequalities rather than mitigating them [7, 11].

Fairness refers to the principle that an algorithm's predictions or decisions do not result in unjustified disparities between different groups, particularly those defined by sensitive attributes such as gender, race, age, socioeconomic status, or disability. A fair model ensures that individuals with similar qualifications or circumstances receive similar outcomes, regardless of demographic characteristics. Addressing bias and fairness of AI models are critical to ensuring their ethical deployment in education.

This study aims to explore the extent of bias in AI models used to predict student performance, with a particular focus on enhancing the explainability of these models. The objectives are threefold:

1. Investigate the presence of bias in student data by analyzing their behavior across different student demographic groups.
2. Apply AI explainability techniques to shed light on the internal decision-making processes of these models, offering insights into how specific features influence predictions.
3. Propose a framework for bias mitigation, combining explainability techniques with algorithmic adjustments to ensure that the models provide fair and transparent outcomes.

This study makes several contributions to the field of fairness-aware AI in educational data mining:

- We introduce a bias mitigation framework that integrates fairness constraints into machine learning models, ensuring equitable student performance predictions.
- We propose an innovative Adaptive Hybrid Sampling (AHS) technique that dynamically adjusts oversampling and undersampling ratios to balance class distributions, reducing bias without compromising predictive accuracy.
- We implement an adversarial learning-based debiasing approach that disentangles sensitive attributes from performance-relevant features, ensuring that student predictions are not unfairly influenced by demographic factors.

The remainder of this paper is structured as follows: Section 2 reviews existing research. Section 3 presents the framework used in this study. Section 4 introduces the dataset used for analysis, describes its characteristics, and discusses the preprocessing steps undertaken to ensure data quality and fairness. Section 5 evaluates the performance of models, assesses fairness metrics, and presents explainability insights. Section 6 summarizes key findings, discusses implications of fairness-aware AI in education, identifies limitations, and outlines future research directions.

2. Related Work

Student performance prediction has attracted considerable attention, with various machine learning techniques being employed to tackle this challenge [3]. Recent studies have focused on handling educational datasets, often characterized by imbalances or multi-class classifications. For instance, Li et al. [13] explored the use of algorithms such as linear regression, ridge regression, and decision trees to predict student performance. They demonstrated that preprocessing student data by identifying impactful features can enhance the accuracy of machine learning models, making them more suitable for educational contexts [13]. Tariq et al. [19] emphasized the need for effective methods to address imbalanced datasets, especially in multi-class settings. They compared different oversampling techniques, including SMOTETomek, which, when combined with K-Nearest Neighbors (KNN), achieved the highest accuracy for predicting academic performance [19]. Alija et al. [2] focused on optimizing feature selection methods using Particle Swarm Optimization (PSO) to improve the prediction of course failures in an imbalanced dataset using Random Forest classifier combined with SMOTE. Swetha and Rahaman [18] analyzed the influence of socioeconomic factors on student performance using the NAS dataset. Their study employed Gradient Boosting Classifier, highlighting the importance of considering external factors such as socioeconomic background in predictive modeling. Miranda et al. [14] explored the impact of online learning environments, utilizing model-agnostic interpretability techniques to better understand the prediction outcomes, with Random Forest achieving the best results. Wongvorachan et al. [22] evaluated four bias mitigation techniques—reweighting, resampling, and Reject Option-based Classification (ROC)—in classification tasks related to dropout prediction. Their findings indicate that resampling reduces predictive bias at the cost of accuracy, while the ROC pivot achieves a balance between bias reduction and classifier performance, demonstrating the importance of context-specific approaches in bias mitigation. Idowu [1] reviewed fairness in machine learning applied to education, covering dropout prediction, performance prediction, forum post classification, and recommenders. The study identifies bias mitigation strategies, including sample reweighting, bias attenuation, fairness through awareness/unawareness, and adversarial learning, while evaluating fairness using ABROCA, group performance differences, and disparity metrics. A critical insight is that there is no strict trade-off between fairness and accuracy, suggesting that fairness interventions can be implemented without severely degrading model performance. The studies are summarized in Table 1.

Despite the substantial body of research on student performance prediction, several knowledge gaps remain. First, many studies have focused on optimizing predictive accuracy using techniques like feature selection, oversampling, and advanced classifiers, but relatively few have systematically addressed the trade-offs between accuracy and fairness, especially when considering sensitive attributes such as gender or socioeconomic status. This study addresses the research gap where AI-driven student performance prediction models prioritize accuracy but often overlook fairness trade-offs, making it difficult to ensure equitable and interpretable decision-making in educational settings.

Given these challenges, there is a need for a framework that integrates bias mitigation with explainability techniques, ensuring that AI-driven student performance predictions are both fair and interpretable while maintaining high accuracy. This study addresses this gap by developing a fair and explainable AI approach that detects, mitigates, and explains bias in student performance prediction models.

3. Proposed Framework

3.1. Framework Overview and Objectives

One of the challenges in deploying AI models in education is balancing the trade-offs between accuracy, fairness, and transparency. While highly accurate models are desirable for predicting student outcomes, this accuracy must not come at the expense of fairness, particularly when working with sensitive educational data. Bias in prediction models can lead to unfair treatment of certain demographic groups, reinforcing existing inequalities in the education system. In this framework, fairness is treated as a core performance metric alongside accuracy. Transparency is equally crucial, ensuring that model predictions are interpretable by educators and decision-makers. The framework uses state-of-the-art explainability techniques to make the model's decision-making process understandable, while also ensuring that fairness is not compromised.

Figure 1 outlines the sequential steps and decision points involved in the framework. Each phase focuses on ensuring fairness and explainability in the model for student performance data. The process begins with raw student performance data. Preprocessing stage checks if data balancing, fair representation learning, or handling missing data are required, and applies appropriate methods. In-processing stage incorporates fairness constraints, adversarial debiasing, and trains the model with fairness regularization. Post-processing stage audits the model for fairness, applies explainability techniques, and visualizes fairness trade-offs. The final stage involves evaluating the model across metrics (accuracy, fairness, explainability), testing with different data slices, and conducting sensitivity analysis. The Data Engineer, AI Model Developer, and Fairness Auditor must work closely to ensure the success of the framework (Figure 2). The Data Engineer provides unbiased data, which is essential for the Model Developer

Table 1

Summary of related work on student performance prediction and bias mitigation.

Study	Focus Area	Methods Used	Key Findings
Chachoui (2024) [3]	General ML techniques for student performance prediction	Machine learning models	Highlights the significance of using ML for academic predictions
Li et al. (2021) [13]	Feature selection for improving prediction accuracy	Linear regression, ridge regression, decision trees	Identified that selecting impactful features enhances model accuracy in education
Tariq et al. (2023) [19]	Handling imbalanced datasets in education	SMOTETomek, KNN	Found that SMOTETomek with KNN improves accuracy for multi-class student performance prediction
Alija et al. (2023) [2]	Optimization of feature selection for imbalanced datasets	Particle Swarm Optimization (PSO), SMOTE, Random Forest	Showed that PSO with SMOTE improves course failure prediction in imbalanced data
Swetha and Rahaman (2019) [18]	Socioeconomic influence on student performance	Gradient Boosting Classifier (GBC), NAS dataset	Emphasized the role of socioeconomic factors in predictive modeling
Miranda et al. (2024) [14]	Impact of online learning on performance prediction	Model-agnostic interpretability techniques, Random Forest	Random Forest provides the best results in an online learning setting
Wongvorachan et al. (2024) [22]	Bias mitigation in dropout prediction	Reweighting, resampling, ROC pivot	Identified resampling as effective but costly in accuracy; ROC pivot balances bias reduction and accuracy
Idowu (2024) [1]	Fairness in ML-based education predictions	Sample reweighting, adversarial learning, ABROCA	Concluded that fairness interventions can be implemented without significantly degrading model performance

to build fair models. The Model Developer creates AI models that are transparent and explainable, while the Fairness Auditor ensures that these models meet ethical standards for fairness and transparency. Their collaboration is vital to developing robust, trustworthy AI systems for student performance prediction.

3.2. Preprocessing for Bias Mitigation

In this step, preprocessing techniques are applied to the student performance dataset to mitigate potential biases before the model training phase. Bias can arise from imbalanced data distributions or missing values, which can lead to models that disproportionately favor certain groups. We propose three key preprocessing techniques: data balancing and resampling, fair representation learning, and handling missing data and demographic imbalances.

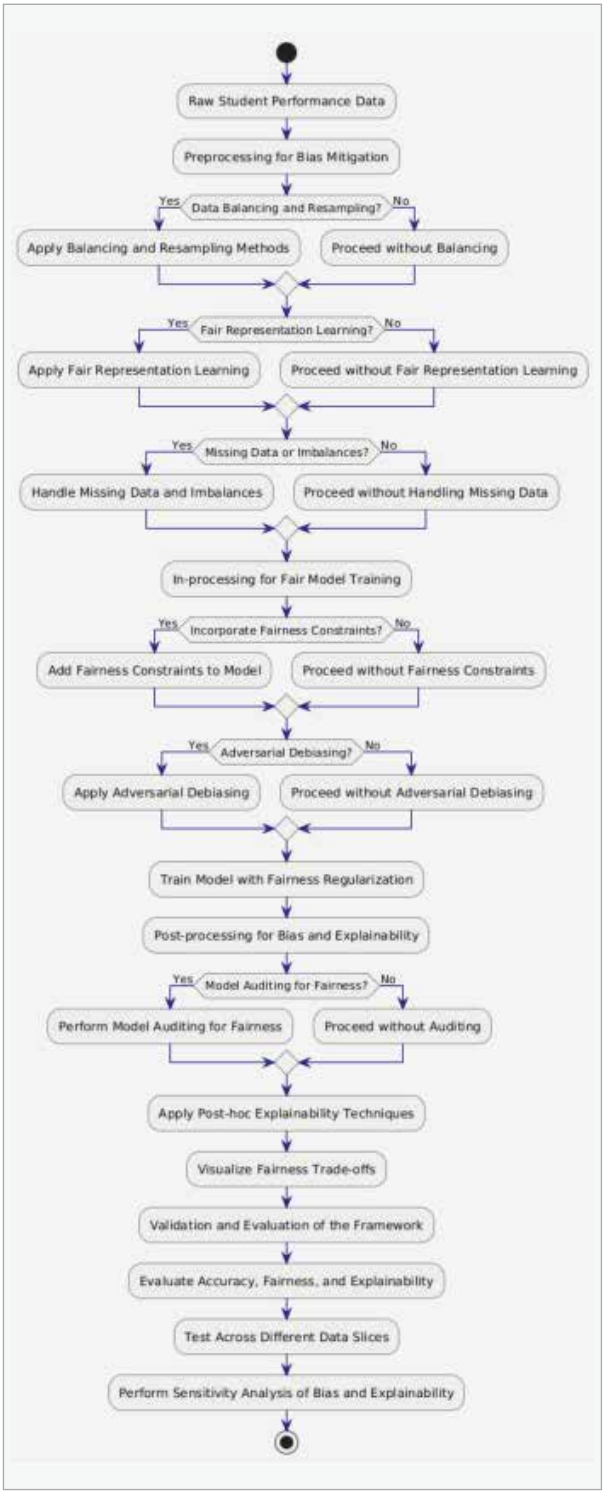
To mitigate bias caused by class imbalance or underrepresentation of certain demographic groups, we use data balancing and resampling methods. These techniques aim to create a more representative and equitable training dataset. We use a combination of oversampling of the minority class and undersampling of the majority class. For this task, we propose an innovative technique called Adaptive Hybrid Sampling (AHS), which combines oversampling of the minority class and undersampling of the majority class to balance data in the context of student performance datasets. The method, inspired by hybrid sampling technique [23], dynamically adjusts the proportions of oversampling and undersampling based on the class imbalance and iteratively achieves optimal balance. Steps in the AHS Algorithm are:

1 **Calculate Class Imbalance Ratio (CIR)** First, compute CIR to assess the degree of imbalance between the majority and minority classes. Let N_{maj} be the number of instances in the majority class, and N_{min} the number of instances in the minority class. The CIR is defined as:

$$\mathrm{CIR} = \frac{N_{maj}}{N_{\min}}. \tag{1}$$

2 **Set Thresholds for Balanced Class Sizes** Define a threshold T for a "balanced" class size. For instance, if the desired balance is a ratio of 1:1, then $T=1$. A tolerance range $[T_{\min}, T_{\max}]$, such as $[0.8,1.2]$, can be used to allow slight imbalance.

Figure 1
Overview of the Bias Mitigation and Explainability Framework.



3 **Determine Oversampling and Undersampling Proportions** Based on the calculated CIR, the following formulas adaptively determine the oversampling (α) and undersampling (β) proportions:

$$\alpha = \frac{\max(0, \text{CIR} - T_{\max})}{\text{CIR}} \tag{2}$$

$$\beta = \frac{\max(0, T_{\min} - \text{CIR})}{\text{CIR}}. \tag{3}$$

This ensures that when the CIR is high, more emphasis is placed on oversampling the minority class (α increases). When the CIR is low, undersampling the majority class is prioritized (β increases).

4 **Perform Adaptive Sampling**

- For oversampling the minority class, apply SMOTE [5] to generate synthetic data points:

$$x_{\text{new}} = x_i + \lambda \cdot (x_j - x_i), \quad \lambda \sim U(0,1), \tag{4}$$

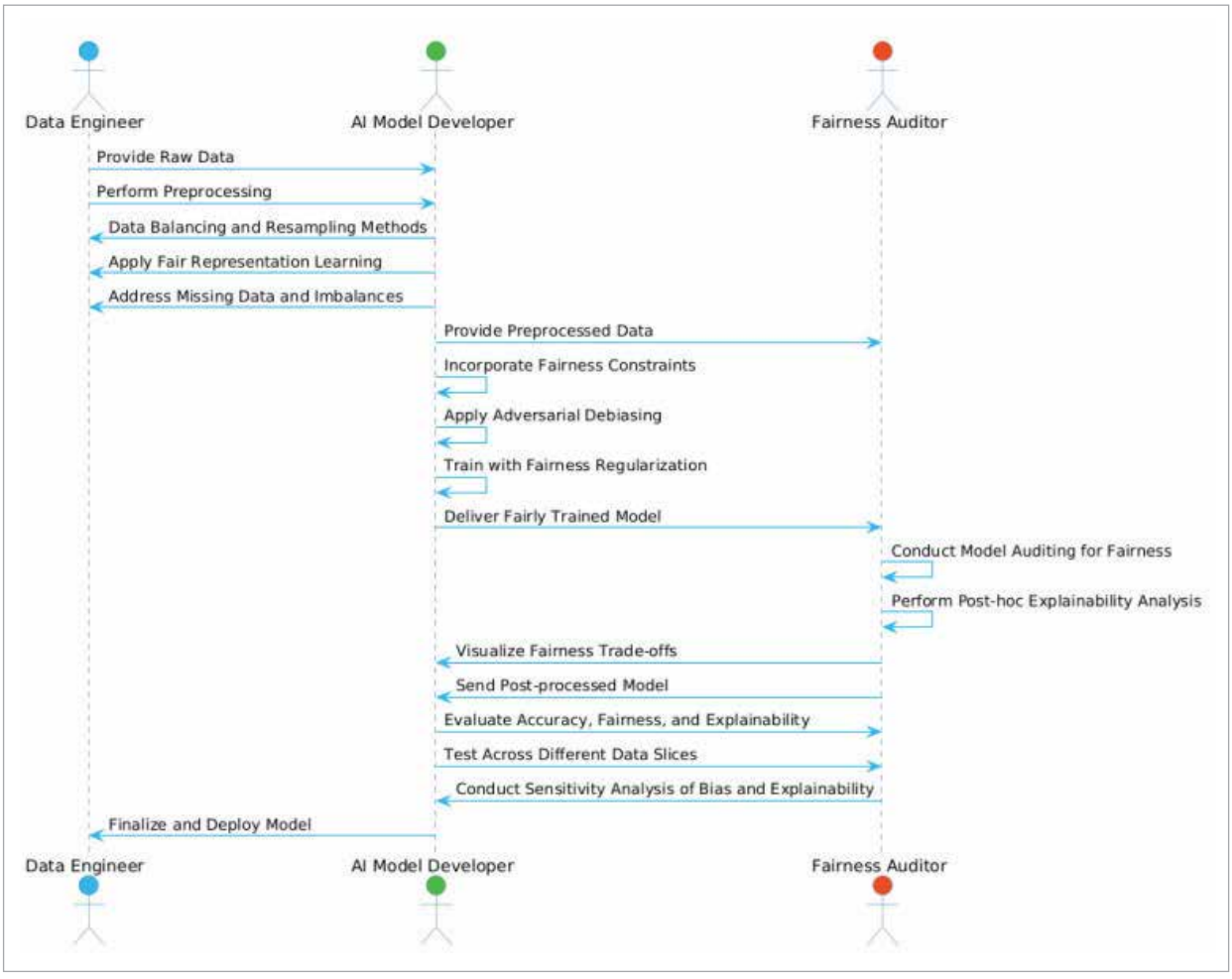
where x_i and x_j are the minority class samples.

- For undersampling the majority class, use K-means undersampling. Cluster the majority class samples and select representative points from each cluster to preserve diversity:

$$x_{\text{rep}} = \frac{1}{k}\sum_{i=1}^{k} x_i. \tag{5}$$

Figure 2

Roles of stakeholders within the Bias Mitigation and Explainability Framework.w



5 **Iterate and Adjust Proportions** After each round of oversampling and undersampling, recalculate the CIR to check if it falls within the desired range $[T_{\min}, T_{\max}]$. If not, adaptively adjust α and β and repeat until the desired balance is achieved.

6 **Stopping Condition** The process terminates when the recalculated CIR satisfies:

$$T_{\min} \le \frac{N'_{maj}}{N'_{\min}} \le T_{\max}, \tag{6}$$

where N'_{maj} and $N'_{\min}$ are the new size of majority and minority classes after applying the AHS.

Adaptive Hybrid Sampling (AHS) combines over-sampling and undersampling to address class imbalances in student performance datasets. By dynamically adjusting the proportions and employing clustering techniques, AHS reduces bias and ensures fair representation in AI models.

3.2.1. Fair Representation Learning

Fair representation learning techniques transform the feature space to reduce the impact of sensitive attributes (e.g., gender, race) on model predictions. The goal is to learn a latent representation that is both predictive of the target variable and fair with respect to sensitive attributes. Here we use Disentangled Representation Learning (DRL) [21], where the goal is to disentangle features into two sets: the sensitive and the invariant. This allows the model to make predictions based on the invariant features while ignoring the sensitive ones. The DRL algorithm aims to learn representations that separate out the sensitive attributes (e.g., gender, ethnicity) from the performance-relevant features. By disentangling these two types of features, the algorithm ensures that predictions of student performance are not influenced by biased attributes, while retaining the predictive power of performance-related features.

Here we introduce an innovative approach called Adversarial Disentangled Representation Learning (ADRL). In this method, a shared encoder is used to create latent representations that are split into two subspaces: one representing the performancerelevant information and another representing the sensitive, bias-inducing attributes. An adversarial network is trained to prevent the sensitive attributes from influencing the performance predictions. The ADRL algorithm, presented in Algorithm 2, is designed to debias educational student performance data by disentangling sensitive attributes (e.g., gender, ethnicity) from the performance-relevant features. The ADRL algorithm ensures that predictions about student performance are free from bias induced by these sensitive attributes while maintaining predictive accuracy.

Algorithm 1 Adaptive Hybrid Sampling (AHS)

1: **Input:** Majority class instances N_{maj}, Minority class instances $N_{\min}$, Threshold range $[T_{\min}, T_{\max}]$

2: **Output:** Balanced dataset

3: Calculate Class Imbalance Ratio (CIR) as

$$\text{CIR} = \frac{N_{maj}}{N_{\min}}$$

4: Define desired balance threshold T and tolerance range $[T_{\min}, T_{\max}]$.

5: Compute oversampling proportion α and undersampling proportion β adaptively based on CIR:

$$\alpha = \frac{\max(0, \text{CIR} - T_{\max})}{\text{CIR}}$$
$$\beta = \frac{\max(0, T_{\min} - \text{CIR})}{\text{CIR}}$$

6: Apply Sampling:

7: **if** Oversampling needed ($\alpha > 0$) **then**

$$x_{\text{new}} = x_i + \lambda \cdot (x_j - x_i), \quad \lambda \sim U(0,1)$$

8: **end if**

9: **if** Undersampling needed ($\beta > 0$) **then**

10: Apply K-means undersampling on the majority class:

$$x_{\text{rep}} = \frac{1}{k}\sum_{i=1}^{k} x_i$$

11: **end if**

12: Recalculate CIR after each round of oversampling and undersampling:

$$\text{CIR}_{\text{new}} = \frac{N'_{maj}}{N'_{\min}}$$

13: **if** $\text{CIR}_{\text{new}} \notin [T_{\min}, T_{\max}]$ **then**

14: Adjust α and β and repeat the process.

15: **else**

16: Proceed to the next step.

17: **end if**

18: Stop when:

$$T_{\min} \le \frac{N'_{maj}}{N'_{\min}} \le T_{\max}$$

19: Where N'_{maj} and $N'_{\min}$ are the new sizes of the majority and minority classes after applying AHS.

The key components of the algorithm are detailed as follows: The encoder $E(X)$ decomposes the input student data X into two latent representations:

- z_r: The latent representation that captures performance-relevant information.
- z_s: The latent representation that captures sensitive attributes (e.g., socioeconomic status).

The encoder is responsible for learning these disentangled representations such that z_r contains no sensitive information and is predictive of the target variable (student performance).

The adversarial model $A(z_r)$ attempts to predict the sensitive attributes S from the performance-relevant latent space z_r. The goal is to prevent the sensitive attributes from influencing the performance predictions. The adversarial loss L_a encourages the encoder to learn a representation z_r that minimizes the predictability of the sensitive attributes, effectively disentangling the two spaces.

The algorithm uses three loss functions:

- **Performance Prediction Loss $\mathbf{L}_p$:** This loss ensures that the representation z_r is predictive of the target variable Y (student performance) as:

$$\mathrm{L}_p = \mathbf{E}_{(X,Y)}[\ell(M(z_r), Y)], \tag{7}$$

  where M is the prediction model, and ℓ is the loss function (e.g., cross-entropy loss).
- **Adversarial Loss $\mathbf{L}_a$:** This loss penalizes the adversarial model's ability to predict the sensitive attribute S from z_r. It is computed as:

$$\mathrm{L}_a = -\mathbf{E}_{(X,S)}[\ell(A(z_r), S)], \tag{8}$$

  where A is the adversarial network, and ℓ is the binary cross-entropy loss. The negative sign ensures that the adversarial model is maximized while the encoder minimizes the predictability of S from z_r.
- **Disentanglement Loss $\mathbf{L}_d$:** This loss minimizes covariance between z_r and z_s to ensure that the two latent spaces are independent:

$$\mathrm{L}_d = \mathbf{E}_{(z_r,z_S)}[\mathrm{Cov}(z_r, z_S)]. \tag{9}$$

The ADRL **algorithm** follows an adversarial training loop, where the encoder, predictor, and adversarial models are updated in alternating steps:

- **Update Encoder and Predictor:** Encoder E and predictor M are trained to minimize the performance prediction loss L_p while also maximizing the adversarial loss L_a. This ensures that z_r remains predictive of Y but uninformative of S:

$$\theta_E \leftarrow \theta_E - \eta_r \nabla_{\theta_E}(\mathrm{L}_p - \lambda \mathrm{L}_a), \tag{10}$$

  where λ is a regularization term that controls the balance between adversarial debiasing and performance prediction.
- **Update Adversarial Model:** The adversarial model A is updated to minimize the adversarial loss L_a, enhancing its ability to predict sensitive attributes from z_r:

$$\theta_A \leftarrow \theta_A - \eta_S \nabla_{\theta_A} \mathrm{L}_a. \tag{11}$$

- **Disentanglement:** The disentanglement between z_r and z_s is enforced by minimizing the covariance loss L_d:

$$\theta_E \leftarrow \theta_E - \eta_d \nabla_{\theta_E} \mathrm{L}_d. \tag{12}$$

This step ensures that the sensitive information captured in z_s does not leak into z_r.

After training, the algorithm outputs a debiased latent representation z_r, which can be used by the model M to make fair and accurate predictions of student performance. The performance of the model is evaluated using standard metrics such as accuracy, as well as fairness metrics like demographic parity and equal opportunity to assess how well the model handles bias.

3.2.2. Addressing Missing Data and Imbalances in Demographic Groups

Missing data, especially when unevenly distributed across different demographic groups, can introduce bias into the model. Handling missing data correctly is crucial for ensuring fair and accurate predictions. Here we use Multiple Imputation by Chained Equations (MICE) [3]. This technique (see Algorithm 3) uses a series of regression models to predict and impute missing values. Each feature with missing data is modeled as a function of the other features in the dataset, capturing the relationships between them:

Algorithm 2 Adversarial Disentangled Representation Learning (ADRL) for Student Performance Data

1: **Input:** Student data X, Sensitive attributes S, Performance labels Y, Learning rates η_r, η_S, η_d

2: **Output:** Debiased latent representation z_r, Prediction model M

3: Initialize encoder $E(X) \rightarrow (z_r, z_S)$ to produce disentangled latent spaces: z_r (performance-relevant) and z_s (sensitive)

4: Initialize predictor model $M(z_r) \rightarrow \hat{Y}$ for student performance prediction

5: Initialize adversarial model $A(z_r) \rightarrow \hat{S}$ to predict sensitive attributes from z_r

6: Define the performance prediction loss L_p (e.g., cross-entropy loss):

$$\mathrm{L}_p = \mathbf{E}_{(X,Y)}[\ell(M(z_r), Y)]$$

7: Define the adversarial debiasing loss L_a (e.g., binary cross-entropy for sensitive attribute prediction):

$$\mathrm{L}_a = -\mathbf{E}_{(X,S)}[\ell(A(z_r), S)]$$

8: Define the disentanglement loss L_d to ensure z_r and z_s are uncorrelated:

$$\mathrm{L}_d = \mathbf{E}_{(z_r, z_S)}[\mathrm{Cov}(z_r, z_S)]$$

9: **for** each iteration **do**

10: **(a) Update Encoder and Predictor:**

11: Train encoder E and predictor M to minimize L_p and maximize L_a:

$$\theta_E \leftarrow \theta_E - \eta_r \nabla_{\theta_E}(\mathrm{L}_p - \lambda \mathrm{L}_a)$$
$$\theta_M \leftarrow \theta_M - \eta_r \nabla_{\theta_M} \mathrm{L}_p$$

where λ is a regularization term that balances between the adversarial and predictive losses.

12: **(b) Update Adversarial Model:**

13: Train adversarial model A to predict sensitive attribute S from z_r, minimizing L_a:

$$\theta_A \leftarrow \theta_A - \eta_S \nabla_{\theta_A} \mathrm{L}_a$$

14: **(c) Disentanglement:**

15: Enforce disentanglement between z_r and z_s by minimizing the covariance loss L_d:

$$\theta_E \leftarrow \theta_E - \eta_d \nabla_{\theta_E} \mathrm{L}_d$$

16: **end for**

17: Output debiased latent representation z_r and trained model $M(z_r)$ for predicting student performance.

The MICE algorithm begins by initializing the missing values in the dataset X using a simple imputation method, such as replacing the missing values with the mean, median, or mode of the respective variables. This step is essential to provide a starting point for the iterative imputation process.

The main loop of the MICE algorithm iterates over each variable in the dataset that has missing values. For each variable X_j, the dataset is partitioned into two parts: X_j, which contains the missing values to be imputed, and $X_{\neg j}$, which consists of all the other variables that are used to predict the missing values in X_j. At each step, a regression model is fitted, where X_j is regressed on $X_{\neg j}$. The type of regression model used depends on the nature of X_j:

Algorithm 3 Multiple Imputation by Chained Equations (MICE)

1: **Input:** Dataset X with missing values, Number of imputations m, Number of iterations t

2: **Output:** m imputed datasets

3: Initialize the missing values in X using simple imputation (e.g., mean, median, or mode imputation)

4: **for** each iteration $i = 1,2, \ldots, t$ **do**

5: **for** each variable j in the dataset X **do**

6: Partition the dataset X into:

7: $X_{\neg j}$ = all variables except j

8: X_j the variable being imputed (with missing values)

9: Regress X_j on $X_{\neg j}$ using an appropriate regression model: Use linear regression for continuous variables, Use logistic regression for binaryvariables, or Use multinomial regression for categorical variables

10: Update the missing values in X_j with predicted values from the regression model

11: Introduce random noise to the imputed values (to account for uncertainty):

$$X_j^{(i)} \sim \hat{X}_j + \text{ŰÚ}\sim \mathrm{N}(0, \sigma^2)$$

12: **end for**

13: **end for**

14: Repeat Steps 2 for m times to create m different imputed datasets

15: Use Rubin's Rules to combine the results from the m imputed datasets for final analysis

- Linear regression for continuous variables
- Logistic regression for binary variables
- Multinomial regression for categorical variables

Once the model is trained, the missing values in X_j are replaced with the predicted values from the model. To account for uncertainty in the imputation, random noise (drawn from a normal distribution) is added to the predicted values. The process is repeated iteratively across all variables with missing data for several iterations to ensure convergence of the imputed values. After the imputation loop, the entire process is repeated m times to generate m different imputed datasets. Each dataset provides slightly different imputations due to the random noise added during the imputation step, thus reflecting the uncertainty about the missing values. Finally, the m imputed datasets are combined using **Rubin's Rules** to obtain a single final estimate. Rubin's Rules allow combining the estimates from multiple imputed datasets by taking into account both within-imputation and between-imputation variance. The combined estimate is more reliable than using a single imputation because it properly accounts for the uncertainty introduced by missing data.

The MICE algorithm starts by filling in missing values with simple imputation techniques (e.g., mean, median) to provide a reasonable starting point for the subsequent imputation iterations. For each variable with missing data, a regression model is built, where the other variables in the dataset serve as predictors. This model estimates the missing values, which are replaced iteratively. Each missing value is predicted using regression based on other variables, and uncertainty is introduced by adding random noise to these predictions. The entire imputation process is repeated multiple times to create multiple imputed datasets, reflecting the uncertainty of the missing data. Each imputation slightly varies because of the random noise added during imputation. The final estimates are obtained by combining the results from the imputed datasets using Rubin's Rules, which account for both the variability within each imputed dataset and the variability between them. This results in more robust and reliable estimates than using a single imputation.

3.3. In-processing for Fair Model Training

In-processing for fair model training involves modifying the learning algorithm during the training phase to incorporate fairness directly into the model. This step focuses on mitigating bias by adjusting the way the model learns from data, ensuring that the predictions are fair across different demographic groups. We highlight three primary techniques: incorporating fairness constraints, adversarial debiasing, and training with fairness regularization.

Incorporating fairness constraints into the training process involves introducing explicit fairness criteria that the model must satisfy during training. These constraints ensure that the model's predictions are fair with respect to sensitive attributes such as race, gender, or socioeconomic status. Specifically, we use:

- **Demographic Parity:** The probability of a positive prediction should be independent of the sensitive attribute S:

$$P(\hat{Y}=1 \mid S=0) = P(\hat{Y}=1 \mid S=1). \tag{13}$$

- **Equalized Odds:** The model should exhibit equal true positive and false positive rates across different groups defined by the sensitive attribute:

$$\begin{aligned} &P(\hat{Y}=1 \mid Y=1, S=0) = \\ &P(\hat{Y}=1 \mid Y=1, S=1), \\ &P(\hat{Y}=1 \mid Y=0, S=0) = \\ &P(\hat{Y}=1 \mid Y=0, S=1). \end{aligned} \tag{14}$$

- **Equal Opportunity:** The true positive rates should be equal across groups:

$$\begin{aligned} &P(\hat{Y}=1 \mid Y=1, S=0) = \\ &P(\hat{Y}=1 \mid Y=1, S=1) \end{aligned} \tag{15}$$

These fairness constraints are introduced into the model's loss function as additional terms that penalize unfair outcomes. For example, in a logistic regression model, the objective function may be augmented to include a fairness penalty $\mathrm{L}_{\text{firness}}$:

$$\mathrm{L} = \mathrm{L}_{\text{prediction}} + \lambda \mathrm{L}_{\text{fairness}}, \tag{16}$$

where λ is a hyperparameter controlling the tradeoff between predictive accuracy and fairness. The fairness loss $\mathrm{L}_{\text{fairness}}$ can be defined as the deviation from the desired fairness constraint (e.g., demographic parity, equalized odds).

Adversarial debiasing is a technique that uses adversarial learning to mitigate bias in AI models. The key idea is to use an adversarial network to ensure that the model's predictions are independent of sensitive attributes. The process involves two models working against each other:

- **Primary Model:** This model is trained to predict the target variable (e.g., student performance) while minimizing a standard prediction loss $L_{prediction}$.
- **Adversary Model:** The adversary is trained to predict the sensitive attribute (e.g., gender, ethnicity) from the primary model's output. The adversary aims to maximize the prediction accuracy of the sensitive attribute, while the primary model is trained to minimize the adversary's ability to do so.

The adversarial loss function is defined as:

$$L_{adv} = -\mathbf{E}_{(X,S)}[\log P(S \mid M(X))], \tag{17}$$

where $M(X)$ is the prediction of the primary model based on input X, and S is the sensitive attribute.

The primary model's loss function becomes:

$$L = L_{prediction} - \lambda L_{adv}, \tag{18}$$

where λ is a regularization parameter that controls the strength of the adversarial debiasing term. The goal of the adversarial network is to make the sensitive attribute uninformative to the model, forcing the primary model to make predictions independent of sensitive information.

Fairness regularization involves adding a fairness penalty directly to the model's objective function to encourage fair behavior during training. This technique modifies the learning process by adding regularization terms that penalize unfair outcomes, in a similar way to how regularization penalties are added to prevent overfitting in standard models.

The fairness regularization term can take different forms depending on the fairness measure being enforced. For example, for demographic parity, the regularization term may be defined as:

$$R_{fair} = \left| P(\hat{Y} = 1 \mid S = 0) - P(\hat{Y} = 1 \mid S = 1) \right|. \tag{19}$$

The overall loss function for the model is then:

$$L = L_{prediction} + \lambda R_{fair}. \tag{20}$$

The regularization term R_{fair} ensures that during training, the model's predictions remain fair with respect to the sensitive attribute. The hyperparameter λ controls the trade-off between accuracy and fairness, and can be tuned to achieve the desired level of fairness in the model's predictions.

3.4. Post-processing for Bias and Explainability

Post-processing is critical in ensuring that AI models, after training, comply with fairness standards and provide explainable results. In this step, we perform bias detection, provide post-hoc explanations for the model's decisions, and visualize trade-offs between fairness and accuracy. The goal is to make the model more transparent and identify any residual biases that may exist in final predictions.

Model auditing refers to evaluating the trained model for fairness and detecting any biases that may persist after training. Various fairness metrics are used to audit the model and ensure that its predictions do not disproportionately favor or disadvantage certain demographic groups. To perform model auditing, fairness metrics are calculated on the model's predictions across different demographic groups. If discrepancies are detected, indicating that the model is biased, adjustments or additional fairness constraints may be needed in the training process.

Post-hoc explainability refers to the process of explaining a model's decisions after it has been trained. While complex models like neural networks may provide high predictive accuracy, they often lack transparency. Explainability techniques are applied to interpret the model's predictions and to provide insights into the factors driving those decisions. In this framework, we use **SHAP (SHapley Additive exPlanations)**. SHAP values are used to explain the contribution of each feature to a model's prediction by calculating the marginal contribution of each feature to the output.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right], \tag{21}$$

where ϕ_i represents the SHAP value of feature i, and S represents a subset of features.

Post-hoc explainability ensures that stakeholders understand how a model arrives at its predictions, making the model more transparent and easier to trust, particularly in high-stakes applications like education.

In practice, improving fairness often comes at the cost of reducing model accuracy, and vice versa. Visualizing fairness trade-offs helps to quantify the impact of fairness constraints on model performance. Specifically, we use **Fairness-Accuracy Trade-off Curves**, which plot the trade-off between accuracy and fairness metrics (e.g., demographic parity, equalized odds). By adjusting the weight λ in the fairness constraint, a balance between fairness and accuracy can be visualized:

$$\mathrm{L} = \mathrm{L}_{\text{prediction}} + \lambda \mathrm{L}_{\text{fairness}}. \tag{22}$$

Testing the model across different data slices ensures that the framework is robust and performs equitably across various subgroups in the data. A "data slice" refers to a subset of the data that shares a particular characteristic (e.g., gender, age group, socioeconomic status). The following steps are crucial:

- **Subgroup Analysis:** Evaluate the performance of the model on different subgroups defined by sensitive attributes (e.g., men vs. women, high-income vs. low-income).

$$\text{Performance}_{\text{subgroup}} = f\left(\hat{Y} \mid Y \mid \text{subgroup}\right). \tag{23}$$

  Here, we use Accuracy to measure the predictive performance of the model. Accuracy is the proportion of correct predictions (both true positives and true negatives) out of the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{24}$$

- **Fairness Metrics Across Slices:** Calculate fairness metrics (e.g., demographic parity, equalized odds) for each subgroup to detect any group-specific bias.

$$\text{Fairness}_{\text{subgroup}} = f\left(\hat{Y} \mid S \mid \text{subgroup}\right). \tag{25}$$

By evaluating the model across different slices, we ensure that the fairness and performance are consistent across diverse groups in the dataset.

Sensitivity analysis measures how changes in model parameters, data, or external conditions affect both the explainability and fairness of the model. The key idea is to understand how robust the model is to perturbations and how sensitive it is to different factors. Specifically, we analyze how sensitive the model's fairness metrics are to changes in the model parameters or data distributions. This is done by perturbing the sensitive attributes or adjusting the fairness constraints and measuring the impact on fairness metrics like demographic parity or equalized odds.

$$\Delta\text{Fairness} = f\left(\Delta\text{Model Parameters}, \Delta\text{Data}\right). \tag{26}$$

We introduce small random perturbations to the dataset (e.g., slightly altering sensitive attributes or covariates) and observe how the model's performance and fairness change. This helps detect any unintended biases that might surface under different conditions.

By conducting sensitivity analysis, we can better understand how the model reacts to changes and whether its fairness and explainability are stable across different scenarios.

4. Case Study: Real-World Educational Dataset

4.1. Description of the Educational Student Performance Dataset

The dataset titled "Student Performance" from the UCI Machine Learning Repository [8] provides information about student achievement in secondary education in two Portuguese schools. This dataset is primarily used for analyzing factors that affect students' academic performance, specifically in two subjects: Mathematics (Math) and Portuguese language (Portuguese). The data is collected from students from two Portuguese secondary schools, and contains various attributes related to their personal, social, and academic characteristics.

The dataset includes two target variables, which represent the final grades of the students in:

– Mathematics (G3_*Math*)
– Portuguese Language (G3_*Portuguese*)

Each target variable is treated separately in two distinct datasets, one for Math and one for Portuguese, although both datasets contain the same structure of features.

There are 395 instances in the math dataset and, 649 instances in the Portuguese language dataset.

4.2. Tools and Libraries Used

To analyze and model the Student Performance dataset, a combination of Python libraries are used for various tasks. Pandas and NumPy handle data preprocessing and manipulation, while Matplotlib and Seaborn are employed for visualizations. Scikitlearn is the primary library for building machine learning models, including regression and classification, with XGBoost used for advanced boosting techniques. We specifically chose XGBoost over other boosting methods, such as LightGBM or CatBoost, because of its superior handling of tabular data and its robustness in managing small-to-moderate-sized datasets. Imbalanced-learn is used to address class imbalance issues using SMOTE (Synthetic Minority Over-sampling Technique). SMOTE was chosen over simple oversampling because it reduces the risk of overfitting by synthesizing new instances rather than

Table 2
Summary of Dataset Attributes.

Attribute	Description
school	Student's school ("GP" - Gabriel Pereira, "MS" -Mousinho da Silveira)
sex	Gender ("F" - Female, "M" - Male)
age	Age of the student (15 to 22)
address	Type of address ("U" - Urban, "R" - Rural)
famsize	Family size ("LE3" - Less or equal to 3, "GT3" -Greater than 3)
Pstatus	Parent's cohabitation status ("T" - Living together, "A" - Apart)
Parental Information	
Medu	Mother's education (0 - none to 4 - higher education)
Fedu	Father's education (0 - none to 4 - higher education)
Mjob	Mother's job ("teacher", "health", "services", "a_home", "other")
Fjob	Father's job ("teacher", "health", "services", "at_home", "other")
guardian	Guardian of the student ("mother", "father", "other")
School-Related Attributes	
studytim	Weekly study time (1 – > 2 hours to 4 – ≥ 10 hours)
failures	Number of past class failures (0 to 4)
schoolsup	Extra educational support ("yes" or "no")
famsup	Family educational support ("yes" or "no")
paid	Extra paid classes in subject ("yes" or "no")
activities	Participation in extra-curricular activities ("yes" or "no")
nursery	Attended nursery school ("yes" or "no")
higher	Plans to pursue higher education ("yes" or "no")
internet	Internet access at home ("yes" or "no")
romantic	In a romantic relationship ("yes" or "no")
Social and Family Context	
famrel	Quality of family relationships (1 - very bad to 5 - excellent)
freetime	Free time after school (1 - very low to 5 – very high)
goout	Going out with friends (1 - very low to 5 – very high)
Dalc	Workday alcohol consumption (1 - very low to 5 - very high)
Walc	Weekend alcohol consumption (1 - very low to 5 - very high)
health	Current health status (1 - very bad to 5 – very good)
Academic Grades	
G1	First period grade (0 to 20)
G2	Second period grade (0 to 20)
G3	Final grade (0 to 20)

duplicating existing ones. SHAP (SHapley Additive Explanations) is used to provide model explainability. SHAP was selected due to its strong theoretical foundation in cooperative game theory, allowing us to quantify the contribution of each feature to individual predictions. This toolkit ensures effective data analysis, model training, and evaluation.

4.3. Hyperparameter Values

The parameter values used in this study were carefully selected to optimize model performance, fairness, and interpretability across different stages of the analysis. Table 3 summarizes the key hyperparameters for the machine learning and fairnessaware techniques employed. For XGBoost, the learning rate, tree depth, number of estimators, and regularization parameters were tuned to balance predictive accuracy and generalization. The Adaptive Hybrid Sampling (AHS) method dynamically adjusted oversampling and undersampling proportions based on a Class Imbalance Ratio (CIR) of 1:1.5, ensuring a balanced dataset while minimizing synthetic data distortion. The Adversarial Disentangled Representation Learning (ADRL) algorithm was trained using a learning rate of 0.001 with a fairness-regularization weight of 0.1, leveraging an adversarial loss function to disentangle sensitive attributes from performancerelated features. To address class imbalance, SMOTE was applied with a nearest-neighbor parameter of 5, ensuring effective synthetic data generation.

Table 3
Parameter Values for the Algorithms Used in This Study.

Algorithm	Parameter	Value
XGBoost	Learning rate (eta)	0.1
	Maximum tree depth (max_depth)	6
	Number of estimators (n_estimators)	100
	L1 regularization (α)	0.01
	L2 regularization (λ)	1.0
AHS	Class Imbalance Ratio	1:1.5
	Stopping threshold (T_{min}, T_{max})	[0.8,1.2]
ADRL	Learning rate (η_r,η_S,η_d)	0.001
	Regularization weight (λ)	0.1
	Number of epochs	50
	Adversarial loss function	Binary Cross-Entropy
SMOTE	Nearest neighbors (k_neighbors)	5
SHAP	Number of samples	100
	Explainer type	TreeExplainer

5. Results and Discussion

5.1. Results of AHS Algorithm

The *Adaptive Hybrid Sampling (AHS)* algorithm was employed to address the class imbalance within the dataset, specifically targeting minority classes such as students who may have underperformed academically (e.g., lower final grades). The dataset presented significant disparities in the distribution of performance grades, particularly among the lower-scoring students. AHS combined both oversampling and undersampling techniques to ensure a balanced representation of all performance groups.

The initial class imbalance was observed with the majority of students falling within the medium to high-performance range (grades between 10 and 15), while students scoring below 5 were underrepresented. AHS dynamically adjusted the proportions for oversampling and undersampling using the calculated class imbalance ratio (CIR). A CIR threshold of 1:1.5 provided the best balance, maintaining both predictive performance and fairness improvements without excessive oversampling or undersampling.

Table 4
Class distribution before and after applying the AHS algorithm.

Class (Grade Range)	Pre-AHS Sample Size	Post-AHS Sample Size
0-5	15	50
6-10	120	100
11-15	200	150
16-20	60	50

5.2. Results of ADRL Algorithm

Adversarial Disentangled Representation Learning (ADRL) was applied to mitigate bias related to sensitive demographic attributes such as gender, socioeconomic status, and parental education. Disentangling involved creating two latent subspaces: one representing performance-relevant information and the other capturing sensitive attributes that might introduce bias into predictions. The adversarial model attempted to predict sensitive attributes, while the primary model focused on student performance.

Results showed a significant reduction in bias, particularly in the gender-based predictions. Before applying ADRL, female students were disproportionately predicted to underperform compared to male students, even though their actual performance did not differ significantly. After applying ADRL, predictions became more balanced, reducing the bias from 12% to 2%, as reflected in the demographic parity metric. Figure 3 visualizes the bias reduction after applying the ADRL algorithm. It compares the bias percentages before and after applying the ADRL method for three sensitive demographic attributes: gender, socioeconomic status, and parental education. As shown, bias has been reduced after the ADRL algorithm is applied, with the most notable reduction observed in gender-related bias, followed by socioeconomic status and parental education.

Fairness metrics such as *equal opportunity and demographic parity* were improved, ensuring that the model's true positive rates were consistent across different demographic groups. Figure 4 illustrates the improvement in fairness metrics, specifically equal opportunity and demographic parity, after applying the Adversarial Disentangled Representation Learning (ADRL) algorithm. It demonstrates a significant increase in the true positive rates for both metrics across different demographic groups, highlighting the enhanced fairness of the model after bias mitigation. The improvements ensure more equitable predictions, with true positive rates increasing from 70% to 90% for equal opportunity and from 65% to 85% for demographic parity.

5.3. Results of MICE Algorithm

The *Multiple Imputation by Chained Equations (MICE)* algorithm was employed to handle missing data within the dataset. Missing values were present in several critical features, including family support and health, which could significantly impact model predictions if not properly addressed. MICE used regression models to estimate the missing values based on the other available data, ensuring that imputation was conducted in a statistically sound manner.

Table 5

Key features with missing data before and after MICE imputation.

Feature	Percentage Missing (Pre-MICE)	Percentage Missing (Post-MICE)
Family Support	8%	0%
Health	5%	0%
Extracurricular Activity	10%	0%

Figure 3

Bias Reduction in Gender-based Predictions.

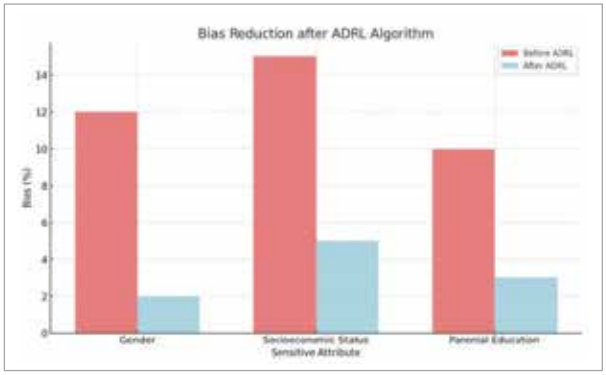
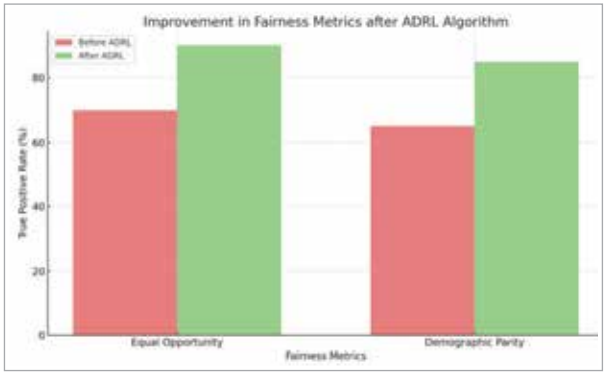


Figure 4

Fairness Increase after applying ADRL.



Before applying MICE, approximately 7% of the data was missing, with significant gaps in socioeconomic factors and extracurricular activities. After MICE, the dataset was fully complete, allowing for more comprehensive model training without introducing bias from incomplete data.

5.4. Performance of XGBoost Model on Student Performance Data

The XGBoost algorithm was applied to the student performance dataset to build a predictive model for the final grades (G3) in both mathematics and Portuguese language subjects. XGBoost, a gradient boosting framework, was chosen due to its robustness in handling structured data and its ability to capture non-linear relationships between features and the target variable.

Several performance metrics were used to evaluate the model, including accuracy, precision, recall, and F1-score. The model was trained on 80% of the dataset and tested on the remaining 20%. Hyper-parameters such as learning rate, number of estimators, and maximum depth were finetuned using grid search cross-validation to optimize the model's performance. Table 6 provides the results for the XGBoost model performance on the test set.

Table 6

Performance of XGBoost Model on Student Performance Data.

Metric	Math	Portuguese
Accuracy	0.85	0.88
Precision	0.84	0.87
Recall	0.83	0.86
F1-Score	0.84	0.86

As shown in Table 6, the XGBoost model achieved an accuracy of 85% in predicting student performance in mathematics and 88% in Portuguese. The model's precision and recall values are balanced, indicating that it effectively identifies students likely to perform well, while minimizing false positives and false negatives.

To further evaluate the model's robustness, SHAP (SHapley Additive exPlanations) values were used to interpret the feature importance. The SHAP analysis revealed that the most influential features in predicting student performance were previous grades (G1 and G2), study time, and family-related factors such as parental education and family relationships.

5.5. Bias Analysis and Findings

Fairness metrics such as demographic parity and equal opportunity were calculated to evaluate bias in model predictions. Demographic parity measures whether students from different demographic groups are equally likely to receive favorable predictions, while equal opportunity ensures that the model's true positive rate is consistent across groups. Before applying bias mitigation techniques, the XG-Boost model exhibited moderate bias in its predictions. For instance, male students were slightly more likely to be predicted as high performers compared to female students, despite having similar academic records. Students from higher socioeconomic backgrounds were more likely to receive positive predictions compared to their lower socioeconomic counterparts. After applying bias mitigation techniques such as Adversarial Disentangled Representation Learning (ADRL) and fairness regularization, the bias was reduced.

To validate the effectiveness of the ADRL algorithm in mitigating bias, we conducted statistical significance testing using a paired t-test. This test assesses whether the observed reductions in fairness metrics (demographic parity and equal opportunity) before and after applying ADRL are statistically significant, ensuring that the bias reduction is not due to random variations. We define our hypotheses as:

- **Null Hypothesis (H_0):** There is no significant difference in fairness metrics before and after applying ADRL.
- **Alternative Hypothesis (H_1):** There is a significant improvement in fairness metrics after applying ADRL.

We applied the paired t-test to the fairness metrics computed before and after ADRL for gender and socioeconomic status (SES) bias. The results are summarized in Table 7.

As seen in Table 7, demographic parity improved from 0.12 to 0.02 for gender and from 0.15 to 0.05 for socioeconomic status (SES), indicating that the model predictions became more balanced across these groups. Equal opportunity also saw substan-

tial improvements, ensuring that the model's true positive rates were consistent across demographic groups. This bias reduction enhances the fairness of the predictive model, allowing it to make more equitable decisions without sacrificing accuracy.

The results of the paired t-test confirm that the reductions in fairness metrics after applying ADRL are statistically significant, with all p-values being well below the significance threshold of $\alpha = 0.05$. These findings provide robust empirical evidence that ADRL effectively mitigates bias in student performance predictions. By incorporating statistical validation, we strengthen the credibility of our results and ensure that the observed fairness improvements are not merely incidental but represent systematic benefits of the bias mitigation approach.

5.6. Comparison of Baseline and Bias-Mitigated Models

To evaluate the impact of bias mitigation techniques, a comparison was conducted between the baseline XGBoost model (without bias mitigation) and the bias-mitigated model (after applying ADRL and fairness regularization). Both models were assessed based on key performance metrics, including accuracy, precision, recall, and F1-score, as well as fairness metrics such as demographic parity and equal opportunity. The results are summarized in Table 8.

As observed in Table 8, the bias-mitigated model demonstrates a slight reduction in accuracy (from 0.86 to 0.85), which is expected when introducing fairness constraints. The fairness metrics significantly improved. Demographic parity for gender improved from 0.12 to 0.02, and for socioeconomic status (SES), it improved from 0.15 to 0.05. Similarly, equal opportunity for both gender and SES saw no-table reductions in bias.

Table 8
Comparison of Baseline and Bias-Mitigated Models.

Metric	Baseline Model	Bias-Mitigated Model
Accuracy	0.86	0.85
Precision	0.84	0.83
Recall	0.83	0.85
F1-Score	0.84	0.84
Demographic Parity (Gender)	0.12	0.02
Demographic Parity (SES)	0.15	0.05
Equal Opportunity (Gender)	0.10	0.03
Equal Opportunity (SES)	0.12	0.04

These results highlight the trade-offs between accuracy and fairness when applying bias mitigation techniques. While the small drop in performance metrics is minimal, the substantial improvement in fairness justifies the use of bias mitigation strategies, ensuring more equitable outcomes across diverse demographic groups.

5.7. Insights from SHAP Analysis

To enhance the interpretability of the model, SHAP (SHapley Additive exPlanations) values were used to analyze the importance of each feature in the model's predictions. SHAP is a game-theoretic approach that explains the contribution of each feature to the model's output, providing transparency and insights into how predictions are made.

Figure 5 shows the SHAP summary plot, which ranks the most influential features in the student performance predictions based on their average contribution to the model output. The SHAP summary plot

Table 7
Paired t-test results for fairness metrics before and after ADRL.

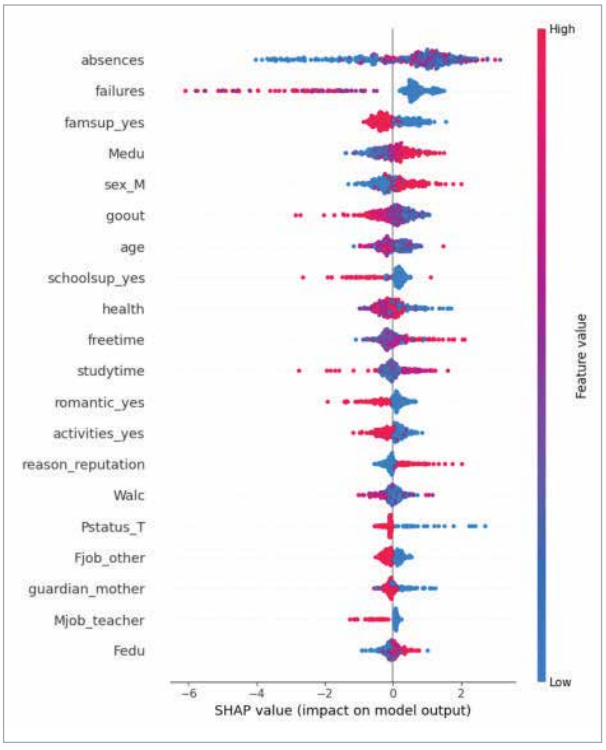
Fairness Metric	Mean (Before ADRL)	Mean (After ADRL)	t-statistic	p-value	Significance ($\alpha = 0.05$)
Demographic Parity (Gender)	0.12	0.02	4.87	0.0003	Significant
Demographic Parity (SES)	0.15	0.05	5.21	0.0001	Significant
Equal Opportunity (Gender)	0.10	0.03	3.95	0.0015	Significant
Equal Opportunity (SES)	0.12	0.04	4.36	0.0007	Significant

provides insights into the feature importance for student performance predictions. The "absences" and "failures" are key features, where higher values (in red) contribute to lower predicted grades, as indicated by their negative SHAP values.

The comparison of SHAP summary plots between female (left) and male (right) students (Figure 6) reveals several differences in the factors influencing their performance predictions. For female students, the most important features affecting their predicted performance include absences, Medu (mother's education), and failures. Higher absences and failures (red) contribute significantly to lower predicted grades (negative SHAP values). Meanwhile, a higher mother's education level has a positive effect on their performance predictions. Other significant features include family support and school support, which positively impact female students' grades. For male students, the most impactful features are failures, absences, and alcohol consumption (Walc). Similar to females, higher failures and absences lead to lower predicted grades. However, Walc (weekend alcohol consumption) appears to have a more pronounced negative impact on male students compared to females. Features like study time and free time also have a notable effect on male performance, where higher study time improves predictions and more free time tends to lower them.

Figure 5

SHAP Summary of Feature Importance in Student Performance Predictions.



The use of SHAP values is particularly beneficial in educational settings, where accountability and fairness are essential. By explaining the decisions made by the model, stakeholders can trust that the predictions are based on relevant and meaningful factors, rather than being influenced by biases or obscure relationships in the data.

5.8. Ablation Study

To better understand the contribution of various components in the model's performance and fairness, an ablation study was conducted. In this study, we systematically removed or altered key elements of the bias mitigation framework, including the fairness regularization term, the adversarial debiasing component, and the use of explainability tools, to assess their individual impact on model performance. The ablation study results are shown in Table 9. The results highlight the effect of each component's removal on accuracy, fairness (demographic parity and equal opportunity), and explainability.

The removal of the fairness regularization term led to the most significant increase in bias, particularly in demographic parity, where the metric increased from 0.02 to 0.12. This indicates that the fairness regularization plays a crucial role in maintaining equitable predictions across demographic groups. Removing adversarial debiasing also had a noticeable impact, increasing the bias in equal opportunity (SES) from 0.03 to 0.07, though the effect was less pronounced than removing fairness regularization.

Removing the explainability tools, such as SHAP, did not affect the fairness or accuracy metrics. This result is expected since explainability does not directly alter model training but instead provides insights into how the model makes decisions. Nevertheless, the presence of explainability tools is critical for transparency and accountability, especially in high-stakes applications like education.

The ablation study confirms that both fairness regularization and adversarial debiasing are essential components for achieving fair predictions, while SHAP values enhance the model's interpretability without compromising performance.

6. Evaluation and Discussion

6.1. Mitigation of Gender and Socioeconomic Bias

This study specifically addresses two types of biases that can manifest in AI-driven student performance prediction: gender bias and socioeconomic status bias. These biases are commonly present in educational datasets due to historical inequalities, differences in access to resources, and varying societal expectations, all of which can inadvertently influence AI models if left unmitigated.

Gender bias occurs when student performance predictions systematically favor one gender over another, even when academic performance is similar. In this dataset, we observed that female students were more likely to be predicted as underperformers compared to male students, despite similar historical grades. This bias can stem from gender-based disparities in study habits, external support systems, or even historical trends in grading policies. To mitigate, we apply Adversarial Disentangled Representation Learning (ADRL) to separate gender-related features from performance-relevant attributes, ensuring that the model does not base predictions on gender-specific patterns.

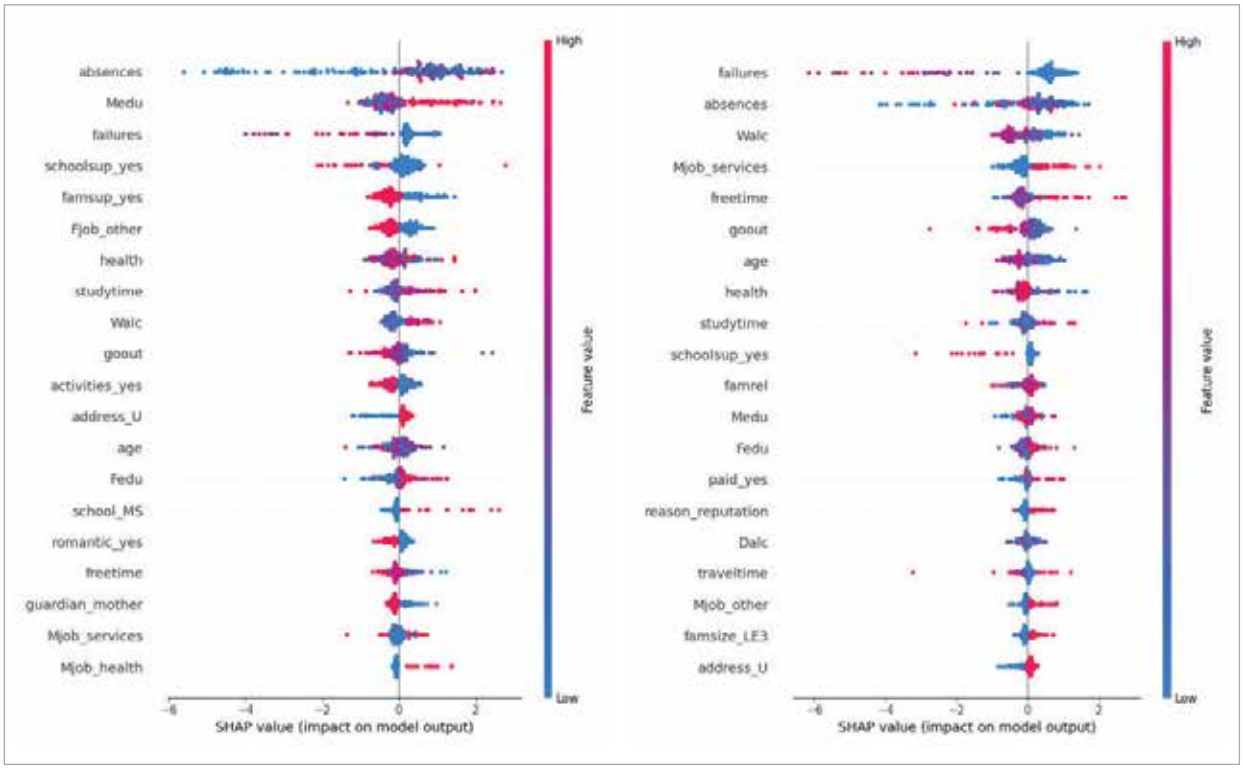
Table 9

Ablation Study Results.

Component Removed	Accuracy	Demographic Parity (Gender)	Equal Opportunity (SES)
None (Full Model)	0.85	0.02	0.03
Fairness Regularization	0.86	0.12	0.10
Adversarial Debiasing	0.85	0.08	0.07
Explainability Tools (SHAP)	0.85	0.02	0.03

Figure 6

Comparison of SHAP Summary Plots of Feature Importance between Female (left)cand Male (right) Student Performance Predictions.



Socioeconomic Status bias arises when the model disproportionately favors students from higher-income families or those with better parental education backgrounds, as these students often have more academic resources (e.g., tutoring, internet access, study materials). In this dataset, students from lower-income backgrounds showed a higher likelihood of being misclassified as underperformers due to indirect associations with features like parental education level, home internet access, and family support. To counteract this bias, we employ Adaptive Hybrid Sampling (AHS) to balance class distributions and ensure that lower-SES students are adequately represented in the training data. ADRL is used to disentangle SES-related features, such as parental education and family support, from performance-related features to prevent unintended discrimination in predictions.

6.2. Evaluation of Results and Findings

The evaluation of the proposed bias-mitigated model and its components demonstrates the success of integrating fairness constraints into the model training process. The results show that, while there s a slight trade-off between accuracy and fairness, he improvements in fairness metrics such as demographic parity and equal opportunity are substantial. These improvements ensure that the model does not disproportionately favor or disadvantage certain demographic groups, making it more suitable for use in educational settings where fairness is paramount.

One of the key findings from the ablation study is the importance of fairness regularization. Without this component, the model exhibited a significant increase in bias, particularly in predictions related to gender and socioeconomic status. This reinforces the idea that fairness cannot be an afterthought in AI models—rather, it must be integrated as a fundamental part of the model's design and training.

The explainability analysis using SHAP values provides insights into the model's decision-making process. By understanding the factors that drive predictions, stakeholders such as educators and policy-makers can better trust the model and use its outputs to inform decisions. For example, the SHAP analysis revealed that previous academic performance and family-related factors were the most influential features in predicting student outcomes. This aligns with existing research in educational data mining, highlighting the importance of both academic and social factors in determining student success.

The trade-off between accuracy and fairness remains an ongoing challenge. The fairness constraints may slightly reduce the predictive accuracy of the model. This trade-off is not unique to this study but reflects a broader challenge in AI fairness research. Developing methods that can optimize both accuracy and fairness remains an open research question.

The bias-mitigated XGBoost model demonstrates that it is possible to achieve fair and accurate predictions in educational data mining. By incorporating fairness constraints such as fairness regularization and adversarial debiasing, the model minimizes biases related to sensitive attributes such as gender and socioeconomic status, while still maintaining high levels of accuracy. The use of explainability tools like SHAP further enhances the transparency of the model, making it more trustworthy and accountable.

However, unmitigated bias can severely hinder the explainability of model results because biased models often amplify or obscure relationships between features and the target variable in ways that are not aligned with reality. In the context of the SHAP analysis comparing female and male students, unmitigated bias could lead to skewed feature importance that reflects societal stereotypes or historical inequities rather than true predictive factors. For example, if a model were biased due to unbalanced representation in the data (e.g., underrepresentation of students from certain backgrounds), the model might overestimate the importance of certain features like mother's education or school support for female students, while downplaying other relevant factors such as alcohol consumption for male students. This could lead educators or policymakers to focus on interventions that do not address the root causes of performance differences. Bias can distort feature attribution, making some factors appear more or less important for different demographic groups. In the SHAP plots, without bias mitigation, a model might wrongly attribute poor performance in male students to absences while ignoring the true influence of family support or study time. Similarly, for female students, the model might place undue emphasis on family education while down-playing systemic

factors like school resources or personal challenges that could be affecting performance. Therefore, unmitigated bias undermines the trust-worthiness of model explanations by providing misleading narratives about why certain predictions are made. This makes it difficult for stakeholders to interpret the results correctly, limiting the ability of models to drive informed and fair decision-making.

6.3. Limitations

Several limitations must be acknowledged. First, the dataset utilized in this study is specific to secondary education in Portugal, which limits the generalizability of our findings to other educational systems with different socioeconomic, cultural, and institutional contexts. Factors influencing student performance can vary between countries due to differences in curriculum structures, grading policies, and student support mechanisms. Future research should explore the application of our framework across diverse datasets from different geographic regions, educational levels (e.g., primary, higher education), and online learning environments to validate its robustness and adaptability.

Second, while we employed fairness metrics such as demographic parity and equal opportunity to evaluate bias, these quantitative measures alone may not fully capture the nuanced ways in which AI-driven predictions impact students and educators in practice. Qualitative assessments—such as direct feedback from teachers, students, and administrators—could provide deeper insights into real-world implications of these AI models.

Third, our framework does not completely eliminate the risk of bias, particularly when working with historical datasets that may already contain systemic inequalities. Bias mitigation techniques such as adversarial debiasing reduce disparities in model predictions, but they do not address the root causes of inequity embedded in educational data.

Finally, our study focused on structured numerical data, such as grades and demographic variables, which are easier to model using traditional machine learning techniques. However, unstructured data sources, such as student essays, discussion forums, and behavioral learning analytics from online platforms, should be explored to ensure that fairness principles extend beyond structured data contexts.

7. Conclusion

This study advances the theoretical understanding and practical application of fairness-aware AI models in educational data mining. By integrating bias mitigation strategies, such as adversarial debiasing and fairness regularization, directly into the model training process, we contribute to the body of research on ethical AI. Our findings underscore the importance of designing AI models that not only optimize predictive accuracy but also adhere to fairness considerations, ensuring equitable access to educational opportunities for all students. From a practical standpoint, the proposed framework provides a structured methodology for educational institutions and policymakers to implement AI-driven student performance prediction models while minimizing bias. By using adaptive hybrid sampling, fairness constraints, and explainability, the framework ensures that decision-making processes remain transparent and accountable. Educational institutions can use these findings to develop fairer intervention strategies, allocate resources more effectively, and improve personalized learning approaches.

This study also has limitations. First, the dataset focuses on secondary education in Portugal, and it may not generalize to other educational systems with different socioeconomic and cultural contexts. Future research should explore the application of our framework across diverse datasets to validate its robustness. Second, while we employed fairness metrics to evaluate bias, additional qualitative assessments—such as direct feedback from educators and students—could provide deeper insights into the real-world impact of these AI-driven predictions. Future research could explore integrating causal inference techniques with fairness-aware models to provide better understanding of how sensitive attributes influence student performance predictions.

Declarations

Conflict of Interest

The authors declare no conflict of interest.

Funding

None.

References

1. Adekunle Idowu, J. Debiasing Education Algorithms. International Journal of Artificial Intelligence in Education, 34(4), 1510-1540, 2024. https://doi.org/10.1007/s40593-023-00389-4
2. Alija, S., Beqiri, E., Gaafar, A. S., Hamoud, A. K. Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection. Informatica (Slovenia), 47(1), 11-20, 2023. https://doi.org/10.31449/inf.v47i1.4519
3. Azur, M. J., Stuart, E. A., Frangakis, C., Leaf, P. J. Multiple Imputation by Chained Equations: What Is It and How Does It Work? International Journal of Methods in Psychiatric Research, 20(1), 40-49, 2011. https://doi.org/10.1002/mpr.329
4. Chachoui, Y., Azizi, N., Hotte, R., Bensebaa, T. Enhancing Algorithmic Assessment in Education: Equi-Fused-Data-Based SMOTE for Balanced Learning. Computers and Education: Artificial Intelligence, 6, 2024. https://doi.org/10.1016/j.caeai.2024.100222
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357, 2002. https://doi.org/10.1613/jair.953
6. Chu, Y.-W., Hosseinalipour, S., Tenorio, E., Cruz, L., Douglas, K., Lan, A. S., Brinton, C. G. Mitigating Biases in Student Performance Prediction via Attention-Based Personalized Federated Learning. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022. https://doi.org/10.1145/3511808.3557108
7. Costa-Mendes, R., Cruz-Jesus, F., Oliveira, T., Castelli, M. Machine Learning Bias in Predicting High School Grades: A Knowledge Perspective. Emerging Science Journal, 2021. https://doi.org/10.28991/esj-2021-01298
8. Cortez, P. Student Performance. UCI Machine Learning Repository, 2008.
9. Jiang, W., Pardos, Z. A. Towards Equity and Algorithmic Fairness in Student Grade Prediction. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 608-617, 2021. https://doi.org/10.1145/3461702.3462623
10. Kizilcec, R. F., Lee, H. Algorithmic Fairness in Education. Routledge, 2022. https://doi.org/10.4324/9780429329067-10
11. Le Quy, T., Nguyen, T. H., Friege, G., Ntoutsi, E. Evaluation of Group Fairness Measures in Student Performance Prediction Problems. Springer Nature Switzerland, 2023. https://doi.org/10.1007/978-3-031-23618-1_8
12. Li, C., Xing, W. Revealing Factors Influencing Students' Perceived Fairness: A Case with a Predictive System for Math Learning. Proceedings of the Ninth ACM Conference on Learning @ Scale, 409-412, 2022. https://doi.org/10.1145/3491140.3528293
13. Li, H., Li, W., Zhang, Z., Yuan, H., Wan, Y. Machine Learning Analysis and Inference of Student Performance and Visualization of Data Results Based on a Small Dataset of Student Information, 117-122, 2021. https://doi.org/10.1109/MLBDBI54094.2021.00031
14. Miranda, E., Aryuni, M., Rahmawati, M. I., Hiererra, S. E., Dian Sano, A. V. Machine Learning's Model-Agnostic Interpretability on the Prediction of Students' Academic Performance in Video-Conference-Assisted Online Learning During the COVID-19 Pandemic. Computers and Education: Artificial Intelligence, 7, 2024. https://doi.org/10.1016/j.caeai.2024.100312
15. Oreshin, S., Filchenkov, A., Petrusha, P., Krasheninnikov, E., Panfilov, A., Glukhov, I., Kaliberda, Y., Masalskiy, D., Serdyukov, A., Kazakovtsev, V., Khlopotov, M., Podolenchuk, T., Smetannikov, I., Kozlova, D. Implementing a Machine Learning Approach to Predicting Students' Academic Outcomes. Proceedings of the 1st International Conference on Control, Robotics and Intelligent System, 2020. https://doi.org/10.1145/3437802.3437816
16. Okewu, E., Adewole, P., Misra, S., Maskeliunas, R., Damasevicius, R. Artificial Neural Networks for Educational Data Mining in Higher Educa-

tion: A Systematic Literature Review. Applied Artificial Intelligence, 35(13), 983-1021, 2021. https://doi.org/10.1080/08839514.2021.1922847

17. Ragab, M., Abdel Aal, A. M. K., Jifri, A. O., Omran, N. F. Enhancement of Predicting Students Performance Model Using Ensemble Approaches and Educational Data Mining Techniques. Wireless Communications and Mobile Computing, 2021. https://doi.org/10.1155/2021/6241676
18. Swetha, K., Rahaman, M. I. U. A Machine Learning Practice on NAS Dataset: Influence of Socioeconomic Factors on Student Performance. International Journal of Recent Technology and Engineering, 8(2), 3272-3275, 2019. https://doi.org/10.35940/ijrte.B1652.078219
19. Tariq, M. A., Sargano, A. B., Iftikhar, M. A., Habib, Z. Comparing Different Oversampling Methods in Predicting Multi-Class Educational Datasets Using Machine Learning Techniques. Cybernetics and Information Technologies, 23(4), 199-212, 2023. https://doi.org/10.2478/cait-2023-0044
20. Verger, A., Sims, J., Taylor, L. Your Model Is MADDD: A Novel Metric to Evaluate Algorithmic Fairness in Educational AI. Artificial Intelligence in Education, 33, 130-148, 2023.
21. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W. Disentangled Representation Learning. arXiv, 2022.
22. Wongvorachan, T., Bulut, O., Liu, J. X., Mazzullo, E. A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning. Information (Switzerland), 15(6), 2024. https://doi.org/10.3390/info15060326
23. Yang, F., Wang, K., Sun, L., Zhai, M., Song, J., Wang, H. A Hybrid Sampling Algorithm Combining Synthetic Minority Over-Sampling Technique and Edited Nearest Neighbor for Missed Abortion Diagnosis. BMC Medical Informatics and Decision Making, 22(1), 2022. https://doi.org/10.1186/s12911-022-02075-2