Applying DebiasEd: A Package for Mitigating Unfairness in Educational Data

Frank Stinar
University of Illinois Urbana–Champaign
fstinar2@illinois.edu

Jade Mai Cock
EPFL
jade.cock@epfl.ch

René F. Kizilcec
Cornell University
kizilcec@cornell.edu

Tanja Käser
EPFL
tanja.kaeser@epfl.ch

ABSTRACT

Educational data mining has augmented learning processes and improved outcomes within and outside of traditional classrooms. However, due to the diversity of learners, computational methods could be biased against different (often protected) groups of learners. Due to the biases, researchers have become increasingly interested in understanding and reducing the bias within their educational data and models. In this tutorial, we present and discuss how to use several modern unfairness and bias mitigation techniques encapsulated within the ***DebiasEd*** Python package with a focus on ease of implementation and generalizability for education researchers.

Keywords

bias mitigation, , educational data, pre-processing, in-processing, post-processing, reproducibility

1. INTRODUCTION

Educational data mining has proven to be a transformative field through harnessing data to improve many dimensions of education. [4, 7]. Due to the success of educational modeling, predictive models have become core to many modern educational systems [24, 2]. However, these models have also brought with numerous concerns about the fairness of the models are the biases that are within the data being used to train the models [20, 18]. To address these concerns, researchers in related fields and within education have developed methods to evaluate and mitigate unfairness within computational systems [14, 15, 3].

The machine learning and artificial intelligence research communities have developed many unfairness mitigation techniques to handle different types of biases within data, models, and model outcomes [21]. These techniques are considered to be either pre-processing (e.g., transforming training data), post-processing (e.g., threshold model outcomes), or in-processing (e.g., implementing multiple loss functions during training). These techniques have all proven their success in their respective domains; however, their usefulness within education is underexplored. Focusing on the unique challenges of education data (e.g., sensitivity to human-computer interaction [13], multimodality [11], non-representative populations [22], etc.), the tutorial will teach participants how to use a subset of modern unfairness mitigation techniques that are generally applicable to educational data.

This tutorial teaches participants how to use state-of-the-art approaches to mitigating bias in educational models using either predetermined or personal datasets. These approaches focus on transforming training data, changing outcomes, or implementing new objectives when training educational models. The tutorial focuses on the different ways to harness the approaches that are encapsulated within the ***DebiasEd*** Python package.

2. BACKGROUND

The tutorial focuses on implementing multiple techniques on reducing unfairness within educational data and educational machine learning modeling pipelines. These techniques are split based on where they are implemented into a machine learning pipeline as either pre-processing, in-processing, or post-processing techniques.

Pre-processing: Pre-processing techniques attempt to mitigate biases within the machine learning pipeline by transforming and modifying the data before training the educational models [21]. Specifically, from an original set of data and labels, a pre-processing method produces a new set of data and labels wherein some type of bias has been mitigated. Then, the newly created dataset is used to train models. Only transforming the data has advantages to learning scientists, as the techniques can be applied irrelevant to which machine learning model is used afterwards. Pre-processing techniques generally either modify or massage labels [17, 1], sample datasets differently [8, 10, 16], or transform the features to reduce different types of biases within the data [26, 19].

In-processing: In-processing techniques replace the original machine learning model with an alternative that is debiased.

Jade Cock, Frank Stinar, Rene Kizilcec, and Tanja Kaser. Applying DebiasEd: A Package for Mitigating Unfairness in Educational Data. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 695–698. International Educational Data Mining Society (2025).

© 2025 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
https://doi.org/10.5281/zenodo.15870322

Table 1: Tutorial Schedule

Part	Description	Timing
I	Conceptual Overview	0:00 - 0:30
II	Walk through a toy example with well known open source educational dataset	0:30 - 1:15
III	Showcasing the different ways of using the package, tailored to the needs of the participants	1:15 - 2:00
	Break	
IV	Exploration and application of the tool to researchers' own data	2:30 - 3:15
V	Presentation of results in groups	3:15 - 3:55
VI	Closing remarks	3:55 - 4:00

In-processing methods are often model-specific and many techniques exist such as imposing fairness constraints alongside loss functions, training an adversarial model alongside the original, or training an ensemble of classifiers for each group within the data to name a few [15].

Post-processing: Post-processing techniques mitigate bias after the machine learning model has been trained. These techniques are valuable to learning scientists (similarly to many pre-processing techniques) since they can be implemented within already existing educational pipelines. These techniques can also be adjusted for specific definitions of fairness [5], and post-hoc criteria to align with educational researchers' goals [23, 25].

3. GOALS

At the end of the tutorial, participants will be able to:

- Identify the potential types of biases that may be present in their data, and how it may propagate through their models
- Apply a range of mitigation techniques using our ***DebiasEd*** package to their own data
- Analyse the impact of bias mitigation techniques through different algorithmic fairness metrics [21]
- Present results in terms of fairness performances

4. PLAN

We will start the tutorial with an overview of how historical and societal biases can affect learners' education, how the models we develop as learning scientists can propagate these biases, and how using ***DebiasEd*** can help mitigating these biases. The walk through will showcase a specific open source dataset (the student performance dataset [12]), while the rest of the tutorial will focus on applying ***DebiasEd*** to the participants' own dataset, or alternatively a range of open source EDM datasets.

4.1 Software

We developed an open source Python package in which users can feed their data through our graphical interface, or through our API which contains the standalone mitigation techniques (API), or the entire evaluation pipeline, similar to GridSearchCV in Scikit-learn [6]. We will provide a step-by-step guide on how to install ***DebiasEd***, even if Python has not been installed prior to the workshop. Additionally, we will set up Jupyter notebooks on Google Collab in the cases where participants would run into installation problems.

4.2 Data Sets

The first part of the tutorial will showcase the Student Portuguese Performance dataset (SPP) [12] in which student performances in two Portuguese high schools were tracked [9]. Specifically, the SPP dataset contained grade information from 649 students taking Portuguese classes. The dataset contained 33 features related to academic scores or demographics (e.g., sex, age, and familial education). Specifically, 383 of the students were female and 266 were male. We considered the female sex as the protected group in this dataset.

In the second part of the tutorial, participants will be encouraged to integrate ***DebiasEd*** to their own datasets, models, and/or pipelines.

4.3 Tutorial Organization

Part I: Conceptual Overview The tutorial will start with the presentation of cases in which historical biases were propagated through algorithms, and affected learners' academic journey. We will then share an overview of the different types of biases there exists in education as shown through both learning science and machine learning literature. Finally, we will present ***DebiasEd***. Specifically, we will summarize the types of bias mitigation techniques we implemented, as well as the different ways the package can be used.

Part 2: Walk through We will demonstrate how to analyze the data prior to training our models, to identify ahead of time the type of biases there may exist in the data using ***DebiasEd***, and to select what mitigation techniques to start with. We will then show how to use graphical interface to retrieve a deployable model, and analyze its fairness performances.

Part III: Showcase For those who are more at ease with programming/want to implement mitigation techniques directly into their own ecosystem, we will demonstrate how to use the pipeline through our own evaluation cross validation pipeline, or as standalone pieces in participants' own code.

Part IV: Exploration We will support participants in applying diverse mitigation techniques to their own data sets using ***DebiasEd***, and observe what effect it has on their own model's fairness performances.

Part V: Presentation We will make lightning presentations of 5 minutes about what type of data participants usually work with, the types of biases they are more prone to run into, and what type of mitigation techniques worked best for them.

Part VI: Closing remarks We will put an emphasis on how important it is to consider algorithmic fairness throughout the development of learner models, as well as discuss the differences between equality and equity.

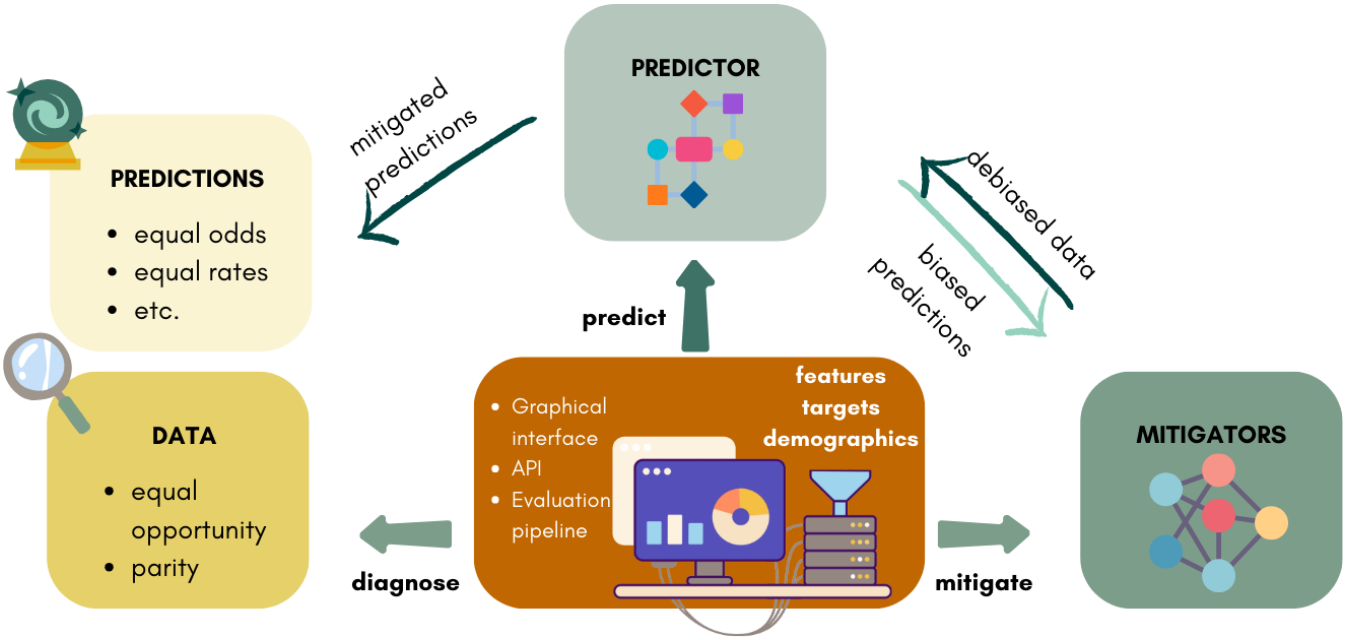

Figure 1: Package Structure. The only input it requires are the features, the targets (for classification purposes), and the demographic attributes. You can upload a CSV directly into our graphical interface, or use python directly to feed it into *Bias in the classroom*'s pipeline, or use the mitigation techniques individually through our API. Using the graphical interface and/or the pipeline will output a deployable model, as well as its classification and fairness performances. Using the pipeline or the API will enable you to retrieve the production/deployable model, the parameters of these models, and the predictions of these models in a self contained way.

5. REFERENCES

[1] I. Alabdulmohsin, J. Schrouff, and O. Koyejo. A reduction to binary approach for debiasing multiclass datasets. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[2] C. J. Arizmendi, M. L. Bernacki, M. Raković, R. D. Plumley, C. J. Urban, A. T. Panter, J. A. Greene, and K. M. Gates. Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, 55(6):3026–3054, Aug. 2022.

[3] R. S. Baker and A. Hawn. Algorithmic Bias in Education. preprint, EdArXiv, Mar. 2021.

[4] R. S. Baker and P. S. Inventado. Educational Data Mining and Learning Analytics. In J. A. Larusson and B. White, editors, *Learning Analytics: From Research to Practice*, pages 61–75. Springer New York, New York, NY, 2014.

[5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[6] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[7] C. Romero and S. Ventura. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, Nov. 2010.

[8] J. Chakraborty, S. Majumder, and T. Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 429–440, 2021.

[9] P. Cortez and A. Silva. USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE.

[10] D. Dablain, B. Krawczyk, and N. Chawla. Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. *Discover Data*, 2(1):4, Apr. 2024.

[11] S. K. D'mello and J. Kory. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Computing Surveys*, 47(3):1–36, Apr. 2015.

[12] D. Dua and C. Graff. Student performance data set, 2014. Accessed: 20.02.2025.

[13] S. Finkelstein, E. Yarzebinski, C. Vaughn, A. Ogan, and J. Cassell. The Effects of Culturally Congruent Educational Technologies on Student Achievement. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik,

editors, *Artificial Intelligence in Education*, volume 7926, pages 493–502. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Lecture Notes in Computer Science.

[14] B. Friedman and C. College. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3), 1996.

[15] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.*, 1(2), June 2024.

[16] V. Iosifidis and E. Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. In *Bias in Information, Algorithms, and Systems*, CEUR Workshop Proceedings, pages 24–29. CEUR Workshop Proceedings, 2018. Funding information: The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA under grant No. 339233 and by the German Research Foundation (DFG) project OSCAR (Opinion Stream Classification with Ensembles and Active leaRners) No. 317686254.; 2018 International Workshop on Bias in Information, Algorithms, and Systems, BIAS 2018 ; Conference date: 25-03-2018 Through 25-03-2018.

[17] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[18] R. F. Kizilcec, O. Viberg, I. Jivet, A. Martinez Mones, A. Oh, S. Hrastinski, C. Mutimukwe, and M. Scheffel. The Role of Gender in Students' Privacy Concerns about Learning Analytics: Evidence from five countries. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, LAK2023, pages 545–551, New York, NY, USA, 2023. Association for Computing Machinery. event-place: Arlington, TX, USA.

[19] P. Lahoti, K. P. Gummadi, and G. Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345, 2019.

[20] C. Li, W. Xing, and W. L. Leite. Do Gender and Race Matter? Supporting Help-Seeking with Fair Peer Recommenders in an Online Algebra Learning Platform. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, pages 432–437, New York, NY, USA, 2022. Association for Computing Machinery. event-place: Online, USA.

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35, July 2022.

[22] L. Paquette, J. Ocumpaugh, Z. Li, A. Andres, and R. Baker. Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, 12(3):1–30, Oct. 2020. Section: EDM 2020 Journal Track.

[23] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5684–5693, Red Hook, NY, USA, 2017. Curran Associates Inc.

[24] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett. Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2):249–255, Apr. 2007.

[25] P. Snel and S. van Otterloo. Practical bias correction in neural networks: a credit default prediction case study. *Computers and Society Research Journal*, (3), 2022.

[26] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.