

Promoting Fairness in Academic Risk Prediction: An Evaluation of the Reweight Technique in Higher Education

Augusto Schmidt¹, Tiago Primo¹, Vinicius Ramos², Roberto Muñoz³, Andrea Bordin², Thales Vieira⁶, Seiji Isotani⁷, Emanuel Queiroga⁵, Virginia Rodéz-Paragarino⁴, Cristian Cechinel²

¹Programa de Pós-Graduação em Computação – Universidade Federal de Pelotas, Rua Gomes Carneiro, 1 – 96010-610 – Pelotas – RS – Brasil.

²Programa de Pós-Graduação em Tecnologias da Informação e Comunicação - Universidade Federal de Santa Catarina – Rod. Gov. Jorge Lacerda, 3201 – 88906-072, Araranguá – SC – Brasil.

³Universidad de Valparaíso, Blanco 951, Valparaíso, Chile.

⁴Institute for the Future of Education Europe – Tecnológico de Monterrey – Av. Eugenio Garza Sada 2501 – 64700 – Monterrey – Nuevo Leon – México.

⁵Instituto Federal de Educação Ciência e Tecnologia Sul-rio-grandense – Praça Vinte de Setembro, 455 – 96015-360 – Pelotas – RS – Brasil.

⁶Núcleo de Excelência em Tecnologias Sociais (NEES) – Universidade Federal de Alagoas (UFAL) – Alagoas, Brasil.

⁷Graduate School of Education – University of Pennsylvania – 3700 Walnut Street – Philadelphia – PA – USA.

{augustgs, emanuelmqueiroga}@gmail.com,
{v.ramos, andrea.bordin}@inf.ufrgs.br, tiago.primo@inf.ufpel.edu.br,
roberto.munoz@uv.cl, thales.vieira@nees.ufal.br, sisotani@upenn.edu,
virginia.rodéz@ifeeurope.es, contato@cristiancechinel.pro.br

Abstract. *This study evaluates algorithmic fairness in academic risk prediction using data from 1,796 Latin American undergraduates. Across 80 model and imbalance combinations, AUC-ROC rises from 0.87 to 0.96 as information increases. reweight reduces DI and AOD disparities across sensitive groups while bootstrap IC95% analyses indicate significant fairness gains without predictive-performance loss.*

Keywords— *Learning Analytics, Academic Risk Prediction, Algorithmic Fairness*

Resumo. *Este estudo avalia a justiça algorítmica na previsão de risco acadêmico utilizando dados de 1.796 estudantes universitários latino-americanos. Ao longo de 80 combinações de modelos e cenários de desequilíbrio, a AUC-ROC aumenta de 0,87 para 0,96 à medida que a quantidade de informações cresce. A técnica de reweight reduz as disparidades de DI e AOD entre grupos sensíveis, enquanto análises de bootstrap IC95% indicam ganhos significativos de equidade sem perda de desempenho preditivo.*

Palavras-chave— *Análise de Aprendizagem, Predição de Risco Acadêmico, Justiça Algorítmica.*

1. Introduction

The advances in digital technologies have led to a transformation in the field of higher education, one of the fastest-growing areas being Learning Analytics (LA). Through educational data, it is possible to apply machine learning models to predict academic trajectories and identify at-risk students (Jayaprakash et al. 2014). According to Cechinel et al. (2020), the Latin American scenario is one of the regions with the greatest prominence in research promoting the use of LA in education for regional development. This emphasis occurs because the countries that make up the region have similar cultures and problems, such as income accumulation and low investment in science and technology.

Despite the significant advances in Learning Analytics and its use in predictive tools and models for making academic decisions, most studies still focus exclusively on predictive efficiency, neglecting the aspects of algorithmic fairness that permeate the educational context. However promising these technologies may be, they can also, unintentionally, end up reinforcing or even amplifying existing disparities. This ultimately directly impacts groups that already face inequalities, such as ethnic minorities, women, international students, or first-generation students.

In this study, the reweight technique was selected as the primary bias mitigation strategy because it is a model-agnostic preprocessing method that can be applied before classifier training without modifying the internal learning algorithm or its decision threshold (Friedler et al., 2019; Kamiran and Calders, 2012). This characteristic is particularly suitable for our experimental design, which compares multiple classifiers and class-imbalance handling strategies under identical data splits. By adjusting the contribution of training instances according to the joint distribution of sensitive attributes and the target variable, the reweight technique enables the effect of a single fairness intervention to be examined consistently across different data-availability scenarios while retaining all observations. Previous studies have also reported that the reweight technique can reduce group disparities while maintaining predictive performance, including in educational settings (Blow et al., 2024; Raftopoulos et al., 2025). The objective of this study is not to establish that the reweight technique is superior to in-processing or post-processing approaches, but to isolate and systematically evaluate the effects of an established preprocessing intervention under different levels of academic data availability.

Given this, we investigate two research questions: **RQ1**: How does predictive performance change before and after applying a fairness intervention? **RQ2**: To what extent does the reweight technique mitigate bias in academic risk prediction in higher education?

By obtaining answers to these questions, this study aims to demonstrate that it is possible to combine predictive performance with fairness criteria, thus contributing to the development of more ethical and sensitive Learning Analytics practices regarding inequalities in the data attributed to predictive models.

This study is organized as follows. Section 2 reviews related work. Section 3 lays out the methodology. Section 4 presents and discusses the results, and Section 5 wraps things up with final remarks and directions for future work.

2. Related work

Research on academic risk prediction within Learning Analytics has traditionally leaned heavily on improving predictive accuracy, and fairness often got left in the background. More recently, though, a growing line of work has started to ask whether those predictive gains come with uneven consequences for different student groups, especially those already facing structural disadvantages (Kizilcec and Lee, 2022). This shift has opened room for studies that explicitly examine fairness in educational prediction tasks, instead of treating it as a side note.

Most fairness-aware work in education has focused first on diagnosing disparities rather than actively mitigating them. These studies typically inspect how errors and outcomes differ across demographic groups (e.g., gender, international status, first-generation), and they do a solid job of showing that models can behave quite differently depending on who's being evaluated (Saxena et al., 2019). The problem is that many of them remain largely descriptive, either stopping short of interventions or evaluating fairness under a single modeling setup.

A smaller but more recent body of research goes one step further by bringing mitigation strategies into educational prediction pipelines. Here, fairness is treated as something that can be influenced by design choices rather than merely reported after it appears. These works explore practical interventions such as data-centric adjustments and bias mitigation techniques; they also highlight that interventions can introduce trade-offs and operational constraints, particularly when sensitive attributes are involved or when institutions rely on partially observed student information (Baker et al., 2023).

Among mitigation strategies, the reweight technique tends to stand out because it's simple, model-independent, and easy to slot into existing pipelines. The core idea comes from earlier work on pre-processing methods that adjust instance weights to reduce dependence between sensitive attributes and outcomes (Kamiran and Calders, 2012). Comparative evidence suggests that the reweight technique can reduce group disparities while often keeping performance reasonably stable, especially when contrasted with more invasive approaches (Friedler et al., 2019). More recent empirical studies further support its practical value, including evidence across domains and, increasingly, educational settings (Blow et al., 2024).

Still, the way fairness is evaluated in prior work often comes with some assumptions. A recurring one is treating student information as if it were fully available at training and prediction time. In practice, however, educational institutions typically operate under varying levels of data completeness, particularly when the goal is early intervention. Hlosta et al. (2017), for example, demonstrated that at-risk students can be identified even in the absence of legacy course data by progressively using information generated during the current course. Another common limitation is scoping fairness evaluation to a single configuration, which can hide how disparities shift when the feature set, training setup, or operational setting changes (Baker et al., 2023). In short, we've learned about whether disparities exist, but we still know little about how fairness performance trade-offs behave under realistic, evolving data conditions.

Against this background, the present study positions itself at the intersection of fairness-aware Learning Analytics and realistic academic risk prediction. Instead of proposing a new mitigation technique, it takes a well-established approach and evaluates it systematically across multiple data availability scenarios that reflect different stages of a student's academic journey (Raftopoulos et al., 2025). By jointly analyzing predictive performance and group-based fairness metrics using a standardized evaluation toolkit (Bellamy et al., 2019), this work extends prior research toward a more operational understanding of what fairness improvements look like under conditions that resemble real institutional deployments.

3. Methodology

The remainder of this section is organized as follows. Section 3.1 describes the institutional dataset, the construction of the target variable, and the preprocessing procedures. Section 3.2 introduces the four data-availability scenarios. Section 3.3 presents the sensitive attributes, intersectional groups, fairness metrics, and statistical procedures. Finally, Section 3.4 details the paired experimental design, classifiers, class-imbalance handling techniques, and evaluation pipeline.

This study adopts an experimental methodology, using Cross-Industry Standard Process for Data Mining (CRISP-DM) as a guiding framework, aimed at understanding how predictive performance and algorithmic fairness interact under realistic conditions of academic risk prediction. Rather than assuming an idealized setting with complete information, the proposed approach mirrors how data actually becomes available in higher education environments. The overall pipeline combines data preparation, scenario-based modeling, fairness assessment, and bias mitigation, enabling a structured comparison of baseline and fairness-aware models.

Within this experimental setup, the analysis is organized around two complementary research questions. RQ1 looks at potential performance-fairness tradeoffs by jointly examining predictive discrimination (AUC-ROC) and group-based fairness metrics, namely Disparate Impact and Average Odds Difference. In practical terms, this allows us to see whether gains in predictive performance come at the cost of unequal outcomes or error distributions across groups.

RQ2 then shifts the focus to mitigation, asking whether the application of the reweight technique actually reduces observed disparities. Here, the analysis contrasts model behavior before and after the reweight technique across sensitive groups defined by gender, locality, international background, and first-generation status. By relying on consolidated fairness metrics such as Disparate Impact and Average Odds Difference, this question directly evaluates the extent to which a fairness-oriented preprocessing strategy affects sensitive group disparities.

3.1. Dataset and preprocessing

The experiments were provided by the Monterrey Institute of Technology and Higher Education (Tecnológico de Monterrey, 2024). This institute is a private university, one of the most prestigious in Latin America, with a strong emphasis on technology, innovation, and entrepreneurship. It offers undergraduate and graduate programs in various fields, including Engineering, Administration, Social Sciences, Humanities, and Medicine, among others. The dataset analyzed in this research is part of the Educational Innovation

collection from the Institute for the Future of Education (IFE) at Tecnológico de Monterrey's Research Data Hub (Institute for the Future of Education (IFE), 2024).

However, access to these data is restricted¹, as they were used under a signed Terms of Use agreement for the current study and are therefore not publicly available. The dataset includes student-level information collected across academic trajectories, combining demographic attributes, admission records, engagement indicators, psychometric assessments, and academic performance measures. Each instance corresponds to a student-term observation, and the prediction task focuses on identifying students with an unfavorable academic status at the end of the corresponding academic period. The target variable, `status_numerico`, was derived from the institutional attribute `student.status_academic_desc`, which records the student's academic status. Students classified as "Regular student" were assigned the value 1, representing a satisfactory academic status. All other statuses, including academic warning, failure, and dropout, were assigned the value 0 and treated as an unsuccessful academic outcome. Therefore, in this study, a student is considered academically at risk when their end-of-period institutional status differs from "Regular student." Accordingly, $Y=1$ denotes a regular academic status, considered the favorable outcome, whereas $Y=0$ denotes an at-risk academic status, considered the unfavorable outcome.

The original dataset contained 22,718 records. Before model training, standard preprocessing procedures were applied to ensure consistency and avoid data leakage: inconsistent records and predictors not available at prediction time were removed, categorical variables were encoded, and numerical features were scaled when required. After preprocessing, the final dataset comprised 1,796 unique students and 20,461 student-term records. As is common in educational settings, the outcome is highly imbalanced, with approximately 8% of observations corresponding to unsuccessful academic outcomes.

3.2. Data availability

Four data availability scenarios were defined to reflect different stages of a student's academic trajectory. Table 1 summarizes the availability of performance-related attributes across scenarios. **The First Scenario** includes only demographic and admission-related data (early-stage students). **The Second Scenario** adds current-term Grade Point Average (GPA). **The Third Scenario** includes cumulative program-level GPA, without the current-term value. The Fourth Scenario represents a fully informed setting, where all available attributes are included.

This scenario-based design allows us to analyze how predictive performance and fairness evolve as additional information becomes available.

¹ Data are however available from the IFE Data Hub upon reasonable request at <https://doi.org/10.57687/FK2/UFM2QH>, (accessed on 08 July 2024).

Table 1. Scenario-based availability of performance-related attributes

Feature Name	Scenario	Description
student.term_gpa	2nd, 4th	Current-term GPA
gpa_label	2nd, 4th	Categorical label derived from the current-term GPA.
student.term_gpa_label	3rd, 4th	Cumulative GPA at the program level.
gpa_program_label	3rd, 4th	Categorical label derived from the program-level GPA.

3.3. Fairness assessment

Algorithmic fairness was assessed using Disparate Impact (DI) and Average Odds Difference (AOD) with respect to four binary sensitive attributes commonly associated with educational inequalities: gender (male/female), locality (rural/urban), international background (no/yes), and first-generation status (no/yes). To capture intersectional effects, sensitive groups are defined as the full combination of these attributes, resulting in 16 distinct groups.

In our analysis, all combinations of the four sensitive attributes were considered, resulting in 16 intersectional groups and 240 possible pairwise contrasts. Across models and data availability scenarios, some patterns kept emerging, and one contrast consistently stood out as showing the strongest evidence of disparity according to DI and AOD, with the privileged group being male, urban, international, non-first-generation students versus the unprivileged group as female, rural, non-international, first-generation students. Importantly, this pairing was not predefined but surfaced directly from the observed fairness metrics, suggesting that the intersection of gender, territorial context, migration background, and family educational history concentrates meaningful inequalities in model behavior.

Fairness was evaluated using two complementary group-based metrics: Disparate Impact (DI) and Average Odds Difference (AOD), computed using the AI Fairness 360 (AIF360) framework (Bellamy et al., 2019). These metrics allow us to assess whether predictive outcomes and error rates differ systematically between groups defined by sensitive attributes.

Disparate Impact, shown in Equation 1, measures the ratio of favorable outcomes assigned to unprivileged and privileged groups. Let $Y \in \{0,1\}$ denote the model prediction, where $Y=1$ represents the favorable outcome, namely a predicted regular academic status. Let G denote membership in the sensitive-group contrast, with $G=\text{unprivileged}$ representing the unprivileged group and $G=\text{privileged}$ representing the privileged group. Disparate Impact is defined as:

$$DI = \frac{P(Y=1 | G=unprivileged)}{P(Y=1 | G=privileged)} \quad (1)$$

$P(Y=1|G=unprivileged)$ is the probability that a member of the unprivileged group receives a favorable prediction, whereas $P(Y=1|G=privileged)$ is the corresponding probability for the privileged group. Following established empirical practice, we use the interval $[0.8, 1.2]$ as a practical reference for potential disparate impact (Feldman et al., 2015).

Average Odds Difference, represented in Equation 2, evaluates whether the model produces similar true-positive and false-positive rates across the privileged and unprivileged groups. For each group g , the true positive rate (TPR_g) is the probability that the model predicts a regular academic status among students who are actually regular. The false positive rate (FPR_g) is the probability that the model predicts a regular academic status among students who are actually classified as at risk. AOD is then defined as:

$$AOD = \frac{1}{2} [(TPR_{unpriv} - TPR_{priv}) + (FPR_{unpriv} - FPR_{priv})] \quad (2)$$

In Equation 2, TPR_{unpriv} and FPR_{unpriv} refer to the unprivileged group, whereas TPR_{priv} and FPR_{priv} refer to the privileged group. An AOD value equal to zero indicates that the two groups have equal average true-positive and false-positive rates.

In line with prior empirical studies, we consider $|AOD| < 0.2$ as a conservative low-disparity threshold, reflecting the absence of a universal standard while maintaining alignment with existing fairness evaluations in learning analytics (Bellamy et al., 2019).

To assess the statistical robustness of fairness metrics, we estimated 95% confidence intervals for Disparate Impact and Average Odds Difference using non-parametric bootstrap resampling (1,000 iterations) on the test set, recomputing metrics before and after applying the reweight technique (Hastie et al., 2009).

The two metrics have different reference values and directions. For DI, a value equal to 1 indicates parity in favorable prediction rates. Values below 1 indicate that the unprivileged group receives favorable predictions less frequently than the privileged group, whereas values above 1 indicate that the unprivileged group receives favorable predictions more frequently. In this study, the interval $[0.8, 1.2]$ is used as a practical reference band rather than as a universal fairness criterion. For AOD, a value equal to 0 indicates parity in average true-positive and false-positive rates. Positive values indicate that these rates are, on average, higher for the unprivileged group, whereas negative values indicate that they are lower. Because higher TPR and higher FPR have different practical meanings, the sign of AOD alone should not be interpreted as an advantage or disadvantage. Therefore, the magnitude $|AOD|$ is used to represent the degree of disparity, with values closer to zero indicating fairer behavior.

3.4. Experimental design, evaluation and pipeline

For each data availability scenario, supervised machine learning models were trained to predict academic risk. We adopted a paired experimental design: for each combination of scenario, classifier, and sampling strategy, two models were trained using the same data split (a baseline model and a fairness-aware model). In the fairness-aware configuration,

the reweight technique was applied to the training data to reduce statistical dependence between sensitive attributes and predicted outcomes (Kamiran and Calders, 2012).

Predictive performance was evaluated using AUC-ROC, while fairness was assessed using Disparate Impact and Average Odds Difference. This design enables a direct comparison of predictive performance and fairness before and after mitigation under identical experimental conditions, isolating the effects of both data availability and bias mitigation.

Across scenarios, we evaluated a diverse set of commonly used classifiers in Learning Analytics, including Random Forest, AdaBoost, XGBoost, Gradient Boosting Machine, Histogram-based Gradient Boosting, Decision Tree, Naive Bayes, and CatBoost. Given the strong class imbalance, imbalance handling techniques were applied exclusively to the training data, including ADASYN, SMOTE, SMOTEENN, NearMiss, and random over-sampling. For reporting, we highlight the best-performing configuration per scenario (CatBoost + SMOTEENN), which consistently achieved the highest discrimination for the at-risk class.

The set of classifiers considered in this study was intentionally selected to balance methodological diversity with what is already well established in Learning Analytics research. These models are particularly effective at capturing nonlinear relationships, handling heterogeneous feature spaces, and accommodating complex interactions between variables, all of which are common characteristics of academic datasets (Prokhorenkova et al., 2018).

Simpler models, including Decision Trees and Naive Bayes, were deliberately included as baseline references. While they may not always deliver state-of-the-art predictive performance, their transparency and interpretability make them a recurring presence in educational data mining studies (Romero et al., 2016). In practice, these models help ground the analysis, making it easier to contextualize both performance gains and fairness-related trade-offs observed in more complex approaches.

Fig 1 summarizes the pipeline: preprocessing, feature selection, imbalance handling, model training, and evaluation of performance and fairness before and after reweight. This workflow isolates the effects of data availability and bias mitigation under realistic institutional conditions.

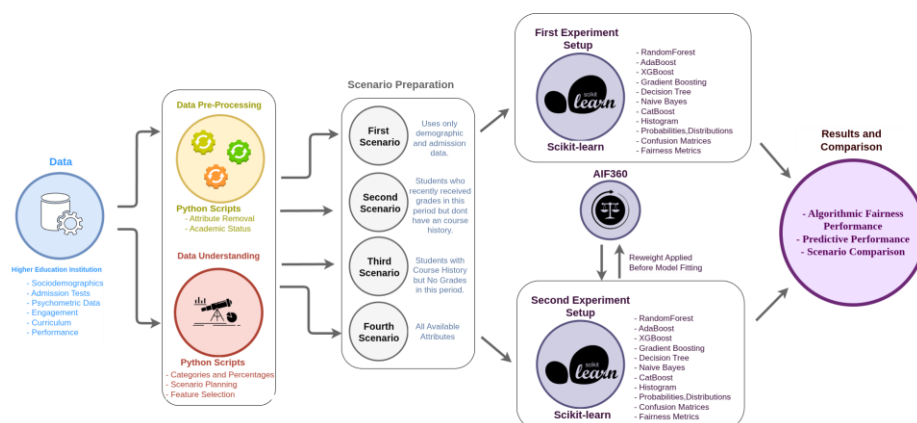


Figure 1. Overview of the experimental pipeline with baseline and fairness-aware models.

4. Results and Discussion

This section presents the empirical results and discusses their implications. Section 4.1 reports predictive performance across the four data-availability scenarios and compares the baseline and fairness-aware configurations. Section 4.2 presents the fairness results before and after the reweight technique, including the bootstrap analysis. Section 4.3 interprets the findings in relation to previous research, the institutional context, and their practical implications for academic risk prediction.

4.1. Predictive performance results

Fig 2 presents the predictive performance of the best-performing model configuration in each data-availability scenario, before and after the reweight technique. The selected configuration, CatBoost combined with SMOTEENN, consistently achieved the strongest discrimination for the at-risk class across the evaluated scenarios.

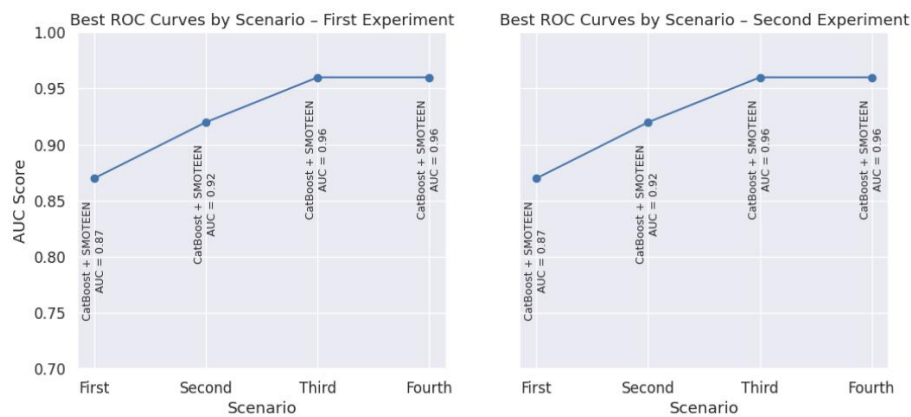


Figure 2. Overview of the best models per scenario with and without the reweight technique.

The models highlighted in Fig 2 were selected based on their ability to better predict the negative class, corresponding to students with an unsuccessful academic outcome. In both experiments across all data availability scenarios, CatBoost combined with SMOTEENN consistently achieved the strongest performance for this group. For each data availability scenario, we emphasize the configuration achieving the strongest performance for this class, as accurately identifying students at academic risk is central to supporting early educational interventions. Fig 2 contrasts two experimental settings: a baseline experiment without bias mitigation and a fairness-aware experiment in which the reweight technique is applied.

Predictive performance in both experiments improves as more academic information becomes available. In the most constrained setting (First Scenario), models already achieve reasonable discrimination, with AUC-ROC values around 0.85–0.87, suggesting that early-stage risk prediction is feasible even before grades are observed. Performance increases noticeably once partial academic information is introduced (Second and Third Scenarios) and reaches its upper bound in the fully informed setting (Fourth Scenario), with AUC-ROC values approaching 0.96. However, the Fourth Scenario should be interpreted with caution, as it assumes that the model has access to information that would typically only be available at the end of the academic trajectory. Overall, the differences in AUC-ROC between the baseline and fairness-aware

configurations were small across all scenarios, indicating that the fairness intervention did not materially affect predictive discrimination.

4.2. Fairness results

This subsection examines whether the reweight technique reduced disparities between the selected privileged and unprivileged intersectional groups. Fig 3 and Fig 4 compare Disparate Impact and Average Odds Difference before and after mitigation, whereas Table 2 and Table 3 report the corresponding bootstrap confidence intervals.

Before any mitigation is applied, fairness diagnostics reveal systematic disparities across sensitive groups. As shown in Fig 3, Disparate Impact (DI) frequently deviates from the parity threshold.

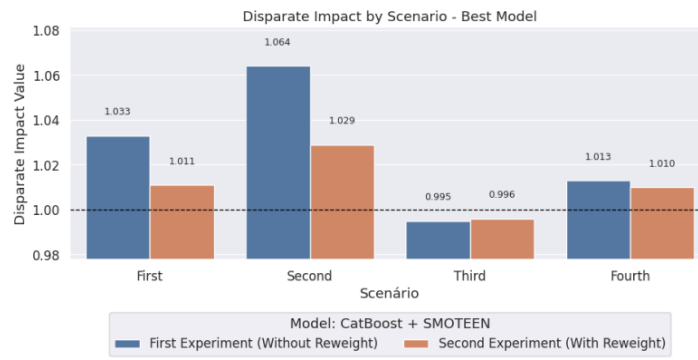


Figure 3. Disparate Impact (DI), before and after the reweight technique across data availability scenarios.

Fig 4, shows that Average Odds Difference (AOD) indicates non-negligible gaps between privileged and unprivileged groups across all data availability scenarios.

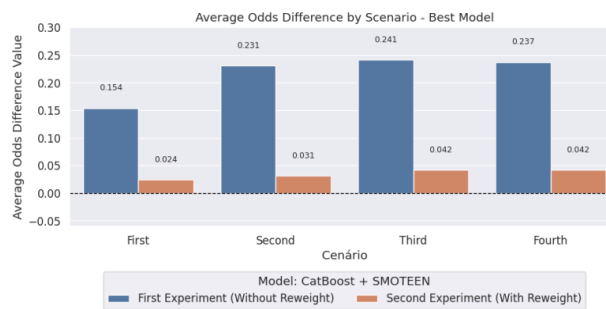


Figure 4. Average Odds Difference (AOD) before and after the reweight technique across data availability scenarios.

Overall, in the experiments where the reweight technique was applied, DI (Disparate Impact) values tend to move closer to parity, while AOD (Average Odds Difference) values consistently shrink toward zero, indicating more balanced error rates between privileged and unprivileged groups.

To formally assess whether the reweight technique mitigates these disparities, the analysis is framed as a set of within-scenario hypothesis tests comparing fairness metrics before and after mitigation.

To evaluate whether the reweight technique moves the metric closer to this threshold, we consider the absolute deviation from parity represented in Equation 3.

$$D_{DI} = |DI - 1| \quad (3)$$

For each scenario s , the following hypotheses, represented in Equation 4 and Equation 5, are tested:

$$H_0^{DI}: D_{DI,after}^{(s)} - D_{DI,before}^{(s)} = 0 \quad (4)$$

$$H_1^{DI}: D_{DI,after}^{(s)} - D_{DI,before}^{(s)} < 0 \quad (5)$$

Average Odds Difference is a difference-based metric, represented in Equation 6 and 7, with a reference value equal to zero, where non-zero values indicate disparities in true-positive rate and false positive rates. For each scenario s , we test:

$$H_0^{AOD}: AOD_{AOD,after}^{(s)} - AOD_{AOD,before}^{(s)} = 0 \quad (6)$$

$$H_1^{AOD}: AOD_{AOD,after}^{(s)} - AOD_{AOD,before}^{(s)} < 0 \quad (7)$$

Statistical inference is conducted using percentile bootstrap confidence intervals (95%, $n = 1000$). The null hypothesis is rejected when the confidence interval of the metric difference excludes the null value.

Table 2. Bootstrap 95% confidence intervals for Disparate Impact (DI)

Scenario	Model + Technique	Before (IC95%)	After (IC95%)	Signif.
1	CatBoost + SMOTEENN	1.033 [1.010; 1.056]	1.011 [0.990; 1.032]	Yes
2	CatBoost + SMOTEENN	1.064 [1.041; 1.087]	1.029 [1.010; 1.048]	Yes
3	CatBoost + SMOTEENN	0.995 [0.977; 1.013]	0.996 [0.979; 1.013]	No
4	CatBoost + SMOTEENN	1.013 [0.995; 1.031]	1.010 [0.992; 1.028]	No

Table 2 and Table 3 report the bootstrap confidence intervals for DI and AOD, respectively. For Disparate Impact, statistically significant reductions toward parity are observed primarily in scenarios with limited or partial academic information. In contrast, Average Odds Difference exhibits statistically significant reductions across all scenarios.

Table 3. Bootstrap 95% confidence intervals for Average Odds Difference (AOD)

Scenario	Model + Technique	Before (IC95%)	After (IC95%)	Signif.
1	CatBoost + SMOTEENN	0.154 [0.123; 0.185]	0.024 [-0.003; 0.051]	Yes
2	CatBoost + SMOTEENN	0.231 [0.201; 0.261]	0.031 [0.004; 0.058]	Yes
3	CatBoost + SMOTEENN	0.241 [0.208; 0.274]	0.042 [0.009; 0.075]	Yes
4	CatBoost + SMOTEENN	0.237 [0.206; 0.268]	0.042 [0.015; 0.0693]	Yes

4.3. Discussion and institutional implications

The results indicate that the reweight technique improved group fairness without materially changing predictive discrimination. This pattern is consistent with previous evidence showing that instance-weighting approaches can reduce group disparities while preserving predictive performance (Blow et al., 2024; Friedler et al., 2019; Raftopoulos et al., 2025).

However, the effects were not identical across fairness metrics. Improvements in Average Odds Difference were larger and more consistent than improvements in Disparate Impact, suggesting that the reweight technique had a stronger effect on disparities in error rates than on the overall distribution of favorable predictions. This distinction is important because DI (Disparate Impact) and AOD (Average Odds Difference) capture different dimensions of fairness. DI (Disparate Impact) evaluates whether groups receive favorable predictions at similar rates, whereas AOD evaluates whether true-positive and false-positive rates are balanced across groups.

Therefore, a model may present a DI (Disparate Impact) value close to parity while still producing unequal error patterns. This helps explain why the third and fourth scenarios showed limited changes in DI (Disparate Impact) but statistically significant reductions in AOD (Average Odds Difference).

Regarding RQ1, the results show that applying the reweight technique did not produce meaningful changes in AUC-ROC across the four data-availability scenarios. Predictive discrimination increased primarily as additional academic information became available, rather than as a consequence of the fairness intervention.

Regarding RQ2, the results provide evidence that the reweight technique reduced algorithmic disparities, particularly those related to differences in error rates. AOD (Average Odds Difference) moved substantially closer to zero in all four scenarios, whereas DI (Disparate Impact) moved significantly closer to parity in the first and second scenarios.

5. Final remarks

This study examined whether fairness-aware preprocessing can be realistically integrated into academic risk prediction pipelines without undermining their practical usefulness. Across multiple data availability scenarios, the results consistently show that fairness does not need to come at the expense of predictive performance. Even in early-stage settings with limited information, models with and without the reweight technique achieved comparable performance, while the fairness-aware configuration led to statistically significant reductions in Disparate Impact and Average Odds Difference in most scenarios. Together, these findings reinforce the practical viability of incorporating fairness-aware preprocessing into academic risk prediction in higher education.

Beyond the numerical results, these findings carry important implications for educational practice. They suggest that predictive systems can support early and more equitable interventions, helping institutions identify students at academic risk without disproportionately disadvantaging specific populations. In this sense, algorithmic fairness emerges not as a purely technical concern, but as an educational and ethical requirement, particularly in contexts where predictive models may influence access to support, resources, or institutional attention.

Rather than treating fairness as a post hoc diagnostic, this work frames it as an operational property that evolves alongside data availability. The scenario-based design highlights how both performance and fairness shift as academic information accumulates, offering a more realistic view of how learning analytics systems behave in practice.

At the same time, this study has clear limitations. The analysis is based on data from a single institution, which may limit the generalizability of the findings. In addition, fairness was evaluated using a specific set of sensitive attributes and a mitigation strategy, and other definitions of fairness or alternative interventions may yield different dynamics.

Moreover, this study focuses on a single preprocessing intervention. Although this design enables us to isolate the effects of the reweight technique across scenarios, it does not provide a direct comparison with alternative preprocessing, in-processing, or post-processing approaches. Future studies should compare these mitigation families under the same data splits, classifiers, sensitive-group definitions, and evaluation criteria.

These limitations open several directions for future work. Further studies could replicate this analysis across multiple institutions, explore additional sensitive attributes, or compare different fairness-aware techniques under similar scenario conditions. Longitudinal analyses and qualitative investigations may also help clarify how fairness-aware predictions translate into actual educational interventions. Taken together, the evidence presented here supports a practical approach for the field, as fairness-aware design choices can and should be incorporated into academic risk prediction systems from the outset, rather than treated as an afterthought.

Acknowledgments

The authors would like to acknowledge the Living Lab & Data Hub of the Institute for the Future of Education, Tecnológico de Monterrey, Mexico, for the data used in this work and provided through the Call "Encouraging Research on Higher Education Students' Profiles with Social Commitment". This work was supported by the National Council for Scientific and Technological Development (CNPq, Brazil) under Grants 445016/2024-8 and 308745/2025-6.

References

- Baker, R.S., Esbenshade, L., Vitale, J., Karumbaiah, S.: Using demographic data as predictor variables: a questionable choice. *Journal of Educational Data Mining* 15(2), 22–52 (2023). DOI: 10.5281/zenodo.7702628.
- Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al.: AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63(4/5), 4:1–4:15 (2019). DOI: 10.1147/JRD.2019.2942287.
- Blow, C.H., Qian, L., Gibson, C., Obiomon, P., Dong, X.: Comprehensive validation on reweighting samples for bias mitigation via AIF360. *Applied Sciences* 14(9), 3826 (2024). DOI: 10.3390/app14093826.
- Cechinel, C., Ochoa, X., Lemos dos Santos, H., Carvalho Nunes, J.B., Rodés, V., Marques Queiroga, E.: Mapping Learning Analytics initiatives in Latin America. *British Journal of Educational Technology* 51(4), 892–914 (2020). DOI: 10.1111/bjet.12941.
- Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 259–268. Association for Computing Machinery, Sydney, Australia (2015). DOI: 10.1145/2783258.2783311.
- Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. pp. 329–338. Association for Computing Machinery, New York, NY, USA (2019). DOI: 10.1145/3287560.3287589.
- Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer, New York (2009). DOI: 10.1007/978-0-387-84858-7.
- Hlosta, M., Zdrahal, Z., Zendulka, J.: Ouroboros: Early identification of at-risk students without models based on legacy data. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*. pp. 6–15. Association for Computing Machinery, Vancouver, Canada (2017). DOI: 10.1145/3027385.3027449.
- Institute for the Future of Education (IFE): Educational Innovation Dataset Collection. Tecnológico de Monterrey Research Data Hub, Monterrey, Mexico (2024). DOI: 10.57687/FK2/UFM2QH. Last accessed 2026/06/24.

- Jayaprakash, S.M., Moody, E.W., Lauría, E.J.M., Regan, J.R., Baron, J.D.: Early alert of academically at-risk students: an open source analytics initiative. *Journal of Learning Analytics* 1(1), 6–47 (2014). DOI: 10.18608/jla.2014.11.3.
- Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1), 1–33 (2012). DOI: 10.1007/s10115-011-0463-8.
- Kizilcec, R.F., Lee, H.: Algorithmic fairness in education. In: Holmes, W., Porayska-Pomsta, K. (eds.) *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*, pp. 174–202. Routledge, New York (2022). DOI: 10.4324/9780429329067-10.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems* 31 (NeurIPS 2018). pp. 6639–6649 (2018). DOI: 10.5555/3327757.3327770.
- Raftopoulos, G., Davrazos, G., Kotsiantis, S.: Evaluating fairness strategies in educational data mining: a comparative study of bias mitigation techniques. *Electronics* 14(9), 1856 (2025). DOI: 10.3390/electronics14091856.
- Romero, C., Cerezo, R., Bogarín, A., Sánchez-Santillán, M.: Educational process mining. In: *Data Mining and Learning Analytics: Applications in Educational Research*, chap. 1, pp. 1–28. Wiley (2016). DOI: 10.1002/9781118998205.ch1.
- Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., Liu, Y.: How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. pp. 99–106. Association for Computing Machinery, Honolulu, USA (2019). DOI: 10.1145/3306618.3314248.
- Tecnológico de Monterrey: Monterrey Institute of Technology and Higher Education. Monterrey, Mexico (2024). Last accessed 2026/06/22.