Evaluation of group fairness measures in student performance prediction problems

Tai Le Quy[1], Thi Huyen Nguyen[1], Gunnar Friege[2], and Eirini Ntoutsi[3]

[1] L3S Research Center, Leibniz University Hannover, Hanover, Germany
{tai, nguyen}@l3s.de
[2] Institute for Didactics of Mathematics and Physics, Leibniz University Hannover, Hanover, Germany
friege@idmp.uni-hannover.de
[3] Institute of Computer Science, Free University Berlin, Berlin, Germany
eirini.ntoutsi@fu-berlin.de

Abstract. Predicting students' academic performance is one of the key tasks of educational data mining (EDM). Traditionally, the high forecasting quality of such models was deemed critical. More recently, the issues of fairness and discrimination w.r.t. protected attributes, such as gender or race, have gained attention. Although there are several fairness-aware learning approaches in EDM, a comparative evaluation of these measures is still missing. In this paper, we evaluate different group fairness measures for student performance prediction problems on various educational datasets and fairness-aware learning models. Our study shows that the choice of the fairness measure is important, likewise for the choice of the grade threshold.

Keywords: fairness · fairness measures · student performance prediction · machine learning · educational data mining

1 Introduction

Educational data mining (EDM) applies data mining, artificial intelligence (AI), and machine learning (ML) to improve academic experiences. In recent years, AI-infused technologies have been widely studied and deployed by many educational institutions [3, 19]. One of the most important tasks in EDM that attract great attention is student performance prediction. The early estimation of student learning outcomes can help detect and notify students at risk of academic failure. Besides, it supports institutional administrators in identifying key factors affecting students' grades and providing suitable interventions for outcome improvement. The performance prediction process relies on historical academic records and trains ML algorithms on labeled data to predict students' performance. Various datasets [11,26,37] and approaches [16,24,41] have been proposed for the purpose. With the widespread use and benefits of AI systems, fairness has become a crucial criterion in designing such systems.

Non-discriminative ML models have been a topic of increasing importance and growing momentum in education. Despite advances and superior accuracy

of recent ML models, some studies have shown that ML-based decisions can be biased to protected attributes such as gender or race due to historical discrimination embedded in the data [28,32]. Endeavoring to reduce biases is important and decisive in the applicability of an ML model in education. As an example, a recent study has proposed approaches that aim at predicting calculated grades of students in England as a replacement for actual grades due to the cancellation of exams during COVID-19 [5]. However, the proposal could not be applied as a consequence of some exposed historical biases.

A large variety of fairness measures have been introduced in ML area. However, choosing proper measures can be cumbersome due to the dependence of fairness on context. There are more than 20 different fairness measures introduced in the computer science research area [28,36]. In fact, no metric is universal and fits all circumstances [15, 28, 36]. Model developers should explore various fairness measures to decide the most appropriate notions for the context. Fairness is a fundamental concept of education, whereby all students must have an equal opportunity in study or be treated fairly regardless of their household income, assets, gender, or race [29]. Fairness definitions in education, hiring, and ML in the 50-year history have been discussed in the research of [20]. However, no previous work exists on the efficiency of different fairness metrics and how to choose them in educational settings.

In this paper, we provide a comprehensive study to evaluate the sufficiency of various fairness metrics in student performance prediction. We consider a group of the most prevalent fairness notions in ML. Various experiments are conducted on diverse educational datasets and evaluated using different fairness metrics. Our experiments provides users a broad view of unfairness from diverse aspects in an educational context. Besides, the results also guide the selection of suitable fairness measures to evaluate students' grade predictive models. We believe our contributions are crucial to alleviate the burden of choosing fairness measures for consideration and motivate further studies to improve the accuracy and fairness of student performance prediction models.

The rest of the paper is organized as follows. In Section 2, we present some closely related work on fairness-aware ML and student performance prediction. Section 3 describes the most popular group fairness measures in ML. Next, we conduct quantitative evaluations of predictive models on educational datasets and discuss the choice of suitable fairness metrics in Section 4. Finally, we conclude the paper in Section 5.

2 Related work

Extensive research efforts have been conducted to provide useful insights into students' performance analysis and prediction [38]. Various ML models were tested on different problem settings. Cortez et al. [11] presented an early study to predict the grades of secondary students in Portuguese and Mathematics classes. Their results showed that good predictive accuracy could be achieved when previous school period grades are available. Similarly, Berhanu et al. [7] employed Decision Tree to predict students' performance using the agriculture

college dataset. Some studies [25, 41] proposed diverse approaches to forecast students' grades in higher education. Besides, many other studies were reviewed in multiple surveys [1, 31, 33, 33, 34]. They pointed out the most common techniques such as Decision Tree, Naive Bayes, Support Vector Machines, and neural networks and dominant factors impacting predictive outcomes (i.e., Cumulative Grade Point Average, previous grades, classroom attendance, etc.).

There are more than 20 fairness notions introduced for classification [28, 36]. One of the most well-known fairness measure is *demographic parity*, so-called *statistical parity*. It requires an equal probability of positive predictions in protected and non-protected groups. However, Dwork et al. [13] argued that the metric fails to ensure individual fairness. To avoid this, Hardt et al. [18] proposed *equalized odds* metric. It measures whether a classifier predicts labels equally well for all values of attributes. Besides, many other popular metrics were introduced and used in fairness ML studies such as *predictive parity*, *predictive equality* [9], *treatment equality* [8], etc. Despite a substantial number of fairness measures, there is no metric that fits all circumstances [28, 36].

Following the evolution of fairness measures, recent studies have attempted to evaluate fairness in an educational context [17, 22, 39]. Anderson et al. [6] conducted two post-hoc fairness assessments for existing student graduation prediction models. Renzhe et al. [39] studied different combinations of student data sources for building highly predictive and fair models for predictions of college success. Jiang et al. [23] proposed several strategies to mitigate bias in the LSTM grade prediction model. They report experimental results on the true positive rate (TPR), true negative rate (TNR), and accuracy.

3 Fairness measures

Table 1: An overview of group fairness measures

Measures	Proposed by	Published year	#Citations
Statistical parity	[13]	2012	2,367
Equal opportunity	[18]	2016	2,575
Equalized odds	[18]	2016	2,575
Predictive parity	[9]	2017	1,430
Predictive equality	[10]	2017	878
Treatment equality	[8]	2018	626
Absolute Between-ROC Area	[17]	2019	84

This section presents the most prevalent group fairness notions used in ML. The list of notions[4] is summarized in Table 1. To simplify, we consider the student performance prediction problem as a binary classification task, which is formalized as below:

Let $\mathcal{D}$ be a binary classification dataset with class attribute $Y = \{+, -\}$, e.g., $Y = \{pass, fail\}$. S is a binary protected attribute, $S \in \{s, \overline{s}\}$, e.g., S = "gender", $S \in \{female, male\}$. In which, s is the discriminated group (*protected group*), e.g., "*female*", and $\overline{s}$ is the non-discriminated group (*non-protected*

[4]The number of citations is reported by Google Scholar on 1^{st} August 2022.

group), e.g., "*male*". The predicted outcome is denoted as $\hat{Y} = \{+, -\}$. The notions s_+ (s_-), $\overline{s}_+$ ($\overline{s}_-$) are used to denote the protected and non-protected groups for the positive (negative, respectively) class.

We use a confusion matrix (Fig. 1) to demonstrate the group fairness measures with an example of a dataset with 100 instances, class $Y = \{pass, fail\}$. The protected attribute is "gender", and the protected group is "female"; the distribution of "female":"male" is 46:54. Examples of fairness measures in the following sub-sections are computed based on this confusion matrix.

		Predicted class	
		Positive +	**Negative -**
Actual class	**Positive +**	True Positive (TP) $TP_{prot} + TP_{non-prot}$ **70** (32:38)	False Negative (FN) $FN_{prot} + FN_{non-prot}$ **10** (4:6)
	Negative -	False Positive (FP) $FP_{prot} + FP_{non-prot}$ **9** (4:5)	True Negative (TN) $TN_{prot} + TN_{non-prot}$ **11** (6:5)

Fig. 1: The confusion matrix with an example

3.1 Statistical parity

Statistical parity (denoted as *SP*) is a well-known group fairness measure [13], whereby the output of any classifier satisfies statistical parity if the difference (bias) in the predicted outcome ($\hat{Y}$) between any two groups under study (i.e., s and $\overline{s}$) is up to a predefined tolerance threshold ϵ:

$$P(\hat{Y}|S = s) - P(\hat{Y}|S = \overline{s}) \leq \epsilon. \tag{1}$$

We use the violation of statistical parity [27, 35, 40] to measure the bias of a classifier:

$$SP = P(\hat{Y} = +|S = \overline{s}) - P(\hat{Y} = +|S = s). \tag{2}$$

The value range: $SP \in [-1, 1]$, with $SP = 0$ indicating no discrimination, $SP \in (0, 1]$ designating that the protected group is discriminated, and $SP \in [-1, 0)$ standing for *reverse discrimination* (the non-protected group is discriminated). In our example (Fig. 1), this measure shows the proportion of "*pass*" students between the two demographic subgroups. $SP = \frac{38+6}{54} - \frac{32+4}{46} \approx 0.0322$.

3.2 Equal opportunity

Equal opportunity (denoted as *EO*) is proposed by Hardt et al. [18], whereby a binary predicted outcome $\hat{Y}$ satisfies equal opportunity w.r.t. the protected attribute S and the class attribute Y if:

$$P(\hat{Y} = +|S = s, Y = +) = P(\hat{Y} = +|S = \overline{s}, Y = +). \tag{3}$$

In other words, the protected and non-protected groups should have equal true positive rates (TPR) [28, 36], $TPR = \frac{TP}{TP+FN}$ (i.e., the classifier should give similar results for students of both genders with actual "*pass*" class). A classifier with equal false negative rates (FNR), $FNR = \frac{FN}{TP+FN}$, will also have equal TPR [36]. The equal opportunity can be measured by:

$$EO = |P(\hat{Y} = -|Y = +, S = \overline{s}) - P(\hat{Y} = -|Y = +, S = s)|. \quad (4)$$

The value range: $EO \in [0, 1]$; with 0 standing no discrimination and 1 indicating maximum discrimination. In our example, $EO = |\frac{38}{38+6} - \frac{32}{32+4}| \approx 0.0253$.

3.3 Equalized odds

A predictor $\hat{Y}$ is satisfied *equalized odds* (denoted as *EOd*) w.r.t. the protected attribute S and class label Y, if "$\hat{Y}$ and S are independent conditional on Y" [18]. Specifically, predicted true positive and false positive probabilities should be the same between male and female student groups.

$$P(\hat{Y} = +|S = s, Y = y) = P(\hat{Y} = +|S = \overline{s}, Y = y), \quad y \in \{+, -\}. \quad (5)$$

Therefrom, we can measure the equalized odds as the following [21, 27]:

$$EOd = \sum_{y \in \{+,-\}} |P(\hat{Y} = +|S = s, Y = y) - P(\hat{Y} = +|S = \overline{s}, Y = y)|. \quad (6)$$

The value range: $EOd \in [0, 2]$; with 0 standing for no discrimination and 2 indicating the maximum discrimination. In our example, $EOd = |\frac{32}{32+4} - \frac{38}{38+6}| + |\frac{4}{4+6} - \frac{5}{5+5}| \approx 0.1253$.

3.4 Predictive parity

Predictive parity [9] (denoted as *PP*) is satisfied if both protected and non-protected groups have an equal positive predictive value (PPV) or *Precision*, $PPV = \frac{TP}{TP+FP}$, i.e., the probability of a student predicted to "*pass*" actually having "*pass*" class should be the same, for both male and female students.

$$P(Y = +|\hat{Y} = +, S = s) = P(Y = +|\hat{Y} = +, S = \overline{s}). \quad (7)$$

Therefore, we report the predictive parity measure as:

$$PP = |P(Y = +|\hat{Y} = +, S = s) - P(Y = +|\hat{Y} = +, S = \overline{s})|. \quad (8)$$

where $PP \in [0, 1]$, with 0 standing for no discrimination and 1 indicating the maximum discrimination. $PP = \frac{32}{32+4} - \frac{38}{38+5} \approx 0.0052$, in our example.

3.5 Predictive equality

Predictive equality [10] (denoted as PE), also referred as false positive error (FPR) rate balance [9] ($FPR = \frac{FP}{TN + FP}$), aims to the equality of decision's accuracy across the protected and non-protected groups. In detail, the probability of students with an actual "*fail*" class being incorrectly assigned to the "*pass*" class should be the same for both male and female students.

$$P(\hat{Y} = +|Y = -, S = s) = P(\hat{Y} = +|Y = -, S = \overline{s}). \tag{9}$$

In practice, researchers report predictive equality measure by the difference of $FPRs$ [21]:

$$PE = |P(\hat{Y} = +|Y = -, S = s) - P(\hat{Y} = +|Y = -, S = \overline{s})|. \tag{10}$$

The value range: $PE \in [0, 1]$, 0 and 1 indicate no discrimination and maximum discrimination, respectively. $PE = |\frac{4}{6+4} - \frac{5}{5+5}| = 0.1$, in our example.

3.6 Treatment equality

Treatment equality [8] (denoted as TE) is satisfied if the ratios of false negatives and false positives are the same for both protected and non-protected groups.

$$\frac{FN_{prot.}}{FP_{prot.}} = \frac{FN_{non-prot.}}{FP_{non-prot.}}. \tag{11}$$

In our paper, we report the treatment equality by the difference between two ratios described in Eq.11. The metric becomes unbounded if $FP_{prot.}$ or $FP_{non-prot.}$ is zero[5]. In our example, $TE = -0.2$, because the ratios of FN and FP are 1 and 1.2 for female and male groups, respectively.

3.7 Absolute Between-ROC Area

Absolute Between-ROC Area (ABROCA) [17] is based on the Receiver Operating Characteristics (ROC) curve. It measures the divergence between the protected (ROC_s) and non-protected group ($ROC_{\overline{s}}$) curves across all possible thresholds $t \in [0, 1]$ of FPR and TPR. The absolute difference between the two curves is measured to capture the case that the curves may cross each other.

$$\int_0^1 \mid ROC_s(t) - ROC_{\overline{s}}(t) \mid \, dt. \tag{12}$$

The value range: $ABROCA \in [0, 1]$. The lower value indicates a lower difference in the predictions between the two groups and, therefore, a fairer model.

[5] https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-post-training-bias-metric-te.html

4 Evaluation

In this section, we evaluate the performances of predictive models w.r.t. accuracy and fairness measures on five datasets and investigate the effect of choosing grade threshold on fairness measures.

4.1 Datasets

We evaluate the fairness measures on popular educational datasets [27, 30, 38], which are summarized in Table 2. All datasets are imbalanced, as shown in the imbalance ratio (IR) column.

Table 2: An overview of educational datasets

Datasets	#Instances	#Instances (cleaned)	#Attributes	Protected attribute	Class label	IR (+:-)
Law school	20,798	20,798	12	Race	Pass the bar exam	8.07:1
PISA	5,233	3,404	24	Gender	Reading score	1.35:1
Studden academics	131	131	22	Gender	ESP	3.70:1
Student performance	649	649	33	Gender	Final grade	5.49:1
xAPI-Edu-Data	480	480	17	Gender	Grade level	2.78:1

Law school. The Law school dataset[6] contains the law school admission records from 163 law schools in the US in 1991. The target is to predict whether a candidate would pass the bar exam or not. The protected attribute is "race" $= \{white, non-white\}$, where *"non-white"* is the protected group.

PISA dataset. The PISA dataset[7] contains information on the performance of American students [14] taking the exam in 2009 from the Program for International Student Assessment (PISA). The grade threshold ("readingScore" attribute) is chosen at 500 to compute the class label $= \{low, high\}$ since the mean reading score is 497.6. The experiments are performed on the cleaned version of this dataset with 3,404 instances after removing missing values.

Student academics performance dataset. The student academics performance dataset[8] [19] consists of socio-economic, demographic, and academic information of students from three different colleges in India with 22 attributes. The class label is ESP (end semester percentage). In this paper, we encode class label as a binary attribute with values { *"pass", "good-and-higher"*}, where *"good-and-higher"* is a positive class.

Student performance dataset. The student performance dataset[9] [11] was collected in two Portuguese schools in 2005 - 2006. It contains 33 features describing demographics, grades, social and school-related information of students. "gender" is considered the protected attribute. The target is to predict the final outcome. The class label $= \{pass, fail\}$ is computed based on the final grade (attribute "G3") as $\{<10, \geq 10\}$ [11, 27].

[6] https://github.com/tailequy/fairness_dataset/tree/main/Law_school

[7] https://www.kaggle.com/econdata/pisa-test-scores

[8] https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance

[9] https://archive.ics.uci.edu/ml/datasets/student+performance

Students' academic performance dataset (xAPI-Edu-Data). xAPI-Edu-Data[10] [4] contains 480 student records described by 17 attributes collected from *Kalboard 360* learning management system. We encode the class label as a binary attribute as $\{Low, Medium - High\}$ corresponding to $\{L,\ M\ or\ H\}$ in the original dataset. The positive class is *"Medium-High"*.

4.2 Predictive models

We select four prevalent classifiers used for student performance prediction problems based on the survey of Xiao et al. [38], and two well-known fairness-aware classifiers, namely Agarwal's [2] and AdaFair [21]. In which, Agarwal's method reduces the fair classification to a sequence of cost-sensitive classification problems with the lowest (empirical) error subject to the desired constraints, and AdaFair is based on AdaBoost that further updates the weights of the instances in each boosting round. In brief, the predictive models are: 1) Decision Tree (DT); 2) Naive Bayes (NB); 3) Multi-layer Perceptron (MLP); 4) Support Vector Machines (SVM); 5) Agarwal's; 6) AdaFair. In our experiments, we use 70% of data for training and 30% for testing (single split). Predicted models are implemented and executed with default parameters provided by Scikit-learn and Iosifidis et al. [21]. Agarwal's method is implemented in the AI Fairness 360 toolkit[11].

4.3 Experimental results

Law school dataset. The results are presented in Table 3. AdaFair is the best predictive model w.r.t. fairness measures, although its balanced accuracy is significantly lower than that of other models. Besides, the fairness measures show a quite large variation across the classification methods, as demonstrated in Fig. 7-a. Furthermore, the shape and position of the ROC curves, as visualized in Fig. 2, have been changed across the predictive models, which indicates the change in the performance of models w.r.t. each value in the protected attribute. Because the datasets are imbalanced, we report the performance of predictive models on both accuracy and balanced accuracy measures.

PISA dataset. The interesting point is SVM and DT show their superiority in terms of fairness measures, although AdaFair still has very good results on fairness metrics and accuracy (Fig. 3 and Table 4). Furthermore, fairness measures have the least variability in this dataset, as shown in Fig. 7-b.

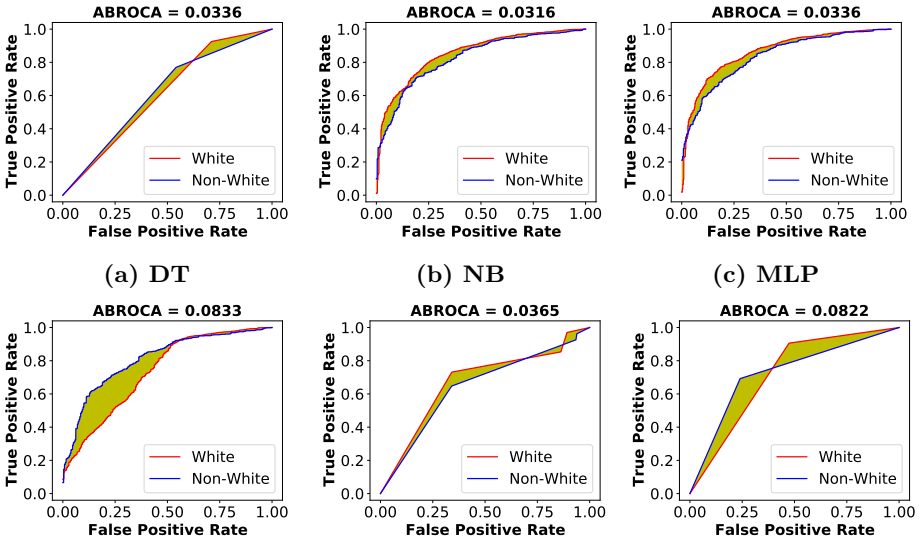
Student academics performance dataset. The AdaFair outperforms other models w.r.t. fairness measures, however, the balanced accuracy is decreased considerably (Table 5). Besides, all fairness measures have significant variation across predictive models (Fig. 4 and 7-c).

Student performance dataset. In general, all models show good accuracy (balanced accuracy) on predicting students' performance (Table 6). MLP and AdaFair models fairly guarantee the fairness of results on most measures.

[10] https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data
[11] https://github.com/Trusted-AI/AIF360

Table 3: Law school: performance of predictive models

Measures	DT	NB	MLP	SVM	Agarwal's	AdaFair
Accuracy	0.8458	0.8191	**0.9042**	0.8926	0.7952	0.8921
Balanced accuracy	0.6301	**0.7784**	0.6596	0.5029	0.5848	0.5
Statistical parity	0.1999	0.5250	0.2367	0.0052	0.0326	**0.0**
Equal opportunity	0.1557	0.4665	0.1237	0.0014	0.0202	**0.0**
Equalized odds	0.3253	0.8105	0.5501	0.0169	0.0953	**0.0**
Predictive parity	0.1424	**0.0130**	0.0754	0.1857	0.1802	0.1885
Predictive equality	0.1696	0.3440	0.4265	0.0154	0.0751	**0.0**
Treatment equality	-0.0667	22.440	0.7770	0.0039	-1.9676	**0.0**
ABROCA	0.0336	**0.0316**	0.0336	0.0833	0.0365	0.0822



(a) DT **(b) NB** **(c) MLP**

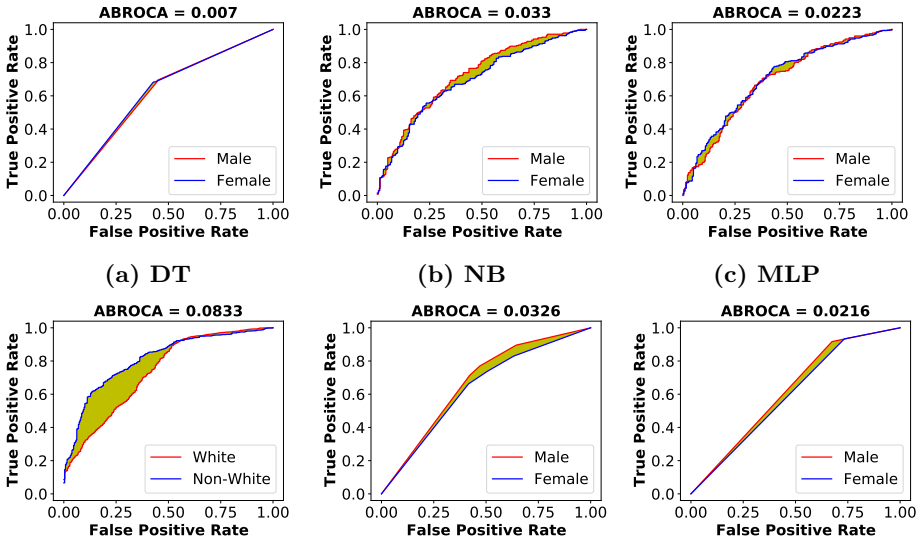
(d) SVM **(e) Agarwal's** **(f) AdaFair**

Fig. 2: Law school: ABROCA slice plots

Table 4: PISA: performance of predictive models

Measures	DT	NB	MLP	SVM	Agarwal's	AdaFair
Accuracy	0.6360	0.6624	0.6526	0.6096	0.6614	**0.6810**
Balanced accuracy	0.6224	**0.6379**	0.5732	0.5026	0.6340	0.6130
Statistical parity	-0.0200	-0.0316	-0.0771	**-0.0022**	-0.0096	-0.0573
Equal opportunity	**0.0019**	0.0262	0.0330	0.0043	0.0414	0.0164
Equalized odds	0.0165	0.0709	0.1398	**0.0068**	0.0548	0.0752
Predictive parity	0.1012	**0.0683**	0.0826	0.1108	0.0785	0.0868
Predictive equality	0.0146	0.0446	0.1067	**0.0024**	0.0134	0.0588
Treatment equality	0.5642	0.3855	-0.0251	**-0.0033**	0.4609	0.0260
ABROCA	**0.0070**	0.0330	0.0223	0.0844	0.0326	0.0216

Besides, the values of fairness measures also do not vary significantly across predictive models (Fig. 7-d), although the ABROCA slices are quite different in shape (Fig. 5).



(a) DT (b) NB (c) MLP

(d) SVM (e) Agarwal's (f) AdaFair

Fig. 3: PISA: ABROCA slice plots

xAPI-Edu-Data dataset This is a surprising dataset because the traditional classification methods show a better performance not only in terms of accuracy/balanced accuracy measures but also w.r.t. fairness measures (Table 7). In addition, variation in the values of fairness measures across the predictive models is not significant, as shown in Fig. 7-e, except for the ABROCA measure with a noticeable change in the shape (Fig. 6).

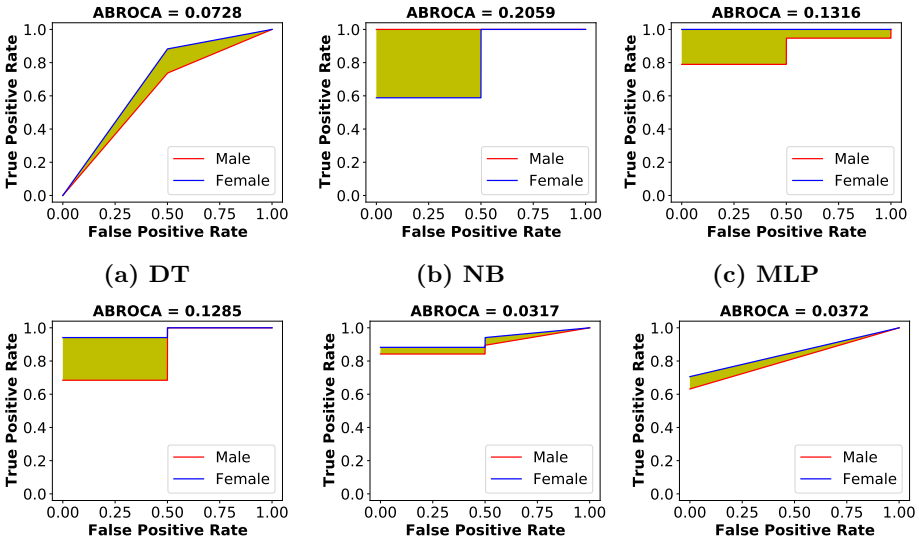
Regarding the *treatment equality* measure, this measure is entirely different from all other measures with an extensive range of values, which is visualized in Fig. 7-f[12]. On the *PISA* datasets, this TE measure shows the best values across predicted models, followed by *Law school* and *Student Academics* datasets.

Summary of results: In general, *ABOCA* is the measure with the lowest variability across predictive methods and datasets. It also clearly presents the ML model's accuracy variation over each value of the protected attribute. *Equal*

Table 5: Student academics: performance of predictive models

Measures	DT	NB	MLP	SVM	Agarwal's	AdaFair
Accuracy	0.7750	0.8750	0.8750	**0.9250**	0.8750	0.9
Balanced accuracy	0.6528	**0.8194**	**0.8194**	0.6250	**0.8194**	0.5
Statistical parity	-0.1278	-0.1328	-0.1328	0.0526	0.0677	**0.0**
Equal opportunity	0.1455	0.0991	0.2105	**0.0**	0.0123	**0.0**
Equalized odds	0.1455	0.5991	0.7105	0.5	0.5124	**0.0**
Predictive parity	**0.0042**	0.0588	0.0552	0.0397	0.0556	0.01
Predictive equality	**0.0**	0.5	0.5	0.5	0.5	**0.0**
Treatment equality	-3.0	*N/A*	*N/A*	**0.0**	*N/A*	**0.0**
ABROCA	0.0728	0.2059	0.1316	0.1285	**0.0317**	0.0372

[12]We use the abbreviations of the fairness measures and datasets in Fig. 7



(a) DT (b) NB (c) MLP

(d) SVM (e) Agarwal's (f) AdaFair

Fig. 4: Student academics: ABROCA slice plots

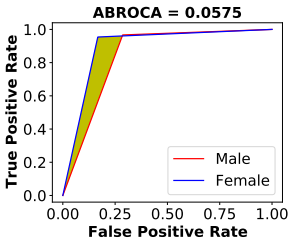
Table 6: Student performance: performance of predictive models

Measures	DT	NB	MLP	SVM	Agarwal's	AdaFair
Accuracy	0.9333	0.8974	0.9077	0.9231	0.8923	**0.9487**
Balanced accuracy	**0.8639**	0.8595	0.7840	0.7441	0.8565	0.8240
Statistical parity	-0.0382	-0.0509	-0.0630	**0.0151**	-0.0209	-0.0255
Equal opportunity	0.0125	0.0174	0.03	0.0183	0.0176	**0.0092**
Equalized odds	0.1316	0.2198	**0.1252**	0.3279	0.2200	0.1877
Predictive parity	**0.0456**	0.0591	0.0601	0.0944	0.0577	0.0639
Predictive equality	0.1190	0.2024	**0.0952**	0.3095	0.2024	0.1786
Treatment equality	2.0	7.5	**0.3333**	0.5	9.75	**0.3333**
ABROCA	0.0575	0.0686	0.0683	**0.0231**	0.0762	0.0887

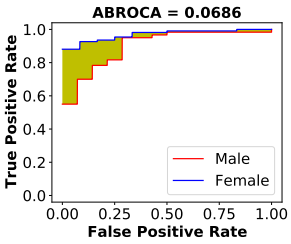
Table 7: xAPI-Edu-Data: performance of predictive models

Measures	DT	NB	MLP	SVM	Agarwal's	AdaFair
Accuracy	0.8333	0.8750	0.8750	0.8611	**0.8681**	0.8056
Balanced accuracy	0.8	**0.8970**	0.8545	0.8505	0.8859	0.8162
Statistical parity	**-0.1274**	-0.2608	-0.2112	-0.2209	-0.2505	-0.2292
Equal opportunity	0.0282	0.0974	0.0654	**0.0308**	0.0974	0.0538
Equalized odds	0.1329	0.1954	**0.1262**	0.2706	0.1684	0.3207
Predictive parity	0.0752	0.0074	0.0654	0.0088	0.0122	**0.0057**
Predictive equality	0.1047	0.0980	**0.0608**	0.2399	0.0709	0.2669
Treatment equality	1.0667	-8.0	**0.0**	-0.2667	-2.0	-1.1667
ABROCA	0.0665	**0.0216**	0.0263	0.0796	0.0293	0.1065

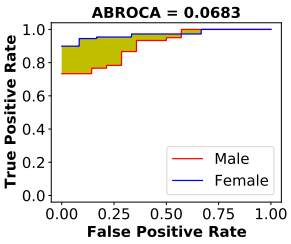
opportunity and *predictive parity* also have a slight variation across methods and datasets. *Equalized odds*, to some extent, can represent two measures *equal opportunity* and *predictive equality* as it is the sum of the other two metrics.



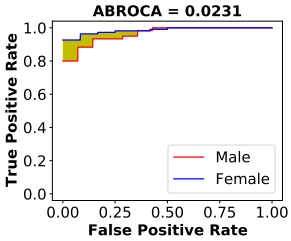
(a) DT



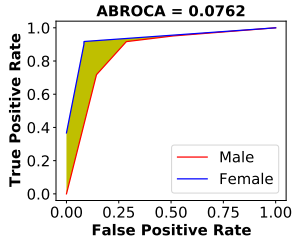
(b) NB



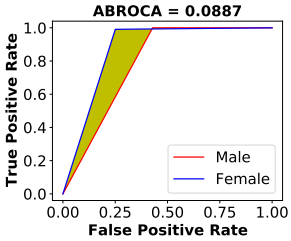
(c) MLP



(d) SVM

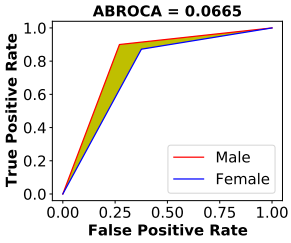


(e) Agarwal's

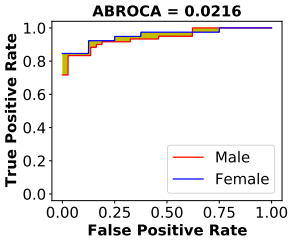


(f) AdaFair

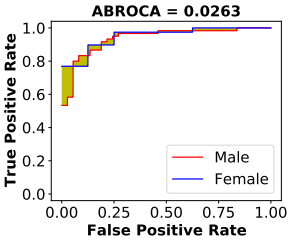
Fig. 5: Student performance: ABROCA slice plots



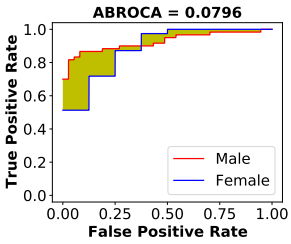
(a) DT



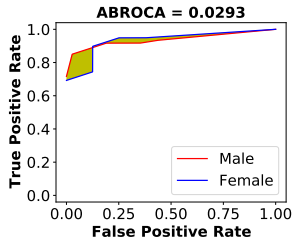
(b) NB



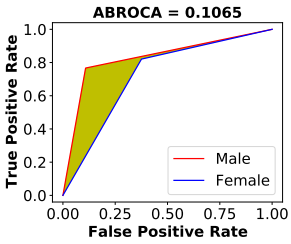
(c) MLP



(d) SVM



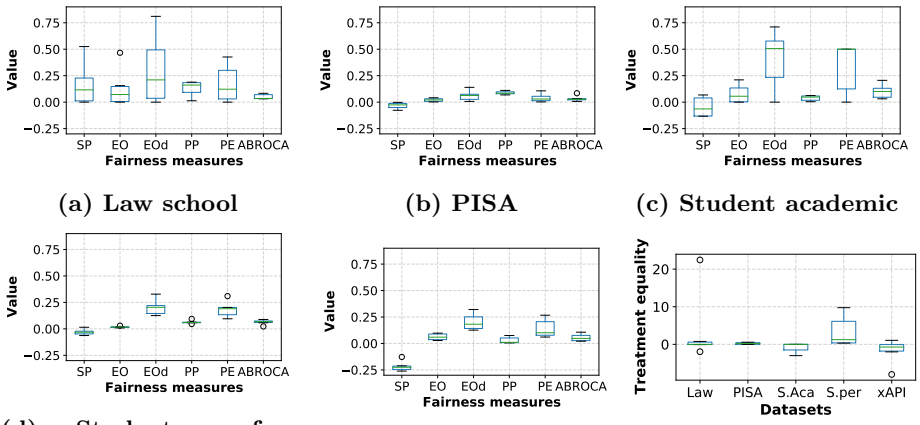
(e) Agarwal's



(f) AdaFair

Fig. 6: xAPI-Edu-Data: ABROCA slice plots.

Furthermore, *treatment equality* has a very wide range of values (sometimes the value may not be bounded), making it difficult to compare and evaluate.



(a) Law school **(b) PISA** **(c) Student academic**

(d) Student performance **(e) xAPI-Edu-Data** **(f) TE measure**

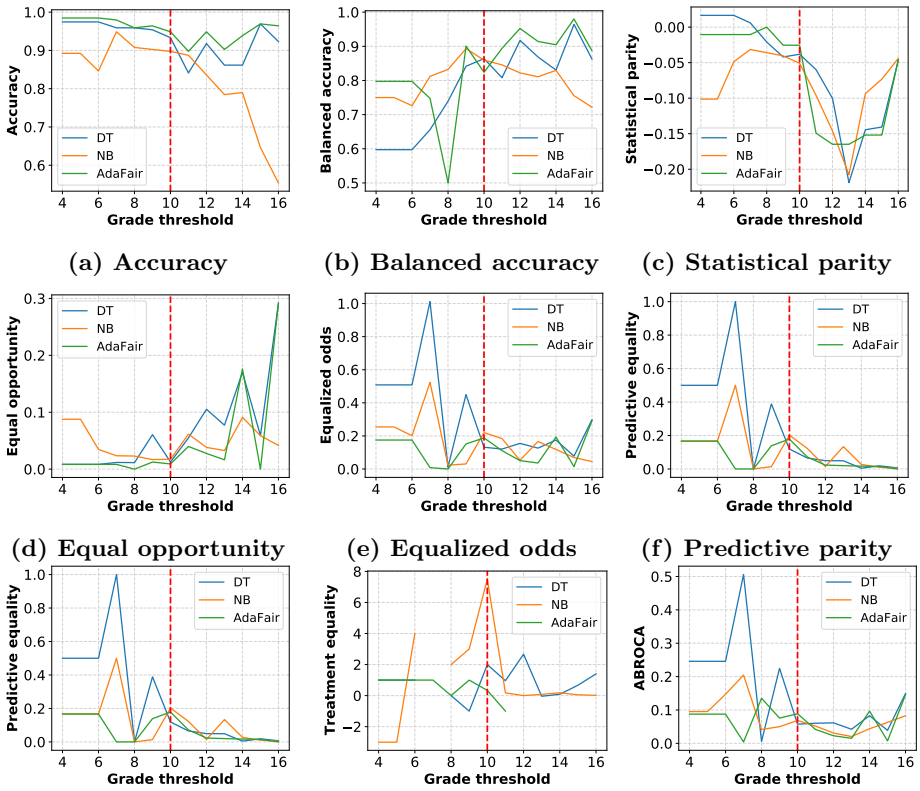
Fig. 7: Variation of fairness measures

4.4 Effect of varying grade threshold on fairness

Grade thresholds are often chosen as a basis for determining whether a candidate passes or fails an exam. In the student performance dataset, 10 (out of 20) is selected as the grade threshold [11, 27]. However, the selection of a threshold can affect the fairness of the predictive models, as shown in the IPUMS Adult dataset [12]. Hence, we investigate the effect of grade threshold on fairness by varying the threshold in a range of [4, 16], corresponding to 25% to 75% of the maximum grade (20). The results in Fig. 8 show that all fairness measures are affected by the grade threshold. When the grade threshold is gradually increased, the predictive models tend to be fairer (shown on the measures: equalized odds, predictive equality, and ABROCA). The opposite trend is observed in the remaining measures (except the treatment equality measure). Regarding the balanced accuracy, two models (DT and AdaFair) tend to predict more accurately. The NB model has a decreasing accuracy after the threshold is increased.

5 Conclusion and outlooks

In this work, we evaluate seven popular group fairness measures for student performance prediction problems. We conduct experiments using four traditional ML models and two fairness-aware ML methods on five educational datasets. Our experiments reflect variations and correlations of fairness measures across datasets and predictive models. The results provide a overview picture for the selection of fairness measure in a specific case. Besides, we investigate the effect of varying grade thresholds on the accuracy and fairness of ML models. The preliminary results suggest that choosing the threshold is an important factor contributing to ensuring fairness in the output of the ML models. In the future, we plan to extend our evaluation of fairness w.r.t. more than one protected attribute, such as gender and race, and further explore the correlation between groups of fairness notions.



(a) Accuracy (b) Balanced accuracy (c) Statistical parity

(d) Equal opportunity (e) Equalized odds (f) Predictive parity

(g) Predictive equality (h) Treatment equality (i) ABROCA

Fig. 8: Accuracy and fairness interventions with varying grade threshold on Student performance dataset (Decision Tree method).

Acknowledgments

The work of the first author is supported by the Ministry of Science and Culture of Lower Saxony, Germany, within the Ph.D. program "LernMINT: Data-assisted teaching in the MINT subjects". The work of the second author is funded by the German Research Foundation (DFG Grant NI-1760/1-1), project "Managed Forgetting".

References

1. Abu Saa, A., Al-Emran, M., Shaalan, K.: Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. Technology, Knowledge and Learning **24**(4), 567–598 (2019)
2. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: ICML. pp. 60–69. PMLR (2018)
3. Alvero, A., Arthurs, N., Antonio, A.L., Domingue, B.W., Gebre-Medhin, B., Giebel, S., Stevens, M.L.: AI and holistic review: Informing human reading in college admissions. In: AIES. p. 200–206. ACM (2020). https://doi.org/10.1145/3375627.3375871

4. Amrieh, E.A., Hamtini, T., Aljarah, I.: Preprocessing and analyzing educational data set using x-api for improving student's performance. In: AEECT. pp. 1–5. IEEE (2015). https://doi.org/10.1109/AEECT.2015.7360581
5. Anders, J., Dilnot, C., Macmillan, L., Wyness, G.: Grade expectations: How well can we predict future grades based on past performance? CEPEO Working Paper No. 20-14 (2020)
6. Anderson, H., Boodhwani, A., Baker, R.S.: Assessing the fairness of graduation predictions. In: EDM (2019)
7. Berhanu, F., Abera, A.: Students' performance prediction based on their academic record. International Journal of Computer Applications **131**(5), 0975–8887 (2015)
8. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021). https://doi.org/10.1177/0049124118782533
9. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
10. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: KDD. pp. 797–806 (2017)
11. Cortez, P., Silva, A.M.G.: Using data mining to predict secondary school student performance (2008), `https://hdl.handle.net/1822/8024`
12. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: New datasets for fair machine learning. NeurIPS **34**, 6478–6490 (2021)
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: ITCS. pp. 214–226 (2012). https://doi.org/10.1145/2090236.2090255
14. Fleischman, H.L., Hopstock, P.J., Pelczar, M.P., Shelley, B.E.: Highlights from pisa 2009: Performance of us 15-year-old students in reading, mathematics, and science literacy in an international context. nces 2011-004. National Center for Education Statistics (2010)
15. Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., Lane, J.: Big data and social science: A practical guide to methods and tools. crc Press (2016)
16. Francis, B.K., Babu, S.S.: Predicting academic performance of students using a hybrid data mining approach. Journal of medical systems **43**(6), 1–15 (2019). https://doi.org/10.1007/s10916-019-1295-4
17. Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: LAK19. pp. 225–234 (2019). https://doi.org/10.1145/3303772.3303791
18. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
19. Hussain, S., Dahan, N.A., Ba-Alwib, F.M., Ribata, N.: Educational data mining and analysis of students' academic performance using weka. Indonesian Journal of Electrical Engineering and Computer Science **9**(2), 447–459 (2018)
20. Hutchinson, B., Mitchell, M.: 50 years of test (un) fairness: Lessons for machine learning. In: FAT. pp. 49–58 (2019). https://doi.org/10.1145/3287560.3287600
21. Iosifidis, V., Ntoutsi, E.: AdaFair: Cumulative fairness adaptive boosting. In: CIKM. pp. 781–790 (2019). https://doi.org/10.1145/3357384.3357974
22. Jiang, W., Pardos, Z.A.: Towards equity and algorithmic fairness in student grade prediction. In: AIES. pp. 608–617. ACM (2021). https://doi.org/10.1145/3461702.3462623
23. Jiang, W., Pardos, Z.A.: Towards equity and algorithmic fairness in student grade prediction. In: AIES. pp. 608–617 (2021). https://doi.org/10.1145/3461702.3462623

24. Khan, A., Ghosh, S.K.: Student performance analysis and prediction in classroom learning: A review of educational data mining studies. Education and information technologies **26**(1), 205–240 (2021). https://doi.org/10.1007/s10639-020-10230-3
25. Khan, N.A.U., Khan, I.U., Alamri, L.H., Almuslim, R.S.: An improved early student's academic performance prediction using deep learning. International Journal of Emerging Technologies in Learning (iJET) (2021)
26. Kuzilek, J., Hlosta, M., Zdrahal, Z.: Open university learning analytics dataset. Scientific data **4**(1), 1–8 (2017). https://doi.org/10.1038/sdata.2017.171
27. Le Quy, T., Roy, A., Vasileios, I., Wenbin, Z., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. WIREs Data Mining and Knowledge Discovery **12**(3) (2022). https://doi.org/10.1002/widm.1452
28. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021). https://doi.org/10.1145/3457607
29. Meyer, K.: Education, justice and the human good: Fairness and equality in the education system. Routledge (2014)
30. Mihaescu, M.C., Popescu, P.S.: Review on publicly available datasets for educational data mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **11**(3), e1403 (2021). https://doi.org/10.1002/widm.1403
31. Namoun, A., Alshanqiti, A.: Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences **11**(1), 237 (2020). https://doi.org/10.3390/app11010237
32. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al.: Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(3), e1356 (2020). https://doi.org/10.1002/widm.1356
33. Saleem, F., Ullah, Z., Fakieh, B., Kateb, F.: Intelligent decision support system for predicting student's e-learning performance using ensemble machine learning. Mathematics **9**(17), 2078 (2021). https://doi.org/10.3390/math9172078
34. Shahiri, A.M., Husain, W., et al.: A review on predicting student's performance using data mining techniques. Procedia Computer Science **72**, 414–422 (2015)
35. Simoiu, C., Corbett-Davies, S., Goel, S.: The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics **11**(3), 1193–1216 (2017). https://doi.org/10.1214/17-AOAS1058
36. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7 (2018). https://doi.org/10.23919/FAIRWARE.2018.8452913
37. Wightman, L.F.: LSAC national longitudinal bar passage study. LSAC research report series. (1998)
38. Xiao, W., Ji, P., Hu, J.: A survey on educational data mining methods used for predicting students' performance. Engineering Reports **4**(5), e12482 (2022). https://doi.org/10.1002/eng2.12482
39. Yu, R., Li, Q., Fischer, C., Doroudi, S., Xu, D.: Towards accurate and fair prediction of college success: Evaluating different sources of student data. In: EDM (2020)
40. Žliobaitė, I.: On the relation between accuracy and fairness in binary classification. In: FAT/ML 2015 workshop at ICML. vol. 15 (2015)
41. Zohair, A., Mahmoud, L.: Prediction of student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education **16**(1), 1–18 (2019). https://doi.org/10.1186/s41239-019-0160-3