

Fairness of MOOC Completion Predictions Across Demographics and Contextual Variables

Sébastien Lallé¹, François Bouchet¹, Mélina Verger¹, and Vanda Luengo¹

¹Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

{sebastien.lalle,francois.bouchet,melina.verger,vanda.luengo}@lip6.fr

Abstract While machine learning (ML) has been extensively used in Massive Open Online Courses (MOOCs) to predict whether learners are at risk of dropping-out or failing, very few work has investigated the bias or possible unfairness of the predictions generated by these models. This is however important, because MOOCs typically engage very diverse audiences worldwide, and it is unsure whether the existing ML models will generate fair predictions to all learners. In this paper, we explore the fairness of ML models meant to predict course completion in a MOOC mostly offered in Europe and Africa. To do so, we leverage and compare ABROCA and MADD, two fairness metrics that have been proposed specifically in education. Our results show that some ML models are more likely to generate unfair predictions than others. Even in the fairest models, we found biases in their predictions related to how the learners' enrolled as well as their country, gender, age and job status. These biases are particularly detrimental to African learners, which is a key finding as they are an understudied population in AI fairness analysis in education.

Keywords: Fairness · MOOCs · Machine Learning · Demographics.

1 Introduction

In AI research, the growing usage of ML techniques has risen concerns about their fairness towards certain groups of people, especially the less privileged ones [4]. In recent years, there has been similar concerns in the field of AI in Education (AIED), where ML techniques are increasingly used to support a variety of decisions, such as providing learners with dedicated support and scholarships, which can strongly impact their learning and academic outcomes [1,10]. However, so far, only a few studies have focused on formally assessing the fairness of ML techniques used in education, leading AIED researchers to call for extending and broadening our understanding of ML fairness in this context [1,10].

In this paper, we study the fairness of ML models trained to predict whether learners will fail to earn certification in a MOOC. This is a common ML task in research on MOOCs [14], because it is known that only a fraction of the enrolled students typically pass, due to issues with dropping-out, time management and lack of help, among others. The ability to predict as early as possible that a learner will fail can enable early personalized interventions to prevent it,

e.g., [21]. Thanks to the ability of MOOCs to engage thousands and facilitate data collection, there have been several studies about building ML models for predicting MOOC completion, allowing us to better understand the most useful ML models and features for this task (see overview in [14]). However, in contrast, there is very little understanding about how fair these models are against their diverse groups of learners, as pointed out by Baker & Hawn [1], raising questions about whether everybody can fully trust and benefit from these models in MOOCs. Moreover, there seems to be discrepancies in how underrepresented groups complete MOOCs compared to historically privileged populations, thus risking to perpetuate existing differences as MOOCs tend to be a good entry point for adults to learn about emerging fields in a lifelong learning context [13].

We explore this issue by augmenting MOOC log data with a survey collecting key demographics and other educational-related information about the learners. Next, we trained ML models to predict at the end of each week which learners will fail the course, and we use the survey answers to analyze in-depth how fair the models are for various groups of learners. We do so using two fairness metrics previously used in education (ABROCA and MADD), to get a more comprehensive understanding of potential fairness issues in this context.

Our work contributes to existing research in three ways. First, despite the existence of extensive ML research on MOOCs, so far only two studies explored the fairness of ML models in MOOCs, solely regarding gender [7,17]. A few papers highlighted this gap [1,7], calling for broadening fairness analysis in MOOCs to take advantage of their large, international and diverse audience. Here, we study this gap by exploring possible unfairness in a MOOC, using key variables other than gender, such as learners' country of residence and socio-economic context.

Second, beyond MOOCs, we broaden the set of variables leveraged for ML fairness analysis of education. Indeed, existing work has focused primarily on demographic information such as gender and ethnicity [1]. Here, we do leverage standard demographics for the sake of comparison (namely gender, native language, parental education, country), but we also explore additional demographic variables that could impact the learners' time and motivation, such as their job and family status. We also consider non-demographic variables relevant in MOOCs, namely prior experience with MOOCs and whether learners enrolled on their own or as part of a university degree.

Third, we leverage a dataset where about half of the learners live and study in Africa. This is important because so far, fairness studies in education have been mostly limited to the USA, and to a lesser extent Western Europe and Eastern Asia [1,8], whereas ML fairness in other parts of the world have not been widely studied and thus remains a key gap in the literature. This is even more concerning in MOOCs as learners can typically come from anywhere in the world, but the ML models used in this context might not treat everybody fairly.

Overall, our work aims at broadening our understanding of ML fairness in AIED, by targeting a unique pool of learners and novel demographic and contextual variables in an understudied type of learning environments in ML fairness analysis (MOOCs), a much needed avenue of research for AIED.

2 Related Work

ML models have been extensively studied in MOOCs for the tasks of predicting as early as possible whether a learner will drop-out / successfully pass the course. For example, two recent systematic reviews [14,6] covered several dozens of such research papers, highlighting the importance of this line of research. Such predictions are seen as key to detect at-risk learners early on and intervene in the hope of remediating the learners’ difficulties whenever possible.

However, as said above, to the best of our knowledge, only two studies [7,17] explored the fairness of these ML models in MOOCs, focusing on gender biases of ML models meant to predict drop-out in several North-American and Chinese MOOCs. They both found evidence of gender-biased ML predictions, especially depending on the gender imbalance in each MOOC, and no direct link among the fairness and performance of the ML models. This provides preliminary evidence of unfair predictions in MOOCs, and justifies further fairness analysis in this context, as we do in this paper for gender and beyond.

Besides MOOCs, several studies have focused on the fairness of ML models trained for educational purposes, including predicting university course completion from Learning Management Systems (LMS)’s data [9,22], dropout in an e-learning platform for writing homework [16], and knowledge mastery in Intelligent Tutoring Systems (ITS) [24]. However, other work pointed out that these studies are often limited in terms of demographics (mostly gender and ethnicity) and learners’ origin (mostly USA), both key issues when it comes to fairness [1,8]. Still, some work have started to fill these gaps. In terms of demographics, research has studied the fairness of automatic scoring and dropout models towards the learners’ country of origin and/or native language [3,15,16,20]. Other work focused on biases in LMS, e-learning and ITS due to parental education and/or levels of poverty [22,9,16,23,24]. Compared to this line of work, we target a different learning environment (MOOC). We also reuse the aforementioned demographic variables to provide further understanding of their impact on fairness, and broaden them to include as well variables more specific to the asynchronous nature of MOOCs and the fact that they target populations larger than university students. This includes family and professional situations, as well as whether the learners attend a MOOC for the first time and how they enrolled.

Some fairness studies in education have considered learners beyond the USA, however, few of them explicitly included African countries. In [2], they focus on the fairness of ML models for automatic essay scoring, and include a pool of learners from Nigeria among 15 other countries. However, the amount of Nigerian’s essays considered in the study was extremely low (less than 0.03%, versus over 75% from the USA), making it hard to draw strong conclusions as for the African learners. We follow this line of work, but with about half of the learners in our dataset living in Africa, allowing us to more formally compare the fairness of ML models between Africa and other continents. African countries are particularly interesting to consider as previous studies have already highlighted the specificity of their context, including strong differences in urban and rural experience [11], and reliance on Open Digital Spaces to compensate for the lack

of physical infrastructures [18]. In [12], they study the fairness of ML predictors of academic achievement using the 2015 PISA dataset. This dataset includes learners from 65 countries, but only two from Africa (Algeria and Tunisia), for which specific results were not explicitly reported. In contrast, our dataset includes learners from 24 African countries¹, representing 53% of the learners in this dataset, and we explicitly study the fairness of our models towards them.

3 Dataset

The data used in this paper came from a French MOOC on Project Management² ran on an OpenEdX platform. It has been held twice a year since 2013, being one of the first French MOOCs to deliver a certificate. Over 22 editions, more than 250,000 persons have signed up to the MOOC (45,000 of whom have obtained a certificate). The main section of the MOOC which needs to be followed to obtain the basic certificate lasts four weeks, with a final exam being made available at the beginning of week 5 and remaining available for four weeks. The grade on this final exam represents 80% of the final grade, with the remaining 20% coming from weekly quizzes embedded throughout the MOOC. A grade of 50% or more is required to obtain the basic certificate at the end of the MOOC, which we use to define whether a learner passed the MOOC or not. In this paper, we focus more specifically on data from the 19th edition of the MOOC which took place in Spring 2022 for which $N = 9,752$ learners were enrolled, and 1,892 obtained the certificate. In addition to the raw OpenEdX logs that register every actions performed by learners while connected to the MOOC, we rely on data coming from a non-mandatory research survey we made available to learners at the beginning of the first week of the MOOC. Participants provided informed consent for their data to be used for research purposes. All data are fully anonymized and stored on secured servers with limited access for research purpose only, and learners can request their data to be deleted during and after the MOOC, in accordance with GDPR European regulations.

Out of all learners, we retain those who filled in entirely the survey and passed all attention checks (i.e., items where students are asked to choose a predefined answer to limit the risk of random or automated survey filling). We also discard learners who reported not wanting to earn the certification, as they might be inactive or curious for reasons others than learning. As a result, we obtained a dataset comprised of 1007 learners, among whom 383 (38%) successfully passed the course (i.e., earned the certification).

4 Analysis

We leverage the aforementioned MOOC’s dataset to build classifiers that can predict binary labels of whether the learners passed the course. To simulate the

¹ Algeria, Benin, Burkina-Faso, Burundi, Cameroon, DR Congo, Ivory Coast, Djibouti, Gabon, Guinea, Guinea-Bissau, Madagascar, Mali, Morocco, Mauritania, Mozambique, Niger, Central African Republic, Republic of the Congo, Rwanda, Senegal, Tchad, Togo, Tunisia.

² <https://mooc.gestiondeprojet.pm>

Table 1. List of OpenEdX’s events considered for feature engineering^a.

Component Events	
Video	Load, Play, Pause, Stop, Seek, Change speed, Show/hide transcript
Quiz	Load, Save, Check, Graded, Show answers
Forum	Create thread or response, View, Search, Vote
Platform	Go to, Next page, Previous page, Close

^a Edx research guide: Events in the tracking logs. https://edx.readthedocs.io/projects/devdata/en/stable/internal_data_formats/tracking_logs.html.

predictions over time, the classifiers are trained and tested at the end of each of the four first weeks of the MOOC, until the final exam becomes available. The features and code of the ML and fairness analysis, described next, are available in open source at https://github.com/Mocahteam/GdPMOOC_Fairness.

4.1 Features Engineering

We process the MOOC’s logs data to generate a set of predictive features based on the events reported in Table 1. Specifically, for each event, we generate one feature corresponding to the count of that event performed by the learner, for a total of 20 features per learner. These events include most of the events relevant for our MOOC, and they are meant to capture the extent to which the learners interacted with each of the four main components of traditional MOOCs, namely the videos, the forum, the quizzes, and the platform. We focus on these four components because they have been the most used in similar prediction tasks to build features in MOOCs [14]. Indeed, our goal here is not to engineer novel features, but instead understand the fairness of the existing ML models typically trained to predict MOOC completion.

As said above, we train and test the classifiers for each of the first four weeks of the MOOC. This means that, at the end of each week, the features (i.e., *counts of events*) are generated using all of the available data so far from the current and past weeks³, thus producing four different weekly feature sets.

4.2 Machine Learning Setup

We train classifiers using a set of 6 machine learning (ML) techniques on the aforementioned sets of features to predict whether the learners passed the course. Specifically, we chose the six most successful ML techniques used for prediction tasks in MOOCs, as reported in the systematic review from Moreno-Marcos et al. [14]. We do so both because these models are more likely to generate accurate predictions on our data based on this previous work, and because our main goal is to measure the fairness of the models actually being the most commonly used in MOOCs. These techniques include: Logistic Regression (LR), Support Vector Machine (SVM), Random Forests (RF), Naive Bayes (NB), Stochastic

³ An alternative approach is to compute the features solely for the current week, thus “forgetting” the previous ones, however in our case, we found that both approaches yield similar results.

Gradient Boosting (SGB) and Neural Network. For the later, we opted for a Multi-Layer Perceptron (MLP) with one hidden layer of 10 neurons (half the number of input features). This means that we train a total of 4 (weeks) $\times$ 6 (ML techniques) = 24 classifiers, implemented with the scikit-learn Python library (<https://scikit-learn.org>). We compare the classifiers against a majority-class baseline, which always predicts that the learners will fail.

We train and evaluate the classifiers using 5-folds stratified cross-validation and a standard 0.5 decision threshold. Stratification means that the proportion of the labels (“pass”, “not pass”) is kept consistent across the folds. To remember, 38% of the learners passed in our dataset. We perform feature selection within the train set by discarding highly correlated features to avoid issues with colinearity. We also attempted to balance the data with SMOTE and fine-tune the hyperparameters of the classifiers using a grid search approach (within the train sets only), however, this did not significantly impact the performances of the classifiers. Hence, we retain the default values of all hyperparameters in the rest of the paper for simplicity and to facilitate reproducibility⁴. We report classification performance in terms of the F1 score averaged over the five folds, as the F1 metric is suitable for imbalanced data.

4.3 Fairness Analysis

In this part, we explore the possible unfairness of the classifiers using a set of demographics and contextual variables (called *protected variables* from now on) listed in Table 2. For each of the variables, we bin the learners into two groups G_0 and G_1 , as shown in Table 2. This stems from the two fairness metrics we have chosen (ABROCA and MADD, see below) which have been designed to compare only two groups. Although pairwise comparisons of more than two groups is possible, it significantly complicates the interpretation. Eight of these variables are demographics, as follows:

- **Age**, **Gender** and **First Language** (**Lang** for short) have been already used in AIED fairness analysis (see section 2). We leverage them to assess whether they generate similar (un)fair predictions in our data. Also, only **Gender** was used in the context of a MOOC, hence studying **Age** and **Lang** is meant to provide additional insights about fairness in MOOCs. In particular, **Age** might play a role due to the digital divide among generations, and **Lang** might be key due to the fact that MOOCs are typically served in one language only (in our case French), which might not be the first language of all learners.

- The country of residence (**Country**) generates in our data an almost even split between African and Western European learners. Namely, 24 African countries are represented in our dataset, all of them having a Human Development Index (HDI) on the low to medium side (< 0.7). We also lump 38 Haitian learners with the African ones, as Haiti’s HDI is similar to these countries. In contrast, the rest of the learners come from economically privileged countries with very

⁴ The grid search space and default values are documented in our Github repository linked above.

Table 2. Definition and size (N) of the binary groups for each protected variable.

Variable	Definition of G_0	Definition of G_1	N(G_0)	N(G_1)
Age	Age $\geq 35^a$	Age < 35	519	488
Gender ^b	Female	Male	569	436
Lang	Not french	French	283	724
Country	Africa + Haiti	OECD countries	536	471
JobStatus	Employed, business owner	Unemployed, students, retired	452	555
HasChildren	Yes	No	446	561
StudentEdu	Less than a master degree ^a	Master’s degree and above	485	522
ParentsEdu ^b	Less than a bachelor degree ^a	Bachelor degree and above	456	400
FirstMOOC	Yes	No	314	693
InCohort	No	Yes	766	241

^a Corresponds to the median. ^b Only 2 learners reported a gender other than male and female, which is too few for a meaningful analysis, and 151 learners reported not knowing their parents’ education. This leaves us with 1005 learners for Gender and 856 learners for ParentsEdu. For the other variables, we do keep the full sample of 1007 learners.

high HDI (≥ 0.9), mostly in Western Europe. This unique split of learners allow us to conduct a formal analysis of fairness across continents.

- **JobStatus** and **HasChildren** are two variables that we leverage to capture how much time the learners can dedicate to learning. Indeed, MOOCs’ learners tend to be older than in traditional high school and university contexts (in our data, age ranges from 21 to 74, median of 35), and thus generate unique learning experiences. Here, nearly half of the learners declare having a family and children, and are committed to a job, which can all impact the time and energy they are able to dedicate to the MOOC compared to full-time single university students. These features are novel in AIED fairness analysis at large.

- The highest levels of education of the learners (**StudentEdu**) and of their parents (**ParentsEdu**) capture the access to and success in education, which can arguably strongly impact learning with a MOOC where mastering self-regulated learning skills is critical to succeed. While similar variables have been used in previous fairness analysis, as said in section 2, it was not in a MOOC context.

In addition, we leverage two contextual variables that are more specific to MOOCs but can both strongly impact the behaviors of the learners and which are novel in fairness studied in education, so far focused solely on demographics:

- Whether this is the very first MOOC attended to by the learners (**FirstMOOC**), as these learners might struggle more with this form of remote learning.

- Whether the learners registered on their own, versus were enrolled by their university as part of their curriculum (**Cohort**), as we can expect that the later will have more (external) motivation to pass the MOOC and more opportunity to interact face-to-face with peers, a key factor as socialization helps in limiting drop-out. In our case, 10 universities in Europe and Africa enrolled 241 learners.

To assess the fairness of the trained ML classifiers, we leverage two fairness metrics that have been proposed in AIED research, namely ABROCA [7] and MADD [19]. Similarly to the F1 score, these metrics are computed for each cross-validation fold, and we report their average over all folds. ABROCA measures the

areas between the ROC curve of two groups of learners (e.g., female versus male) to identify a model’s ability to perform similarly for both groups. ABROCA ranges from 0 (fairest) to 1 (most unfair). MADD leverages the histograms of probabilities outputted by a model for two groups to compute the areas where the histograms do not overlap, i.e., behave differently. MADD ranges from 0 (fairest) to 2 (most unfair). These two metrics are complementary, as ABROCA focuses on the difference in predictive performances across two groups given the ground truth, whereas MADD assesses how differently the model behaves towards the two groups regardless of the ground truth. In fairness analysis, this corresponds to two important and complementary concepts: separation (classification performance should be equal within groups, i.e., ABROCA) and independence (the predictions should not be related to any sensitive variable, i.e., MADD) [5]. This is an important distinction, because the ground truth used by separation metrics allows measuring how accurate the model is across groups, but can be biased itself [5]. Both ABROCA and MADD are publicly available in open-source Python packages⁵, which is important for reproducibility.

As a final metric, we build an aggregated indicator (**F1+FAIR**) of the performance metric (F1) and the fairness ones (ABROCA and MADD), rescaled in a same interval $[0, 1]$. We then compute their weighted average as follows:

$$F1+FAIR = \alpha \times F1_{scaled} + \beta \times \frac{\beta_1 \times (1 - ABROCA_{scaled}) + \beta_2 \times (1 - MADD_{scaled})}{(\beta_1 + \beta_2)}$$

Such as $\alpha + \beta = 1$ meaning **F1+FAIR** ranges from 0 (worst) to 1 (best). Here we give equal weights ($\alpha = \beta = 0.5$) to performance and fairness, and equal weight to ABROCA and MADD ($\beta_1 = \beta_2 = 1$) to get of sense of which classifiers offer the best compromise, similarly to how F1 balances precision and recall, but the value of these weights could be explored in future research. Here, we mainly use **F1+FAIR** as a quick way to gauge the performance-fairness trade-off.

5 Results

To investigate the performance and fairness of the classifiers, we first compare them on average across all weeks to identify the best/fairest ones (section 5.1). Next, we explore these results per week, to assess whether fair early prediction is possible (section 5.2). Lastly, we delve into the fairness results per protected variable and week, to isolate the specific biases in the ML predictions (section 5.3).

5.1 Classifiers Evaluation

Table 3 provides the results of all metrics reported in section 4 for all classifiers, averaged over the folds and the feature sets (e.g., 0.82 is the average F1 performance of all LR classifiers trained respectively on the data accumulated so far at the end of each of the 4 weeks). For the fairness metrics, we consider here only their average over all protected variables to ease comparison of the classifiers, but we explore the fairness per protected variable later on in section 5.3.

⁵ <https://pypi.org/project/abroca/> and <https://pypi.org/project/maddlib/>

Table 3. Average performance and fairness per classifier.

Metric	Baseline	LR	MLP	NB	RF	SGB	SVM	Statistically best model(s)
F1	0.47	0.82	0.76	0.72	0.79	0.78	0.80	LR $\approx$ SVM $\approx$ RF $\approx$ SGB
ABROCA	- ^a	0.06	0.08	0.07	0.07	0.07	0.08	LR $\approx$ SGB $\approx$ RF $\approx$ NB
MADD	- ^a	0.86	0.71	0.31	0.94	0.76	0.81	NB
F1+FAIR	- ^a	0.62	0.55	0.63	0.54	0.59	0.57	NB $\approx$ LR $\approx$ SGB

^a Studying the baselines’ fairness is irrelevant, because they always generate the same prediction.

Table 3 shows that all ML classifiers achieve substantially higher F1 than the baseline, which is consistent with the literature. However, the trends in terms of fairness are more complex to analyze. We investigate the performance-fairness trade-off by running a statistical analysis to formally compare the classifiers. Namely, for each metric, we run a linear mixed-effect model with that metric at each fold of cross-validation as the dependent variable, classifiers as the factor, and week as the random effect. To account for multiple comparisons, we adjust all p -values with the Benjamini & Hochberg procedure. All mixed models reveal a significant main effect of classifier ($p < 0.001$) with large effect sizes ($\eta^2 > 0.8$). Next, we run adjusted pairwise comparisons to explore these main effects, and report in the last column of Table 3 which classifier(s) significantly outperform all of the others, with no statistical difference among them. The pairwise comparisons reveal three main findings:

1. In terms of performance (F1), the best classifiers are LR, SVM, RF and SGB, with F1 close to 0.8. In contrast, NB is significantly worse than all other ML classifiers (F1 of 0.72), but still significantly outperforms the baseline.
2. As for fairness, NB yields significantly better MADD than all other classifiers, whereas for ABROCA, LR, SGB, RF, and NB are tied for the best.
3. In terms of F1+FAIR, NB, LR and SGB are tied for the best.

All in all, these results show that NB, LR and SGB offer the best compromise between performances and fairness, especially as revealed by the F1+FAIR metric. Specifically, LR exhibits the highest F1, but yields contradicting fairness results (ranked first for ABROCA, but fifth for MADD). NB, on the other hand, is much less accurate than LR, but performs very well on all fairness metrics, suggesting that this model is more fair at the expense of lower performances. Lastly, SGB appears to offer a compromise between LR and NB: it is statistically as good as LR in terms of F1, for decent fairness results (ranked second for ABROCA, third for MADD). Thus, for the rest of the paper, we will focus only on LR, NB and SGB, as we want to ascertain what levels of unfairness remain even among the best and/or most fair classifiers.

Another interesting finding is that the ABROCA values are very low (less than 0.1), whereas the MADD values are overall on the medium side (recall that ABROCA ranges from [0-1] and MADD [0-2], 0 being the best for both). This indicates that the classifiers generate similar performances across groups on average for all protected variables (low ABROCA), but their behaviors are quite impacted by the learners’ group membership (medium MADD).

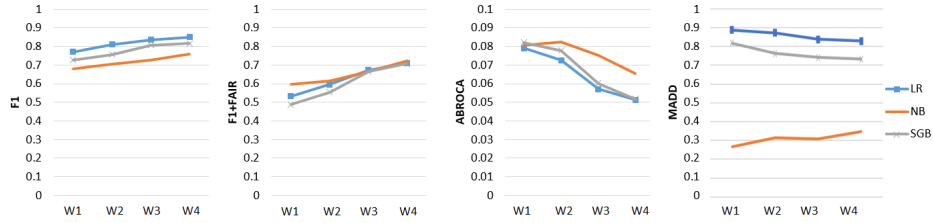


Figure 1. Weekly results for all performance and fairness metrics. Notice that the Y-axis for ABROCA is narrower for readability.

5.2 Feature Sets Evaluation

We now look at the results per feature set, in Figure 1. The figure indicates that all classifiers improve on the F1 metric when data is accumulated over time, which is consistent with existing work. F1+FAIR and ABROCA also get better as more data is accumulated overtime, whereas MADD appears more stable. An ANOVA analysis confirms that there is a significant main effect of feature set on F1, F1+FAIR and ABROCA ($p < 0.005, \eta^2 > 0.38$), but not on MADD ($p = 0.06, \eta^2 = 0.01$). Adjusted pairwise comparisons among the weeks further show that the best F1, F1+FAIR and ABROCA are achieved at Week 3 and Week 4, whereas Week 1 is worse than all feature sets. Overall, these findings do show that accumulating data over at least three weeks is valuable not only in terms of performance, but also ABROCA. They also suggest that MADD is less sensitive to time, perhaps because the classifiers behave similarly across the weeks, in spite of their improved performance.

5.3 Fairness per Protected Variable

We now explore the fairness of the classifiers for each protected variable reported in Table 2. Following up on the previous analysis, we focus on ABROCA and MADD averaged over all feature sets, as well as specifically on the features sets for week 1 and week 4 (the later capturing all logs from week 1-4). We do so for simplicity, to get a sense of the fairness early in the MOOC (Week 1), just before the final exam (Week 4), and overall (average). Table 4 provides these fairness results with ABROCA, and Table 5 with MADD. Please refer to Table 2 for the definition of G_0 and G_1 . In general, Table 4 and Table 5 show that the direction ($\downarrow$ / $\uparrow$) are very inconsistent and often contradictory both among classifiers, and among ABROCA and MADD. One possible explanation is that ABROCA captures very low levels of unfairness in the classifiers as said above, and thus the actual directions of the bias ABROCA captures is not that important, since all classifiers are overall quite fair. As for MADD, NB is overall fairer than LR and SGB, hence the same explanation might hold: the direction of NB’s bias might not be that important when its MADD value is low.

That being said, both ABROCA and MADD yield some of their highest values for **InCohort**. In terms of MADD, all models are biased against learners not

Table 4. ABROCA results ($\downarrow$ means more unfairness towards G_0 , $\uparrow$ towards G_1).

Variable	Week 1			Week 4			Weekly average		
	LR	NB	SGB	LR	NB	SGB	LR	NB	SGB
Age	0.07 $\downarrow$	0.09 $\uparrow$	0.07 $\downarrow$	0.04 $\uparrow$	0.06 $\uparrow$	0.04 $\uparrow$	0.05 $\downarrow$	0.07 $\uparrow$	0.05 $\downarrow$
Gender	0.10 $\uparrow$	0.07 $\uparrow$	0.12 $\uparrow$	0.05 $\uparrow$	0.06 $\uparrow$	0.05 $\uparrow$	0.08 $\uparrow$	0.07 $\uparrow$	0.08 $\uparrow$
Country	0.08 $\downarrow$	0.08 $\uparrow$	0.08 $\downarrow$	0.05 $\downarrow$	0.07 $\uparrow$	0.06 $\downarrow$	0.07 $\downarrow$	0.08 $\uparrow$	0.07 $\downarrow$
Lang	0.08 $\downarrow$	0.08 $\uparrow$	0.09 $\uparrow$	0.05 $\uparrow$	0.04 $\uparrow$	0.05 $\uparrow$	0.06 $\uparrow$	0.07 $\uparrow$	0.07 $\uparrow$
JobStatus	0.10 $\uparrow$	0.08 $\downarrow$	0.09 $\uparrow$	0.04 $\uparrow$	0.06 $\downarrow$	0.05 $\uparrow$	0.06 $\uparrow$	0.08 $\downarrow$	0.07 $\uparrow$
HasChild	0.06 $\downarrow$	0.08 $\uparrow$	0.06 $\downarrow$	0.04 $\uparrow$	0.07 $\uparrow$	0.05 $\downarrow$	0.05 $\downarrow$	0.08 $\uparrow$	0.06 $\downarrow$
LearnerEdu	0.08 $\uparrow$	0.08 $\downarrow$	0.06 $\uparrow$	0.04 $\uparrow$	0.06 $\uparrow$	0.05 $\downarrow$	0.06 $\uparrow$	0.07 $\downarrow$	0.05 $\uparrow$
ParentsEdu	0.09 $\downarrow$	0.09 $\downarrow$	0.09 $\downarrow$	0.06 $\downarrow$	0.07 $\downarrow$	0.08 $\downarrow$	0.08 $\downarrow$	0.09 $\downarrow$	0.09 $\downarrow$
FirstMOOC	0.06 $\uparrow$	0.06 $\downarrow$	0.08 $\uparrow$	0.05 $\uparrow$	0.05 $\uparrow$	0.04 $\uparrow$	0.06 $\uparrow$	0.06 $\uparrow$	0.06 $\uparrow$
InCohort	0.08 $\uparrow$	0.08 $\uparrow$	0.11 $\downarrow$	0.10 $\uparrow$	0.12 $\uparrow$	0.06 $\uparrow$	0.09 $\uparrow$	0.10 $\uparrow$	0.08 $\uparrow$

Table 5. MADD results ($\downarrow$ means more unfairness towards G_0 , $\uparrow$ towards G_1).

Variable	Week 1			Week 4			Weekly average		
	LR	NB	SGB	LR	NB	SGB	LR	NB	SGB
Age	0.71 $\downarrow$	0.16 $\downarrow$	0.69 $\downarrow$	0.63 $\downarrow$	0.27 $\downarrow$	0.59 $\downarrow$	0.67 $\downarrow$	0.22 $\downarrow$	0.62 $\downarrow$
Gender	0.80 $\downarrow$	0.15 $\uparrow$	0.87 $\downarrow$	0.75 $\uparrow$	0.22 $\uparrow$	0.66 $\uparrow$	0.78 $\downarrow$	0.20 $\uparrow$	0.78 $\downarrow$
Country	1.22 $\downarrow$	0.31 $\downarrow$	0.91 $\downarrow$	1.11 $\downarrow$	0.48 $\downarrow$	0.92 $\downarrow$	1.14 $\downarrow$	0.41 $\downarrow$	0.87 $\downarrow$
Lang	0.86 $\downarrow$	0.20 $\uparrow$	0.86 $\downarrow$	0.87 $\downarrow$	0.31 $\downarrow$	0.69 $\downarrow$	0.87 $\downarrow$	0.27 $\downarrow$	0.74 $\downarrow$
JobStatus	0.74 $\uparrow$	0.26 $\uparrow$	0.66 $\uparrow$	0.72 $\uparrow$	0.30 $\uparrow$	0.62 $\uparrow$	0.74 $\uparrow$	0.27 $\uparrow$	0.63 $\uparrow$
HasChild	0.69 $\downarrow$	0.25 $\downarrow$	0.69 $\downarrow$	0.63 $\downarrow$	0.24 $\downarrow$	0.57 $\downarrow$	0.66 $\downarrow$	0.23 $\downarrow$	0.63 $\downarrow$
LearnerEdu	0.74 $\downarrow$	0.16 $\uparrow$	0.66 $\uparrow$	0.67 $\downarrow$	0.20 $\uparrow$	0.56 $\uparrow$	0.72 $\downarrow$	0.18 $\uparrow$	0.62 $\downarrow$
ParentsEdu	0.74 $\uparrow$	0.17 $\uparrow$	0.68 $\downarrow$	0.66 $\downarrow$	0.20 $\uparrow$	0.62 $\downarrow$	0.67 $\uparrow$	0.20 $\uparrow$	0.66 $\downarrow$
FirstMOOC	0.69 $\downarrow$	0.20 $\downarrow$	0.72 $\downarrow$	0.68 $\uparrow$	0.25 $\uparrow$	0.68 $\uparrow$	0.72 $\uparrow$	0.22 $\uparrow$	0.67 $\uparrow$
InCohort	1.69 $\downarrow$	0.81 $\downarrow$	1.44 $\downarrow$	1.56 $\downarrow$	1.01 $\downarrow$	1.40 $\downarrow$	1.59 $\downarrow$	0.90 $\downarrow$	1.43 $\downarrow$

in cohort, whereas ABROCA’s results mostly show the opposite. While difficult to interpret, these results do suggest that the fairness of the classifiers is impacted by **InCohort**, perhaps because these students might have used the MOOC in a more intensive way as part of their degree or because they had a high level of interaction with peer learners outside of the MOOC, which may be less true for other learners. MADD also captures high unfairness for **Country**, showing that all classifiers (especially LR and SGB) are more biased against African and Haitian learners. This confirms how important it is to consider the country of origins in MOOCs’ fairness analysis, as well as the high need to leverage data from understudied continents like Africa to actually measure how trustful our classifiers are. Empirically, from forum posts, we also know that a higher proportion of MOOC users in Africa tend to download some of the resources from the MOOC for offline consultation, which would result in an overall lower number of actions and significantly different interaction patterns.

There are a few more interesting findings for the other variables. Specifically, MADD captures similar trends in all classifiers for **Age** (more biased against older learners), **JobStatus** (more biased against unemployed and students), and

HasChildren (more biased against learners who have children). These trends are expected at least for **Age** and **HasChildren** (cf. section 4.3), as older learners with children are typically less likely to have ample time to study. For **JobStatus**, while we hypothesized that having a job might impend learning, MADD suggests otherwise. The fact that we grouped together unemployed learners and full-time students might explain this finding, which could be addressed in future work.

When comparing our results with [7,17] for **Gender**, we can notice that both ABROCA and MADD yield some of their highest values for this variable, thus suggesting some levels of gender bias in the predictions. However, the direction of this bias differs across the classifiers and metrics, suggesting the need to run further analysis of gender biases on MOOC datasets.

6 Conclusion and Discussion

In this work, we investigated the performance and fairness of ML classifiers in the task of predicting completion of a MOOC, which is important to assess the possibility of fair interventions to all at-risk MOOC learners. We did so by training a set of standard ML classifiers on features capturing the learners’ behaviors in a MOOC. We then explored the possible biases in the predictions based on demographics and contextual information we collected via a survey at the beginning of the MOOC. Our main contribution is addressing the key gap of understanding ML fairness in MOOCs, which so far has been addressed solely by two studies [7,17] for the gender bias. We contribute to their work by providing more insights into biases based on a larger set of protected variables. A second contribution is studying a dataset where more than half of the learners live and learn in Africa, an understudied continent in AIED fairness analysis, despite the many calls for expanding fairness analysis in this context.

Our results do show that biases can arise in MOOCs, especially towards African and Haitian learners, as well as depending on whether the learners have children and enrolled in the MOOC as part of their university’s degree requirement. This highlights the need to explore further these biases and consider, whenever possible, learners from understudied regions of the globe and contextual variables beyond demographics.

Our results also suggest that in general, there is to some extent a trade-off in terms of performance and fairness. Interestingly, we did find that the NB classifiers consistently generate less accurate but fairer predictions, which was already shown in [19] in a different context. Still in line with [19], we also found that LR is very accurate but with varying levels of fairness across protected attributes. It is worthwhile investigating whether these trends still hold in other educational datasets for these models, as it would offer key insights for stakeholders in terms of what classifiers to use (NB for fairness, LR for accuracy).

Lastly, the two fairness metrics we used (ABROCA and MADD) generated contrasted results. Namely, ABROCA assessed the classifiers as much fairer than MADD. The key difference between both metrics is that unlike MADD, ABROCA leverages the ground truth (whether the learners actually failed).

However, the ground truth might be biased itself (e.g., learners with easy access to higher education might be more successful in MOOCs), which is not accounted for by metrics like ABROCA. MADD informs about how independent the classifiers’ output is from all protected variables, regardless of the accuracy of this output. While both metrics can be useful depending on the context and stakeholders, we do show the value of leveraging and carefully comparing both.

There are some limitations in our work, which can be addressed in future work. First, we focus on solely one MOOC, meaning that our results might not generalize to other MOOCs. Hence, we see our study as a valuable preliminary step towards building a stronger understanding of ML fairness in MOOCs. Focusing on one MOOC also allowed us to collect the demographic and contextual information we needed about the learners, as this information is typically not available in MOOCs, and it entailed modifying the MOOC to add a dedicated survey. All of this was possible solely because we could collaborate with the MOOC’s team. This MOOC was also a suitable choice to better understand issues with understudied countries, thanks to its strong popularity in Africa and Haiti. In future work, we plan to replicate our analysis on future sessions of the MOOC, as well as extend it to other MOOCs popular in Africa.

A second limitation is the fact that we binarized the protected variables, both for simplicity given the various scales of each variable, and because ABROCA and MADD can only handle binary groups by default, as they are limited to comparing two ROC curves or two histograms, respectively. While several fairness studies also used binary groups for that reason, e.g., [24,9,22,19], in future work, it could be relevant to examine finer-grained ones. In particular, we plan to study fairness across specific sub-parts of Africa, and by country when possible.

Acknowledgement: This work was supported by the Sorbonne Center for Artificial Intelligence (SCAI) and the *Direction du numérique pour l’éducation* (DNE) of the French Ministry of Education. The opinions expressed in this paper are those of the authors and do not necessarily reflect those of DNE and SCAI. We thank Rémi Bachelet and the MOOC GdP’s team for the data collection.

References

1. Baker, R.S., Hawn, A.: Algorithmic bias in education. *Int. J. Artif. Intell. Educ.* pp. 1052–1092 (2021)
2. Bridgeman, B., Trapani, C., Attali, Y.: Considering fairness and validity in evaluating automated scoring. In: *Proc. of the Annual meeting of the National Council on Measurement in Education*, San Diego, CA (2009)
3. Bridgeman, B., Trapani, C., Attali, Y.: Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Appl. Meas. Educ.* **25**(1), 27–40 (2012)
4. Buolamwini, J., Gebru, T.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Proc. of the 1st Conference on Fairness, Accountability and Transparency*. pp. 77–91 (2018)
5. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., et. al.: A clarification of the nuances in the fairness metrics landscape. *Sci. Rep.* **12**(1), 4209 (2022)

6. Chen, J., Fang, B., Zhang, H., Xue, X.: A systematic review for mooc dropout prediction from the perspective of machine learning. *Interact. Learn. Environ.* pp. 1–14 (2022)
7. Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: *Proc. of the 9th International Conference on Learning Analytics & Knowledge*. pp. 225–234 (2019)
8. Idowu, J.A.: Debiasing education algorithms. *Int. J. Artif. Intell. Educ.* pp. 1–31 (2024)
9. Kai, S., Andres, J., Paquette, L., Baker, R., Molnar, K., Watkins, H., Moore, M.: Predicting student retention from behavior in an online orientation course. In: *Proc. of the Int. Conference on Educational Data Mining*. pp. 250–255 (2017)
10. Kizilcec, R.F., Lee, H.: Algorithmic fairness in education. In: *The Ethics of Artificial Intelligence in Education*, pp. 174–202. Routledge (2022)
11. Lembani, R., Gunter, A., Breines, M., Dalu, M.T.B.: The same course, different access: the digital divide between urban and rural distance education students in south africa. *J. Geogr. High. Educ.* **44**(1), 70–84 (2020)
12. Li, X., Song, D., Han, M., Zhang, Y., Kizilcec, R.F.: On the limits of algorithmic prediction across the globe. *arXiv preprint arXiv:2103.15212* (2021)
13. Littenberg-Tobias, J., Reich, J.: Evaluating access, quality, and equity in online learning: A case study of a mooc-based blended professional degree program. *Internet High. Educ.* **47**, 100759 (2020)
14. Moreno-Marcos, P.M., Alario-Hoyos, C., Muñoz-Merino, P.J., Kloos, C.D.: Prediction in moocs: A review and future research directions. *IEEE Trans. Learn. Technol.* **12**(3), 384–401 (2018)
15. Naismith, B., Han, N.R., Juffs, A., Hill, B., Zheng, D.: Accurate measurement of lexical sophistication in esl with reference to learner data. In: *Proc. of the 11th International Conference on Educational Data Mining*. pp. 259–265 (2018)
16. Rzepka, N., Simbeck, K., Müller, H., Pinkwart, N.: Fairness of in-session dropout prediction. In: *Proc. of the International Conference on Computer Supported Education*. pp. 316–326 (2022)
17. Sha, L., Raković, M., Das, A., Gašević, D., Chen, G.: Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Trans. Learn. Technol.* **15**(4), 481–492 (2022)
18. Sylla, K., Nkwetchoua, G., Bouchet, F.: How does the use of open digital spaces impact students success and dropout in a virtual university? In: *Proc. of the Conference on Cognition and Exploratory Learning in Digital Age*. pp. 251–258 (2022)
19. Verger, M., Lallé, S., Bouchet, F., Luengo, V.: Is your model” madd”? a novel metric to evaluate algorithmic fairness for predictive student models. In: *Proc. of the International Conference on Educational Data Mining*. pp. 91–102 (2023)
20. Wang, Z., Zechner, K., Sun, Y.: Monitoring the performance of human and automated scores for spoken responses. *Lang. Test.* **35**(1), 101–120 (2018)
21. Xing, W., Du, D.: Dropout prediction in moocs: Using deep learning for personalized intervention. *J. Educ. Comput. Res.* **57**(3), 547–570 (2019)
22. Yu, R., Lee, H., Kizilcec, R.F.: Should college dropout prediction models include protected attributes? In: *Proc. of the ACM Conference on Learning@Scale*. pp. 91–100 (2021)
23. Yudelson, M., Fancsali, S., Ritter, S., Berman, S., Nixon, T., Joshi, A.: Better data beats big data. In: *Proc. of the Int. Conference on Educational data mining* (2014)
24. Zambrano, A.F., Zhang, J., Baker, R.S.: Investigating algorithmic bias on bayesian knowledge tracing and carelessness detectors. In: *Proc. of the 14th International Conference on Learning Analytics & Knowledge*. pp. 349–359 (2024)