Unbias me! Mitigating Algorithmic Bias for Less-studied Demographic Groups in the Context of Language Learning Technology

[α]Nathalie Rzepka, [α]Linda Fernsel, [β]Hans-Georg Müller, [α]Katharina Simbeck, & [γ]Niels Pinkwart

[α] Department of Computer Science, University of Applied Science Berlin
[β] Department of German Studies, University Potsdam
[γ] Department of Computer Science, Humboldt-University Berlin

Corresponding author: nathalie.rzepka@htw-berlin.de

Abstract. Algorithms and machine learning models are being used more frequently in educational settings, but there are concerns that they may discriminate against certain groups. While there is some research on algorithmic fairness, there are two main issues with the current research. Firstly, it often focuses on gender and race and ignores other groups. Secondly, studies often find algorithmic bias in educational models but don't explore ways to reduce it. This study evaluates three drop-out prediction models used in an online learning platform to teach German spelling skills. The aim is to assess the fairness of the models for (in part) less-studied demographic groups, including first spoken language, home literacy environment, parental education background, and gender. To evaluate the models, four fairness metrics are used: predictive parity, equalized odds, predictive equality, and ABROCA. The study also examines ways to reduce algorithmic bias by analyzing the models at each stage of the machine learning process. The results show that all three models had biases that affected the fairness of all four demographic groups to varying degrees. However, the study found that most biases could be mitigated during the process. The methods used to mitigate bias differed by demographic group, and some methods improved fairness for one group but worsened it for others. Therefore, the study concludes that reducing algorithmic bias for less-studied demographic groups is possible, but finding the right method for each algorithm and demographic group is crucial.

Rzepka, N., Fernsel, L., Müller, H., Simbeck, K., & Pinkwart, N., (2023) Unbias me! Mitigating Algorithmic Bias for Less-studied Demographic Groups in the Context of Language Learning Technology. Computer-Based Learning in Context, 6 (1), 1-23.

Keywords: Algorithmic Bias, Discrimination, Fairness, Educational Data Mining, Prediction, Dropout Prediction, Language Learning.

1 Introduction

Machine learning algorithms have a significant impact on people's lives, as they make recommendations and aid decision-making. Therefore, it is crucial to ensure the fairness of these algorithms. Mehrabi et al. (2021) define fairness as "*the absence of any prejudice or favoritism towards an individual or group based on their inherent or acquired characteristics*". Unfortunately, algorithms have been found to be unfair in various contexts, including hiring processes (Köchling & Wehner, 2020), medical applications (Chen et al., 2021), and educational software (Baker & Hawn, 2021).

Due to the COVID-19 pandemic, there has been an increase in online learning, resulting in the use of techniques and tools from learning analytics and educational data mining (Goudeau et al., 2021; McClain et al., 2021). However, the move to digital learning has widened the digital divide, making it essential to ensure the fairness of the algorithms and models applied in education to prevent further discrimination (Goudeau et al., 2021; McClain et al., 2021). Biased models may result in certain users receiving incorrect interventions, no interventions when needed, or being prevented from progressing quickly due to incorrect assessments. To ensure fair machine learning models, model fairness must be measured, and models must be examined for causes of unfairness and refined to increase fairness.

Although fairness in educational contexts is already being examined (Anderson et al., 2019; Baker et al., 2020; Gardner et al., 2019; Riazy et al., 2020; Vasquez Verdugo et al., 2022), fairness is mainly determined concerning demographic groups such as gender, race, or migration background (Baker & Hawn, 2021). The fairness metrics that are often used are those with legal sanctions for non-compliance, which vary between countries, leading to different priorities. However, other demographic characteristics play an important role, especially in the education system, making it crucial to consider groups that are less frequently examined, such as parents' educational background.

This study evaluates three machine learning-based in-session dropout prediction models, which predict users' early dropout on an online platform for learning German spelling and grammar. The models are evaluated for four different and less frequently studied demographic groups, including gender, parental education background, first spoken language, and home literacy environment (HLE). To minimize bias and reduce potential harm to users, we apply actions to each possible source of harm following the seven sources of harm described by Suresh and Guttag (2021) in the related works section. Therefore, the research questions for this study are as follows:

RQ1: At what steps in the machine learning lifecycle of educational drop-out prediction models can discrimination arise?
RQ2: How can the fairness of drop-out prediction models be improved?
RQ3: How does the potential for improvement differ among demographic groups?

To be able to answer the research questions, we first summarize related work about fairness measurements, potential sources of discrimination and previous attempts at unfairness mitigation in the educational field. Then, we present our method, the dataset used, the dropout prediction models, and the approach to evaluate and mitigate discrimination. Next, we present the results of both the evaluation with regard to fairness (RQ 1) and the attempts to mitigate bias for each demographic group (RQ 2). Finally, we discuss our findings and compare the potential for improvement among demographic groups (RQ 3). We close the paper with relevant limitations and a conclusion.

2 Review of Related Literature

2.1 **Algorithmic fairness in educational contexts.** Fairness in education has been an area of research for more than 50 years due to the significant impact education can have on a person's life, as noted in Hutchinson and Mitchell's (2019) review. Researchers in the field of learning analytics (LA) use machine learning methods to model students, such as in predicting student success. However, when learning analytics are implemented in online learning environments or platforms, algorithmic discrimination can occur, as observed by Anderson et al. (2019), Baker et al. (2020), Gardner et al. (2019), Riazy et al. (2020), Rzepka et al. (2022a), and Vasquez Verdugo et al. (2022).

One type of learning analytics model is the dropout prediction model, which identifies students who are at risk of not completing a course or program. Dropout prediction models were created in response to high dropout rates in Massive Open Online Courses (MOOCs), allowing teachers to intervene before a student drops out of a course (Dalipi et al., 2018; Kloft et al., 2014; Xing & Du, 2019). Supervised learning algorithms, such as logistic regression (LR), support vector machines (SVMs), and decision trees (DTs), are commonly used to create these models (Dalipi et al., 2018; Liang et al., 2016; Prenkaj et al., 2020). Model features are typically derived from user clickstream data (Dalipi et al., 2018; Sun et al., 2019), and predictions can be made periodically throughout a course or at a significant point in time (Prenkaj et al., 2020).

Fairness audits have been conducted on student success prediction models in various educational contexts (Gardner et al., 2019; Hu & Rangwala, 2020; Riazy et al., 2020; Rzepka et al., 2022a). These audits have shown discrimination against students with declared disabilities (Riazy et al., 2020), performance disparities based on gender (Gardner et al., 2019; Hu & Rangwala, 2020), and home literacy environment (e.g., the number of books in a household) (Rzepka et al., 2022a). Gardner et al. (2019) and Hu & Rangwala (2020) indicate that discrimination is influenced by the specific course and topic to which the model is applied. Other studies have identified representational bias, where a lack of training data for a particular group leads to bias against that group, such as a different gender (Gardner et al., 2019) or students with declared disabilities (Riazy et al., 2020). In contrast, Christie et al. (2019) could not find any differences in their dropout model by race.

Despite the extensive research in this area, some groups of learners and factors have not been well studied, such as Indigenous learners, nonbinary or transgender learners, religion, socioeconomic status, native language, disabilities, age, parental educational background, and parent work that affects student mobility (Baker & Hawn, 2021), personality, and social identity (Belitz et al., 2021). Furthermore, the practice of aggregating several smaller groups into one larger group for testing fairness is being questioned (Baker & Hawn, 2021; Belitz et al., 2021), such as when studies combine all Latin Americans or Asian Americans into one group despite their diversity.

2.2 **Measuring Fairness.** Different types of group fairness can be measured based on the confusion matrix (Verma & Rubin, 2018). Fairness is usually analyzed through slicing analysis (Gardner et al., 2019), that is, by “[b]reaking down performance measures by different dimensions or categories of the data” (Sculley et al., 2018): a chosen quality metric is calculated for a reference group, usually the historically favored group, as well as for a group for which fairness is tested (Verma & Rubin, 2018). A model is considered fair under the definition of fairness when the quality metric is similar enough for both groups (Verma & Rubin, 2018). The similarity is assumed if the difference between the quality metrics for both groups does not exceed a chosen threshold that typically lies between 0.01 and 0.05 (Chouldechova, 2017; Franklin et al., 2022; Riazy & Simbeck, 2019).

We selected four metrics to examine in more detail as part of this study. In doing so, we specifically chose metrics that test for different inequalities and that, if they were to indicate a bias, would have different types of consequences for users. Groups against which the dropout prediction model is biased could receive fewer necessary or more unhelpful interventions. The concrete practical implications related to the dataset of this study are described in the methods under "Fairness Evaluation.". For an extensive overview of fairness definitions and the quality metrics they are based on, the reader can refer to (Verma & Rubin, 2018). Table 1 provides an overview of the fairness metrics, their definitions, measurements and formula (Verma & Rubin, 2018).

Predictive Parity (PP) requires equal positive predictive value (PPV), which describes the precision of the models (Verma & Rubin, 2018). Differences in this metric would suggest that an algorithm has disparate explanatory power for different groups: A lower PPV points to less reliable positive predictions. Equal Opportunity (EO) and Predictive Equality (PE) focus on error-related fairness. EO requires equal false negative rates (FNR), implying that fairness is given if an algorithm misses the same relative number of positive cases per group. PE requires equal false positive rates (FPR), meaning that fairness according to this definition is given if an algorithm produces the same ratio of wrongly assigned positive classes per group. Gardner et al. (2019) introduce ABROCA (“absolute between-ROC area”, AB) as a fairness metric specifically for slicing analysis (SA) of models in educational contexts. A smaller AB signifies greater fairness (Riazy et al., 2020). The ROC curves are defined by the TPR and FPR (Centor, 1991). AB is thus a combination of PE and EO and involves a trade-off between these two metrics. The advantages of conducting a slicing analysis with AB as a fairness metric are that it is independent of classification thresholds and that it adds unfairness toward reference and test groups instead of evening it out (Gardner et al., 2019).

Table 1. Fairness metrics with their definitions, measurement, and formula*.

Fairness Metric	**Definition**	**Measurement**	**Formula**
Predictive Parity (PP)	Groups have an equal positive predictive value (PPV)	Difference of PPV of two groups	$P(Y =1 \mid R =1,S = s1)= P(Y =1 \mid R =1,S = s2)$
Predictive Equality (PE)	Groups have equal false positive rates (FPR)	Difference of FPR of two groups	$P(R =1 \mid Y =0,S = s1)= P(R =1 \mid Y =0,S = s2)$
Equal Opportunity (EO)	Groups have equal false negative rates (FNR)	Difference of FNR of two groups	$P(R =0 \mid Y =1,S = s1)= P(R =0 \mid Y =1,S = s2)$
ABROCA	Groups have equal ratios between true positive rates (TPR) and false positive rates (FPR) over varying positivity thresholds.	Area between the ROC curves of two groups	$\int_0^1 \lvert ROC_b(t) - ROC_c(t) \rvert dt$, Here, ROC_b and ROC_c are two curves of two different models. t is the ROC curve’s threshold

We use the following notations for the formula:* **S *is the protected variable,* **Y** *is the attribute to be predicted, and* **R** *is the prediction outcome.*

2.3 **Sources of Unfairness.** Once bias is discovered through fairness metrics, understanding the kind of bias and where it arises enables effective bias mitigation (Suresh & Guttag, 2021). Early work divided sources of unfairness into three different categories (Friedmann & Nissenbaum, 1996): preexisting bias, technical bias and emerging bias. Preexisting bias is independent of the technical implementation and occurs in “social institutions, practices, and attitudes” (Friedmann & Nissenbaum, 1996). Technical bias can arise from algorithms as well as limited computer tools or imperfections in random number generation (Friedmann & Nissenbaum, 1996). Emergent bias arises when applying computer systems, for example, when a use context was not taken into consideration during the design stage of the system or when the user population changed over time (Friedmann & Nissenbaum, 1996). In our article we focus on bias within the machine learning process (Mehrabi et al., 2021; Mitchell et al., 2021; Silva & Kenney, 2019; Suresh & Guttag, 2021; van Giffen et al., 2022). In the following, we summarize the possible types of biases that can occur in machine learning.

When data are collected, they can accurately represent an existing historical bias, also referred to as a societal bias (Mitchell et al., 2021; van Giffen et al., 2022) or a systemic bias (Schwartz et al., 2022). This type of bias occurs, for example, when a model might learn to associate the word "nurse" with women (Suresh & Guttag, 2021).

Representational bias can occur due to problematic data sampling, e.g., when the model is later used on a different population than the population that was sampled from, when data are observed only for segments of a population (Suresh & Guttag, 2021), or when the population the model is trained and used on changes over time (van Giffen et al., 2022). Sampling too little data from minorities can also cause algorithmic discrimination by disadvantaging historically underrepresented groups (Chawla et al., 2002). This situation is referred to as statistical bias (Mitchell et al., 2021).

Measurement bias can stem from the steps of feature engineering and labeling: data commonly contain information that is measured easily, and some groups can be observed more easily and accurately than others (Suresh & Guttag, 2021). It is also particularly important to define a suitable target variable (Kizilcec & Lee, 2020). Previous work showed that the choice of target variable to use in educational settings comes with limitations and should not be made lightly (DeBoer et al., 2014; Duckworth & Yeager, 2015; Gašević et al., 2015; Schwartz & Arena, 2013).

Optimizing a model for a variable that is not appropriate has tremendous consequences for model validity. Moreover, a model definition contains assumptions regarding the relationship between features and labels. When these assumptions cannot be generalized to all groups of a population, aggregation bias can occur (Suresh & Guttag, 2021). Another type of bias is learning bias (or algorithmic bias) stemming from how a model optimizes, e.g., when the model maximizes overall accuracy at the expense of accuracy for some groups (Mehrabi et al., 2021; Suresh & Guttag, 2021). In addition, evaluation bias can include two factors that can hide or encourage unfairness, that are easily comparable or one-sided evaluation metrics and biased test data (Suresh & Guttag, 2021).

Finally, deployment bias can be introduced (Suresh & Guttag, 2021). For example, a model can be misused in applications to the real world, e.g., its results may be misinterpreted (Suresh & Guttag, 2021), or the model's biased results can influence the next batch of training data for the model, thus perpetuating existing bias in a feedback loop (Mansoury et al., 2020; Mehrabi et al., 2021). Deployment bias falls under Friedmann and Nissenbaum's definition of emergent bias.

2.4 **Bias Mitigation in LA.** Now that the various sources of bias have been presented, actions can be taken to mitigate the bias in a model. Deho et al. (2022) differentiate between preprocessing methods, in-processing methods, and postprocessing methods for unfairness mitigation. Approaches to preprocessing methods, for example, are described by Yu et al. (2021), who compare two models where one model included protected attributes and the other did not. They find no evidence for differences in model performance or fairness. Other recent approaches propose focusing on the feature selection process and integrating model training, evaluation, and optimization, which is called procedural fairness (Belitz et al., 2021). Belitz et al. (2021) propose methods to improve the feature selection process with trade-offs for different fairness metrics and accuracies. Adjustments in the sampling can also be counted among the preprocessing methods: Sha et al. (2022) find that applying class balancing techniques to oversample the minority group does increase the predictive accuracy of the minority group, while the majority group is not negatively affected. If minority groups are underrepresented in the training data, some studies suggest using case weights to increase user influence in the model learning process (Gardner et al., 2019; Ocumpaugh et al., 2014). However, Baker and Hawn (2021) suggest collecting more real data instead of using oversampling techniques.

Most in-processing methods use constraints to optimize fairness within learning algorithms (Calders & Verwer, 2010; Kamishima et al., 2012; Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017; Zemel et al., 2013). Riazy et al. (2020) improve the model fairness of an at-risk prediction model with the use of a prejudice remover (Kamishima et al., 2012) and a margin-based classifier (Zafar, Valera, Rogriguez, & Gummadi, 2017). While the type of model does not seem to influence fairness, discrimination can vary across datasets as well as within different random splits to create training and test data (Friedler et al., 2019). In addition to the progress made with in-processing methods to optimize model fairness, others suggest that the focus should be on more interpretable models so that the results are more transparent and understandable (Conati et al., 2018; Doshi-Velez & Kim, 2017).

Lee and Kizilcec (2020) approach unfair models and achieve EO by postprocessing a random forest by setting protected group-specific classification thresholds. Bias mitigation techniques can lead to a trade-off between fairness and accuracy. However, Zliobaite (2015) notes that the comparison between two models should take into account acceptance rates. However, little research has focused on mitigating bias in models in the learning analytics context.

3 Methodology

This paper aims to evaluate and improve discriminative drop-out prediction models and make them fairer by optimizing all steps in the machine learning lifecycle. We focus on the in-session dropout prediction models described by Rzepka et al. (2022b), which predict the dropout of users on the Orthografietrainer.net platform. The Orthografietrainer.net platform is a platform to acquire German spelling and grammar skills. It is used in Germany, Austria and Switzerland mainly in schools as part of blended classroom scenarios. The platform is suitable for users from the 5th grade and offers exercises in different orthographical areas such as capitalization, comma formation, separated and combined spelling, sounds and letters, or grammar. Most students use the platform between fifth and ninth grade. Usually, assignments are given as mandatory homework. One exercise set consists of 10 sentences. However, with each incorrect answer, some more sentences are added to the exercise set, so that it can easily become a quite long exercise with up to 60 sentences. The spelling or grammar exercises build on each other so that completion of the set is preferred form a didactic point of view. When a user drops out from a training session, he or she is allowed to continue the exercise later in time. If a homework assignment is not done at all, there is usually a reaction from the teacher.

3.1 **In-Session Dropout Prediction Models.** In contrast with dropout prediction models in MOOCs, Rzepka et al. (2022b) present dropout prediction models that predict early dropout within a session. The in-session dropout prediction models provide the opportunity to offer in-session interventions. The models recalculate after each sentence the user provides and include all sentences the users has responded to thus far. Thus, the models adjust to the user's performance during the session.

The models in Rzepka et al. (2022b) were trained with clickstream data from the Orthografietrainer.net platform of more than 52,000 users from more than 181,000 sessions. Those sessions took place between March and April in 2020. The data included 24 one-hot encoded features that can be summarized in three categories: demographic data (e.g., class level, gender, user attribute, how long is the user registered), session information (e.g., date and time, count of pending tasks, count of incorrect answers, count of multiple incorrect answers of the same sentence, count of earlier interruptions), and information about the learning subject (e.g., field of grammar, difficulty). The different models tested included a decision tree classifier (DTE) and k-nearest neighbor (KNN) that are implemented using the sklearn library. Furthermore, a multilayer perceptron (MLP) using the tensorflow library was implemented. The DTE was incorporating entropy as the criterion to measure split quality and a maximum of depth of the tree of 5. KNN used a default of 2 neighbors for n_neighbors parameter.

The first spoken language is the language the user has learned at home as a first language. The data were split by users whose first spoken language was German and users whose first spoken language was not German. Both the level of parents' education and the migration background (indicated by the first spoken language), are known to influence children's literacy (Carroll et al., 2019; Lee & Burkam, 2007; Steinlen & Piske, 2013). Steinlen and Piske (2013), for example, find that children whose first language is not German have poorer results on German language exams.

The HLE variable was obtained during the survey by asking participants about the number of books in the household. This question is often included in questionnaires that measure cultural capital (Noble & Davies, 2009) and can be used to determine the home literacy environment. Previous research linked the literacy skills of children to the HLE feature (Griffin & Morrison, 1997; Sénéchal & LeFevre, 2002). User could choose between the options "less than 10", "11-50", "51-100", "more than 100". Here, we consider edge cases only and split the data by users in which households have fewer than 10 books or users in which households have more than 100 books. Users in households with 10 to 100 books are not considered.

In total, 2749 users answered the survey from March to June 2020. Figure 1 shows the distribution of sessions in the respective demographic categories and by dropout. Table 2 shows the distribution of users among the subgroups.

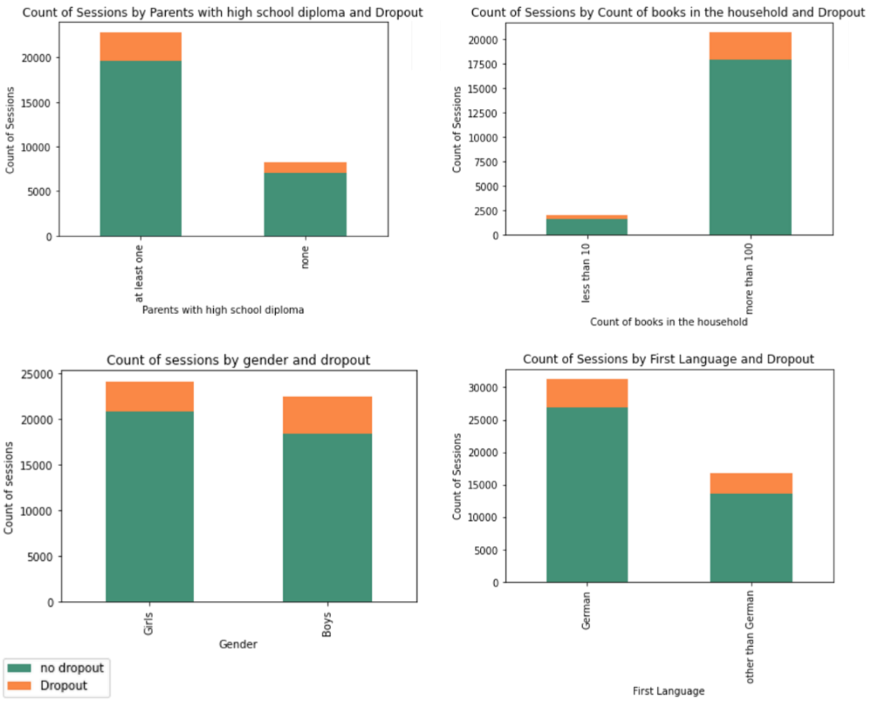

Figure 1. Distribution of session per demographic group and by dropout.

The evaluation was performed for each sentence (n=60), each machine learning model, and four different fairness metrics, which are described in related work and in Rzepka et al. (2022a): equal opportunity (EO), predictive equality (PE), ABROCA (AB), and predictive parity (PP). If a bias is detected by one of these metrics, this has different consequences for the users on the platform: EO describes the probability of missing a user who drops out. If interventions are being implemented for at-risk students, these users would not receive the intervention they need. PE, on the other hand, can be explained by the probability of a false alarm. Users would receive interventions, although they would not be necessary. If interventions hinder users from proceeding on time or if additional simple exercises need to be carried out and the student gets bored, this would impact a user's learning experience negatively. AB describes the area between two ROC curves and should be minimal. The ROC curves are defined by the TPR and FPR, where the TPR can be explained as the probability of detection. It is thus a combination of PE and EO and involves a trade-off between these two metrics. PP describes the precision of the models. Differences in PP would suggest the models have disparate explanatory power.

Table 2. Distribution of users per demographic group.

Demographic variable	**Majority group**	**Count of majority group**	**Minority group**	**Count of minority group**
Gender	Male	1339	Female	1325
Parental education background	At least one parent with high school diploma	1206	No parent with high school diploma	464
First spoken language	German as a first language	1683	Not German as a first language	1066
HLE	More than 100 books estimated	1092	Less than 10 books estimated	149

A model is considered fair if the value of the metric is below a certain threshold. In previous studies, different thresholds were used between 0.01 and 0.05 (Chouldechova, 2017; Gardner et al., 2019; Riazy & Simbeck, 2019). In our previous work we chose 0.03 and 0.05 (Rzepka et al., 2022a). We now decide to define a threshold of |0.05| to delineate fair from unfair to be able to focus on the larger biases to mitigate.

3.2 **Improving Fairness.** In this paper, the goal is to improve the fairness of the in-session dropout prediction models described previously and in Rzepka et al. (2022b). To do so, we analyze the models according to the seven sources of harm by Suresh & Guttag (2021), which point out potential sources of bias during the machine learning lifecycle. Step-by-step, we review the sources of harm and the steps of the ML lifecycle and apply the issue to the machine learning models at hand. To mitigate historical bias, we check whether demographic features are used in the training data of the models. This is the case with gender, which is obtained during the registration process and part of the training data. To improve historical bias, we thus delete gender from the training data. Representational bias arises when minorities are not well represented in the data (Suresh & Guttag, 2021). To reduce this bias, we balance the training data by oversampling minority groups using the synthetic minority oversampling technique (SMOTE), a technique which has been used in previous work (Sha et al., 2022). Measurement bias can arise when model features are determined. As the data do not include manual labels, we are checking only for a correlation between demographic features and other features that are still part of the training data by calculating Pearson's correlation coefficient. There is only one small correlation, which is between gender and the number of pending tasks (=0.11). Therefore, we deleted the number of pending tasks from the training data as well. Aggregation bias assumes that a one-size-fits-all solution is not always the best option but that a separate model for each demographic group is preferrable (Suresh & Guttag, 2021). In our case, we trained separate models for each demographic group. To mitigate learning bias, different values of parameters are tested to compare the impact on model fairness. In our work, we focus on three implementations: DTE, KNN, and MLP. For this, we tested different values of several parameters for each model and then compared the fairness. We trained the models using a 5-fold-cross-validation. Table 3 contains the models and the parameters we tested to improve model fairness.

Table 3. Models and parameters to be optimized during the process*.

Model	**Parameter**	**Description**	**Values**
DTE	Max_depth	Defines the maximum depth of the decision tree.	5,10,15,20,25
	Min_samples_leaf	Defines the minimum number of samples that are required per leaf node.	1,5,10,15,20,25
	Min_samples_split	Defines the minimum number of samples that are required to split an internal node.	1,5,10,15,20,25
KNN	N_neighbors	Defines the number of neighbors.	2,3,4,5,6,7,8,9,10
	weights	Defines the weight function that is used in the prediction.	uniform, distance
MLP	optimizer	Defines the optimizer applied in the model.	Adam, SGD
	loss	Defines the loss function that is applied.	Binary Crossentropy, Mean Squared Error, Hinge
	metrics	Defines the metrics by which model performance is evaluated.	Accuracy, AUC

*The Values column lists the values that were tested during the process.
Bold values are the default values used to train the dropout prediction model.

Evaluation bias arises when the test dataset does not represent the population to which the models are applied. In our case, evaluation bias could exist, as the models are evaluated on a subset of the data. As we can only work with the data at hand, this issue is covered in the Discussion section. When considering deployment bias, the scenario where the models are applied should be discussed. In our case, the models are trained with data from Orthografietrainer.net and will be applied only on this platform. That lowers the risk for deployment bias. However, we will cover deployment bias in the discussion section. Table 4 summarizes the action we took to mitigate bias in the models and improve model fairness. The source code of this study is published online (Rzepka, 2023).

Table 4. Sources of harm proposed by Suresh & Guttag (2021) and actions to mitigate bias.

Sources of Harm	**Mitigations**	**Application**
Historical Bias	Delete the feature of gender from the training data	Only for Gender
Representational Bias	Oversample all minority groups	For all groups
Measurement Bias	Exclude correlating the feature of number of pending tasks	Only for Gender
Aggregation Bias	Build two separate models for each group	For all groups
Learning Bias	Optimize parameters of the models as described in Table 3	For all groups
Evaluation Bias	Discuss the test dataset	In general
Deployment Bias	-	-

4 Results

For each demographic group and each machine learning implementation, we evaluated model fairness as in Rzepka et al. (2022a) and afterward tried to improve fairness by addressing all seven sources of harm described by Suresh & Guttag (2021). We define the threshold of |0.05| to decide whether we consider a model fair or unfair. Each value above the threshold is marked in red. Our results section is structured by demographic group. For each group, we show the evaluation results (i.e., the bias found before applying any mitigation techniques) and the evaluation results after applying the most promising mitigation technique. A mean of every ten sentences was calculated so that each table has six rows. Further we show results from mitigation techniques that are particularly surprising or interesting. All further results are described in the results section as well, however, the tables with the results are shown in the appendix to increase readability.

4.1 **Gender.** Table 5 shows the evaluation results of the fairness metrics for the attribute of gender. Unfairness above the threshold occurs only for the PE and AB metrics and only for higher sentence positions. In the temporary models, discrimination may change over time. As the user continues to perform exercises during the session, the models learn about his or her mistakes and that could improve the models.

Table 5. Evaluation results of the metrics for the attribute of gender (Rzepka et al., 2022a).

	EO			**PE**			**PP**			**AB**		
Model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.022	-0.029	-0.029	-0.036	-0.049	-0.049	0.008	0.008	0.008	-0.001	-0.010	-0.010
10 to 19	-0.011	-0.026	-0.026	0.003	-0.018	-0.018	0.009	0.007	0.007	0.007	0.004	0.005
20 to 29	-0.009	-0.021	-0.021	0.017	-0.011	-0.011	0.009	0.007	0.007	0.007	0.005	0.005
30 to 39	-0.003	-0.009	-0.009	0.008	-0.006	-0.005	0.006	0.006	0.006	0.009	0.001	0.002
40 to 49	-0.006	-0.006	-0.005	-0.079	-0.109	-0.109	-0.001	-0.003	-0.003	-0.003	-0.052	-0.052
50 to 60	0.001	0.003	0.003	-0.149	-0.155	-0.160	-0.003	-0.003	-0.003	-0.031	-0.079	-0.081

Appendix A1-A4 show the results after applying the fixes from historical bias (excluding the feature of gender), representational bias (oversampling), and measurement bias (excluding correlating features) and all those three mitigation techniques combined. As evident by comparing the results to the original evaluation, we could improve fairness with each of those mitigation techniques a little bit. The best results were achieved by combining all three techniques (Appendix A4). In all attempts, AB could be improved more effectively than PE. This outcome can be explained by the fact that the value of AB was only slightly above the threshold (still below |0.1|). Furthermore, these attempts did not influence model performances.

We tried to address aggregation bias by computing two separate models for both male and female users. As shown, when comparing the metrics of the model the threshold is not exceeded anymore (Table 6). Thus, we were able to reduce all discriminatory results. Comparing model performances shows that they are equally accurate.

Table 6. Fairness measurements for the attribute of gender after computing two separate models: one for boys and one for girls.

	EO			**PE**			**PP**			**AB**		
Model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.012	0.000	-0.018	-0.016	-0.004	-0.014	0.012	0.016	0.011	-0.001	-0.002	0.002
10 to 19	-0.006	0.004	-0.009	-0.011	0.003	-0.014	0.008	0.013	0.005	0.001	0.000	-0.002
20 to 29	0.006	0.004	-0.010	0.018	-0.001	-0.021	0.012	0.008	-0.001	0.002	-0.002	-0.006
30 to 39	0.001	-0.019	-0.001	0.019	-0.043	-0.012	0.011	-0.003	0.001	-0.001	-0.012	-0.005
40 to 49	-0.002	0.001	-0.023	0.034	0.041	-0.005	0.012	0.014	0.004	0.001	0.020	0.009
50 to 60	-0.005	0.002	-0.016	0.001	0.003	-0.015	-0.001	-0.001	-0.003	0.007	0.000	0.001

Table 7. Results of the metrics for the attribute of gender after attempting to mitigate learning bias for all three implementations.

Model	**DTE**				**KNN**				**MLP**			
	Max_depth: 25 Min_sample_leaf: 1 Min_sample_split: 2				N_neighbors: 3 Weights: uniform				Loss: Hinge Optimizer: SGD Metrics: Accuracy			
Metrics	EO	PE	PP	AB	EO	PE	PP	AB	EO	PE	PP	AB
Sentence												
2 to 9	-0.032	-0.033	0.006	0.000	-0.035	-0.048	0.007	-0.006	-0.031	-0.049	0.015	0.008
10 to 19	-0.031	-0.011	0.007	0.010	-0.030	-0.029	0.007	0.001	-0.026	-0.022	0.011	0.001
20 to 29	-0.023	-0.010	0.005	0.006	-0.022	-0.026	0.006	-0.002	-0.017	-0.003	0.014	0.008
30 to 39	-0.020	-0.003	0.005	0.008	-0.020	-0.015	0.005	0.002	-0.009	-0.023	0.010	0.000
40 to 49	-0.018	-0.064	-0.001	-0.023	-0.015	-0.046	0.001	-0.016	-0.005	-0.048	0.005	-0.015
50 to 60	-0.006	-0.115	-0.003	-0.055	0.002	-0.198	-0.007	-0.100	0.000	-0.037	0.007	-0.021

Finally, we addressed learning bias. Through parameter optimization we found the best combination of parameters, which is shown in Table 7. The KNN model improved in two fields, while the MLP parameter optimization was able to remove all discriminatory values. The DTE model did not improve significantly.

In conclusion, we found that for the attribute of gender, separate models contain the least discriminatory values while maintaining high accuracy in both models. Furthermore, optimizing parameters improved the models as well, while addressing historical, representational, and measurement bias did not improve model fairness very much.

4.2 **Parental Education Background.** The fairness evaluation of the attribute of parental education background shows that discrimination can be found with regard to the PE and AB metrics in the last ten sentences (Table 8).

Table 8. Evaluation results of the metrics for the attribute of parental education background (Rzepka et al., 2022a).

	EO			**PE**			**PP**			**AB**		
Model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.020	-0.022	-0.022	-0.002	-0.007	-0.007	0.002	0.002	0.002	0.009	0.008	0.008
10 to 19	-0.013	-0.022	-0.022	-0.029	-0.030	-0.030	-0.001	-0.001	-0.001	0.007	-0.004	-0.004
20 to 29	-0.005	-0.010	-0.010	0.030	0.0180	0.018	0.004	0.004	0.004	0.005	0.014	0.014
30 to 39	-0.005	-0.004	-0.004	0.047	-0.010	-0.010	0.007	0.004	0.004	0.004	-0.003	-0.003
40 to 49	-0.004	-0.005	-0.005	0.044	-0.009	-0.009	0.009	0.006	0.006	-0.007	-0.002	-0.002
50 to 60	-0.026	-0.021	-0.021	0.133	0.111	0.111	0.016	0.015	0.015	0.058	0.066	0.066

Appendix B1 shows the model evaluations after applying the SMOTE oversampling technique to balance the dataset and to mitigate representational bias. Compared with the original models, discriminatory values were mitigated for the AB metric, while the PE metric actually worsened. Furthermore, we found a decrease in model accuracy.

As an attempt to mitigate aggregation bias, we computed two separate models for users who have at least one parent with a high school diploma and users whose parents do not have a high school diploma. As shown in Appendix B2, the model biases worsened. The majority of the values are above the threshold, which means that the difference between the two models is large. Moreover, comparing model performances shows that the first models (high school diploma) are slightly better than the other ones (no high school diploma). This finding indicates that addressing aggregation bias by calculating two different models does not improve fairness and actually increases discrimination in favor of the advantaged group.

Addressing the learning bias, we could remove all discriminatory results except for one value in the MLP (Table 9). However, by adapting the parameters to make the models fairer, some of the model performances worsened slightly. The original decision tree model was between 80-85%, while the model at hand was irregular, but most time points were worse than 82%. The KNN model, on the other hand, improved slightly from 77% to more than 80% at some time points. The MLP, however, was the best model in the original evaluation, with 87%, but thereafter stayed below 80%. In summary, while optimizing the model parameters still had the best impact on model fairness, it was nevertheless necessary to contend with weaker model performances. The other attempts, especially constructing two separate models, was not effective for this demographic group.

Table 9. Results of the metrics for the attribute of parental education background after attempting to mitigate learning bias for all three implementations.

Model	**DTE**				**KNN**				**MLP**			
	Max_depth: 5 Min_sample_leaf: 5 Min_sample_split: 2				N_neighbors: 2 Weights: uniform				Loss: MSE Optimizer: Adam Metrics: Accuracy			
Metrics	EO	PE	PP	AB	EO	PE	PP	AB	EO	PE	PP	AB
Sentence												
2 to 9	-0.033	-0.016	0.001	0.009	-0.002	-0.019	-0.002	-0.009	-0.022	-0.006	0.002	0.007
10 to 19	-0.035	-0.010	0.001	0.012	-0.030	-0.008	0.000	0.011	-0.031	-0.023	0.000	0.010
20 to 29	-0.019	-0.009	0.001	0.005	-0.020	-0.010	0.000	0.005	-0.026	-0.047	-0.001	0.000
30 to 39	-0.023	0.040	0.007	0.031	-0.012	0.015	0.003	0.013	-0.020	-0.050	0.002	-0.016
40 to 49	-0.013	0.000	0.007	0.007	0.011	0.034	0.006	0.012	-0.010	-0.045	0.008	-0.023
50 to 60	-0.007	0.024	0.011	0.016	-0.005	0.000	0.006	0.002	-0.011	-0.005	0.010	0.001

4.3 **First Spoken Language.** The first spoken language attribute is an indicator of the users' migration background. The evaluation results show that only for the PE attribute are there some discriminatory values in the middle of the timeline in sentences 10 to 30 (Table 10). Our attempt to mitigate bias regarding the attribute of first spoken language had, in many fields, the same results as the attribute of parental education background. Table 11 shows that oversampling worsened the model evaluations for the minority group, namely, users whose first language was not German. We further found that model accuracies decreased as the model accuracies stayed between 0.75 and 0.80.

Table 10. Evaluation results of the metrics for the attribute of first spoken language (Rzepka et al., 2022a).

	EO			**PE**			**PP**			**AB**		
Model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.032	-0.028	-0.028	-0.042	-0.041	-0.041	0.010	0.013	0.013	0.002	-0.007	-0.007
10 to 19	-0.024	-0.025	-0.025	-0.050	-0.058	-0.058	0.004	0.003	0.003	-0.001	-0.017	-0.017
20 to 29	-0.024	-0.026	-0.026	-0.034	-0.057	-0.057	0.006	0.005	0.005	0.003	-0.016	-0.016
30 to 39	-0.022	-0.016	-0.016	-0.036	-0.030	-0.030	0.007	0.009	0.009	0.007	-0.007	-0.007
40 to 49	-0.019	-0.014	-0.014	-0.044	-0.046	-0.046	0.008	0.008	0.008	0.004	-0.016	-0.016
50 to 60	-0.016	-0.014	-0.014	-0.019	-0.041	-0.046	0.014	0.014	0.014	0.006	-0.013	-0.016

Table 11. Results of the metrics for the attribute of first spoken language after attempting to mitigate representational bias.

	EO			**PE**			**PP**			**AB**		
Model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.037	-0.040	-0.035	-0.027	-0.038	-0.014	0.012	0.009	0.007	0.008	0.001	0.011
10 to 19	-0.038	-0.035	-0.039	-0.038	-0.054	-0.024	0.004	0.002	0.001	0.002	-0.010	0.007
20 to 29	-0.034	-0.032	-0.043	-0.058	-0.045	-0.023	0.004	0.004	0.002	0.002	-0.006	0.010
30 to 39	-0.029	-0.027	-0.041	-0.039	-0.016	-0.015	0.012	0.008	0.003	0.014	0.005	0.013
40 to 49	-0.016	-0.025	-0.039	-0.020	-0.084	-0.062	0.017	0.007	0.000	0.005	-0.030	-0.012
50 to 60	-0.012	-0.026	-0.036	-0.040	-0.138	-0.128	0.020	0.005	-0.002	0.007	-0.056	-0.046

While the original evaluation shows discriminatory results only for the PE metric and the DTE and KNN models, oversampling led to additional discriminatory values for the MLP and the AB metric.

Furthermore, calculating two different models, one for users whose first spoken language was German and one for users whose first spoken language was not German, we again see a large difference between the models and eventually worse model evaluations (Appendix C1). Comparing the model accuracies, we find a slightly better model for users whose first spoken language was German.

By optimizing model parameters, we could improve model fairness, as seen in Table 12. Only two values that are above the threshold remain. The model performance metric from the DTE shows increased irregularity, while the KNN model improved overall, although variability between the time points increased. However, the accuracy of the MLP worsened significantly from 87% to less than 80%.

Table 12. Results of the metrics for the attribute of first spoken language after attempting to mitigate learning bias for all three implementations.

Model	**DTE** Max_depth: 5 Min_sample_leaf: 25 Min_sample_split: 2				**KNN** N_neighbors: 10 Weights: uniform				**MLP** Loss: Hinge Optimizer: SGD Metrics: AUC			
Metrics	EO	PE	PP	AB	EO	PE	PP	AB	EO	PE	PP	AB
Sentence												
2 to 9	-0.040	-0.038	0.009	0.001	-0.042	-0.030	0.010	0.006	-0.039	0.024	0.030	0.042
10 to 19	-0.035	-0.055	0.002	-0.010	-0.041	-0.044	0.003	-0.002	-0.034	0.019	0.020	0.036
20 to 29	-0.036	-0.040	0.004	-0.002	-0.046	-0.035	0.004	0.006	-0.024	-0.020	0.012	0.018
30 to 39	-0.036	-0.018	0.009	0.009	-0.031	-0.025	0.007	0.003	-0.015	-0.006	0.024	0.006
40 to 49	-0.034	-0.020	0.012	0.007	-0.024	-0.015	0.010	0.005	-0.005	-0.010	0.026	-0.017
50 to 60	-0.015	-0.024	0.018	-0.005	-0.008	-0.069	0.013	-0.030	-0.004	-0.008	0.029	-0.013

To conclude, our results regarding the attribute of first spoken language has similarities to that of the attribute of parental education background. While our attempt to improve bias at the representational or aggregation step of the ML lifecycle failed, we were successful in optimizing model parameters.

4.4 **Home literacy environment (HLE).** The fairness evaluation results of the HLE attribute show many values that are above the threshold of |0.05| (Table 13). For the PE metric, all values show discrimination, as do some values for the EO and AB metrics.

Table 13. Evaluation results of the metrics for the HLE attribute (Rzepka et al., 2022a).

	EO			**PE**			**PP**			**AB**		
Model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.065	-0.051	-0.051	-0.102	-0.122	-0.122	0.011	0.010	0.010	-0.008	-0.035	-0.035
10 to 19	-0.046	-0.050	-0.050	-0.133	-0.147	-0.147	0.006	0.002	0.002	-0.005	-0.049	-0.049
20 to 29	-0.042	-0.044	-0.044	-0.079	-0.107	-0.107	0.012	0.010	0.001	-0.003	-0.032	-0.032
30 to 39	-0.034	-0.029	-0.029	-0.067	-0.103	-0.103	0.010	0.009	0.009	0.003	-0.037	-0.037
40 to 49	-0.031	-0.017	-0.017	-0.163	-0.119	-0.119	0.001	0.007	0.007	-0.041	-0.051	-0.051
50 to 60	-0.045	-0.043	-0.043	-0.213	-0.090	-0.097	0.006	0.022	0.021	-0.076	-0.024	-0.027

None of the attempts to mitigate bias for the HLE variable were successful. Table 14 shows the fairness evaluation after applying SMOTE oversampling to the minority group. Here, the metrics worsened significantly. Especially for the EO metric, most values are above the threshold. Further, the MLP model performance decreased and often is below 0.70.

Additionally, computing two different models, one for each group, did not lead to an improvement in the fairness evaluation (Appendix D1). Interestingly, even the PP metric showed many values above the threshold, although this metric is very robust, and values remain below the threshold most of the time. Furthermore, addressing learning bias and optimizing parameters did not result in improvements, as shown in Appendix D2. However, the PP metric is below the threshold again, and the discriminatory values appear mainly in the EO and PE metrics. In conclusion, HLE was the only variable where no attempt to mitigate bias was successful.

Table 14. Results of the metrics for the attribute of home literacy environment after attempting to mitigate representational bias.

	EO			PE			PP			AB		
Model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.081	-0.081	-0.077	-0.095	-0.089	-0.125	0.009	0.008	0.002	-0.009	-0.004	-0.024
10 to 19	-0.077	-0.076	-0.075	-0.137	-0.107	-0.130	0.003	0.002	-0.001	-0.017	-0.015	-0.027
20 to 29	-0.084	-0.068	-0.076	-0.157	-0.103	-0.130	0.003	0.006	-0.001	-0.011	-0.017	-0.027
30 to 39	-0.071	-0.071	-0.070	-0.180	-0.095	-0.123	0.003	0.005	0.002	-0.036	-0.012	-0.026
40 to 49	-0.030	-0.055	-0.045	-0.157	-0.156	-0.267	0.009	0.000	-0.004	-0.072	-0.050	-0.111
50 to 60	-0.035	-0.057	-0.037	-0.209	-0.194	-0.328	0.020	0.004	0.012	-0.097	-0.068	-0.145

5 Discussion

We assess fairness and implement measures to alleviate bias in in-session dropout prediction models in this study by addressing potential sources of harm during the machine learning lifecycle, as suggested by Suresh and Guttag (2021). Our study examines three different machine learning implementations (DTE, KNN, MLP) and evaluates model fairness using four metrics (EO, PE, PP, AB) as outlined in Rzepka et al. (2022a). Fairness evaluation and optimization are conducted four times with distinct demographic groups. For the first group, we examine gender and compare male and female users. The second group pertains to parental educational background, where we compare model fairness for users with at least one parent having a high school diploma and those without. The third group analyzes users' primary language, comparing those whose first language is German to those whose first language is not. Moreover, we investigate the home literacy environment (HLE), which we indicate by estimating the number of books in the household. We compare model fairness for edge cases with fewer than 10 books and over 100 books in the household.

At the outset of our research, we defined three research questions that we sought to address during our study.

RQ1: At what steps in the machine learning lifecycle of educational drop-out prediction models can discrimination arise?
RQ2: How can the fairness of drop-out prediction models be improved?
RQ3: How does the potential for improvement differ among demographic groups?

Concerning RQ 1, we can say that discrimination can arise in all steps of the machine learning lifecycle. We explained how this can happen and how we can mitigate the bias in our specific context. We eventually applied those mitigation techniques to all models and evaluated fairness for all demographic groups and fairness metrics.

With regard to RQ 2 and RQ 3, we were able to mitigate bias in three out of four demographic groups. We could not improve fairness for the HLE attribute. All attempts to mitigate bias for this attribute led to increased discrimination. When we were able to mitigate bias, the approach differs for the different groups. An overview is provided in Table 15. Please note that historical bias and measurement bias was only applied to the attribute gender: mitigating historical bias means to delete the features from the model training. As gender was the only feature that was included in training the model, an attempt to mitigate historical bias was only possible for the gender group. Measurement bias, in our case, was mitigated by dropping features that correlate with a demographic attribute. As this was only the case for gender (correlating with number of pending tasks), we could only apply the attempt to mitigate measurement bias to the gender group.

Table 15. Summarized results of the metrics after attempting to mitigate bias.

	Gender	**Parental Education Background**	**First spoken language**	**HLE**
Historical Bias	0	/	/	/
Representational Bias	0	+	-	-
Measurement Bias	0	/	/	/
Aggregation Bias	++	--	--	--
Learning Bias	+	++	++	--

No change = 0; (strong) reduction in bias = (+)+; (strong) increase in bias = (-)-; no attempt = /.

Representational bias did not work as expected and if, it only improved the AB metric, while other fairness metrics either did not improve or even worsened. Although imbalanced datasets are said to be one of the main reasons for bias in ML models, we could not mitigate the bias by oversampling minority groups.

Building two separate models to mitigate aggregation bias, led to different results. We found that the gender attribute showed the best improvement out of the considered options. All fairness metrics remained below the threshold, and model performances did not suffer for either the model for boys or the model for girls. However, the same effect was not found for the attributes of first spoken language and parental education background. Here, we found that two separate models led to increased discrimination for all implementations and all metrics. Furthermore, the models of the disadvantaged groups (users with parents without high school diplomas and users whose first language is not German) performed more poorly than those of the advantaged group. Our results show that some attempts to mitigate bias can actually increase discrimination. All attempts should thus be carefully evaluated. Comparing the results between different demographic groups shows that an approach that works for one group (in this case, gender) does not automatically work for other groups.

Learning bias mitigation (where different parameters of the model implementations were tested) was the most promising approach to reducing discrimination. For three out of four demographic factors (gender, parental education, first spoken language), it worked equally well for all model implementations. However, this approach results in a slight worsening of the model performances. Subject matter experts thus must make a decision concerning the trade-off between model performance and model fairness.

There was no calculation to be carried out with regard to the deployment bias. However, some considerations should be made: Since the models have been trained and are used on the Orthografietrainer.net platform, the risk of bias through use is relatively low. Nevertheless, the way the orthography trainer is used in a blended classroom scenario can also influence the learning experience. We cannot address this, as it depends heavily on the environment in which the orthography trainer is used. If the models are ever used on another online language learning platform, this bias should be carefully considered. Another point to consider in terms of deployment bias is the impact of model bias on future training data. If some models used on the Orthografietrainer.net platform are biased and some users may receive inappropriate interventions, the training data for future models will also be biased. This can of course be prevented by evaluating the models before they are used and only using those that are not biased.

With regard to the different fairness metrics, some are more often above threshold than others. PP, for example, is not above thresholds most of the time, while PE is above the threshold the most. That means, that the power of explainability (determined by PP) does not differentiate much. However, students more often face the probability of false alarm, which is determined by PE. Furthermore, we observed that historical, representational, and measurement biases improved the AB metric but not the other metrics. An explanation can be that the AB metric derives from two values: TPR and FPR. If an attempt leads to an improvement in the AB metric, but not in the PE metric (which requires equal FPRs), one can assume that the TPR improved. This outcome would mean that historical, representational, and measurement bias mitigation increases the fairness of the probability of detection but not the probability of false positives. This result can, when triggering certain educational interventions, improve learning (e.g., when at-risk students receive additional tips and hints). Depending on the implementation, which fairness metrics should be considered to determine if a model can be deployed or which threshold to select depend on the subject. In an educational context and more specifically, in in-session dropout predictions, interventions offered to users depend on the predicted outcome. Discrimination may occur in three forms: probability of missing (EO, AB), probability of false alarm (PE) and power of explainability (PP). Whether one form is worse than another depends on the interventions that follow with the predictive outcome. Thus, implementation and interventions must be considered to decide which fairness metrics are important in a specific case.

In this work, the three ML implementations did not differ very much. Mostly, the approach worked either for all model implementations or none. Furthermore, we found that model accuracy often decreases after applying a mitigation technique, for example after applying learning bias mitigation to the parental education background group. This is a trade-off that is already discussed in previous work, e.g. by Zliobaite (2015). We can now state that we found the same effect. More insightful results can be obtained by looking at the temporal analysis. Discrimination tends to occur more often later in sessions when model performances improve. Performance mainly improves in our model, because with each time point in a session data availability increase. This could mean, that in our case an increasing amount of data leads to algorithmic bias. The more data the model has on the user, the better the model accuracy becomes, and the more bias occurs.

5.1 **Limitations.** Our study comes with limitations that should be considered by the reader. First, our data were generated by a demographic survey that is voluntary. Thus, not all users of the platform are included in our study, only those who were motivated enough to take part in the survey. There are some limitations to the HLE attribute. First of all, it must be taken into account that there is also criticism of the use of attributes that are examining verbal environments from different socioeconomic backgrounds, e.g. Sperry et al. (2019), who could not confirm that lower-class children are exposed to fewer words than middle-class children. Furthermore, users on the platform are asked to estimate the number of books in the household, which is not exact. Moreover, we considered only edge cases in our calculations, leading to further abstraction of the data and a large number of excluded data points.

Regarding the evaluation bias, one must consider that the models were trained on the complete dataset of Orthografietrainer.net platform while the fairness evaluation only contains data from users who took part in the voluntary survey. Thus, the training data differ from the evaluation data. Furthermore, we did not try to improve the models to be fair for all minority groups simultaneously. In reality, however, the latter should be the goal. Additionally, our calculations lack in studying intersectionality, meaning users who are in more than one disadvantaged group. Moreover, we focused only on the four groups in our study, but there are many potential demographic groups against which a model may discriminate. Finally, our work only considers one language learning platform and its associated dataset. To what extent our results can be replicated on the basis of other learning platforms needs further research.

6 Conclusion

In this work, we focused on three less-studied demographic groups (parental education, first spoken language, and home literacy environment) and another demographic group (gender) to evaluate three ML models with regard to fairness and find ways to mitigate bias. The evaluated models are dropout prediction models for an online platform that promotes the development of German spelling skills that predict users' dropout in a session. The three model implementations (decision tree classifier, KNN, multilayer perceptron) were evaluated using different fairness metrics (EO, PE, PP, AB). To mitigate algorithmic bias, we focused on the seven "sources of harm" as proposed by Suresh and Guttag (2021). Our results showed that for three of four demographic groups, different approaches made it possible to mitigate bias. For gender, the best attempt was to mitigate the aggregation bias and build two separate models, one for boys and one for girls. Through parameter optimization to resolve the learning bias, fairness could be increased for the features of parental education background and first spoken language. For the fourth attribute, HLE, none of the attempts was successful, and we could not improve fairness.

Further research is required to determine whether other types of models that are implemented in educational contexts respond similarly to these approaches. Furthermore, other less-studied groups should be included. Moreover, it is necessary to research intersectional discrimination and whether the approaches to mitigate bias we found can be applied to intersectional discrimination. In addition, mitigating bias for only one group at a time is not practical in reality. Models implemented in real educational contexts must be optimized and fair for all groups at the same time. Another point that goes beyond the scope of the study but still needs to be considered is whether variables such as first spoken language should be used as predictors or demographic variables. As already referenced, Steinlen & Piske (2013) show that there is indeed predictive power behind this. At the same time, our analyses show that there is discrimination potential behind this variable and that it can therefore also be used as a demographic variable to test for fairness. It should thus be discussed to which extent variable like first spoken language should be used as predictors or demographic variables. Ultimately, the transferability of our results to other (language) learning platforms also needs to be explored and replicated.

Acknowledgments

This research is funded by the Federal Ministry of Education and Research of Germany (project nr: 16DHB4002).

References

Anderson, H., Boodhwani, A., & Baker, R. S. (2019). Assessing the Fairness of Graduation Predictions. Proceedings of the *12th International Conference on Educational Data Mining* (EDM 2019). http://radix.www.upenn.edu/learninganalytics/ryanbaker/edm2019_paper56.pdfAgrawal, S., De Smet, A., Poplawski, P., & Reich, A. (2020, January). Beyond hiring: How companies are reskilling to address talent gaps. *McKinsey & Company*.

Baker, R. S., Berning, A., & Gowda, S. M. (2020). Differentiating Military-Connected and Non-Military-Connected Students: Predictors of Graduation and SAT Score. *Center for Open Science.* https://doi.org/10.35542/osf.io/cetxj

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education,* 1–41. https://doi.org/10.1007/s40593-021-00285-9

Belitz, C., Jiang, L., & Bosch, N. (2021). Automating Procedurally Fair Feature Selection in Machine Learning. In M. Fourcade (Ed.), ACM Digital Library, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 379–389). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462585

Centor, R. M. (1991). Signal Detectability: The Use of ROC Curves and Their Analyses. *Medical Decision Making,* 11(2), 102–106. https://doi.org/10.1177/0272989X9101100205

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, 4(1), 123–144.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2), 153–163.

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges. *IEEE Global Engineering Education Conference.*

Franklin, J. S., Bhanot, K., Ghalwash, M., Bennett, K. P., McCusker, J., & McGuinness, D. L. (2022). An Ontology for Fairness Metrics. In V. Conitzer, J. Tasioulas, M. Scheutz, R. Calo, M. Mara, & A. Zimmermann (Eds.), *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 265–275). ACM. https://doi.org/10.1145/3514094.3534137

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. *In Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM.

Goudeau, S., Sanrey, C., Stanczak, A., Manstead, A., & Darnon, C. (2021). Why lockdown and distance learning during the COVID-19 pandemic are likely to increase the social class achievement gap. *Nature Human Behaviour*, 5(10), 1273–1281. https://doi.org/10.1038/s41562-021-01212-7

Hutchinson, B., & Mitchell, M. (2019). 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In FAT* '19, Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 49–58). *Association for Computing Machinery*. https://doi.org/10.1145/3287560.3287600

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC Dropout over Weeks Using Machine Learning Methods. *Association for Computational Linguistics*.

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016). Big Data Application in Education: Dropout Prediction in Edx MOOCs. *In 2016 IEEE Second International Conference on Multimedia Big Data* (BigMM).

McClain, C., Vogels, E. A., Perrin, A., Sechopoulos, S., & Rainie, L. (2021). The Internet and the Pandemic. *Pew Research Center*. https://policycommons.net/artifacts/1811100/the-internet-and-the-pandemic/2547007/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6). https://doi.org/10.1145/3457607

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Computing Surveys*, 53(3), 1–34. https://doi.org/10.1145/3388792

Riazy, S., & Simbeck, K. (2019). Predictive Algorithms in Learning Analytics and their Fairness. 1617-5468. Advance online publication. https://doi.org/10.18420/delfi2019_305

Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments. In *Proceedings of the 12th International Conference on Computer Supported Education* - Volume 1: CSEDU (pp. 15–25). SCITEPRESS.

Rzepka, N. (2023). Bias Mitigation: v1.0.1-bias_mitigation [Computer software]. Zenodo. https://zenodo.org/record/7746108

Sun, D., Mao, Y., Du, J., Xu, P., Zheng, Q., & Sun, H. (2019). Deep Learning for Dropout Prediction in MOOCs. In *2019 Eighth International Conference on Educational Innovation through Technology (EITT*).

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In EAAMO '21, *Equity and Access in Algorithms, Mechanisms, and Optimization. Association for Computing Machinery*. https://doi.org/10.1145/3465416.3483305

Vasquez Verdugo, J., Gitiaux, X., Ortega, C., & Rangwala, H. (2022). Faired: A Systematic Fairness Analysis Approach Applied in a Higher Educational Context. In A. F. Wise (Ed.), ACM Digital Library, Lak22: 12th International Learning Analytics and Knowledge Conference (pp. 271–281). *Association for Computing Machinery*. https://doi.org/10.1145/3506860.3506902

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *In Proceedings of the International Workshop on Software Fairness*. ACM. https://doi.org/10.1145/3194770.3194776

Xing, W., & Du, D. (2019). Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *Journal of Educational Computing Research*, 57(3), 547–570. https://doi.org/10.1177/0735633118757015

Appendix A. Gender.

Table A1. Results of the metrics for the attribute of gender after attempting to mitigate historical bias.

	EO			**PE**			**PP**			**AB**		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.038	-0.027	-0.050	-0.061	-0.048	-0.022	0.005	0.008	0.004	-0.001	-0.011	0.014
10 to 19	-0.028	-0.023	-0.048	-0.037	-0.013	-0.020	0.006	0.010	0.003	0.008	0.005	0.014
20 to 29	-0.027	-0.022	-0.037	-0.019	-0.027	-0.022	0.007	0.007	0.002	0.007	-0.002	0.007
30 to 39	-0.017	-0.014	-0.023	-0.028	-0.017	-0.038	0.004	0.005	0.000	0.008	-0.002	-0.004
40 to 49	-0.008	-0.010	-0.036	-0.113	-0.116	-0.080	-0.002	-0.003	-0.003	-0.002	-0.053	-0.022
50 to 60	-0.003	-0.000	-0.034	-0.171	-0.143	-0.068	-0.004	-0.003	-0.002	-0.023	-0.071	-0.017

Table A2. Results of the metrics for the attribute of gender after attempting to mitigate representational bias.

	EO			PE			PP			AB		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.025	-0.029	-0.050	-0.042	-0.048	-0.025	0.008	0.008	0.004	-0.001	-0.009	0.012
10 to 19	-0.023	-0.024	-0.041	-0.019	-0.017	-0.017	0.008	0.009	0.003	0.008	0.003	0.012
20 to 29	-0.014	-0.021	-0.030	0.009	-0.019	-0.020	0.009	0.008	0.002	0.008	0.001	0.005
30 to 39	-0.007	-0.012	-0.020	-0.001	-0.020	-0.028	0.006	0.006	0.001	0.006	-0.004	-0.004
40 to 49	-0.005	-0.008	-0.023	-0.099	-0.119	-0.070	-0.001	-0.003	-0.003	-0.002	-0.055	-0.024
50 to 60	-0.002	-0.001	-0.028	-0.178	-0.182	-0.057	-0.005	-0.005	-0.001	-0.029	-0.091	-0.014

Table A3. Results of the metrics for the attribute of gender after attempting to mitigate measurement bias.

	EO			PE			PP			AB		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.018	-0.027	-0.034	-0.029	-0.048	-0.016	0.009	0.008	0.005	-0.001	-0.011	0.009
10 to 19	-0.011	-0.023	-0.030	0.003	-0.014	-0.016	0.010	0.010	0.003	0.007	0.005	0.007
20 to 29	-0.005	-0.022	-0.023	0.025	-0.026	-0.013	0.010	0.007	0.003	0.007	-0.002	0.005
30 to 39	-0.002	-0.014	-0.014	0.033	-0.016	-0.032	0.010	0.006	0.000	0.008	-0.001	-0.009
40 to 49	-0.004	-0.010	-0.035	-0.106	-0.120	-0.074	-0.002	-0.003	-0.003	-0.005	-0.055	-0.020
50 to 60	-0.001	-0.000	-0.024	-0.203	-0.142	-0.125	-0.006	-0.003	-0.006	-0.024	-0.071	-0.051

Table A4. Results of the metrics for the attribute of gender after attempting to mitigate historical bias, representational bias, and measurement bias.

	EO			PE			PP			AB		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.037	-0.028	-0.043	-0.062	-0.048	-0.028	0.005	0.008	0.003	-0.002	-0.010	0.007
10 to 19	-0.029	-0.030	-0.046	-0.030	-0.022	-0.028	0.007	0.008	0.002	0.008	0.004	0.009
20 to 29	-0.023	-0.021	-0.039	-0.011	-0.018	-0.012	0.008	0.008	0.003	0.007	0.001	0.013
30 to 39	-0.016	-0.011	-0.036	-0.010	-0.003	-0.034	0.006	0.007	0.000	0.008	0.004	0.001
40 to 49	-0.012	-0.008	-0.038	-0.109	-0.079	-0.046	-0.002	-0.001	-0.001	-0.001	-0.036	-0.004
50 to 60	0.000	0.002	-0.012	-0.133	-0.170	-0.141	-0.003	-0.004	-0.007	-0.025	-0.086	-0.064

Appendix B. Parental Education Background.

Table B1. Results of the metrics for the attribute of parental education background after attempting to mitigate representational bias.

	EO			PE			PP			AB		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.020	-0.033	-0.022	-0.006	-0.017	-0.022	0.002	0.001	0.000	0.008	0.008	0.000
10 to 19	-0.026	-0.035	-0.029	-0.021	-0.013	-0.025	0.000	0.000	-0.001	0.010	0.011	0.002
20 to 29	-0.020	-0.019	-0.018	-0.051	-0.005	-0.016	-0.002	0.002	0.001	-0.001	0.007	0.001
30 to 39	-0.017	-0.018	-0.023	-0.015	0.052	-0.019	0.006	0.008	0.003	0.003	0.035	0.002
40 to 49	-0.001	-0.014	-0.013	-0.066	0.059	-0.012	0.005	0.011	0.008	-0.044	0.037	0.000
50 to 60	-0.009	-0.004	-0.018	0.029	0.082	-0.019	0.032	0.013	0.013	0.009	0.043	-0.001

Table B2. Results of the metrics for the attribute of parental education background after comparing two models.

	EO			PE			PP			AB		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.112	-0.095	-0.068	-0.021	0.008	-0.058	0.092	0.101	0.031	0.029	0.051	0.005
10 to 19	-0.053	-0.075	-0.101	0.075	0.031	-0.010	0.136	0.107	0.067	0.081	0.053	0.046
20 to 29	-0.098	-0.114	-0.115	-0.026	-0.057	-0.070	0.092	0.062	0.022	0.083	0.029	0.022
30 to 39	-0.166	-0.110	-0.113	-0.046	-0.075	-0.116	0.088	0.034	-0.008	0.177	0.017	-0.001
40 to 49	0.034	-0.145	-0.128	0.156	-0.100	-0.190	0.077	0.087	0.010	0.078	0.023	-0.031
50 to 60	0.044	-0.111	-0.042	0.064	-0.325	-0.245	0.165	0.028	0.033	-0.057	-0.107	-0.101

Appendix C. First Spoken Language.

Table C1. Results of the metrics for the attribute of first spoken language after comparing two separate models.

	EO			PE			PP			AB		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	-0.220	-0.188	-0.173	-0.148	-0.091	-0.110	0.049	0.092	0.039	0.023	0.048	0.032
10 to 19	-0.132	-0.116	-0.154	-0.078	-0.061	-0.080	0.096	0.096	0.054	0.047	0.028	0.037
20 to 29	-0.169	-0.122	-0.194	-0.159	-0.068	-0.079	0.176	0.103	0.068	0.053	0.027	0.057
30 to 39	-0.114	-0.064	-0.144	-0.249	-0.121	-0.056	0.080	0.075	0.086	0.026	-0.028	0.044
40 to 49	-0.146	-0.113	-0.182	-0.096	-0.172	-0.123	0.253	0.101	0.059	0.089	-0.029	0.030
50 to 60	-0.089	-0.133	-0.186	-0.064	-0.129	0.039	0.211	0.090	0.112	-0.046	0.002	0.112

Appendix D. Home Literacy Environment.

Table D1. Results of the metrics for the attribute of home literacy environment after comparing two models. Since there were too few data points for the periods from record 40 onward, these data were excluded during the evaluation process.

	EO			PE			PP			AB		
model	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN	MLP	DTE	KNN
Sentence												
2 to 9	0.388	0.409	0.100	0.290	0.277	0.172	-0.121	-0.103	0.012	0.004	-0.066	0.037
10 to 19	0.426	0.392	0.191	0.337	0.259	0.175	-0.204	-0.182	-0.030	-0.023	-0.067	-0.005
20 to 29	0.469	0.265	0.199	0.427	0.208	0.132	-0.256	-0.160	-0.051	-0.047	-0.028	-0.027
30 to 39	0.482	0.154	0.101	0.497	0.236	0.174	-0.053	-0.091	0.007	-0.037	0.041	0.034
40 to 49												
50 to 60												

Table D2. Results of the metrics for the attribute of home literacy environment after attempting to mitigate learning bias for all three implementations.

Model	DTE				KNN				MLP			
	Max_depth: 10 Min_sample_leaf: 1 Min_sample_split: 2				N_neighbors: 10 Weights: distance				Loss: MSE Optimizer: SGD Metrics: AUC			
Metrics	EO	PE	PP	AB	EO	PE	PP	AB	EO	PE	PP	AB
Sentence												
2 to 9	-0.071	-0.077	0.008	-0.003	-0.080	-0.073	0.004	0.003	-0.086	-0.028	0.033	0.027
10 to 19	-0.068	-0.097	0.004	-0.015	-0.076	-0.087	-0.001	-0.005	-0.079	-0.094	0.019	0.000
20 to 29	-0.071	-0.090	0.005	-0.009	-0.070	-0.073	0.003	-0.002	-0.071	-0.106	0.018	-0.015
30 to 39	-0.079	-0.047	0.009	0.016	-0.072	-0.035	0.008	0.018	-0.050	-0.131	0.009	-0.025
40 to 49	-0.046	-0.087	0.004	-0.021	-0.047	-0.156	-0.002	-0.055	-0.022	-0.099	0.023	-0.046
50 to 60	-0.075	-0.213	-0.003	-0.069	-0.043	-0.287	0.000	-0.122	-0.021	-0.190	0.024	-0.032