Journal of Algorithms and Computations

journal homepage: http://jac.ut.ac.ir

# Modified Causal Fairness Algorithms to Enhance Equity in Education

M. Fathian[†1], M. R. Rasouli[1] and S. Oladi[1]

Electronic Commerce Department, Iran University of Science and Technology, Tehran, Iran

## ABSTRACT

Ensuring equity in educational assessment is essential for providing equitable learning opportunities to all students. However, persistent disparities in large-scale educational evaluations indicate that data-driven models may unintentionally amplify existing inequities when trained on biased data. This study evaluates the effectiveness of two causal fairness–modification algorithms in mitigating discrimination within the TIMSS student-achievement dataset. MData applies threshold-based causal label modifications, whereas MF reconstructs group-level distributions by enforcing conditional independence through factorization. Experimental results demonstrate that both algorithms substantially mitigate discrimination across gender, race, and socio-economic status. MData consistently preserves superior predictive accuracy, while MF achieves greater fairness. Together, these findings underscore the potential of causal pre-processing techniques to modify biased datasets and enhance equity in educational analytics. They also highlight the importance of integrating fairness-aware data modification into large-scale pipelines.

*Keywords:* Educational Fairness, Causal Learning, TIMSS dataset, Fairness Modification, Pre-processing Algorithms

AMS subject classification: 91D15, 97C60

## ARTICLE INFO

*Article history:*
Research paper
Received 15, October 2025
Accepted 30, December 2025
Available online 30, December 2025

[†] Corresponding author: M. Fathian

Email: fathian@iust.ac.ir

# 1 Introduction

Education, recognized as one of the most fundamental human rights and a cornerstone of sustainable development, plays an irreplaceable role in reducing poverty, improving quality of life, and fostering social empowerment. Among the most critical dimensions of equity, educational equity stands out as a central concern. In its simplest form, it refers to ensuring equal learning opportunities for all students, regardless of gender, socio-economic status, or race.

Despite notable progress in expanding access to education, substantial disparities persist across many societies. Underlying factors such as poverty, geographic disadvantage, gender discrimination, and ineffective policy frameworks continue to deprive large numbers of children and adolescents of high-quality education. According to UNESCO's 2020 report, approximately 258 million children worldwide remain out of school. Likewise, OECD (2019) findings consistently identify family socio-economic status as one of the strongest predictors of academic achievement; students from lower-income households scored, on average, nearly 90 points lower on PISA assessments than their more affluent peers. In this context, educational equity is not merely an aspirational ideal—it is an essential prerequisite for social justice and sustainable development.

In the contemporary world, where advanced technologies and intelligent algorithms increasingly shape decision-making processes, fairness has become critically important. Decisions related to planning, resource allocation, policy design, and educational interventions are increasingly data-driven and algorithmically mediated. Fairness, in its broadest sense, refers to equal and non-discriminatory treatment of individuals and groups. Within education, identifying capable and talented students who may be overlooked due to sensitive attributes—such as gender, socio-economic background, or place of birth—is vital. Detecting such inequities enables more effective policy interventions and targeted resource allocation. This, however, requires methods capable of identifying and addressing discriminatory and inequitable patterns.

Machine learning models are often perceived as neutral tools; however, in practice, they may reproduce and even amplify existing discrimination when trained on biased data. A primary source of unfairness in machine learning stems from the data itself. Data typically reflect societal biases embedded in the source population, and consequently, models trained on such data may unintentionally generate discriminatory outcomes. In education, such biased data can adversely shape decisions and policies, leading to unequal resource distribution and the perpetuation of existing inequities.

Thus, equity in educational decision-making depends not only on equitable models but also on bias-aware, discrimination-free datasets. Because models learn directly from the underlying data, any form of bias—whether explicit or implicit—can propagate through the model. As these systems are automated and widely deployed, such biases may even become amplified. In other words, the quality and fairness of a model's predictions are inherently constrained by those of the data on which it is trained. When data encodes discriminatory patterns, the model will inevitably reproduce them, resulting in unfair behavior. Traditional classification algorithms primarily focus on maximizing overall prediction accuracy, typically evaluated by the closeness of the predicted outcomes to observed data. However, this optimization often favors majority groups, reducing prediction error for these groups while increasing it for minority populations.

In educational datasets, the majority groups typically include students with non-sensitive attributes, such as male gender, middle- or upper-class socio-economic status, or urban residency in developed regions. When models are trained predominantly on such majority patterns, they implicitly learn and reinforce biased relationships. Consequently, they perform worse and generate higher error rates for underrepresented groups. Machine learning classification models, therefore, face two fundamental fairness challenges [1]:

(1) Elevated prediction error for minority groups due to disproportionate optimization in favor of majority groups;

(2) Reinforcement and reproduction of discriminatory patterns present in the training data, resulting in persistent algorithmic unfairness.

These challenges demonstrate that relying solely on accuracy, without incorporating fairness considerations, can have serious real-world implications—particularly in sensitive domains such as education. To address this issue, growing attention has been directed toward modifying data prior to model training. This approach, broadly categorized under pre-processing fairness, is closely aligned with the emerging paradigm of causal fairness pre-processing. Unlike traditional statistical fairness methods that rely primarily on correlations, causal fairness approaches aim to uncover and correct discriminatory mechanisms embedded within the underlying causal structure of the data. By distinguishing between spurious correlations and genuine causal relationships, causal fairness provides a principled foundation for mitigating discrimination while preserving the integrity of causal dependencies essential for valid inference and policy design.

More broadly, fairness-enhancing techniques proposed in the machine learning literature can be classified into three major categories [2]:

(1) Pre-processing methods, which modify or repair the training data prior to being fed into the model;

(2) In-processing methods, which adjust the model's architecture, constraints, or objective function during training; and

(3) Post-processing methods, which revise the model's predictions after training is completed.

Among these categories, pre-processing methods have gained substantial traction due to their model-agnostic nature, adaptability across learning algorithms, and their applicability in contexts where modifying the model itself is infeasible. Within the pre-processing family, causal fairness has emerged as a powerful and conceptually rigorous approach. By explicitly modeling causal pathways between features, sensitive attributes (e.g., gender, race, socio-economic status), and target variables, causal fairness methods enable the detection and correction of discriminatory causal mechanisms, rather than merely adjusting surface-level correlational patterns.

A variety of causal fairness algorithms have been introduced, each employing different strategies to reconstruct and adjust relationships among data features, sensitive attributes, and target outcomes. However, empirical evidence indicates that no single algorithm performs uniformly well across all datasets. Some methods are highly effective at reducing discrimination, but they do so at the cost of substantial losses in predictive accuracy. Others maintain strong predictive performance, but exhibit limited ability to mitigate bias. This inherent trade-off highlights that the appropriateness of a given fairness intervention is highly context-dependent. It also underscores the need for systematic evaluations of causal fairness methods on real-world, fairness-sensitive datasets.

This study examines the challenge of ensuring fairness in predictive modeling using the Trends in International Mathematics and Science Study (TIMSS) dataset, one of the most widely used international benchmarks in educational research. Administered every four years, TIMSS provides rich information on students' mathematics and science performance, along with detailed demographic and socio-economic characteristics, such as gender, parental education, and socio-economic status [3]. The presence of such sensitive attributes makes the TIMSS dataset particularly vulnerable to embedded biases, raising the risk that predictive models trained on this data may inadvertently favor certain demographic groups. Such biases can influence educational policy decisions, resource distribution, and targeted interventions, thereby reinforcing existing inequalities in education systems.

To mitigate these concerns, two prominent causal fairness pre-processing algorithms—MData and MF—were applied to the TIMSS dataset. MData identifies meaningful causal partitions (block sets) and repairs the dataset accordingly. On the other hand, the MF approach frames fairness repair as a constrained matrix factorization process designed to systematically eliminate discriminatory patterns. Our objective is to evaluate the effectiveness of these methods in reducing unfair disparities associated with sensitive attributes while preserving adequate predictive performance. By systematically examining the fairness–accuracy trade-offs introduced by each algorithm, this study aims to provide empirical guidance for researchers, data scientists, and policymakers seeking appropriate fairness interventions for large-scale educational datasets.

The remainder of this paper is organized as follows. Section 2 reviews the related literature on algorithmic and causal fairness. Section 3 presents the proposed methodology and describes the causal fairness algorithms employed in this study. Section 4 reports the experimental results and provides an in-depth analysis of the fairness–accuracy trade-offs. Finally, Section 5 concludes the paper and discusses potential avenues for future research.

## Related Work

In recent years, fairness has become a central concern in data mining and machine learning research. With the increasing deployment of intelligent systems in educational, social, and economic domains, concerns regarding discrimination and bias have intensified. Many datasets used for training machine learning and artificial intelligence models contain sensitive attributes, such as gender, socio-economic status, or place of birth. These attributes may inadvertently introduce bias into analyses and model predictions. To address these challenges, a wide range of studies has proposed data repair techniques and fairness intervention methods designed to reconstruct or adjust datasets and model outputs in accordance with fairness principles.

Among these approaches, causal fairness has gained particular prominence. Causal machine learning refers to methods that incorporate causal knowledge into the modeling process [1]. Broadly speaking, causal learning involves estimating the causal effect of interventions on an outcome (the target variable) and distinguishing correlations from confounding factors that may bias analytical results [2]. In the literature, causal learning is typically categorized into five branches: (1) supervised causal learning, (2) causal generative modeling, (3) causal explanation, (4) causal fairness, and (5) causal reinforcement learning [3]. This study focuses specifically on causal fairness, which aims to detect, prevent, and correct biases present in data or models by leveraging conditional independence relationships among sensitive attributes, predicted variables,

and target outcomes [5]. In essence, causal fairness is achieved when sensitive attributes (e.g., gender, age, race) do not exert a causal influence on the final prediction or outcome [3].

Methodologically, and as noted in the Introduction, approaches to achieving fairness are generally grouped into three main categories: (1) pre-processing, (2) in-processing, and (3) post-processing [2]. According to Su et al., pre-processing methods modify or refine the data before model training to reduce the impact of sensitive attributes. In-processing approaches incorporate fairness constraints into the learning algorithm itself to ensure discrimination-free behavior of the model. Post-processing approaches adjust model outputs after training to meet fairness criteria. A key insight highlighted in prior research is that discrimination can only be meaningfully claimed when a causal relationship between sensitive attributes and decisions is established. Statistical correlations alone are insufficient to justify claims of discriminatory outcomes [4]. Accordingly, fairness in machine learning is often defined under two primary paradigms: (1) correlation-based fairness, which considers statistical distributions of data and outputs, and (2) causality-based fairness, which focuses on cause-and-effect relationships and the elimination of discriminatory causal pathways [4].

Fairness can furthermore be assessed at both the group and individual levels [5]. Group fairness examines disparities across demographic groups. It includes two subtypes: demographic-aware (examining outcome distributions across groups) and error-aware (examining group-based error rates) [5]. Conversely, individual fairness emphasizes that similar individuals should receive similar predictions regardless of sensitive attributes [5].

Salimi et al. introduced a framework called Interventional Fairness, which evaluates fairness in machine learning models through causal interventions [3]. The core aim of this framework is to ensure that sensitive attributes—such as race or gender—do not influence algorithmic decisions through impermissible causal pathways. Unlike methods relying solely on statistical correlations, interventional fairness employs causal interventions to quantify and mitigate discrimination. This approach does not require a full causal model; instead, it focuses on identifying which variables (acceptable variables) may legitimately mediate the influence of sensitive attributes on outcomes [3]. If a sensitive attribute affects the outcome through impermissible pathways, the resulting decisions are deemed unfair.

To address this, the authors propose two pre-processing algorithms: (1) a MaxSAT-based repair method, and (2) a Matrix Factorization (MF) repair method [3]. The MaxSAT-based method transforms fairness enforcement into a satisfiability optimization problem, while the MF approach ensures conditional independence and removes discriminatory causal effects. These algorithms have been validated on widely used real-world datasets such as Adult and COMPAS, demonstrating substantial fairness improvements without significant losses in model accuracy.

In a related study, Salimi et al. introduced Capuchin, a practical tool for causal data repair designed to achieve fairness without requiring a complete causal model of the domain [6]. Capuchin operates by distinguishing permissible from impermissible variables, enforcing fairness through logical constraints rather than full causal graphs. It applies the same MaxSAT-based repair formulation while also incorporating the previously introduced MF algorithm to enforce conditional independence between sensitive attributes and outcomes with respect to permissible variables [6]. By reconstructing datasets based on partial causal information, Capuchin enables flexible and effective fairness enforcement in varied machine learning contexts [6].

Wu et al. examined fairness in ranking systems, aiming to detect and mitigate discrimination in ranked data [7]. Their work employed a causal graph to model direct and indirect discrimination by mapping ranked positions to continuous scores. The proposed method identifies discriminatory causal pathways and reconstructs rankings to remove both direct and indirect unfairness. This approach addresses the limitations of traditional statistical techniques, which may overlook certain forms of discrimination [7].

Two key algorithms were introduced: one for discrimination detection using the Bradley–Terry model and causal graph analysis, and the other for discrimination removal and fair ranking reconstruction [7].

Zhang, Wu, and Wu presented three foundational algorithms for detecting and eliminating discrimination in datasets [8]. They defined discrimination as the unfair treatment of sensitive groups, with features such as gender, race, religion, and disability treated as sensitive attributes.

The first algorithm is a non-discrimination verification algorithm that checks all meaningful data partitions (block sets) for discriminatory patterns. The second is a graph-modification algorithm that adjusts the causal graph to produce a fairness-compliant dataset while preserving utility. The third is a direct data-modification algorithm that alters specific tuples to satisfy non-discrimination criteria [8].

Qureshi et al. investigated individual-level causal discrimination using contingency tables for each data instance. This approach allows for fine-grained analysis beyond population-level trends [9]. Their study employed the Adult and Crime datasets. They utilized a causal discrimination discovery framework that integrates propensity score estimation, weighting and balancing, causal risk difference measure, and regression tree analysis. This approach isolates causal effects from confounding factors, enabling precise estimation of individual-level discrimination or favoritism [9].

Nilforoshan et al. categorized causal fairness approaches into two primary types. The first type evaluates outcomes when decisions are counterfactually altered, while the second evaluates outcomes when protected attributes are counterfactually altered [10].

The study focused on the context of college admissions and examined both approaches through fairness-oriented decision analysis. The first approach assesses fairness by determining whether altering decisions (e.g., rejection vs. acceptance) yields equitable outcomes across groups. The second evaluates whether decision outcomes remain consistent when sensitive attributes such as gender or race are counterfactually altered. This corresponds to the principle of counterfactual fairness [10].

Islam et al. evaluated thirteen fairness-enhancing classification methods, spanning causal and non-causal approaches. The evaluation considered multiple dimensions, including accuracy, fairness, efficiency, scalability, robustness, model sensitivity, data efficiency, and stability [5].

Their results indicated that causal methods generally outperform non-causal methods in fairness preservation and bias reduction. In particular, the pre-processing approaches proposed by Salimi et al. [3] and Zhang, Wu, and Wu [8] demonstrated superior fairness improvements. However, these methods incurred higher computational costs and faced scalability challenges in high-dimensional settings. Post-processing approaches, while more efficient and scalable, perform less favorably in balancing fairness and accuracy [5].

Despite these methodological advancements, applications of causal fairness to educational datasets—such as TIMSS—remain limited. Sensitive attributes, such as gender and socio-

economic status, can significantly influence academic performance predictions and risk perpetuating educational inequities.
This study addresses this gap by applying the MData and MF pre-processing algorithms to the TIMSS dataset. It systematically evaluates their effectiveness in reducing bias while preserving predictive performance, and provides insights into context-appropriate fairness interventions for educational research.

## 2 Proposed Method

Given the research problem and its alignment with data mining studies, the CRoss Industry Standard Process for Data Mining (CRISP-DM) framework was adopted as the foundational methodology. This process comprises six stages [11]: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. In this study, the methodological workflow integrates the phases of Data Understanding, Data Preparation, Modeling, and Evaluation, as illustrated in Figure 1.

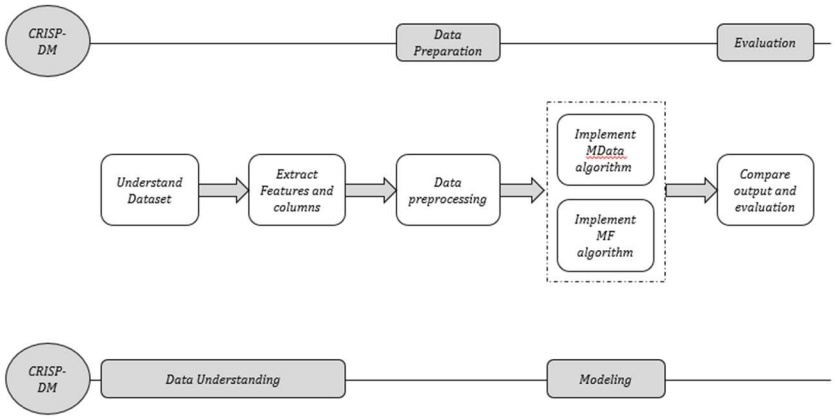


Figure 1: Proposed Method

In the first phase, Data Understanding is conducted to develop an in-depth comprehension of the dataset and to extract the required features and variables. The dataset used in this research is TIMSS, which evaluates fourth- and eighth-grade students from 64 countries in mathematics and science. It consists of multiple questionnaires encompassing 3,384 features.
In the second phase, based on prior research on the TIMSS dataset as well as the objectives of this study, the necessary features were extracted. Given the high dimensionality of the dataset, selecting influential variables is critically important. These variables should not only enhance fairness within causal learning but also reflect the multifaceted nature of educational quality.
Based on the reviewed literature, variables such as gender, race, and socio-economic status were identified as sensitive or protected attributes. These variables were derived from the ASG student questionnaire dataset, completed across 64 countries, and were used to construct the base dataset. The integrated dataset comprises 373,832 rows and 268 columns, collected from 807 schools.
From the 268 ASG questionnaire features, the selected sensitive attributes included gender (ASBG01), race (a composite of ASBG06A, ASBG06B, and ASBG07), and socio-economic status (a composite of ASDGHRL, ASDG05S, and ASBG09B).

Additional variables were selected based on the literature review [12-18]. The selection focused on (1) student background, (2) resources available to the student, and (3) the educational and learning environment, in alignment with the research objectives.
The study examines fairness in mathematics achievement. Since the dataset contains five plausible values for mathematics, the target variable was defined as the mean of these five scores. Details of the selected variables are presented in Table 1 [19]:

Table 1: Definition and Value Range of Selected Features

| Category | | Code | Definition | Values |
|---|---|---|---|---|
| Features | | ASDGSB | Student Bullying | 1: Never or Almost Never; 2: About Monthly; 3: About Weekly |
| | | ASDGSCM | Students Confident in Mathematics | 1: Very Confident; 2: Somewhat Confident; 3: Not Confident |
| | | ASDGSLM | Students Like Learning Mathematics | 1: Very Much Like; 2: Somewhat Like; 3: Do Not Like |
| | | ASBM03B | The Teacher Is Easy to Understand | 1: Agree a lot; 2: Agree a little; 3: Disagree a little; 4: Disagree a lot |
| | | ASDGDML | Disorderly Behavior during Math Lessons | 1: Few or No Lessons; 2: Some Lessons; 3: Most Lessons |
| | | ASBG09A | How Often\Tired | 1: Every day; 2: Almost every day; 3: Sometimes; 4: Never |
| | | ASDGSSB | Students' Sense of School Belonging | 1: High; 2: Some; 3: Little |
| Protected Features | Gender | ASBG01 | Sex of Student | 1: Girl; 2: Boy; 3: <Other> |
| | Race | ASBG06A | Were Your <Parents/Guardian A> Born In | 1: Yes; 2: No; 3: I don't know; 4: Not applicable |
| | | ASBG06B | Were Your <Parents/Guardian B> Born In | 1: Yes; 2: No; 3: I don't know; 4: Not applicable |
| | | ASBG07 | Were You Born In | 1: Yes; 2: No |
| | socio-economic status | ASDGHRL | Home Resources for Learning | 1: Many; 2: Some; 3: Few |
| | | ASDG05S | Number of Home Study Supports | 1: Neither Own Room nor Internet Access; 2: Either Own Room or Internet Access; 3: Both Own Room and Internet Access |
| | | ASBG09B | How Often\Hungry | 1: Every day; 2: Almost every day; 3: Sometimes; 4: Never |

In the second phase, the selected columns identified during the Data Understanding stage were extracted from the initial dataset. This resulted in a new dataset that required pre-processing. The pre-processing stage consisted of the following steps:

- **Replacing invalid values and removing incomplete rows:** According to dataset guidelines, values such as "omitted or invalid," "not administered," "logically not applicable," and numeric placeholders like 9, 99, 999, 9999, 99999, 999999, 9999999, 99999999, 9999999999, ".", ".A", "Sysmis", and blank entries are considered invalid and must be handled appropriately [19]. In this study, all such entries were removed.
- **Creating a binary target variable by averaging five mathematics scores:** Each student has five plausible values representing mathematics performance. Therefore, the average of these scores was calculated and used as the target variable. Due to the requirements of the algorithms employed (explained later), the target variable must be defined as binary.
- **Labeling the target variable:** Students with an average score above the overall mean were labeled as high-performing (value = 1), while the remaining students were labeled as low-performing (value = 0).
- **Combining sensitive features composed of multiple columns (race and socio-economic status):** Race and socio-economic status each consisted of three variables, which were first merged to form composite variables. For race, if either parent or the student was not born in the country under study, the resulting value was set to 1; otherwise, it was set to 0. For socio-economic status, a new variable was constructed from household resources, hunger, and academic support, weighted according to their importance. Students were then categorized into two groups: 1 for high socio-economic status and 0 for low socio-economic status.
- **Mapping and labeling sensitive features as binary:** For the composite race and socio-economic status features, binary mapping was performed following their combination. For gender, entries corresponding to the "other" category were removed prior to applying binary mapping.

In the third step, modeling was performed using two algorithms, MData and MF, which were briefly introduced in the Related Work section.

In the MData algorithm, the dataset was partitioned into subpopulations based on a selected subset of features and their corresponding values, as originally described by Zhang, Wu, and Wu [8]. Segmentation refers to dividing the dataset into mutually exclusive subgroups defined by these features. This process is essential for detecting and quantifying discrimination with respect to a sensitive attribute and a decision attribute [8].

A block set is defined as a subset of features that isolates the causal influence of the sensitive attribute on the decision attribute when used for segmentation. Using a valid block set ensures that any observed disparity in decision outcomes can be attributed directly to the sensitive attribute, without interference from confounding variables [8].

The authors demonstrated that discrimination is present when the sensitive attribute exerts a causal effect on the decision variable. To assess this, the study computes, for each subpopulation s, the probability difference [8]:

$$\Delta P|\boldsymbol{s} = \Pr(e^+|c^+, \boldsymbol{s}) - \Pr(e^+|c^-, \boldsymbol{s}) \quad (1)$$

where $e$ denotes the decision variable, with $e^+$ indicating a favorable outcome. Similarly, c denotes the sensitive attribute, where $c^+$ represents the privileged group and $c^-$ the unprivileged group (e.g., male/female in gender-based analyses). If the absolute value of this difference exceeds a user-defined threshold, the corresponding subpopulation is labeled as discriminatory [8].

The MData algorithm takes the dataset, the sensitive attribute, the decision variable, and the discrimination threshold as input, and outputs a repaired dataset [8].

First, the Non-Discrimination Certification Algorithm determines whether the dataset satisfies the fairness criterion. If unfairness is detected, the algorithm evaluates each partition produced during certification.

For any subpopulation in which the probability difference exceeds the threshold, a specified number of tuples are randomly selected for modification. If the disparity favors the privileged group, tuples belonging to the unprivileged group with unfavorable outcomes are flipped to positive. Conversely, if the disparity favors the unprivileged group, tuples with positive outcomes are flipped to negative [8]. The modified tuples are then reintegrated into the dataset, producing an updated, fairness-adjusted version [8].

Building on this original framework, the present study implements an extended fairness-analysis pipeline tailored to the TIMSS dataset.

The proposed system incorporates four main algorithmic components, each adapted from or inspired by the procedures described in the MData paper [8]. These components were redesigned to accommodate the structure, dimensionality, and scale of the TIMSS dataset:

(1) Data Pre-processing and Label Construction.
(2) Subpopulation-Based Fairness Assessment (ΔP Computation).
(3) Enhanced MData Repair.
(4) Utility and Stability Evaluation.

Together, these components form a comprehensive fairness pipeline that identifies discrimination, repairs the dataset, and evaluates the consequences of the repair process. The complete workflow of the proposed methodology is in Table 2.

Table 2: Enhanced MData

| **Input: Dataset, Q-columns, plausible-value mathematics scores (target feature), protected attribute**<br>**Output: Modified dataset, Fairness metrics** |
|---|
| **Do the following steps:**<br>1- **Data Pre-processing and Label Construction.** Standardize missing values, derive the protected attribute, compute the average plausible-value mathematics score, and construct the binary decision label. Fix the Q-columns based on Table 1.<br>2- **Subpopulation Segmentation and Fairness Assessment.** Partition the dataset by Q-columns, compute ΔP for each subpopulation, and identify discriminatory groups whose absolute disparity exceeds the threshold.<br>3- **Fairness Repair (Enhanced MData).** Compute the required number of label flips, modify selected decision labels, and generate the fairness-adjusted dataset.<br>4- **Utility and Stability Evaluation.** Compute the distribution distance and evaluate predictive performance using logistic regression. |

In addition to the MData algorithm, the second modeling approach employed in this study is the Matrix Factorization (MF)–based repair algorithm.

The MF algorithm, originally introduced in the context of enforcing conditional independence constraints [3], operates on a bag $B$ (i.e., a multiset of tuples) defined over the attribute set $V = XYZ$. The goal of the algorithm is to ensure the saturated conditional independence relation $X \perp\!\!\!\perp Y|Z$.

Given the input bag $B$, the algorithm constructs a contingency matrix for each value of $z$. This matrix enumerates all combinations of the values of $X$ and $Y$ conditioned on $Z = z$ [3]. Next, a factorization function decomposes the contingency matrix into two non-negative components that approximate the marginal distributions of $X$ and $Y$ given $Z = z$ [3]. Then, the marginal components are recombined. They are scaled by $|B_Z|$, the size of the sub-bag corresponding to $Z = z$, to generate a reconstructed matrix whose joint distribution satisfies the conditional independence constraint [3]. Finally, the algorithm assembles these reconstructed matrices across all values of $z$, producing the repaired bag $B'$, which is guaranteed to satisfy conditional independence [3].

In summary, the two approaches differ both conceptually and operationally. The MData algorithm focuses on local, targeted modifications of decision outcomes within discriminatory subpopulations, making it efficient and precise for direct bias removal. However, it requires predefined segmentation features and may not capture indirect or proxy discrimination [3, 8].

By contrast, the MF algorithm operates at the distributional level, adjusting joint probability structures to enforce independence across attributes. This approach offers greater flexibility for complex datasets with latent correlations, but is computationally intensive and may introduce broader changes that affect utility. Overall, MData is well-suited for controlled environments with well-defined causal structure, while MF provides a more general mechanism for fairness enforcement at the cost of potentially reduced predictive performance.

Recognizing the limitations of applying the original algorithms directly to large-scale, real-world educational data, the present study incorporates several key modifications.

For **MData**, the maximum allowable percentage of label changes is user-configurable, enabling practitioners to balance fairness and data fidelity. Moreover, the block set $Q$ is fixed based on the features reported in Table 1 to maintain methodological consistency across all experiments.

For **MF**, a minimum group size is set to maintain statistical reliability, and dynamic group resizing ensures scalability across heterogeneous TIMSS subpopulations.

To operationalize the MF modification strategy within the TIMSS context, the structured procedure shown in Table 3 is adopted.

Table 3: Enhanced MF

| **Input: Dataset, admissible attribute(s), selected inadmissible attributes, decision label, protected attribute**<br>**Output: Modified dataset, Fairness metrics** |
|---|
| **Do the following steps:**<br>1- **Data Pre-processing and Feature Selection.** Standardize missing values, construct the decision label, derive the protected attribute, and encode the selected inadmissible attributes. Compute mutual information to select the most informative inadmissible attributes and determine the best admissible attribute.<br>2- **Group-Wise Contingency Construction and Matrix Factorization.** Partition the dataset based on the selected admissible attribute. For each group, construct a contingency matrix of the decision label versus the inadmissible attributes and decompose it using Non-Negative Matrix Factorization to obtain smoothed marginal components.<br>3- **Fairness-Constrained Reconstruction.** Adjust the factorized components to satisfy fairness thresholds (demographic parity and equalized odds), reconstruct the repaired joint distribution, and generate synthetic row counts proportional to the repaired density for each group.<br>4- **Dataset Reconstruction and Evaluation.** Aggregate the repaired groups, normalize counts to match the size of the original dataset, and compute fairness and utility metrics, including accuracy, demographic parity difference, equalized odds difference, and subgroup-level disparities. |

In the final stage of the analysis, the MData and MF algorithms are applied to the dataset with respect to the three sensitive attributes—gender, socio-economic status, and race—and the corresponding results are reported. Student performance is predicted using a logistic regression classifier, while fairness is assessed using the ROD and DP metrics, as defined in Formulas (2) and (3) [6, 20].

The Ratio of Observational Discrimination (ROD) quantifies disparities in favorable outcomes between protected and unprotected groups, conditional on admissible attributes. Demographic Parity (DP) measures group-level equality by comparing the probability of receiving a favorable outcome across sensitive and non-sensitive groups. For both metrics, a value of zero indicates perfect fairness.

In this study, the outcome variable corresponds to the target label derived from the average mathematics score of fourth-grade students.

$$ROD = \frac{|P(Y=1|A=0,S=0) - P(Y=1|A=0,S=1)|}{P(Y=1)} \tag{2}$$

$$\mathrm{DP} = \Pr(O=1|S=1) - \Pr(O=1|S=0) \tag{3}$$

## 3 Results

This study investigates discrimination in predicting fourth-grade students' mathematics performance in TIMSS 2019 with respect to three sensitive attributes—gender, race, and socio-economic status—using the MData and MF fairness-repair algorithms. The objective is to reduce algorithmic bias while preserving predictive accuracy and achieving an effective trade-off between fairness and utility. Logistic regression is used as the predictive model, and fairness is assessed using ΔP (the discrimination measure used in MData), ROD, and DP.

The baseline accuracy of the logistic regression model, prior to applying any fairness interventions, was 63.19% for gender, 63.93% for race, and 57.93% for socio-economic status, reflecting

substantial variability across sensitive attributes. The following sections present the results of applying the MData and MF algorithms.

## 3.1 Results of MData Algorithm

The Non-Discrimination Certification step in MData [8] identified initial ΔP discrimination values of 0.1351 (gender), 0.1646 (race), and 0.1374 (socio-economic status), indicating that all three sensitive attributes exhibit measurable discrimination in the original dataset.

Table 2 summarizes the results of applying the MData algorithm across four discrimination thresholds τ (0.03, 0.05, 0.07, 0.10), with the maximum percentage of allowable modifications capped at 30%.

Table 4: Results of the MData Algorithm

| Protected Attribute | MData (τ) | ΔP | Accuracy | ROD | DP |
|---|---|---|---|---|---|
| **Gender** | **0.03** | **2.60%** | **68.46%** | **1.34%** | **1.16%** |
| | 0.05 | 3.45% | 68.41% | 1.64% | 1.32% |
| | 0.07 | 4.71% | 68.36% | 1.97% | 1.51% |
| | 0.10 | 6.71% | 68.29% | 2.11% | 1.59% |
| **Race** | **0.03** | **2.22%** | **64.44%** | **5.46%** | **2.97%** |
| | 0.05 | 3.35% | 64.37% | 6.16% | 3.03% |
| | 0.07 | 4.57% | 64.37% | 6.74% | 3.07% |
| | 0.10 | 6.35% | 64.35% | 7.46% | 3.06% |
| **Socioeconomic Status** | **0.03** | **2.27%** | **64.60%** | **17.50%** | **10.77%** |
| | 0.05 | 3.32% | 64.60% | 17.95% | 10.96% |
| | 0.07 | 4.62% | 64.61% | 18.62% | 11.18% |
| | 0.10 | 6.55% | **64.64%** | 19.85% | 11.57% |

The results of the MData algorithm reveal a clear and consistent reduction in discrimination across all sensitive attributes. Overall, the algorithm successfully lowers ΔP well below its initial values and within the predefined thresholds, with the strongest improvement observed at $\tau = 0.03$, where ΔP decreases to ≈2–3% for gender and race and to ≈2.27% for socio-economic status. This confirms that the block-set segmentation framework effectively eliminates causal discrimination.

In addition, predictive accuracy remains stable or even improves slightly after repair—for example, increasing from 63.19% to ≈68.4% for gender, from 63.93% to ≈64.4% for race, and from 57.93% to ≈64.6% for socio-economic status. This improvement is likely due to noise reduction through label corrections and the creation of more consistent decision boundaries.

However, fairness metrics such as DP and ROD display more heterogeneous behavior. Although ΔP is consistently reduced, gender shows the strongest fairness improvement with very low DP (≈1%) and ROD (≈1–2%); race exhibits moderate improvement but retains residual disparities, especially in ROD (≈5–7%); and socio-economic status remains the most challenging attribute, with DP and ROD staying relatively high (≈10–20%) despite substantial reductions in ΔP. These differences arise because MData directly optimizes ΔP rather than group-fairness metrics like DP or ROD, meaning that residual disparities may persist even when causal discrimination has been removed.

The comparatively weaker performance for socio-economic status is explained by several underlying factors, including its more imbalanced distribution, deeper and more fragmented block-set segmentation, and the 30% upper bound on permissible label modifications, which limits the

degree of correction achievable. Relaxing this constraint would likely yield further improvements for socio-economic status.

## 3.2 Results of the MF Algorithm

The Matrix Factorization (MF) algorithm was applied to the same three sensitive attributes—gender, race, and socio-economic status—using minimum group sizes of 5,000, 10,000, and 15,000. Table 3 summarizes the resulting accuracy and fairness metrics. Overall, MF exhibits behavior distinct from MData, reflecting its distinct objective of enforcing conditional independence by restructuring joint distributions rather than directly correcting decision labels. According to the results, MF generally achieves lower accuracy than MData across all sensitive attributes, with particularly pronounced declines for race and socio-economic status at smaller group sizes. However, the algorithm demonstrates notable improvements in certain fairness metrics, especially DP and ROD, under specific parameter settings.

Table 5: Results of the MF Algorithm

| Protected Attribute | Metric | Initial | Batch = 5000 | Batch = 10000 | Batch = 15000 |
|---|---|---|---|---|---|
| Gender | Accuracy | 63.19% | 62.28% | **63.84%** | 63.43% |
| | DP | 6.09% | 2.75% | **1.21%** | 4.66% |
| | ROD | 13.35% | 7.70% | **3.77%** | 19.21% |
| Race | Accuracy | 63.93% | 50.62% | 52.36% | **58.34%** |
| | DP | 6.98% | 2.27% | **0.79%** | 4.24% |
| | ROD | 14.60% | 11.17% | **2.47%** | 14.04% |
| Socio-economic Status | Accuracy | 57.93% | 50.18% | **52.09%** | **52.09%** |
| | DP | 11.02% | **1.90%** | 17.51% | 17.51% |
| | ROD | 27.54% | **7.10%** | 30.32% | 30.32% |

The MF algorithm exhibits a distinct performance pattern compared to MData. For the sensitive attribute of gender, a minimum group size of 10,000 yields the most balanced outcome, with accuracy (63.84%) closely matching the baseline while fairness metrics—particularly DP (1.21%) and ROD (3.77%)—show substantial improvement over the initial dataset. Notably, the ROD value at this configuration is markedly lower than the ROD achieved by the MData algorithm at $\tau = 0.03$. This indicates that MF provides superior fairness performance for gender when the batch size is sufficiently large.

For race, however, the MF algorithm introduces a noticeable drop in accuracy, especially at smaller group sizes. Accuracy decreases to 50.62% at a minimum group size of 5,000. Despite this reduction, fairness metrics improve dramatically at a group size of 10,000, where DP falls to 0.79% and ROD to 2.47%. These values represent a substantial reduction compared to both the initial dataset and the MData results. This reflects the MF algorithm's strength in mitigating structural and proxy-based biases by enforcing conditional independence, albeit at the cost of reduced predictive performance.

The socio-economic status attribute demonstrates the greatest challenge for MF. At a minimum group size of 5,000, a favorable balance is achieved, with reasonable accuracy (50.18%) and a notable reduction in DP (1.90%) and ROD (7.10%) compared to the original dataset. However, at larger group sizes (10,000 and 15,000), fairness performance for SES declines sharply, with DP and ROD increasing substantially (≈17–30%). This suggests that MF becomes unstable for attributes with high variability and imbalance when the minimum group size is too large. This

instability likely occurs because enforcing conditional independence in such sparsely populated subspaces distorts the distribution excessively.
Overall, the MF results indicate that the minimum group size plays a critical role in shaping the trade-off between fairness and accuracy. A group size of 10,000 produces the most consistent results for gender and race, while 5,000 yields more reliable outcomes for socio-economic status. The stability of results between group sizes 10,000 and 15,000 for gender and race also suggests a saturation effect, indicating that additional increases in group size do not lead to further fairness improvements.

## 3.3 Comparison of Results

A comparative analysis of the two fairness-repair algorithms—MData and MF—shows that both approaches are capable of reducing discrimination across the sensitive attributes of gender, race, and socio-economic status. However, their behavior and performance differ substantially due to their underlying mechanisms. As summarized in Tables 2 and 3, the MData algorithm consistently achieves higher predictive accuracy than MF for all sensitive attributes. This pattern is clearly visible in Figure 2, where MData maintains accuracy levels above 64% for gender and race and approximately 64% for socio-economic status, regardless of the discrimination threshold $\tau$. In contrast, the MF algorithm exhibits a notable decline in accuracy for race and socio-economic status at smaller minimum group sizes (5,000 and 10,000), with only partial recovery at the 15,000-group configuration.

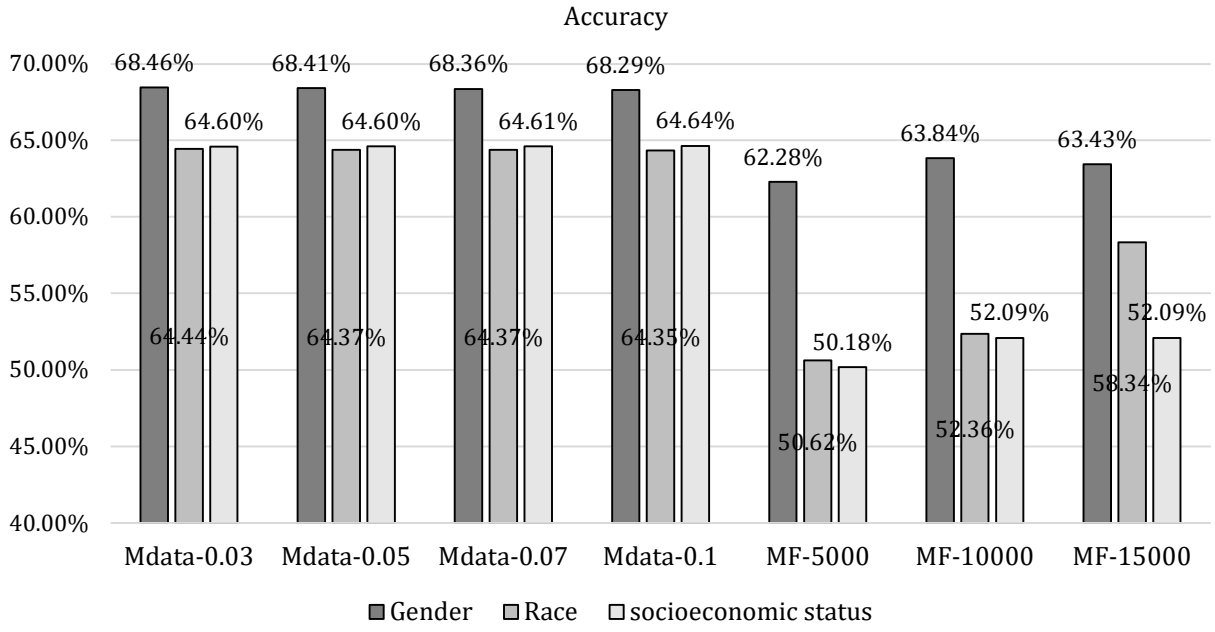


Figure 2: Accuracy comparison of the algorithms

With respect to fairness, the comparison reveals more nuanced behavior. As shown in Figure 3, MData produces consistently low ROD values for gender and race across all thresholds—especially at $\tau = 0.03$, where ROD falls below 2% for gender and remains under 6% for race—while for socio-economic status, ROD values remain higher (≈17–20%). This indicates that although MData effectively reduces $\Delta P$, group-level disparities persist due to SES imbalance and the depth of block-set segmentation.

By contrast, MF demonstrates stronger fairness improvement for gender and race at a minimum group size of 10,000. It achieves the lowest ROD values among all configurations (3.77% for gender and 2.47% for race). These findings highlight MF's ability to correct structural and proxy-related biases through distributional reconstruction and conditional independence enforcement.

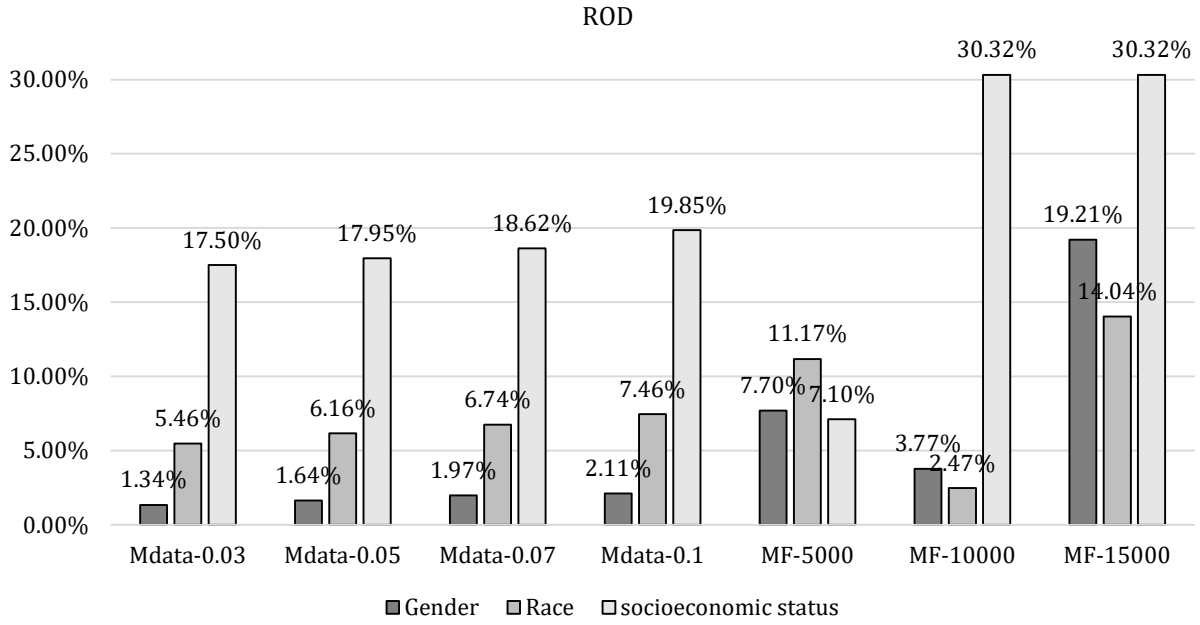


Figure 3. ROD comparison of the algorithms

For socio-economic status, MF displays mixed performance. At a minimum group size of 5,000, MF substantially reduces ROD from 27.54% to 7.10%, outperforming MData in fairness for this attribute, although at the cost of reduced accuracy. At larger group sizes (10,000 and 15,000), SES fairness deteriorates, with ROD rising to ≈30%, suggesting that SES requires smaller, more granular grouping for stable factorization-based repair.

DP results exhibit a similar pattern. As shown in Figure 4, the DP values achieved by MData for gender at $\tau = 0.03$ closely align with those produced by MF at a minimum group size of 10,000. This indicates comparable group-level parity under optimal configurations. For race and socio-economic status, MF again provides stronger fairness improvements at minimum group sizes of 10,000 and 5,000, respectively.

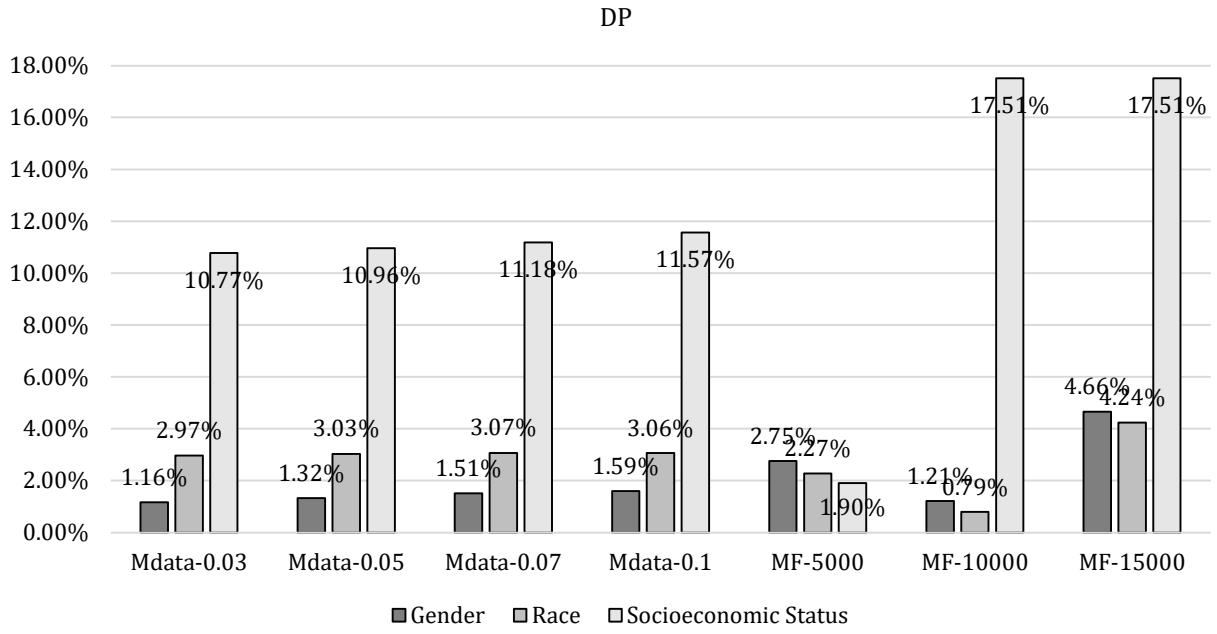


Figure 4. DP comparison of the algorithms

Overall, both algorithms successfully reduce discrimination while maintaining acceptable accuracy relative to the original dataset. The MData algorithm reduces fairness metrics for gender and race to values close to zero through threshold-based label modifications. However, its ability to improve fairness for socio-economic status is constrained by the maximum modification limit. In contrast, MF achieves substantial fairness improvements for gender and race at a group size of 10,000. It also produces reasonable results for socio-economic status at a group size of 5,000, though often accompanied by lower accuracy.

A key observation is that MData maintains a stable accuracy range between 64% and 69%, frequently surpassing the baseline. The MF algorithm, however, exhibits noticeable accuracy reductions for race and SES due to the statistical complexity of distribution-level reconstruction.

Taken together, these findings suggest that MData is the preferred algorithm when maintaining predictive accuracy is the primary objective, particularly when gender is the sensitive attribute of interest. Conversely, MF becomes advantageous when the primary goal is stronger fairness correction—especially for race and socio-economic status—provided that a moderate decrease in accuracy is acceptable.

## 4 Conclusions and Future Work

The MData algorithm effectively eliminates the causal component of discrimination (ΔP) while maintaining stable—or in some cases improved—predictive accuracy. Among the three sensitive attributes examined, gender shows the strongest fairness gains, followed by race, whereas socio-economic status constitutes the most challenging dimension due to demographic imbalance and constraints on the allowable modification rate. Overall, MData demonstrates strong performance as a structured fairness-repair method, particularly when discrimination thresholds and maximum modification rates are carefully calibrated.

Applying the MData and MF algorithms to the TIMSS dataset confirms their effectiveness in mitigating discrimination in predictive models of student performance. This is particularly evident for biases associated with gender, race, and socio-economic status. Initial ΔP values, obtained using the Non-Discrimination Certification Algorithm within MData, were 0.1351 for gender, 0.1646 for race, and 0.1374 for socio-economic status. As shown in Table 2, MData substantially reduced these ΔP values for gender and race to below the user-specified thresholds of 0.03, achieving near-optimal fairness. For socio-economic status, a slightly higher threshold of 0.04 was necessary to preserve accuracy, resulting in a modest trade-off. Predictive performance remained satisfactory, with accuracy levels of ≈68% for gender and 64% for both race and socio-economic status. Moreover, MData successfully lowered the secondary fairness metrics—ROD and DP—close to zero for gender and race. Socio-economic status, however, continued to exhibit weaker fairness outcomes due to limitations on the maximum allowable modification percentage, suggesting that further parameter tuning may be required to enhance performance for this attribute. The MF algorithm, which employs matrix factorization, exhibited more heterogeneous results across the sensitive attributes. While its overall accuracy was lower than that of MData—particularly for race and socio-economic status—it achieved stronger fairness improvements. Notably, MF produced substantially lower ROD values for gender at a minimum group size of 10,000. For socio-economic status, it achieved a balanced combination of accuracy and fairness at a minimum group size of 5,000. For gender and race, MF attained an acceptable fairness–accuracy trade-off at a group size of 10,000, after which the fairness gains plateaued, indicating diminishing returns.

These findings highlight the complementary strengths of the two algorithms. MData maintains higher predictive accuracy, especially for gender, whereas the MF algorithm delivers superior fairness across all attributes when appropriately configured. Both methods significantly reduce discrimination relative to the unmodified dataset while retaining acceptable predictive accuracy, demonstrating their practical utility for fairness enhancement in educational data such as TIMSS.

In summary, MData is best suited for scenarios where maintaining high predictive accuracy is prioritized, particularly when gender is the primary sensitive attribute. In contrast, MF is preferable when broad fairness improvements are required for gender, race, and socio-economic status. Collectively, these results underscore the potential of causal pre-processing techniques to address inequities in educational assessment.

For future work, one promising direction is the integration of the two algorithms to jointly optimize both predictive accuracy and fairness. A sequential approach—applying MData to reduce causal discrimination, followed by MF to further refine group-level fairness metrics—could leverage the strengths of both methods.

An alternative avenue is reinforcement learning, which can develop an adaptive agent to learn optimal policies for selecting and applying fairness-repair algorithms. Such an agent could operate at the row level or partition level, dynamically balancing fairness and accuracy. The authors are currently investigating this direction and intend to present a full methodological framework in subsequent work.

Additionally, examining the combined influence of multiple sensitive attributes could uncover more complex discrimination patterns, since current algorithms treat secondary sensitive attributes as regular features and handle only one sensitive dimension per execution. In more intricate fairness scenarios—such as socio-economic status with multiple categories or target variables defined on continuous or multi-range scales—future research should develop methods capable of effectively addressing multi-valued sensitive attributes and non-binary targets.

# References


[1] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal machine learning: A survey and open problems," *arXiv preprint arXiv:2206.15475,* 2022.

[2] N. McJames, A. Parnell, Y. C. Goh, and A. O'Shea, "Bayesian causal forests for multivariate outcomes: Application to Irish data from an international large scale education assessment," *arXiv preprint arXiv:2303.04874,* 2023.

[3] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Interventional fairness: Causal database repair for algorithmic fairness," in *Proceedings of the 2019 International Conference on Management of Data*, 2019, pp. 793-810.

[4] C. Su, G. Yu, J. Wang, Z. Yan, and L. Cui, "A review of causality-based fairness machine learning," *Intelligence & Robotics,* vol. 2, no. 3, pp. 244-274, 2022.

[5] M. T. Islam, A. Fariha, A. Meliou, and B. Salimi, "Through the data management lens: Experimental analysis and evaluation of fair classification," in *Proceedings of the 2022 International Conference on Management of Data*, 2022, pp. 232-246.

[6] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Capuchin: Causal database repair for algorithmic fairness," *arXiv preprint arXiv:1902.08283,* 2019.

[7] Y. Wu, L. Zhang, and X. Wu, "On discrimination discovery and removal in ranked data using causal graph," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2536-2544.

[8] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in data release," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1335-1344.

[9] Qureshi, B., F. Kamiran, A. Karim, and S. Ruggieri, "Causal discrimination discovery through propensity score analysis," Information Sciences, vol. 12, pp. 1–15, 2016.

[10] H. Nilforoshan, J. D. Gaebler, R. Shroff, and S. Goel, "Causal conceptions of fairness and their consequences," in *International Conference on Machine Learning*, 2022: PMLR, pp. 16848-16887.

[11] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science,* vol. 181, pp. 526-534, 2021.

[12] H. Munir, B. Vogel, and A. Jacobsson, "Artificial intelligence and machine learning approaches in digital education: A systematic revision," *Information,* vol. 13, no. 4, p. 203, 2022.

[13] L. N. Glassow, K. Y. Hansen, and J.-E. Gustafsson, "Does socioeconomic sorting of teacher qualifications exacerbate mathematics achievement inequity? Panel data estimates from 20 years of TIMSS," *Studies in Educational Evaluation,* vol. 77, p. 101255, 2023.

[14] A. Elouafi, I. Tammouch, S. Eddarouich, and R. Touahni, "Evaluating various machine learning methods for predicting students' math performance in the 2019 TIMSS," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 34, no. 1, pp. 565-574, 2024.

[15] P. Bäckström, "Peer Effects in Education: Theoretical Synthesis of Frame Factor Theory and Opportunity to Learn using TIMSS 2011," *International Journal of Educational Research,* vol. 121, p. 102229, 2023.

[16] N. Teig and T. Nilsen, "Profiles of instructional quality in primary and secondary education: Patterns, predictors, and relations to student achievement and motivation in science," *Studies in Educational Evaluation,* vol. 74, p. 101170, 2022.

[17] A. Inoue and R. Tanaka, "The rank of socioeconomic status within a class and the incidence of school bullying and school absence," *Economics of Education Review,* vol. 101, p. 102545, 2024.

[18] E. Salinas, A. Stancel-Piątak, and I. Nicaise, "Who caters for the teacher's needs? The role of teachers' working conditions for students' achievement and motivation in selected TIMSS 2015 countries," *International Journal of Educational Research Open,* vol. 3, p. 100196, 2022.

[19] B. Fishbein, P. Foy, and L. Yin, "TIMSS 2019 user guide for the international database," *Hentet fra https://timssandpirls. bc. edu/timss2019/international-database,* 2021.

[20] B. Salimi, B. Howe, and D. Suciu, "Data management for causal algorithmic fairness," *arXiv preprint arXiv:1908.07924,* 2019.