JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 12(3), 47–65. https://doi.org/10.18608/jla.2025.8761

Balancing Act: Early, Fair, and Accurate Identification of At-Risk Students

Alison Cheng[1], Bo Pei[2] and Cheng Liu[3]

Abstract
Machine learning algorithms have been widely used for identifying at-risk students. Current research focuses on timeliness and accuracy of the predictions, leading to a heavy reliance on demographic data, which introduces severe bias issues. This study develops fairness-aware machine learning models to identify at-risk students in high school Advanced Placement (AP) statistics, where student performance is closely linked to demographic background. We evaluated the predictive performance and bias mitigation strategies of various machine learning algorithms. To determine the optimal time for accurate and fair identification of at-risk students, we divided the dataset into three stages corresponding to the course's progress. At each stage, we examined model performance and fairness across groups defined by race, gender, and eligibility for free/reduced-price lunch. Our findings suggest that at Stage 1 (i.e., up to the first unit review assignment), the models effectively identified at-risk students while maintaining fairness across demographic groups. We discovered that incorporating more learning activity data reduced the potential bias caused by overreliance on demographic information. We also examined the impact of different bias mitigation approaches as well as the exclusion of the sensitive features on predictive accuracy and fairness. We further discuss their implications for designing more context-specific solutions in educational settings.

Notes for Practice

- This study investigates the process of developing fairness-aware machine learning models to balance timing, predictive accuracy, and fairness while identifying at-risk students.
- Balancing the dataset does not always effectively mitigate bias in predictions, suggesting that imbalances in the distribution of samples are not the sole source of bias.
- Removing sensitive features in the dataset during the model training process may reduce direct bias, but it does not necessarily eliminate bias, as models can capture the influence from the correlated variables.

Keywords: Algorithmic bias, identifying at-risk students, machine learning, learning analytics, K–12 students, AP statistics

Submitted: 04/11/2025 — **Accepted:** 25/06/2025 — **Published:** 26/11/2025

[1]Email: ycheng4@nd.edu Address: Department of Psychology, University of Notre Dame, 390 Corbett Family Hall, Notre Dame, IN 46556, USA.
Corresponding author [2]Email: bpei@usf.edu Address: Department of Educational and Psychological Studies, College of Education, University of South Florida, 4110 USF Apple Dr., Tampa, FL 33620, USA. ORCID iD: https://orcid.org/0000-0002-6328-6929
[3]Email: cliu7@nd.edu Address: Department of Psychology, University of Notre Dame, 390 Corbett Family Hall, Notre Dame, IN 46556, USA; Lucy Family Institute for Data and Society, University of Notre Dame, 384E Nieuwland Science Hall & 610 Flanner Hall, Notre Dame, IN 46556 USA. ORCID iD: https://orcid.org/0000-0002-2787-1653

1. Introduction

Fairness issues in the applications of machine learning (ML) algorithms have been widely discussed, especially in high-stakes fields such as healthcare (Chen et al., 2022), criminal justice (Angwin et al., 2016), and hiring (Gu et al., 2014), where biased predictions can lead to significant consequences. In educational settings, addressing fairness issues is also necessary, since the use of predictive models could profoundly impact students, instructors, institutions, and society as a whole (Kizilcec & Lee, 2021). For example, AI algorithms were found to have different levels of discrimination towards female and minority students when used in college admission settings (Marcinkowski et al., 2020). Moreover, Jiang and Pardos (2021) also highlighted that AI algorithms can produce biased predictions for students from different race groups. Such discrimination is usually unintentional and can be challenging to detect, often remaining unnoticed without deliberate application of specific methods designed for its identification.

Despite the extensive research on ensuring fairness in teaching and learning settings (Gardner et al., 2019; Baker & Hawn, 2021; Jiang & Pardos, 2021; Kizilcec & Lee, 2021; Yu et al., 2021), there appears to be limited exploration in fairness-aware predictive models for early identification of at-risk students. There are two unique problems in early identification, the first being that the methods can be particularly prone to fairness issues. In the context of early identification, AI algorithms are trained on incomplete learning processes and data, which can lead to more bias because of the unintentional emphasis on demographic information. Second, algorithms to identify at-risk students should emphasize the reduction of false negatives to ensure that all at-risk students are supported rather than merely pursuing an overall high predictive accuracy. Therefore, there is a clear need for a study in this unique context that examines machine learning models on different accuracy and fairness performance metrics at various stages of a course. Such research is vital for illuminating specific factors that should be taken into account in the pursuit of data-driven decision making in this area.

To meet this need, we conducted a comprehensive investigation into balancing accuracy and fairness in early identification of at-risk students. Our study will particularly focus on an Advanced Placement (AP) statistics course, given its essential role in a student learning process. Enrolling in an AP course allows students to earn college credits, prepare for college enrollment, and explore different subject areas. Most importantly, the AP program tends to increase college admission, since an increasing number of colleges are going test-optional/-blind, which places more weight on high school transcripts and GPAs, with performance in AP courses playing a significant role in these evaluations (Ober et al., 2023). The focus on statistics is further driven by increasing enrollments and high demand in the labour market. Especially from 1990 to 2019, there has been a significant increase in the number of students taking statistics and probability (1% vs. 16.7%) over those taking calculus (7.2% vs. 15.8%; NCES, 2022. And for the job market, the U.S. Department of Labor Statistics projects a 31% increase in statistics-related employment. In addition, various studies indicated that student learning performance in AP Statistics is significantly associated with their demographic background, including race, gender, eligibility for free/reduced-price lunch, etc., suggesting potential bias (Ober et al., 2023).

As such, this study aims to answer the following research question:

RQ1: How do we build machine learning models that can accurately identify at-risk students at an early stage while minimizing potential biases?

Since data balancing and removal of sensitive features are some of the most intuitive approaches to bias mitigation, as suggested by Baker et al. (2023), we investigated two subsequent questions:

RQ2: What are the effects of data balancing on model accuracy and fairness levels in identifying at-risk students?

RQ3: To what extent does the inclusion or exclusion of sensitive features influence model predictive performance and fairness levels?

To answer these questions, we trained four models including LGBMClassifier, StackingClassifier, AdaBoostClassifier, and CatBoostClassifier at three stages in the course: Demog Stage or Stage 0 (i.e., using only demographic information), Stage 1 (i.e., using demographic information and the learning activities up to the first unit review assignment), and Stage 2 (i.e., using the demographic information and the learning activities up to the second unit review assignment). At each stage, we explicitly examined the accuracy and fairness levels of each model across subgroups of students based on race, gender, and eligibility for free/reduced-price lunch. In this process, predictive performance and fairness levels were compared to imbalanced data and balanced data. Furthermore, we also investigated the impacts of inclusion and exclusion of the demographic information on each model's predictive performance and fairness level.

2. Literature Review

2.1. Algorithmic Fairness in Education

Fairness in education has been a longstanding issue even before the application of technologies in teaching and learning settings (Holmes & Tuomi, 2022). Early studies mainly centred on the inequalities and inequities of educational processes and outcomes, with particular attention on segregation and achievement gaps. Socioeconomic status has been identified as one of the significant factors related to bias in education. Kao and Thompson (2003) explicitly investigated a combination of factors — such as residential area, community influences, and in-school dynamics — that could produce disparities in learning outcomes from different student groups.

At the same time, algorithms can perpetuate biases, especially if they were trained on datasets tainted by historical prejudices. Kordzadeh and Ghasemaghaei (2022) defined algorithmic bias as the case in which "a model's predictive performance (however defined) unjustifiably differs across disadvantaged groups along social axes such as race, gender, and class" (p. 1). Within the educational realm, given that demographic data is closely linked to learning outcomes, biased algorithms will inherently classify students from historically disadvantaged groups as lower-performing, further exacerbating existing biases. Apart from this, as Kizilcec and Lee (2021) highlighted, algorithmic bias is "unintentionally or potentially harmful" and difficult to identify unless rigorously examined. Bias issues can exist at every stage from model building to predictive result interpretations. To fully quantify potential bias, we must understand several critical aspects: 1) whether the

dataset used to train AI models represents the diversity of the population; 2) whether the definition of the problem itself is discriminatory against specific groups of students; and 3) whether the interpretations and applications of the predictive results can lead to discrimination.

Additionally, Baker and Hawn (2021) provided a comprehensive literature review highlighting the importance of implementing strategies to identify and mitigate algorithmic bias in educational settings. The authors proposed measurement-related bias assessment approaches, including examining the representativeness of the dataset to uncover inherent bias, monitoring prediction behaviour across different groups to identify unintentional bias, and exploring equity focused metrics in addition to the common performance-based metrics. Instead of merely focusing on fairness statistically, fairness metrics in real educational settings should incorporate more equity perspectives, highlighting the differences in the learning needs of students with different backgrounds. Achieving this requires transparency and interpretability in the AI decision-making process, which in turn facilitates the engagement of domain experts and educators.

2.2. Fairness Issues in Identifying At-Risk Students

Literature on fairness issues around identifying at-risk students is scarce. Most research focuses on building highly accurate models that can detect at-risk students early on in order to provide timely and targeted interventions. For example, Adnan et al. (2021) built various machine learning (ML) and deep learning (DL) models to predict the college student dropout rate at different stages of a course (e.g., 0%, 20%, 40%, ... of the course length). The study identified factors like student assessment scores and engagement intensity as crucial determinants of the student learning performance. Similarly, Marbouti et al. (2016) extensively examined performance in identifying at-risk students in courses with standards-based grading criteria. The best performing model was found to have over 72% accuracy in identifying at-risk students as early as week 5. In another study, Hu and Rangwala (2020) proposed a multiple co-operative classifier model (MCCM) that incorporates two different classifiers, each corresponding to a sensitive feature (e.g., male/female). The authors constructed an objective function that considers differences in both predictive accuracy and fairness levels. Particularly, fairness is measured using KL-divergence, indicating the difference in predicted probabilities when a student's sensitive attribute is altered. The model was evaluated on a dataset (collected from five majors across ten years at George Mason University) by comparing the performance with four other classifiers that are easy to control for fairness levels (e.g., logistic regression, Rawlsian fairness, learning fair representation, and adversarial learned fair representation). The experimental results demonstrated the MCCM can reduce both individual and group-level bias while maintaining relatively high predictive accuracy. In a recent study conducted by Ober and colleagues (2023), a learning-sequence-aware predictive model was built to identify at-risk students in foundational courses at the college level.

However, none of these studies investigated fairness issues considering the balancing between time and accuracy in the process of identifying at-risk students. Specifically, since the identification of at-risk students is always required at the early stage in the learning process, models are often built on partial student learning data, which leads to a higher reliance on students' demographic information. As such, there is a need to conduct systematic evaluations of how algorithms perform under different fairness metrics in identifying at-risk students at different time points and how effective the bias mitigation approaches are in balancing the trade-offs between accuracy and fairness.

2.3. Measurement and Mitigation Approaches in Education

Recently, various metrics have been proposed to address bias issues in AI models in educational settings (Bellamy et al., 2018; Kizilcec & Lee, 2021). Deho et al. (2022) investigated the effectiveness of several fairness metrics, particularly in the context of learning analytics, using a comparative evaluation approach with three-year program dropout data from an Australian university. Their findings indicated that bias issues were not always from datasets and further suggested that ensuring fairness can lead to enhanced utility under specific circumstances. The notion of fairness can be various under different educational contexts, requiring the careful selection of appropriate fairness metrics that align with the specific goals, stakeholders, and contextual constraints. In Cohausz et al. (2024), the authors introduced several case studies to extensively discuss how the choice of different fairness measurement strategies are related to the underlying data generation mechanism, the potential application, and the normative beliefs. In addition to investigating the effectiveness of existing fairness metrics in educational settings, researchers also implemented additional metrics to address educational issues. For example, Gardner et al. (2019) proposed a fairness metric, ABROCA, evaluating unfairness in predictive student models by investigating the differential accuracy between student subgroups. The model was evaluated on 44 unique MOOCs over four million learners and the analytical results indicated effectiveness of the approach in evaluating how the predictive models impact different student groups disproportionately. In another study, Lalor et al. (2024) proposed the FAIR-Frame to model fairness across multiple protected attributes in terms of both representational and allocational harm in the model building and result interpretation process. The framework provides significant practical implications regarding designing adaptive, fairness-aware models in various AI application settings.

In addition to fairness evaluation, approaches to mitigate the identified bias have also been extensively studied. Generally, mitigation algorithms can be divided into three categories — pre-processing, in-processing, and post-processing — according to the potential sources of bias in the machine learning models (Bellamy et al., 2018). Pre-processing mainly focuses on approaches for better representing the samples in the dataset, such as Re-weighing, Optimized pre-processing, and so on. In-processing attempts to optimize the training process to build models that take into account the known biases in the dataset. Post-processing approaches are developed to calibrate the outcomes so that the predictions are more in line with the real-world situation. Such approaches were designed to adjust predictions as indicated by fairness metrics (i.e., demographic parity) while not sacrificing predictive performance with performance metrics (i.e., balanced accuracy score).

Although these fairness evaluation and mitigation approaches are well-established and evaluated in the machine learning literature, they are still underexplored in educational settings. Existing studies like Gardner et al. (2019) provide insights into the implementation of metrics for educational settings, but the authors did not focus on the context of identifying at-risk students. This context is more prone to bias against historically disadvantaged groups given its focus on early action based on less learning activity data, resulting in heavier dependence on demographic data. While later stages allow models to access more learning activity data, it is too late to provide effective interventions for at-risk students. As such, we must not only consider the trade-offs between accuracy and fairness, but must also find an optimal time to conduct the analysis for timely intervention. Given the demands of explainability in identifying and providing interventions for at-risk students, our study mainly employs approaches from the pre-processing and post-processing stages.

3. Methods

3.1. Research Context and Dataset

This study examined 215 students who enrolled in AP Statistics from seven high schools in the State of Indiana during the 2019–2020 academic year. To be included in the study, students must have provided consent/assent, completed self-report surveys including their demographic information, completed at least two end-of-unit review assignments, and not withdrawn from the class (Pei & Xing, 2022; Ober et al., 2023). The background information includes Gender (Male: 50.2%, Female: 49.8%), Race/Ethnicity (White/Asian: 88.4%, Other: 11.6%), Eligibility for free/reduced-price lunch (Yes: 15.4%, No: 84.6%), Math classes previously taken, and self-predicted grades in the course. Other than background information, we also recorded student practice activities for completing each of four end-of-unit assignments online as well as the corresponding performance. Because of the outbreak of COVID 19, the last assignment was dropped, so we have student learning practices for three assignments during the whole school year. The first assignment began around November 3rd, 2019, and ended around November 20th, 2019; the second assignment began on December 6th, 2019, and ended at around December 19th, 2019; and the third assignment began on February 23rd, 2020, and ended around March 13th, 2020. For each assignment, we extracted and derived learning features such as the number of times practised, the number of unique questions answered, the earliest time the student began to practise, the time of the last practice, the time span between the earliest and latest practice, the average response time for the answered questions, and the average performance over all the answered questions. Student learning outcomes in the AP Statistics course were measured by the AP exam score with 1 indicating the lowest and 5 indicating the highest performance. The outcome variable was derived by dichotomizing students' AP exam scores: at-risk (AP score < 3) and others (AP score > = 3). This is because colleges typically require an AP exam score of 3 or higher (sometimes 4 or higher) for college credits.

3.2. Analytical Pipeline

Figure 1 shows the proposed analytical framework consisting of three stages of analysis:

- Stage 0 (Demog stage): Identifying at-risk students with demographic information only
- Stage 1: Identifying at-risk students with demographic information and student learning activities for the first assignment
- Stage 2: Identifying at-risk students with demographic information and student learning activities for the first and second assignments

For each stage, we trained four machine learning models (i.e., LGBMClassifier, CatBoostClassifier, StackingClassifier, and AdaBoostClassifier), and evaluated the predictive performance as well as the fairness levels on three protected features (i.e., gender, race, free/reduced-price lunch eligibility). Finally, we applied bias mitigation strategies (i.e., ThresholdOptimizer, a commonly used post-processing mitigation approach) to each of the models at each stage and measured the changes in the predictive performance to investigate the trade-offs between performance and fairness.

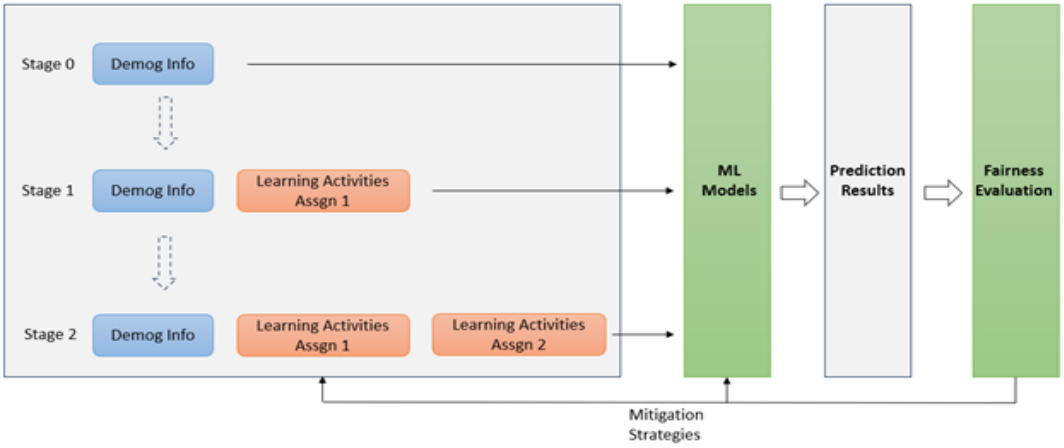


Figure 1. The overall analytical pipeline in this study.

3.3. Different Machine Learning Models

This study examined four machine learning models in each stage in order to compare their accuracy and fairness levels:

1. **LGBMClassifier**: LightGBM classifier is one of the most popular machine learning algorithms and has been widely used in real-world applications due to its high speed and effectiveness (Ayubkhan et al., 2023). It can handle both categorical and numerical features, making it suitable for the purpose of this study. Moreover, it has the potential to capture and reveal complex relationships among various factors in large-scale educational datasets, which can better support educators to optimize the teaching and learning experience.
2. **CatBoostClassifier**: Similar to LGBMClassifier, CatBoostClassifier falls into the category of gradient boosting techniques (Joshi et al., 2021). It can also deal with categorical attributes in educational settings. Moreover, this classification approach offers transparency and interpretability through feature importance scores and model diagnostics, which makes it a valuable tool to gain deeper insights into the associations between learning behaviours and performance.
3. **StackingClassifier**: StackingClassifier is an ensemble machine learning technique that combines the predictions of multiple base classifiers to produce more accurate and robust predictions (Książek et al., 2020). In this classifier, the base classifiers are trained on the same dataset and produce predictions independently. By combining the predictions of these base classifiers, the StackingClassifier can mitigate the weakness of each individual classifier and further improve the overall predictive accuracy and robustness.
4. **AdaBoostClassifier**: AdaBoostClassifier is another ensemble machine learning technique with multiple base classifiers (An & Kim, 2010). Different from the StackingClassifier, this classification technique sequentially trains a set of weak classifiers with multiple iterations. During each iteration, the weights are assigned and updated for each base classifier based on their performance. The overall predictions of the classifier are based on the outputs of each classifier with a weighted voting mechanism.

3.4. Performance and Fairness Evaluation Framework

This study employed multiple metrics to evaluate the predictive performance of these models, including balanced accuracy, accuracy, recall, and precision. During each stage, the average performance of the four ML models on each of these metrics was calculated to indicate the feasibility of identifying at-risk students at the corresponding stage. Similarly, various measures were constructed to evaluate the fairness levels. In the context of early identification of at-risk students, the models must not only perform well overall but must also operate equally well across different demographic groups. As such, it is essential to investigate specific prediction rates or error types, such as true positive rates (TPRs) and false negative rates (FNRs), to ensure that we do not overlook those students who really need help. In this case, a false negative means that a student who was actually at-risk is not flagged, which might result in missed opportunities for timely interventions and lead to long-term harm regarding potential academic opportunities. Therefore, fairness metrics in this situation should provide specific information about prediction rates for students from different groups, highlighting the prediction disparities across the underrepresented and majority groups. This focus also aligns with the principle of educational equity, which emphasizes that students from underrepresented groups can receive timely and personalized learning support (Holstein et al., 2020). As such, we employed the following four fairness measurement strategies in the context of identifying at-risk students.

3.4.1. Demographic Parity Differences

Demographic Parity Differences (Demog_Parity_Diffs) require similar positive prediction rates (e.g., at-risk student identification) from different demographic groups, regardless of their actual academic status (Chen et al., 2023). In other

words, this metric investigates whether the AI models disproportionately flagged at-risk students merely based on their demographic backgrounds. Analyzing demographic parity differences can be a significant starting point in educational practice since large disparities across underrepresented and majority groups can signal systemic bias in either the training data or the AI model's predictive process. It is particularly useful in the context of resource allocation and providing interventions, where flagging students disproportionately occurs more easily, which can lead to over-surveillance of specific groups while overlooking other groups.

While this measurement may not guarantee that students were flagged in perfect alignment with their actual learning status, it is still the essential step for identifying the structural imbalance that might go unnoticed through outcome-based metrics alone. As such, in this paper, we include Demographic Parity Differences as a fundamental measurement to examine the disparities in number of at-risk students identified across different groups. According to the study of Feldman et al. (2015), we set the acceptable fairness level of this measurement as less than 0.20. The formula for Demog_Parity_Diffs is as follows:

$$Demog\_Parity\_Diffs = |(P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = b))| \quad (1)$$

where P ($\hat{Y}$ = 1 | A = a) indicates the probability of the model predicting a positive outcome (e.g., flagging a student as at-risk) for group A = a; Correspondingly, P ($\hat{Y}$ = 1 | A = b) indicates the model predicts a positive outcome for group A = b.

3.4.2. Average Odds Differences

Average Odds Differences (Avg_Odds_Diffs) provide a more balanced assessment of prediction disparities across two groups by considering both the true positive rates (TPRs) and false positive rates (FPRs) across the two groups at the same time. Combining these two aspects into one measurement, the Average Odds Differences measurement emphasizes the equalities of the models in identifying at-risk students across groups. In addition, FPR differences reflect the proportion of students from both groups who were unnecessarily flagged as at-risk students. For example, if a model consistently misidentifies students from a certain group as at-risk students, even when they are performing well, it may inadvertently reinforce the existing learning disparities by repeatedly offering them unnecessary learning materials.

The Average Odds Differences ensure that models not only perform well overall but also treat students equally across groups, making it suited to high-stakes situations like identifying at-risk students. The commonly used acceptable fairness level of Average Odds Differences is below 0.20, with a value of 0 indicating perfect fairness. The formula for Avg_Odds_Diffs is as follows:

$$Avg\_Odds\_Diff = \frac{1}{2}(|TPR_a - TPR_b| + |FPR_a - FPR_b|) \quad (2)$$

where TPR_a and FPR_a are the true positive rates and false positive rates for group A = a, and TPR_b and FPR_b are the corresponding rates for group A = b.

3.4.3. Equal Opportunity Differences

Equal Opportunity Differences (Eq_Opp_Diffs) measure the differences in TPR between the two groups, which is a particularly relevant fairness measurement in the context of identifying at-risk students. The focus on true positives aligns well with our goal of correctly flagging students at-risk regardless of group membership to ensure that they have "equal opportunity" to receive the support they need. Unlike the broader measures, such as Demographic Parity Difference, the Equal Opportunity Difference addresses a core ethical concern in education, ensuring that students who really need help receive it. A larger Equal Opportunity Difference indicates that the model is more likely to correctly identify at-risk students from a certain group than others, which could raise concerns about fair access to academic support and learning opportunities.

In this study, we use Equal Opportunity Differences to evaluate whether our models provide equitable interventions to students across different demographic groups. This measurement emphasizes the need to correctly and equally identify the at-risk students from both groups and provide the required interventions to support their learning. Minimizing the Equal Opportunity Differences can significantly increase the fairness, trust, and inclusivity of AI algorithms to support educational decision-making. We set the acceptable fairness level threshold at below 0.20 under the Equal Opportunity Differences metric (Hardt et al., 2016). The formula for the Eq_Opp_Diffs is as follows:

$$Eq\_Opp\_Diffs = |TPR_a - TPR_b| \quad (3)$$

where TPR_a and TPR_b are true positive rates for group A = a and A = b, respectively.

3.4.4. Predictive Parity Differences

Predictive Parity Differences (PPDs) measure consistency in the precision of positive predictions across two groups. In the context of identifying at-risk students, this measurement evaluates whether the difference in students flagged as at-risk is equally likely across the two groups. More specifically, the Predictive Parity Difference assesses the AI model's trustworthiness and reliability in identifying at-risk students for different groups. This measurement is important in high-stakes situations, such as providing personalized learning interventions, allocating critical learning resources, and making decisions

that may influence a student's academic path. Focusing on the precision of a model's positive predictions, Predictive Parity Differences can help capture the potential bias that affects educator interpretations of predictions and subsequent decisions made for supporting at-risk students.

In this study, we employed Predictive Parity Differences to assess whether the predictions of at-risk students are equally distributed across groups. Ensuring this consistency helps promote fair, data-driven decision making in AI-powered educational practices. Similarly, the acceptable fairness level threshold is set at below 0.20. The formula for PPDs is as follows:

$$PPDs = \left| \left(P\left(Y = 1 | \hat{Y} = 1, A = a\right)\right) - \left(P\left(Y = 1 | \hat{Y} = 1, A = b\right)\right) \right| \quad (4)$$

where $P\left(Y = 1 | \hat{Y} = 1, A = a\right)$ and $P\left(Y = 1 | \hat{Y} = 1, A = b\right)$ is the precision for group A = a and A = b, respectively.

3.5. Approaches for Mitigating Bias

To mitigate the identified bias, we employed a combination of pre-processing and post-processing mitigation strategies. As part of our pre-processing mitigation strategy, we employed a dataset balancing technique to address the potential bias introduced by the imbalanced group representation in the training data. Particularly, we employed the RandomOverSampler approach to oversample the students from the underrepresented groups to ensure that the models were trained on a more demographically representative dataset (Wang et al., 2021).

Post-processing mitigation strategies focus on calibrating the predicted outcomes, adjusting them to fulfill the specified parity constraints (Small et al., 2024). In the context of identifying at-risk students, where the number of flagged cases is typically small, post-processing mitigation strategies are more practical since they can provide specific insights for instructors to closely examine and validate the predictions and thus implement more appropriate strategies. In addressing the challenge of identifying at-risk students, it is imperative to acknowledge that overlooking these individuals could lead to a significant shortfall in providing necessary support. Therefore, our post-processing methodology emphasizes minimizing the false negative rate while simultaneously aiming to achieve a high level of balanced accuracy. Because the pre-processing and post-processing techniques are applied independently, we applied post-processing mitigation strategies on the balanced dataset to further enhance fairness in model predictions.

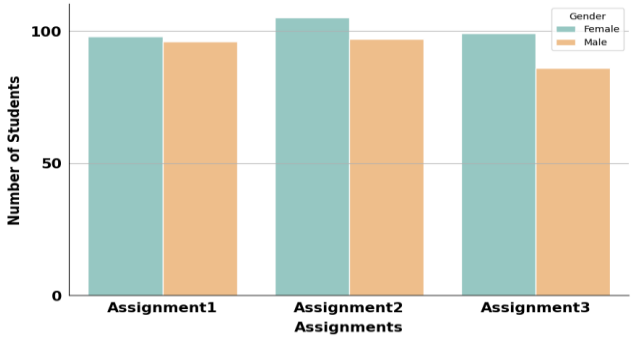
Moreover, the effectiveness of removing sensitive features in mitigating algorithmic bias is also a subject of ongoing discussion (Barocas et al., 2023; Holmes et al., 2022). Removing the sensitive features while training predictive models can prevent direct discrimination caused by these features. It does not eliminate bias, since models could still learn bias from the correlated variables (Haeri & Zweig, 2020). As a result, indirect discrimination might still exist, and the removal of sensitive features in the training process could even make it more difficult to detect and address bias. As such, our study also investigated these mitigation strategies both with and without the inclusion of sensitive features to provide practical implications in designing strategies for promoting accurate and fair identification of at-risk students.

In the Results section, we present a comprehensive analysis with comparisons about model performance and fairness level at each stage under different experimental conditions. We begin by reporting the results of pre-processing the dataset, including student distribution by each demographic indicator, and distributions of student learning behaviours at each stage. Next, we introduce detailed configurations of each predictive model to ensure the reproducibility of our experiments. We then compare and investigate the predictive performance and fairness level of each model, focusing on the race feature across four experimental conditions: 1) the original imbalanced dataset, 2) the balanced dataset, 3) the balanced dataset with post-processing bias mitigation strategies applied, and 4) the balanced dataset with both post-processing bias mitigation strategies applied and sensitive features removed. Detailed analysis regarding gender and lunch status as sensitive features are provided in Appendix A.

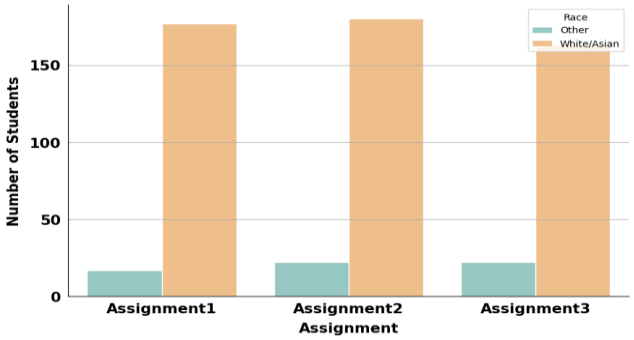
4. Results

4.1. Data Pre-Processing

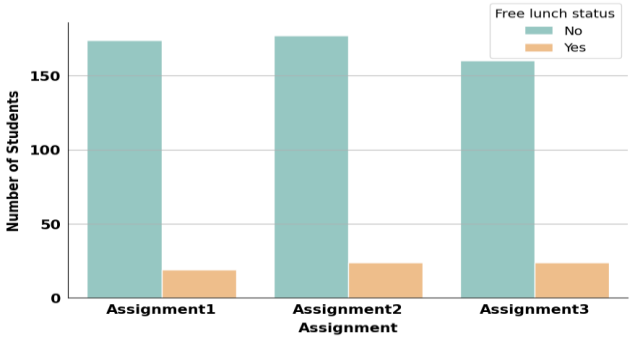
We first examined the distributions of students based on different demographic variables — race, gender, and eligibility for free/reduced-price lunch — within each assignment. Based on each demographic feature, we classified students into two distinct groups, guided by recommendations from the National Science Foundation (NSF): overrepresented and underrepresented (URM) groups. For example, White and Asian students were considered to be overrepresented groups while others were underrepresented. As shown in Figure 2, there was a balanced distribution of students based on gender but a significantly imbalanced distribution based on race and free/reduced-price lunch eligibility. To some extent, this suggests that the resulting ML models may generate biased predictions towards students who are not White or Asian as well as those eligible for free/reduced-price lunch. As such, data balancing techniques were applied to mitigate bias, the effectiveness of which will be discussed in detail in the following sections.



(a) Student distribution based on gender



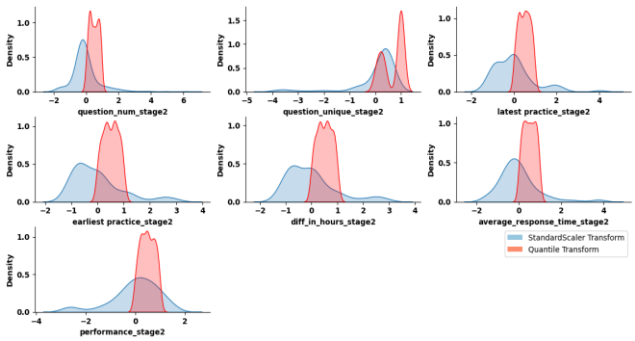
(b) Student distribution based on race



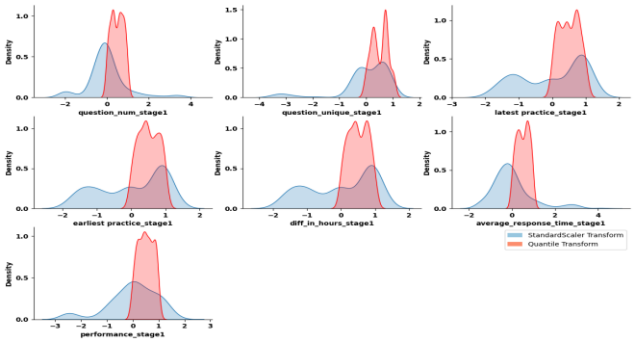
(c) Student distribution based on lunch status

Figure 2. Student distributions in each assignment based on demographic information.

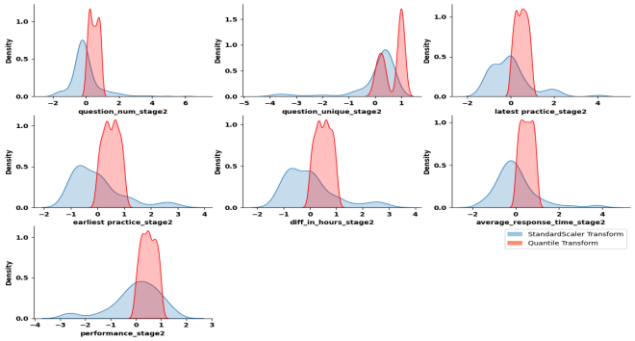
We further conducted a two-step feature transformation for the numerical variables used to indicate the student learning and practice patterns in each assignment, as shown in Figure 3. We first conducted StandardScaler transformation (Raju et al., 2020) to standardize these features into a distribution with zero mean and unit variance (shown as the light blue areas in the figure). Some of them are highly skewed. As such, we further conducted quantile transformation (Peng et al., 2007) to provide a uniform distribution (as shown in the light red areas). After the quantile transformation, most features fall into the range of [0,1], which eliminates the potential influence of scale on predictions.



(a) Assignment 1



(b) Assignment 2



(c) Assignment 3

Figure 3. Distributions of features indicating student learning behaviour in each assignment.

4.2. Model Training and Configuration

To ensure the performance and generalizability of our trained models, we followed a systematic approach that included data processing, model training, hyperparameter fine-tuning, and performance evaluation. At the data processing stage, we employed one-hot encoding and StandardScaler to transform categorical and numerical features. The transformed dataset was then split into training (75%) and testing (25%) sets, stratified by the target variable and sensitive features to ensure balanced representation. Below, we outline the model training, hyperparameter fine-tuning, and performance measurement processes for each model in detail:

1. For training **LGBMClassifier**, we employed a grid search to optimize the parameters — such as number of leaves ([10, 31]), regularization factors (lambda_l1 [0, 1], lambda_l2 [0, 1]), and min_data_in_leaf values ([30, 50, 100, 300, 400]) — using balanced accuracy and 5-fold cross validation.
2. For training **CatBoostClassifier**, the grid search approach was used to optimize the parameters, such as learning rate ([0.03, 0.1, 0.3]), depth ([4, 6, 10]), and L2 leaf regularization ([1, 3, 5, 7, 9]). The model was further trained with cross-entropy loss over 20 iterations evaluated with balanced accuracy.
3. For training **StackingClassifier**, we trained four base models (i.e., Logistic Regression, Decision Tree, Random Forest, and SVC), fine-tuned them with grid search, and combined them into a final estimator of the XGBoost model with five-fold cross-validation.
4. For training **AdaBoostClassifier**, we used the decision tree as the base estimator with 50 estimators and a learning rate of 0.9. Hyperparameters were fine-tuned with grid search to ensure model generalization and to prevent overfitting. The balanced accuracy metric was further employed to evaluate the model performance handling class imbalance issues in the dataset.

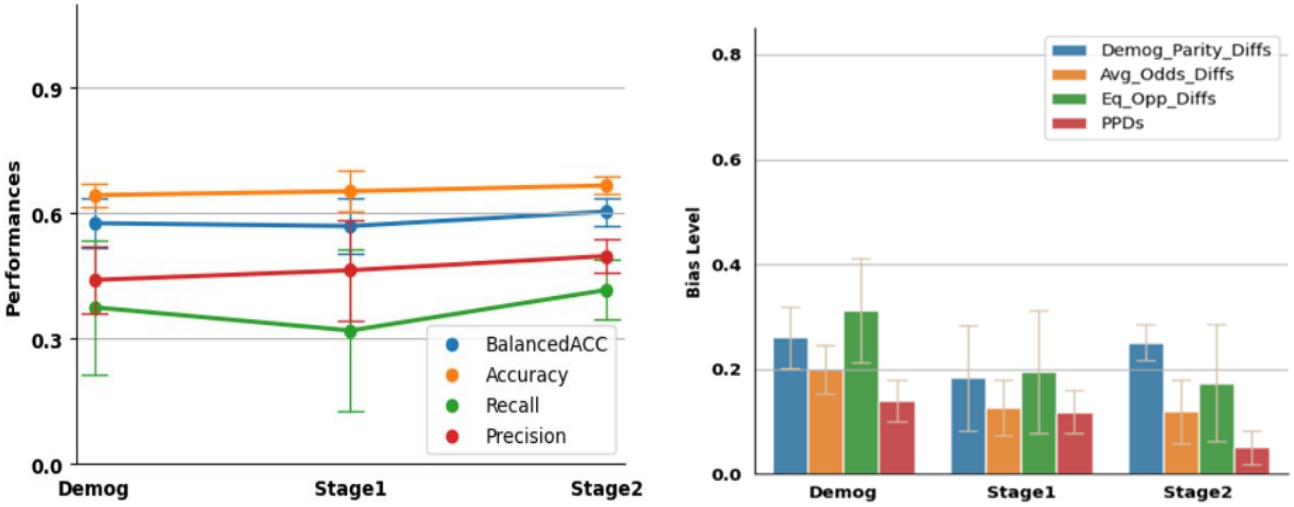
4.3. Model Performance and Bias Level on the Imbalanced Dataset

On the imbalanced dataset based on race, after going through the pre-processing steps, we trained and fine-tuned the four ML models at each stage. Figure 4(a) and Table 1 show the average performance of these models on each metric at each stage.

Table 1. Average Performance of 4 ML Models at Each Stage

Stages	Balanced Accuracy	Accuracy	Recall	Precision
Demog	0.576±0.052	0.644±0.024	0.375±0.052	0.440±0.068
Stage 1	0.569±0.056	0.653±0.042	0.319±0.168	0.464±0.056
Stage 2	0.604±0.029	0.667±0.019	0.417±0.062	0.498±0.034

Comparing the Demog Stage with Stage 2, which contains more learning activity data, there were no significant differences in predictive performance. The comparisons of Stage 2 vs Demog Stage are as follows: Balanced Accuracy (p-value = 0.460, Cohen's d = 0.570), Accuracy (p-value = 0.238, Cohen's d = 0.934), Recall (p-value = 0.658, Cohen's d = 0.338), and Precision (p-value = 0.258, Cohen's d = 0.920). The effect size on most of the metrics exceeds the moderate levels (0.50) suggested in Funder and Ozer (2019), which indicates that with more data, predictive performance would have great potential to improve in practical settings.



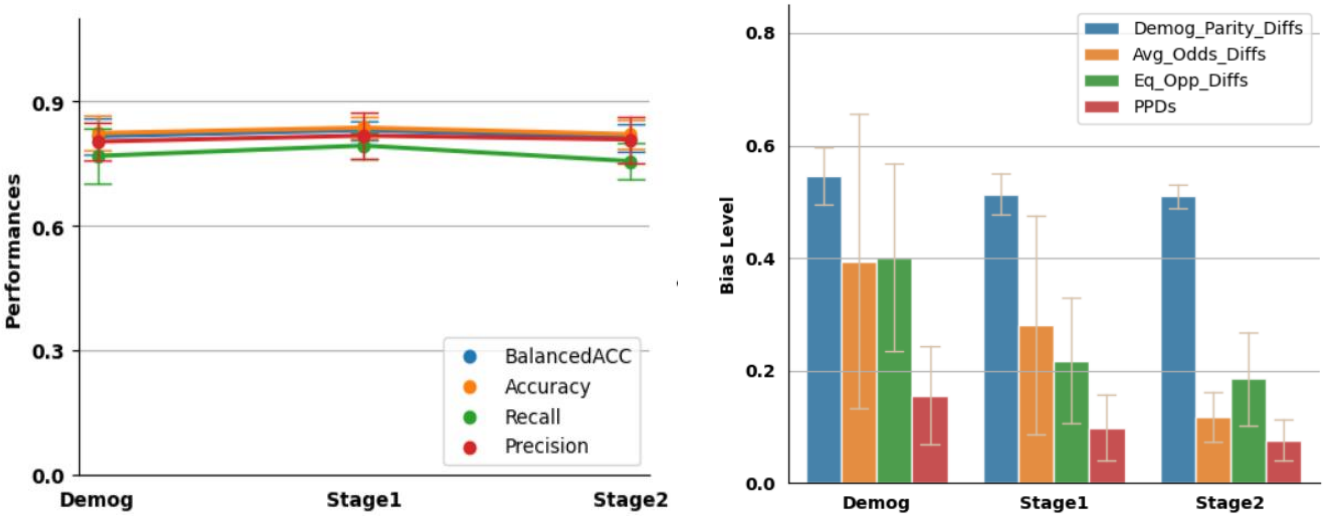
(a) Average performance of models across stages (b) Model fairness level at different stages

Figure 4. Models' average performance and fairness levels at the different stages on the imbalanced dataset.

Next, the bias levels of these different models were examined at each stage using the four fairness evaluation metrics mentioned above: 1) Demog_Parity_Diffs, 2) Avg_Odds_Diffs, 3) Eq_Opp_Diffs, and 4) PPDs, as shown in Figure 4(b). The error bars indicate the variability of model fairness levels under each metric within each stage. Notably, the models at Stage 1 have the least bias, on average, compared to the Demog Stage and Stage 2, with all values falling below the 0.20 threshold. Although the errors bars in Stage 1 seem a bit wider on some metrics compared to the other two stages, this variability reflects the inherent uncertainty in fairness when predictions are made at the early stage with limited data. In addition, the models at Stage 1 also have a comparable performance with that of Stage 2 and the Demog Stage, which means that early predictions can achieve similar levels of accuracy to those using more data at a later stage. This is evident when combining insights from Figure 4(a) (predictive performance) and Figure 4(b) (fairness). As well as timeliness, Stage 1 seems to be an optimal point for identifying at-risk students.

4.4. Model Performance and Bias Level on the Balanced Dataset

The above analytical results from Figure 4(a) suggests that increasing the amount of data has the potential to improve the average model performance. The dataset balancing technique is one of the commonly used strategies to improve model predictive performance since it helps equalize the distributions of underrepresented groups within the original dataset by increasing their corresponding sample sizes. In this section, we first employed the RandomOverSampler technique (Mesquita et al., 2021) to balance the dataset based on race. Then, we applied pre-processing approaches to transform the balanced dataset into the appropriate format for training the four ML models at each stage. As shown in Figure 5(a), significant improvement in performance was found on the models trained on the balanced dataset compared with that of the imbalanced dataset in Figure 4(a) at each of the corresponding stages. This partially demonstrated the effectiveness of balancing the dataset to improve model performance.



(a) Models' average performance across stages

(b) The fairness levels of models at different stages

Figure 5. Models' average performance and bias levels at different stages on balanced dataset.

Furthermore, we evaluated the improvements in performance of the models trained on the balanced and imbalanced datasets on each metric at each stage. Positive t-statistic values indicate that models trained on the balanced dataset have a higher performance compared with those trained on the imbalanced dataset. Detailed statistics on each metric are shown in Table 2. Apart from the significant differences in performance under most metrics, we also found a large Cohen's *d* for each metric at each stage (i.e., larger than 0.80), indicating strong effect sizes. This further suggests that performance improvements achieved from balancing the dataset were not only statistically significant but also meaningful, indicating the practical impact of applying data balancing techniques.

Table 2. Differences in Models Trained on Balanced and Imbalanced Datasets Regarding Performance on 4 Metrics across 3 Stages

Stage	Metrics	p-value	t-statistics	Cohen's *d*
Demog	Balanced Accuracy	0.000	6.388	**4.517**
	Accuracy	0.000	7.209	**5.098**
	Recall	0.010	4.563	**3.226**
	Precision	0.001	7.921	**5.601**
Stage 1	Balanced Accuracy	0.002	7.569	**5.352**
	Accuracy	0.002	6.640	**4.695**
	Recall	0.015	4.816	**3.405**
	Precision	0.005	5.313	**3.757**
Stage 2	Balanced Accuracy	0.000	8.760	**6.194**
	Accuracy	0.001	7.627	**6.194**
	Recall	0.001	8.137	**5.754**
	Precision	0.000	9.104	**6.438**

For bias level, however, compared with models trained on the imbalanced dataset shown in Figure 4(b), we found more bias after training based on the balanced dataset shown in Figure 5(b). The differences in fairness level across models trained on the imbalanced and balanced datasets under each metric at each stage are presented in Table 3.

Table 3. Differences in Fairness Level between Models Trained on Balanced and Imbalanced Datasets

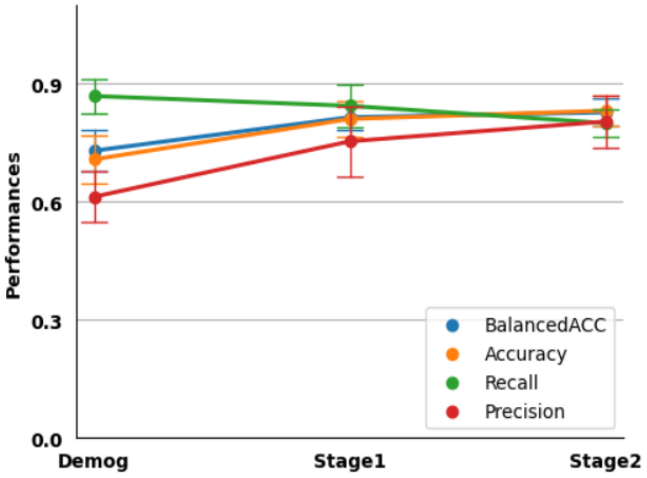
Stage	Metrics	p-value	t-statistics	Cohen's *d*
Demog	Demog_Parity_Diffs	**0.000**	7.365	**5.208**
	Avg_Odds_Diffs	0.232	1.471	**1.040**
	Eq_Opp_Diffs	0.405	0.912	0.645
	PPDs	0.747	0.344	0.243
Stage 1	Demog_Parity_Diffs	**0.004**	**6.176**	**4.367**
	Avg_Odds_Diffs	0.211	1.536	**1.086**
	Eq_Opp_Diffs	0.789	0.280	0.198
	PPDs	0.318	1.090	**0.771**
Stage 2	Demog_Parity_Diffs	**0.000**	12.964	**9.167**
	Avg_Odds_Diffs	0.967	–0.043	0.030
	Eq_Opp_Diffs	0.864	0.179	0.126
	PPDs	0.318	1.090	**0.771**

Table 3 shows a significant increase in bias under the Demog_Parity_Diffs (p-value < 0.005) metric for all stages. All other metrics — except Stage 2 where a slight decrease in bias was found (t-statistics = –0.043, Cohen's d = 0.030) on Avg_Odds_Diffs — showed different levels of increase in bias across all stages. Although no significant increase was found, notable medium to large effect sizes were found on Avg_Odds_Diffs at the Demog Stage (Cohen's d = 1.040) and Stage 1 (Cohen's d = 1.086), and PPDs at Stage 2 (Cohen's d = 0.771). This suggests that the balanced dataset technique alone may be insufficient to effectively mitigate bias in the context of identifying at-risk students at an early stage, indicating that the dataset is not the only source for generating bias. To ensure equitable outcomes, it is necessary to incorporate more appropriate bias mitigation strategies to address bias beyond balancing the datasets.

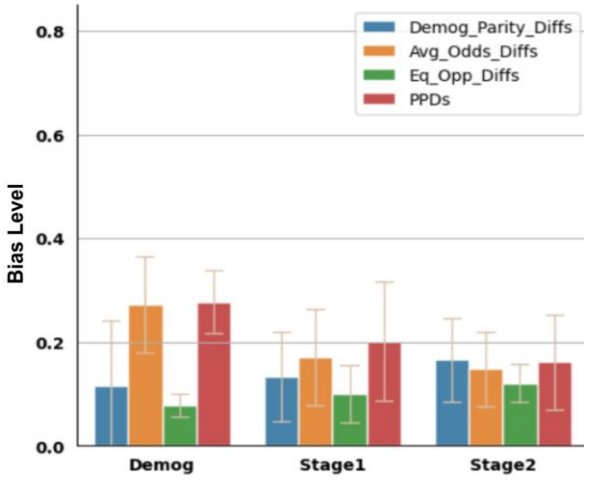
4.5. Mitigated Model Performance and Bias Level on the Balanced Dataset

From the above analysis, we found a significant increase in predictive performance, and various levels of increase in bias when training the models with the balanced dataset. In this section, we further apply post-processing strategies to mitigate bias and examine the changes in model performance and fairness level. Specifically, the ThresholdOptimizer was applied to find the optimal threshold to satisfy both performance objectives and fairness constraints (Weerts et al., 2023). Figure 6(a) shows the average performance of the four ML models after applying post-processing strategies under each metric at each stage. Compared with Figure 5(a), Figure 6(a) has a slight drop in performance on some metrics (e.g., Precision, Accuracy) at the Demog Stage and Stage 1. This suggests a trade-off between predictive performance and fairness level — improving predictive fairness across groups might compromise some prediction accuracy, which is particularly relevant when the model predictions have a high reliance on demographic information. For example, comparing Stages 1 and 2 in Figure 6(a), there have been larger variances in the Demog stage where the predictions were based on demographic information. While comparing model

performance at Stage 1 with that of Stage 2, no significant differences were found in terms of either variance or performance level. This means that with more learning data available at the later stages, model performance becomes stable, further indicating the feasibility of identifying at-risk students at Stage 1.



(a) Models' average performance across stages



(b) Fairness level of models at the different stages

Figure 6. Mitigated models' average performance and fairness levels at the different stages on the balanced dataset.

Accordingly, the fairness levels of the mitigated models were analyzed at each stage, as shown in Figure 6(b). Compared with Figure 5(b), there is an obvious decrease in bias at each stage for Demog_Parity_Diffs, Avg_Odds_Diffs, and Eq_Opp_Diffs, with a slight increase for PPDs. Although the error bars in Figure 6(b) exhibit higher variability on Demog_Parity_Diffs and PPDs compared to those in Figure 5(b), the significant decrease in the average bias level for Demog_Parity_Diffs and others still highlights the effectiveness of this post-processing mitigation strategy. In addition, the lower variability of bias level for Avg_Odds_Diffs and Eq_Opp_Diffs further indicates the consistency and stability of bias levels across the three stages after the post-processing strategy.

Table 4. Differences in Fairness Level Between the Unmitigated and Mitigated Models Using the Balanced Dataset

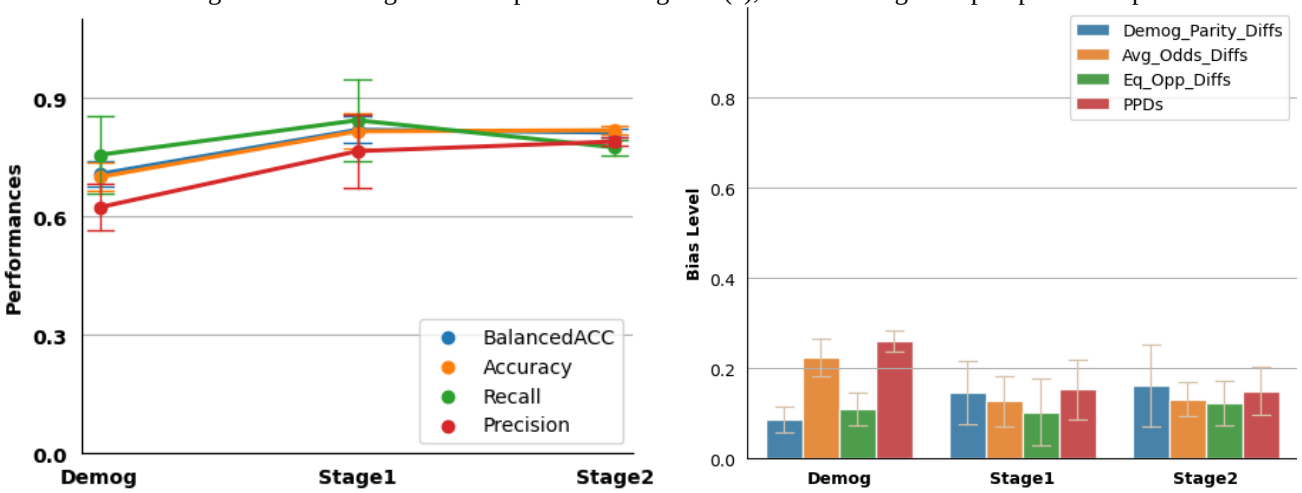
Stage	Metrics	p-value	t-statistics	Cohen's *d*
Demog	Demog_Parity_Diffs	**0.003**	**–6.378**	4.510
	Avg_Odds_Diffs	0.433	**–0.876**	0.620
	Eq_Opp_Diffs	0.029	**–3.845**	2.719
	PPDs	0.068	2.275	1.609
Stage 1	Demog_Parity_Diffs	**0.001**	**–8.136**	5.753
	Avg_Odds_Diffs	0.355	**–1.037**	0.733
	Eq_Opp_Diffs	0.125	**–1.895**	1.340
	PPDs	0.175	1.611	1.139
Stage 2	Demog_Parity_Diffs	**0.002**	**–8.258**	5.839
	Avg_Odds_Diffs	0.506	**0.716**	0.506
	Eq_Opp_Diffs	0.224	**–1.424**	1.007
	PPDs	0.162	1.725	1.220

Table 4 provides a fine-grained comparison regarding the changes in bias level between the mitigated and unmitigated models. As discussed above, a significant decrease in bias level on Demog_Parity_Diffs has been observed at all stages (p-value < 0.005), with very large effect size at the Demog Stage (Cohen's $d = 4.510$), Stage 1 (Cohen's $d = 5.753$), and Stage 2 (Cohen's $d = 5.839$). For Avg_Odds_Diffs, there is no significant decrease in bias level at all stages (p-value > 0.3). However, the statistical differences and the moderate effect size at each stage suggest noticeable though less consistent fairness grains, particularly in the Demog stage (t-statistics = –0.876, Cohen's $d = 0.620$), Stage 1 (t-statistics = –1.037, Cohen's $d = –0.733$), and Stage 2 (t-statistics = 0.716, Cohen's $d = 0.506$). Similarly for Eq_Opp_Diffs, although there is no significant decrease in bias level, we found a large effect size (> 0.80) at each stage, specifically at the Demog Stage (t-statistics = –3.845, Cohen's $d = 2.719$), Stage 1 (t-statistics = –1.895, Cohen's $d = 1.340$), and Stage 2 (t-statistics = –1.424, Cohen's $d = 1.007$). In contrast, we observed increased bias levels in PPDs across stages with positive t-statistics and a large effect size, which further suggests independence across different fairness evaluation approaches.

Overall, combining Figures 6(b) and 5(b) suggests that post-processing strategies can reduce prediction biases under certain metrics and simultaneously maintain predictive accuracy reasonably well. Looking at Eq_Odds_Diffs at Stage 1, which we emphasize most in our study, it is clear that fairness also reached an acceptable level with a bias around 0.10. Combining these analytical insights, we suggest that Stage 1 is appropriate for identifying at-risk students because it offers more time for educators to provide support.

4.6. Mitigated Model Performance and Bias Level on the Balanced Dataset with Sensitive Features (i.e., Race) Removed

In response to concerns that the inclusion of sensitive features might cause bias issues, in this section we examine changes in the predictive performance and fairness levels of the mitigated models after removing sensitive features (i.e., race) from the balanced dataset. In removing race, we used gender, eligibility for free/reduced-price lunch, math classes taken overall, and self-predicted grades to train the models at the Demog Stage. Figure 7(a) presents the average performance of models under each metric at each stage after removing race. Compared with Figure 6(a), there is a slight drop in predictive performance.



(a) Models' average performance across stages (b) The fairness levels of models at different stages

Figure 7. Mitigated models' average performance and fairness levels at the different stages on the balanced dataset after sensitive feature (race) removed.

Table 5. Differences in Performance of Mitigated Models Trained on the Balanced Dataset After and Before Removing Race on the 4 Metrics across the 3 Stages

Stage	Metrics	p-value	t-statistics	Cohen's *d*
Demog	Balanced Accuracy	0.497	0.730	0.516
	Accuracy	0.833	0.223	0.158
	Recall	0.103	2.092	1.489
	Precision	0.817	–0.242	0.171
Stage 1	Balanced Accuracy	**0.851**	**–0.196**	**0.139**
	Accuracy	**0.871**	**–0.169**	**0.120**
	Recall	**1.000**	**0.000**	**0.000**
	Precision	**0.860**	**–0.184**	**0.130**
Stage 2	Balanced Accuracy	0.466	0.814	0.576
	Accuracy	0.541	0.678	0.480
	Recall	0.277	1.225	0.866
	Precision	0.704	0.416	0.294

Table 5 presents detailed statistics regarding performance differences with post-processing mitigation strategies both before and after removing the sensitive feature of race. Given that Stage 1 has been identified as the optimal point for identifying at-risk students, we will mainly interpret the results from this stage. As seen in Table 5, it is clear that the average model performance had a slight drop on most of the metrics, particularly Balanced Accuracy (t-statistics = –0.196, Cohen's d = 0.139), Accuracy (t-statistics = –0.169, Cohen's d = 0.120), Recall (t-statistics = 0.000, Cohen's d = 0.000), and Precision (t-statistics = –0.184, Cohen's d = 0.130). The relatively small effect size indicates that removing race had little impact on predictive performance at Stage 1. Overall, this result indicates that race partially contributes to predictive power at Stage 1,

but its exclusion did not cause a significant drop in model performance. To a certain extent, this finding can also be explained by the availability of learning activity data, which likely mitigates reliance on sensitive features.

Table 6. Differences in Model Fairness Levels Trained on the Balanced Dataset After and Before Removing Race

Stage	Metrics	p-value	t-statistics	Cohen's *d*
Demog	Demog_Parity_Diffs	0.797	0.273	0.193
	Avg_Odds_Diffs	0.082	–2.171	1.535
	Eq_Opp_Diffs	0.061	2.355	1.665
	PPDs	0.001	–6.996	4.947
Stage 1	Demog_Parity_Diffs	**0.110**	**–1.915**	**1.354**
	Avg_Odds_Diffs	**0.460**	**–0.797**	**0.563**
	Eq_Opp_Diffs	**0.330**	**1.068**	**0.756**
	PPDs	**0.056**	**–2.629**	**1.859**
Stage 2	Demog_Parity_Diffs	0.363	–0.998	0.706
	Avg_Odds_Diffs	0.892	–0.144	0.102
	Eq_Opp_Diffs	0.185	1.502	1.062
	PPDs	0.188	–1.487	1.062

Figure 7(b) shows model fairness levels after removing race. Compared with Figure 6(b), no significant differences are observed on bias level at Stage 1. This is further indicated by the statistics in Table 6 (p-value > 0.05) with a slight decrease in bias after removing race for Demog_Parity_Diffs (t-statistics = –1.915, Cohen's *d* = 1.354), Avg_Odds_Diffs (t-statistics = –0.797, Cohen's *d* = 0.563), and PPDs (t-statistics = –2.629, Cohen's *d* = 1.859), with a medium to large effect size. This finding suggests a meaningful improvement in the fairness level by removing race. For Eq_Opp_Diffs, however, we found a slight increase in bias (t-statistics = 1.068, Cohen's *d* = 0.756) with a medium effect size. This implies that removing race might affect predictive precision, which requires close investigation before applying predictions in real settings.

5. Discussion

5.1. The Balance of Timing, Predictive Accuracy, and Fairness in Identifying At-Risk Students

Identifying at-risk students at an early stage enables educators to provide timely interventions before students fall too far behind academically, allowing for designing more effective and personalized support. However, predictions made at early stages are generally based on limited training data and learning context, which risks mislabeling at-risk students and increases the potential of providing biased learning recommendations, especially for those in underrepresented groups. This highlights that identifying at-risk students at early stages involves a careful balance with timing, fairness, and accuracy. Our analytical results indicate that Stage 1 (after the first assignment) is the optimal point for identifying at-risk students. This is particularly indicated by the fact that in our study, there were no significant changes in predictive performance after Stage 1 (see Figures 4(a), 5(a), 6(a), 7(a), and the corresponding results in Appendix A for gender and free/reduced lunch status). As well, Stage 1 showed acceptable fairness levels, with most values falling below the 0.20 threshold.

Findings from this study have several important practical implications. First, they demonstrate the possibility of providing fair, accurate, early stage identification for at-risk students. The implementation of these algorithms enables educators to provide timely, personalized support while considering fairness across different demographic groups. Second, the findings also support the practice of incorporating fairness evaluation mechanisms into the process of identifying at-risk students — a practice that traditionally focuses more on early intervention and accuracy with less practical attention to fairness (Rasooli et al., 2021). By showing the effectiveness of incorporating fairness measurements, this study improves educator awareness of ethical issues while providing learning interventions for at-risk students. Moreover, this study highlights the importance of providing correct, context-aware interpretations of prediction results. Identifying at-risk students is determined with limited learning data, which requires educators to interpret the predictions with caution and contextual understanding. It is important to recognize that these predictions can be impacted by incomplete learning data and systematic bias or stereotypes associated with demographic background (Nezami et al., 2024).

5.2. Demographic Background Plays a Significant Role in Student Learning Performance

We can see how demographic background impacts model performance at the Demog stage (see Figures 7(a) and Figures A3 and A7 in Appendix A). Strong predictive performance in the Demog stage suggests that demographic information accounts for a considerable portion of variation in predictive outcomes, reflecting the persistent associations among demographic backgrounds with learning AP Statistics. Furthermore, this insight is corroborated by Becker and Luthar (2002) who demonstrated that student performance disparities in AP Statistics have a strong association with family characteristics as well

as school factors. Multiple studies have investigated the causes of these disparities, highlighting that, on the one hand, students from marginalized groups always have limited access to learning resources (Milner, 2013). On the other hand, current instructional strategies are also shaped by stereotypes about students' demographic backgrounds, which often overlook each student's personalized learning needs and reinforce existing assumptions rather than promoting equitable, personalized learning support (Hajisoteriou et al., 2019).

This finding partially reflects broader issues in current educational practices that result in students from underrepresented groups being more likely to have difficulties in learning AP Statistics. This is because students from rural communities or economically disadvantaged households often have limited opportunities to engage with relevant resources to gain advanced mathematics knowledge prior to high school. This is particularly highlighted by findings in Lee and LaHaye (2024) based on their analysis of students' math learning status in Alabama, Arkansas, and Mississippi. Moreover, the high predictive accuracy of demographic information in supporting early identification of at-risk students further suggests a need for a more customized, context-sensitive definition of at-risk students. Current definitions of at-risk students are often based on the metrics related to general academic performance, which has a significant relationship with demographic characteristics (Matz et al., 2023). Exploring definitions focusing more on the differences between each individual's learning needs can help reduce the impact of demographic background while promoting personalized, equity-driven educational practices.

5.3. Data Balancing Techniques are Not Always Effective for Mitigating Bias in Predictive Models

In educational settings, there is a higher chance of imbalanced datasets than in other fields, particularly when separating students based on their demographic backgrounds such as gender, race, and so on. For example, in most universities and colleges, male students tend to dominate the enrollments in Engineering, while female students tend to dominate Social Sciences and Arts (Wrigley-Asante et al., 2023). In order to equalize the samples from different groups, data balancing techniques have been identified as an effective approach to improve the predictive performance and fairness levels of ML models (Sha et al., 2021). However, the analytical results from our study did not indicate the effectiveness of this approach in mitigating predictive bias. In particular, our results indicate significant improvement in predictive performance after balancing the dataset; at the same time, results also exhibit a decrease in fairness level for certain metrics, such as Demog_Parity_Diffs (see Figures 4 and 5). This is more obvious when comparing fairness levels before and after balancing the dataset based on free lunch status, shown in Figures A6(a) and A8(a) in Appendix A.

On the other hand, these results also shed light on potential disadvantages associated with dataset balancing techniques. Balanced datasets might obscure the true nature of the underlying distribution, further causing generalization problems when applying the models in real settings. Especially in educational settings, balancing datasets without further investigation can lead to the misallocation of learning resources, as well as misguided pedagogical interventions. This is because insights derived from models trained on balanced datasets might fail to capture the real distributions in real-world settings, potentially overlooking the learning patterns of students in underrepresented groups. Furthermore, the additional fairness issues introduced by the data balancing techniques observed in our study were also discussed in Wang et al. (2019) which explicitly investigated how balancing a dataset based on a specific feature (e.g., race) can inadvertently affect the distribution of the dataset based on other correlated features, like gender. Specifically, if balancing by race results in a higher proportion of female students within a particular racial group, the model will be more likely to label female students as the corresponding class, which could generate new bias issues.

5.4. Excluding Sensitive Variables Does Not Lead to Significant Changes in Fairness

Incorporating sensitive features or not in training ML models has garnered increasing discussion. Baker et al. (2023) suggests that demographic variables should be used in the validation process rather than training stages to minimize the potential bias inherent in such data. In other studies, such as Yu et al. (2020) and Kleinberg et al. (2016), the authors argue that using demographic data as predictors provides a more comprehensive view of the student learning process, which can improve model performance. However, our study indicates that excluding sensitive features did not improve the fairness level nor did it decrease the model's performance. In our case, the effects of demographic background were moderated by incorporating student learning activity data in the later stages. Moreover, the focus of fairness evaluations can be different in different research contexts, which further requires specific analysis regarding whether excluding a certain feature will be helpful or not for the specific tasks.

6. Conclusion

In this study, we investigated trade-offs among timing, accuracy, and fairness in the practice of early identification of at-risk students, specifically exploring the constraints and measures of algorithmic fairness in the context of high school learners in AP Statistics courses. In particular, we explored and evaluated both the performance and fairness levels of four machine learning models in identifying at-risk students at three different stages in the student learning process: Demog stage (i.e., only using student demographic information), Stage 1 (i.e., using student demographic information plus the learning activity data

up to the first assignment), and Stage 2 (i.e., using student demographic information plus the learning activity data up to the second assignment). Furthermore, at each stage, we examined the changes between the models' predictive performance and fairness levels for students grouped by race, gender, and free lunch status. Finally, we also examined the effects of certain techniques, including balancing datasets and excluding sensitive features in training to assess their impact on model performance and fairness levels.

Overall, there are three key takeaways from this study. First, demographic information was identified as having a significant relationship with student performance in AP Statistics at early stages, which can be seen in the small changes in model performance at the Demog stage compared to later stages. Second, it is necessary to examine disparities in the dataset before balancing it, especially for identifying at-risk students. As our results show, the influence of demographic background on a student's AP Statistics performance might be mitigated by student engagement and learning activities in the subsequent stages. To this end, employing balancing strategies to equalize the number of at-risk students across different demographic backgrounds might inadvertently introduce additional issues in predictions. Third, excluding a specific sensitive feature might not yield significant improvement in predictive performance and fairness. The effectiveness of including or excluding a specific sensitive feature in building an ML model and evaluating its impact on fairness largely depends on the specific application context and the nature of the data involved.

7. Limitations and Future Directions

Fairness issues related to AI in education have been longstanding concerns. In this study, various fairness evaluation strategies and analytical approaches have been employed to investigate the critical interplay among timing, accuracy, and fairness in identifying at-risk students. Although the overall analytical approaches and proposed framework have indicated high scalability, the relatively small dataset in our study does not support the practical application of our analytical results to broader contexts. Another limitation is related to statistical testing about the effectiveness of bias detection and mitigation approaches. As indicated in Kobayashi and Nakao (2020), there is no "one-size-fits-all" approach to mitigate algorithmic bias since such biases can differ depending on the specific task, data, and context where an algorithm is applied. Future studies will focus on expanding datasets to include more diverse samples in different learning scenarios, on conducting more robust and comprehensive fairness investigations, and on investigating whether model fine-tuning can improve fairness in predictions. Moreover, more advanced and context-aware statistical testing approaches will be implemented for conducting more rigorous evaluations regarding the effectiveness of different bias mitigation strategies in different contexts.

Declaration of Conflict of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, *9*, 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446

An, T.-K., & Kim, M.-H. (2010). A new diverse AdaBoost classifier. In Q. Li, M. Chen, H. Deng, & Y. Gao (Eds.), *2010 international conference on artificial intelligence and computational intelligence* (Vol. 1, pp. 359–363). IEEE. https://doi.org/10.1109/AICI.2010.82

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ayubkhan, S. A. H., Yap, W.-S., Morris, E., & Rawthar, M. B. K. (2023). A practical intrusion detection system based on denoising autoencoder and LightGBM classifier with improved detection performance. *Journal of Ambient Intelligence and Humanized Computing*, *14*(6), 7427–7452. https://doi.org/10.1007/s12652-022-04449-w

Baker, R. S., Esbenshade, L., Vitale, J., & Karumbaiah, S. (2023). Using demographic data as predictor variables: A questionable choice. *Journal of Educational Data Mining*, *15*(2), 22–54. https://doi.org/10.5281/zenodo.7702628

Baker, R. S., & Hawn, A. (2021). *Algorithmic bias in education*. EdArXiv. https://doi.org/10.35542/osf.io/pbmvz

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunity*. MIT Press. https://fairmlbook.org/

Becker, B. E., & Luthar, S. S. (2002). Social-emotional factors affecting achievement outcomes among disadvantaged students: Closing the achievement gap. *Educational Psychologist*, *37*(4), 197–214. https://doi.org/10.1207/S15326985EP3704_1

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. arXiv. http://arxiv.org/abs/1810.01943

Chen, R. J., Chen, T. Y., Lipkova, J., Wang, J. J., Williamson, D. F. K., Lu, M. Y., Sahai, S., & Mahmood, F. (2022). *Algorithm fairness in AI for medicine and healthcare*. arXiv. https://arxiv.org/abs/2110.00603v2

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

Cohausz, L., Kappenberger, J., & Stuckenschmidt, H. (2024). What fairness metrics can really tell you: A case study in the educational domain. In S. Joksimovic & A. Zamecnik (Eds.), *LAK '24: Proceedings of the 14th learning analytics and knowledge conference* (pp. 792–799). ACM Press. https://doi.org/10.1145/3636555.3636873

Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L., & Duy Le, T. (2022). How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology*, *53*(4), 822–843. https://doi.org/10.1111/bjet.13217

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). *Certifying and removing disparate impact*. arXiv. http://arxiv.org/abs/1412.3756

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In D. Azcona & R. Chung (Eds.), *LAK '19: Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 225–234). ACM Press. https://doi.org/10.1145/3303772.3303791

Gu, J., McFerran, B., Aquino, K., & Kim, T. G. (2014). What makes affirmative action-based hiring decisions seem (un)fair? A test of an ideological explanation for fairness judgments. *Journal of Organizational Behavior*, *35*(5), 722–745. https://doi.org/10.1002/job.1927

Haeri, M. A., & Zweig, K. A. (2020). The crucial role of sensitive attributes in fair classification. In H. K. Singh (Ed.), *2020 IEEE symposium series on computational intelligence (SSCI)*, 2993–3002. IEEE. https://doi.org/10.1109/SSCI47803.2020.9308585

Hajisoteriou, C., Maniatis, P., & Angelides, P. (2019). Teacher professional development for improving the intercultural school: An example of a participatory course on stereotypes. *Education Inquiry*, *10*(2), 166–188. https://doi.org/10.1080/20004508.2018.1514908

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning.* arXiv. http://arxiv.org/abs/1610.02413

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, *57*(4), 542–570. https://doi.org/10.1111/ejed.12533

Holstein, K., Aleven, V., & Rummel, N. (2020, June). A conceptual framework for human–AI hybrid adaptivity in education. *Artificial intelligence in education: 21st international conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I* (pp. 240–254). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-52237-7

Hu, Q., & Rangwala, H. (2020). *Towards fair educational data mining: A case study on detecting at-risk students*. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *Proceedings of the 13th international conference on educational data mining (EDM 2020)* (pp. 431–437). International Educational Data Mining Society. https://eric.ed.gov/?id=ED608050

Jiang, W., & Pardos, Z. A. (2021). Towards equity and algorithmic fairness in student grade prediction. In M. Fourcade, B. Kuipers, S. Lazar, & D. Mulligan (Eds.), *AIES '21: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 608–617). ACM Press. https://doi.org/10.1145/3461702.3462623

Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D., & Khanna, A. (2021). CatBoost: An ensemble machine learning model for prediction and classification of student academic performance. *Advances in Data Science and Adaptive Analysis*, *13*(03n04), Article 2141002. https://doi.org/10.1142/S2424922X21410023

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, *29*, 417–442. https://doi.org/10.1146/annurev.soc.29.010202.100019

Kizilcec, R. F., & Lee, H. (2021). *Algorithmic fairness in education*. arXiv. https://doi.org/10.48550/arXiv.2007.05443

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv. https://doi.org/10.48550/arXiv.1609.05807

Kobayashi, K., & Nakao, Y. (2020). *One-vs.-one mitigation of intersectional bias: A general method to extend fairness-aware binary classification*. arXiv. https://arxiv.org/abs/2010.13494v1

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Książek, W., Hammad, M., Pławiak, P., Acharya, U. R., & Tadeusiewicz, R. (2020). Development of novel ensemble model using stacking learning and evolutionary computation techniques for automated hepatocellular carcinoma detection. *Biocybernetics and Biomedical Engineering*, *40*(4), 1512–1524. https://doi.org/10.1016/j.bbe.2020.08.007

Lalor, J. P., Abbasi, A., Oketch, K., Yang, Y., & Forsgren, N. (2024). Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, *42*(4), Article 99. https://doi.org/10.1145/3641276

Lee, J., & LaHaye, C. (2024). Unequal access to the mathematics course ladder for rural students in the southern states. *Journal of Advanced Academics*, *35*(4), 671–697. https://doi.org/10.1177/1932202X241241355

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *FAT* '20: Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 122–130). ACM Press. https://doi.org/10.1145/3351095.3372867

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1), Article 5705. https://doi.org/10.1038/s41598-023-32484-w

Mesquita, F., Maurício, J., & Marques, G. (2021). Oversampling techniques for diabetes classification: A comparative study. In *2021 international conference on e-health and bioengineering (EHB)* (pp. 1–6). IEEE. https://doi.org/10.1109/EHB52898.2021.9657542

Milner, H. R., IV. (2013). Analyzing poverty, learning, and teaching through a critical race theory lens. *Review of Research in Education*, *37*(1), 1–53. https://doi.org/10.3102/0091732X12459720

Nezami, N., Haghighat, P., Gándara, D., & Anahideh, H. (2024). Assessing disparities in predictive modeling outcomes for college student success: The impact of imputation techniques on model performance and fairness. *Education Sciences*, *14*(2), Article 136. https://doi.org/10.3390/educsci14020136

Ober, T. M., Cheng, Y., Carter, M. F., & Liu, C. (2023). Disruptiveness of COVID-19: Differences in course engagement, self-appraisal, and learning. *Aera Open*, *9*. https://doi.org/10.1177/23328584231177967

Pei, B., & Xing, W. (2022). An interpretable pipeline for identifying at-risk students. *Journal of Educational Computing Research*, *60*(2), 380–405. https://doi.org/10.1177/07356331211038168

Peng, B., Yu, R. K., DeHoff, K. L., & Amos, C. I. (2007). Normalizing a large number of quantitative traits using empirical normal quantile transformation. *BMC Proceedings*, *1*(Suppl. 1), Article S156. https://doi.org/10.1186/1753-6561-1-S1-S156

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the influence of normalization/transformation process on the accuracy of supervised classification. In *2020 third international conference on smart systems and inventive technology (ICSSIT)* (pp. 729–735). IEEE. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Rasooli, A., Razmjoee, M., Cumming, J., Dickson, E., & Webster, A. (2021). Conceptualising a fairness framework for assessment adjusted practices for students with disability: An empirical study. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 301–321. https://doi.org/10.1080/0969594X.2021.1932736

Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gašević, D., & Chen, G. (2021). Assessing algorithmic fairness in automatic classifiers of educational forum Posts. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial intelligence in education: 22nd international conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, proceedings, part I* (pp. 381–394). Springer Cham. https://doi.org/10.1007/978-3-030-78292-4_31

Small, E., Sokol, K., Manning, D., Salim, F. D., & Chan, J. (2024). Equalised odds is not equal individual odds: Post-processing for group and individual fairness. In S. Rezapour & A. Asudeh (Eds.), *FAccT '24: Proceedings of the 2024 ACM conference on fairness, accountability, and transparency* (pp. 1559–1578). ACM Press. https://doi.org/10.1145/3630106.3658989

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *11*, Article 24039. https://doi.org/10.1038/s41598-021-03430-5

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In E. Mortensen & J. Lim (Eds.), *2019 IEEE/CVF international conference on computer vision (ICCV)* (pp. 5309–5318). IEEE. https://doi.org/10.1109/ICCV.2019.00541

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023). *Fairlearn: Assessing and improving fairness of AI systems*. arXiv. https://doi.org/10.48550/arXiv.2303.16626

Wrigley-Asante, C., Ackah, C. G., & Frimpong, L. K. (2023). Gender differences in academic performance of students studying science technology engineering and mathematics (STEM) subjects at the University of Ghana. *SN Social Sciences*, *3*(1), Article 12. https://doi.org/10.1007/s43545-023-00608-8

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes? In C. Meinel, M. Pérez-Sanagustín, M. Specht, & A. Ogan (Eds.), *L@S '21: Proceedings of the eighth ACM conference on learning @ scale* (pp. 91–100). ACM Press. https://doi.org/10.1145/3430895.3460139

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards accurate and fair prediction of college success: Evaluating different sources of student data. In A. N. Rafferty, J. Whitehill, C. Romero, & V. Cavalli-Sforza (Eds.), *Proceedings of the 13th international conference on educational data mining (EDM 2020)* (pp. 292–301). International Educational Data Mining Society. https://eric.ed.gov/?id=ED608066

Appendix A: Supporting Information

This appendix consists of individual analyses for evaluating machine learning model performance and bias level at various stages on gender and the free lunch status, respectively.

Evaluating model performance and bias level at various stages separating students based on gender

As described in the paper, there were similar numbers of female students (108) and male students (107) in the dataset. We first trained the machine learning models on the original dataset that is already balanced, separated based on gender. And the average model performance on each metric at each stage is shown in Figure A1(a). Then, after applying the post-processing mitigation approach to the trained models, the average performance on each metric at each stage is shown in Figure A1(b). Notably, compared with the performance of original models shown in Figure A1(a), the performance of mitigated models in Figure A1(b) had obvious improvement at both Stage 1 and Stage 2.

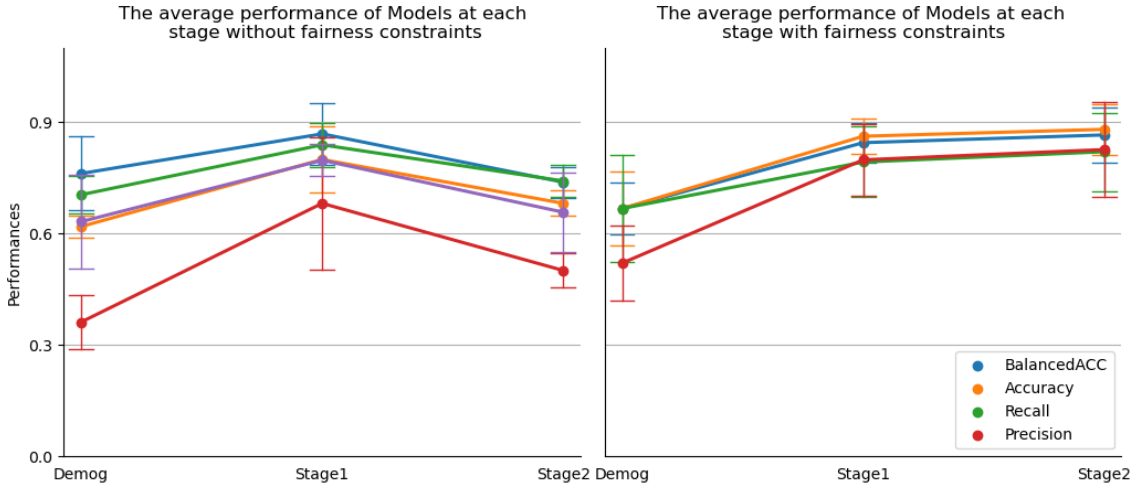


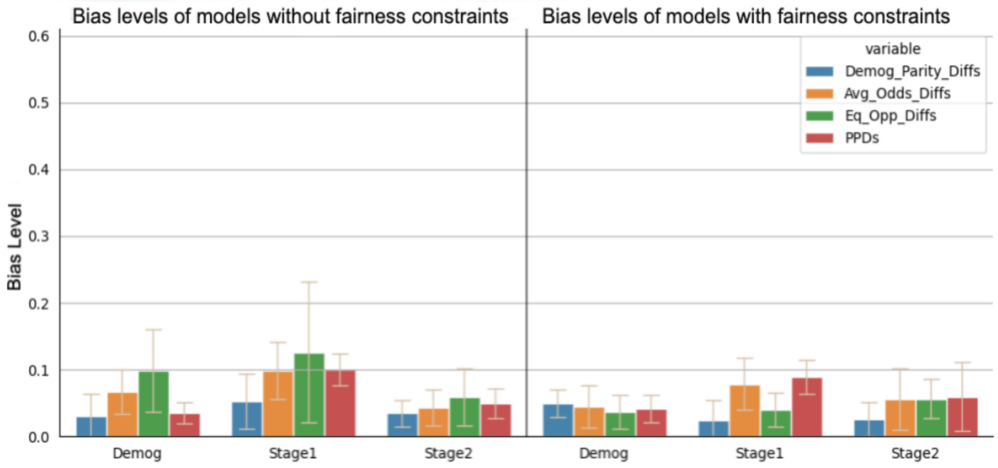
Figure A1. The performance of original models and mitigated models on each metric at each stage.

Table A1. Comparison of Average Model Performance at Each Stage

	Stages	Balanced Accuracy	Accuracy	Recall	Precision
Original Models	Demog	0.618	0.704	0.361	0.631
	Stage 1	0.799	0.838	0.681	0.796
	Stage 2	0.681	0.741	0.500	0.657
Mitigated Models	Demog	0.667	0.667	0.667	0.520
	Stage 1	0.844	0.861	0.798	0.792
	Stage 2	0.865	0.880	0.819	0.825

From Figure A1, it is obvious that Stage 1 is most appropriate for the early identification of at-risk students, given both the high average model performance and the relatively early stage in the course. However, comparing the average performances of the original models and mitigated models at Stage 1, there is no significant difference found on each of the metrics: Balanced Accuracy (Cohen's d = 0.615, t-Statistics = 0.869, p-value = 0.425), Accuracy (Cohen's d = 0.434, t-Statistic = 0.614, p-value = 0.563), Recall (Cohen's d = 0.779, t-Statistic = 1.102, p-value = 0.325), and Precision (Cohen's d = 0.016, t-Statistic = 0.022, p-value = 0.983).

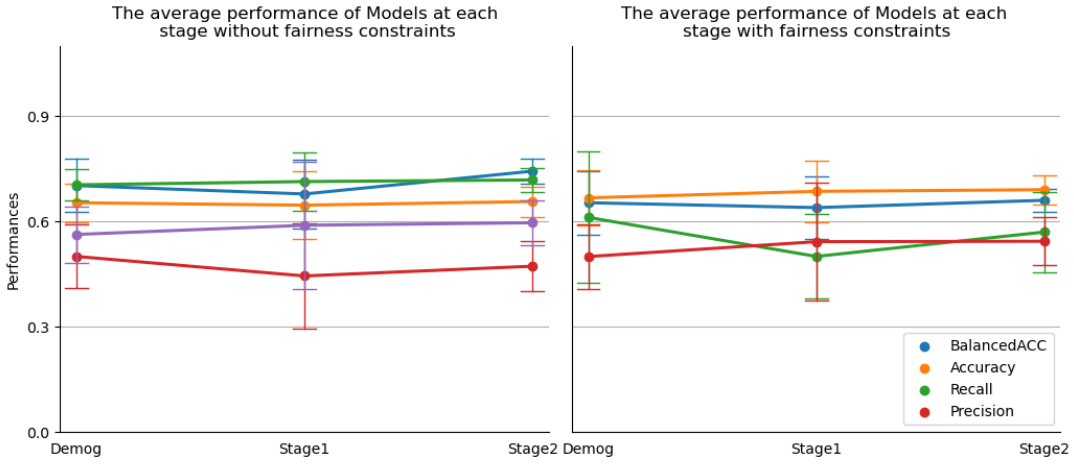
Figure A2 examines the differences in fairness levels on fine-tuned and mitigated models. Similarly, the comparison of Figures A2(a) and (b) did not indicate significant improvement in fairness levels of mitigating models with the post-processing approach. Although Stage 1 does not demonstrate the least bias on either the fine-tuned models and mitigated ones, its fairness level remains within the acceptable range. As such, when considering both performance and fairness, we can conclude that Stage 1 offers a reliable point at which we can identify at-risk students with both high accuracy and fairness while separating by gender.



(a) (b)

Figure A2. The bias levels of various machine learning models.

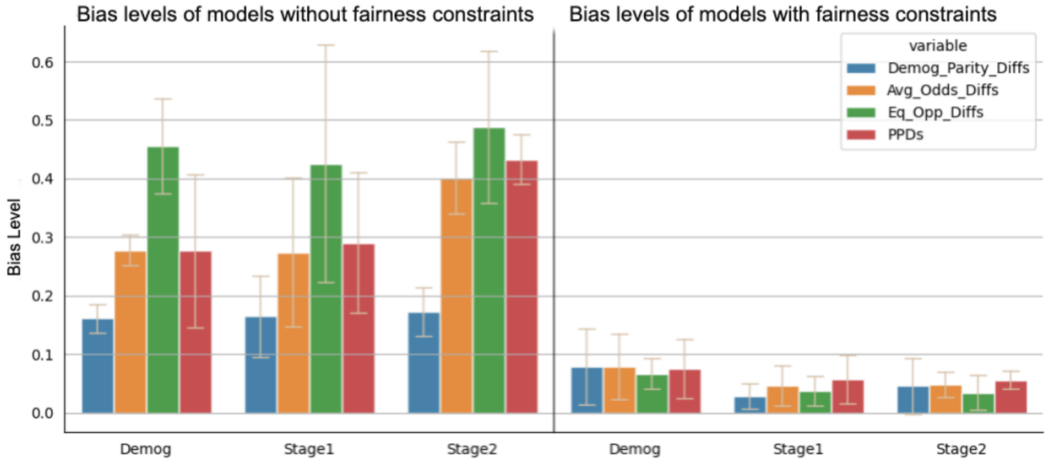
Evaluating the changes in performance and bias level after excluding the gender feature in training machine learning models, Figure A3 shows the performance of the fine-tuned models and mitigated models at different stages.



(a) (b)

Figure A3. The performance of fine-tuned and mitigated models on the dataset after gender was removed.

Comparing Figures A3 and A1, there were obvious drops in model performance after excluding gender. Moreover, in Figure A3, we can also find no significant changes in performance across stages both within and across the original and mitigated models.



(a) (b)

Figure A4. Bias levels of machine learning models after excluding gender.

ISSN 1929-7750 (online).

Figure A4(a) shows the bias levels of fine-tuned models after excluding gender, which indicates an obvious increase in bias compared with Figure A3(a). However, the post-processing mitigation strategies seem effective in improving model bias, as shown in Figure A4(b). Even though Figure A4(b) demonstrates higher fairness levels than Figure A2(b), the bias levels under both conditions are within the acceptable range. Considering the drop in performance after removing gender, we conclude that, in our case, removing the sensitive features in the training process is not always the effective choice.

Evaluating performance and bias level of the fine-tuned and mitigated models on the imbalanced dataset based on free lunch status

Similar to race, there is an imbalanced distribution of students based on their free lunch status (Yes: 33, No: 182). As shown in Figure A5, both the fine-tuned and mitigated models had a relatively low performance when trained on the imbalanced dataset at each stage.

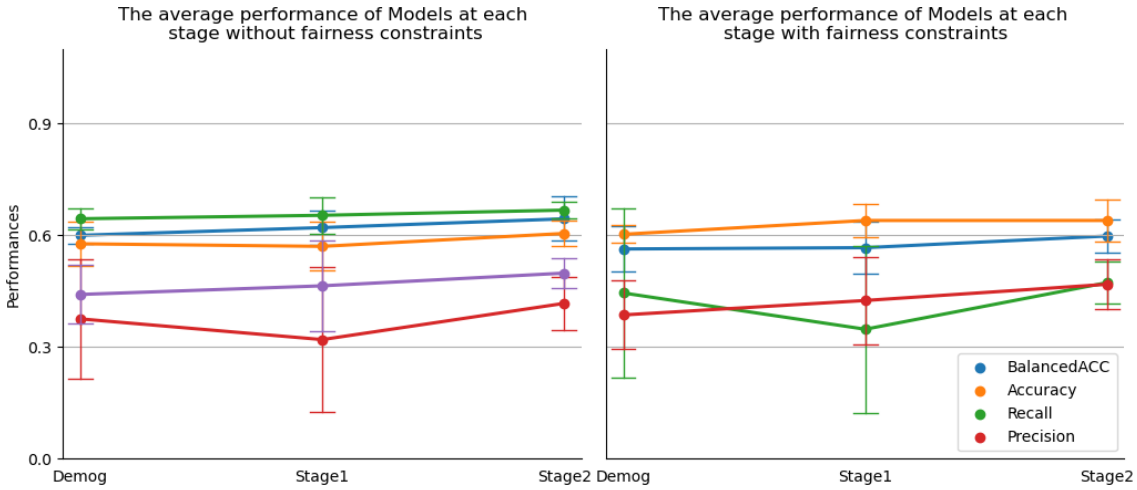


Figure A5. The performance of fine-tuned and mitigated models at each stage regarding free lunch status.

Figure A6 shows the bias levels of models trained on the imbalanced dataset. The bias levels at Stage 1 in both the fine-tuned and mitigated models are roughly within the acceptable range, except for Demog_Parity_Diffs (around 0.2) in the fine-tuned models.

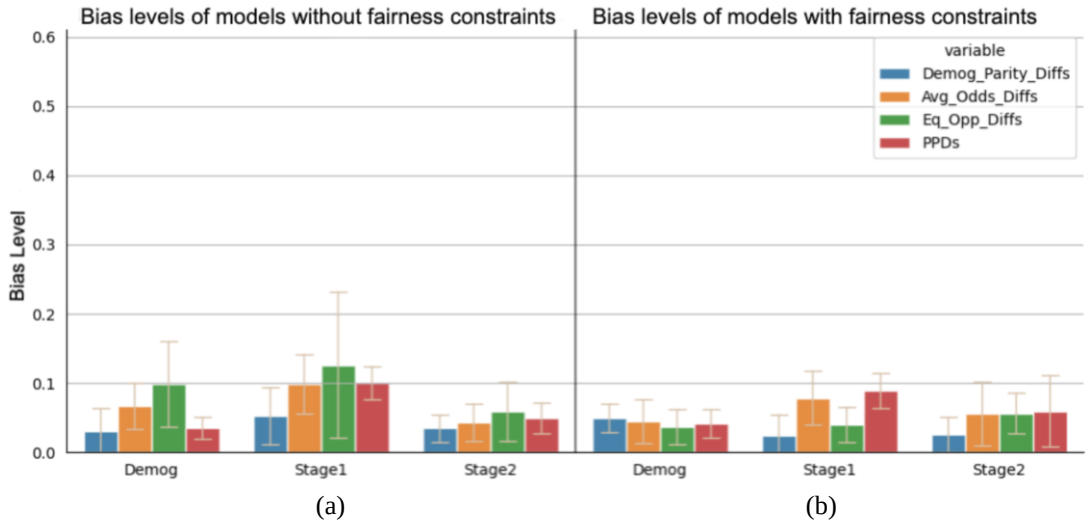


Figure A6. Bias of machine learning models trained on imbalanced datasets divided by free lunch status.

Next, we investigated the model performance and fairness levels on the balanced dataset as well as on the dataset that excluded the free lunch status.

Figure A7 compares the model performance at different stages under various conditions. Figures A7(a) and (b) compare performance of the model trained on the balanced dataset and mitigated with the post-processing strategy. Figures A7(c) and (d) compare the model performance trained on the balanced dataset with the free lunch status removed and mitigated with the post-processing strategy. Overall, Figure A7 indicates that balancing the dataset can improve model performance, and excluding the free lunch status decreases the model performance slightly.

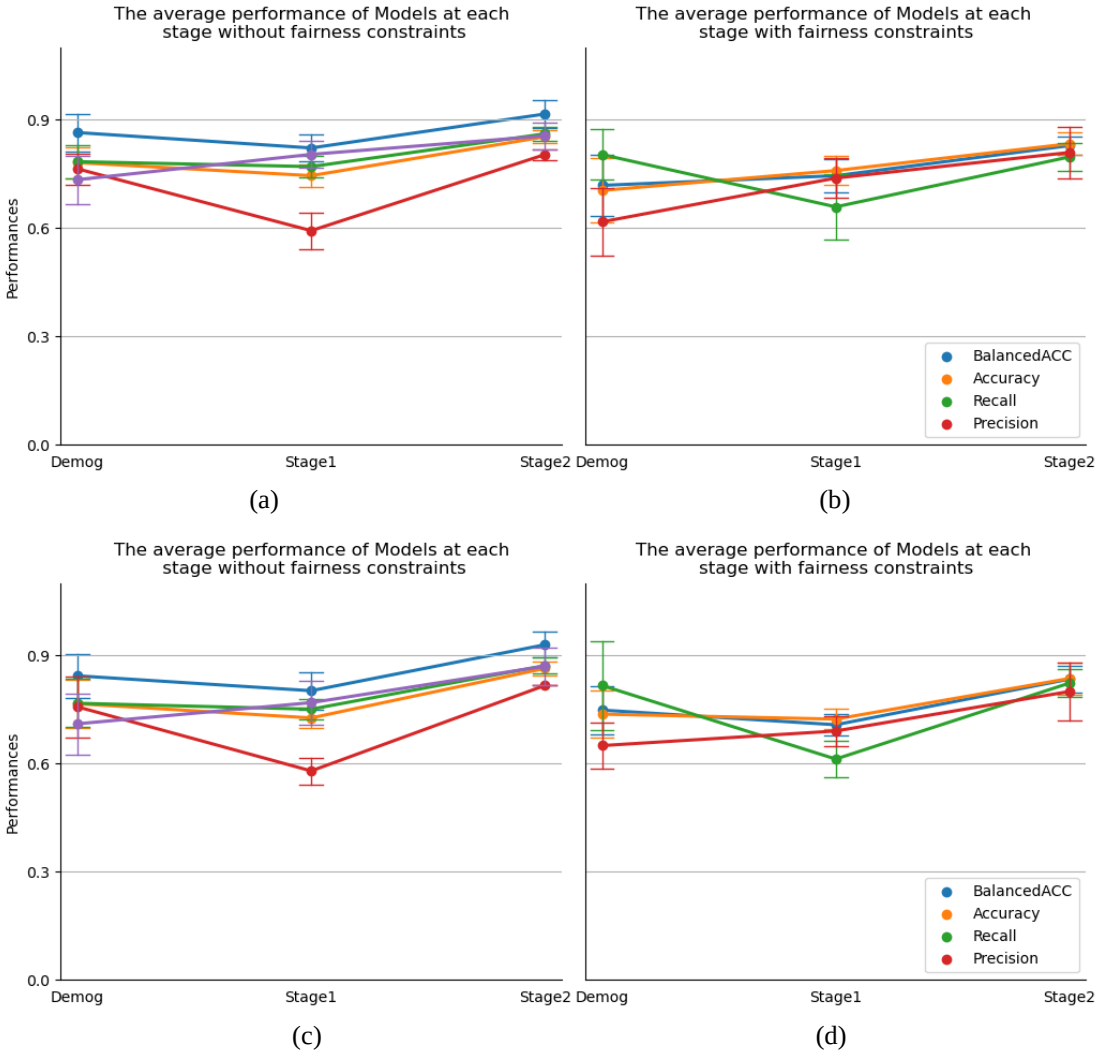


Figure A7. Model performance at different stages regarding free lunch status.

Comparing Figures A7(a) and (c), focusing on Stage 1, it can be found that excluding the free lunch status can slightly reduce the differences in model performance across the groups. Comparing the differences in performance between the fine-tuned and mitigated models at Stage 1, we find small differences under the two conditions. This further indicates that the fine-tuned models can also produce small differences across groups.

Figure A8 compares the bias levels of the fine-tuned and mitigated models trained on the balanced dataset and with the free lunch status excluded. Comparing Figures A8(a) and (c), it can be found that removing the free lunch status causes more bias in predictions. Comparing Figures A8 and A6 overall, we found that balancing the dataset and excluding the feature are not always effective strategies for improving predictive fairness.

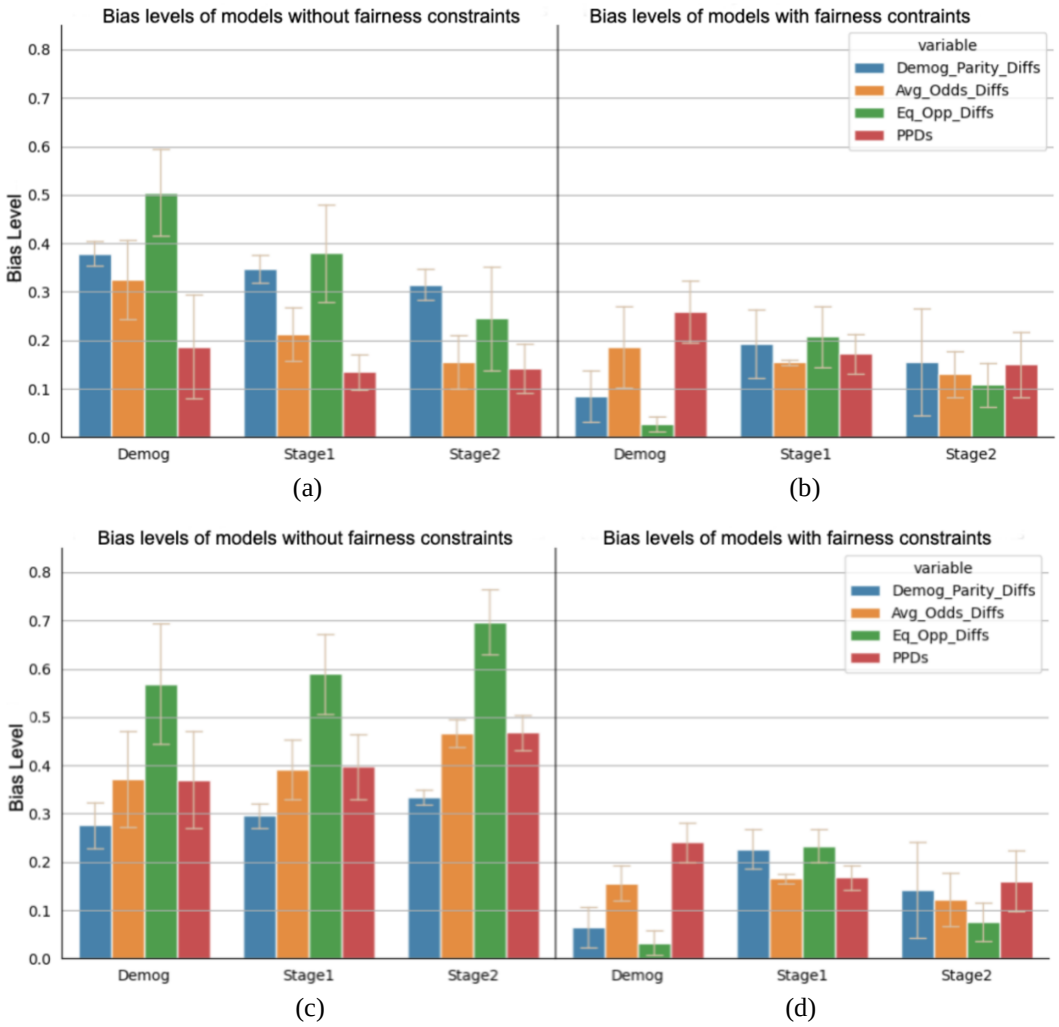


Figure A8. Model bias levels under different conditions.