Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments

Shirin Riazy, Katharina Simbeck and Vanessa Schreck
Hochschule für Technik und Wirtschaft, Berlin, Germany

Keywords: Learning Analytics, At-risk Prediction, Moocs, Fairness.

Abstract: While the current literature on algorithmic fairness has rapidly expanded over the past years, it has yet to fully arrive in educational contexts, namely, learning analytics. In the present paper, we examine possible forms of discrimination, as well as ways to measure and establish fairness in virtual learning environments. The prediction of students' course outcome is conducted on a VLE dataset and analyzed with respect to fairness. Two measures are recommended for the prior investigation of learning data, to ensure their balance and fitness for further data analysis.

1 INTRODUCTION

Following the outcry after the release of an article criticizing racial bias in the prediction of recidivism of criminal offenders (Larson et al., 2016), the debate on algorithmic fairness has increased steadily over the last years. However, even though a large number of fairness measures has been developed (Verma and Rubin, 2018), these measures often only provide idealized notions of fairness (Dwork et al., 2012; Hardt et al., 2016) instead of tailored instructions. Furthermore, they have mostly been tested and developed on the same datasets[1], leading to very limited coverage of the different areas of machine learning applications.

The field of learning analytics may be defined as the "measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and environments in which it occurs." (Siemens, 2013). According to (Ebner and Ebner, 2018) and (Grandl et al., 2017), five main goals of learning analytics should be mentioned:

- **Predictions and Interventions:** predictions of student performance are made to provide them with adequate interventions.
- **Recommendations:** on the basis of analytics, recommendations for interventions are formulated.
- **Personalization and Adaptation:** based on the learning activity, the learners are able to adapt learning environments.
- **Reflexion and Iteration:** the learners should be able to reflect on their learning process.
- **Benchmarking:** student performance is analyzed in order to evaluate methods or learning content.

In order to make predictions, recommendations or adaptations to the learning environments, it is common for either data mining or machine learning tools to be used on student data (Ochoa and Merceron, 2018). However, many machine learning algorithms are used as black-boxes and have yet to be thoroughly examined in terms of fairness. Furthermore, the fairness of an algorithm depends on its context, as the decision to be made determines what possible disadvantages might arise (Chouldechova et al., 2018; O'Neill, 2016).

A large percentage of recent contributions to major learning analytics conferences (Learning Analytics and Knowledge, 2019; Ochoa and Merceron, 2018) deals with the prediction of student performance, specifically whether a student will pass a course. The students predicted to fail, called *at-risk* students, can subsequently be provided with interventions, recommendations and possibilities of adapting their learning environment.

The fairness and validity of these algorithms is essential because predictions of student performance have an impact on student success (Rosenthal and Rubin, 1978) and may thus perpetuate or even increase biases in the data (O'Neill, 2016). Furthermore, a

[1] The most popular examples are the ProPublica (COMPAS) data, the Ricci dataset, the Adult Income dataset and the German Credit dataset (Friedler et al., 2019).

Riazy, S., Simbeck, K. and Schreck, V.
Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments.
DOI: 10.5220/0009324100150025
In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU 2020) - Volume 1*, pages 15-25
ISBN: 978-989-758-417-6

student being predicted a low performance and being transferred to a low-level learning environment might be stripped of their chances to perform well. Similarly, a student who is challenged by the learning environment might not be able to thrive without adaptations.

In the following, we are going to review measures and tactics of fairness regarding their suitability for learning analytics (LA) contexts. To do so, we aim to answer (at least in part) the following leading questions:

1. What types of discrimination may be distinguished and measured with respect to LA?
2. How can fairness be measured in LA?
3. How can fairness be established in LA?

First, we want to review the existing fairness literature on the basis of predictions of the course outcome of students in virtual learning environments. Using this knowledge, we will replicate different models for student at-risk prediction on an openly available dataset and evaluate our research questions in light of the experiment.

2 BACKGROUND

In order to derive answers to our leading questions, we will review the current literature on ways of detecting fairness as well as approaches to establishing algorithmic fairness.

2.1 Types of Discrimination

In many cases, European and US-American law forbids discriminatory action (Ellis and Watson, 2012). The US Civil Rights Act of 1964 (Brown, 2014) prohibits discrimination in areas such as voting, public accomodations, facilities and education. Specifically, the Equal Employment Opportunity Commission (Davila and Bohara, 1994) has outlined guidelines for the detection of discrimination. The *80% rule* represents the permittable differences in rates of hiring, promotion or other employment decision, between members of certain races, sexes or ethnic groups. This is called an *adverse* or *disparate impact* and has been picked up in algorithmic fairness literature (Zafar et al., 2015; Calmon et al., 2017). A common definition of disparate impact (Friedler et al., 2019) for binary classification problems given two groups is

$$\mathbf{DI} = \frac{\mathbb{P}(\text{ Positive classification} \mid \text{Group 1})}{\mathbb{P}(\text{ Positive classification} \mid \text{Group 2})}, \qquad (1)$$

$$\mathbf{DI}, \mathbf{DI}^{-1} \overset{!}{\geq} 0.8. \qquad (2)$$

It is to be noted that even though the jurisdiction penalizes discrimination, the attempts of defining or detecting implicit forms of discrimination have yet to be refined.

In current research, three types of discrimination are often distinguished (Barocas and Selbst, 2016; Kamishima et al., 2012):

1. **Direct Discrimination/Disparate Treatment:** intentionally discriminating actions on the basis of protected group membership. In many cases illegal in European countries and the US (Feldman et al., 2015).
2. **Disparate Impact/Indirect Discrimination:** intentionally or unintentionally discriminating actions without knowledge of group membership. Not illegal in most cases (Barocas and Selbst, 2016). A rigorous mathematical definition of disparate impact as well as different examples of it (redlining, negative legacy, self-fulfilling prophecy) are listed in (Feldman et al., 2015).
3. **Underestimation:** Kamishima and colleagues describe a third source of discrimination, which specifically occurs in algorithmic settings. Here, the sample size of a learning algorithm is too small (or unequally distributed), leading to discriminatory suggestions, even if the setup of the algorithm is otherwise fair (Kamishima et al., 2012).

All three types of discrimination could possibly occur in a learning context. In learning analytics, direct discrimination is very unlikely to appear, as this would mean for makers of algorithms or interventions to include rule-based decisions that disadvantage certain groups. Also, this type of discrimination would be easily traceable and forbidden in many countries (Ellis and Watson, 2012).

The second and third type of discrimination are much more subtle. Machine learning algorithms have been shown to produce outcomes with questionable fairness (Adler et al., 2018; O'Neill, 2016). In learning analytics, these very algorithms are used for predictions, often of student performance. A simple form of performance prediction is the binary classification of a student course outcome (Strecht et al., 2015; Yadav et al., 2012).

2.2 Measurements of Fairness

The debate on algorithmic fairness has proposed a wide range of fairness measures, applicable to different scenarios and settings (Verma and Rubin, 2018). In order to find out which ways of measuring fairness are applicable to learning analytics, we will review the current literature while splitting measures of fairness into three categories:

1. **Measurement of Balance in the Training Sets:** Calmon and colleagues (Calmon et al., 2017) provide a probabilistic framework for the detection (and removal) of bias in data. Moreover, information-theoretic measurements have been introduced in order to measure the connection between group-affiliation and target labels (Kamishima et al., 2012; Williamson and Menon, 2019).
2. **Measurement of Fairness as an Algorithmic Trait:** In order to understand and explain the traits of algorithms – specifically deep learning methods – explanations have focused on the *processing* and *representation* of the data (Gilpin et al., 2018).
3. **Outcome-based Fairness Measurement:** Several authors propose measures to certify fairness after classification (Kusner et al., 2017). These measures usually rely on a confusion matrix (see Section 2.2.3) and are especially in line with the Equalized Odds definition of fairness (Hardt et al., 2016).

Since, in learning analytics, any of these settings might be relevant, we have picked out measures of fairness from each of these categories in order to compare them. In order to set the stage for these measures of fairness, we will introduce our basic definitions.

Given a dataset $(x_i, y_i) \in X \times Y$ with $i = 1, \ldots, D$, where the $x_i \in X$ represent features of instances (such as student data in a virtual learning environment) and $y_i \in Y$ target labels (such a final score in a course), predictions are made to infer labels of previously unseen instances. Often we may assume $X = \mathbb{R}^n$ and $Y \subseteq \mathbb{R}$, where Y is either continuous (regression problems) or discrete (classification problems). Predictions may be written as parametrized maps $p_\Theta : X \to Y$, which may be qualitatively investigated using loss functions $l(\Theta) = l\left(\Theta, [x_i, y_i]_{i=1,\ldots,D}\right)$, where $l(\Theta)$ measures how well p_Θ fits the data. Usually, one tries to find a prediction function that minimizes the loss of a prediction.

Furthermore, there have recently been suggestions to add measures of fairness which additionally quantify the fairness of a prediction with respect to a (protected) group G. Most of the published fairness measures focused on binary classification (see Section 2.2.3). In order to extend these measures, we will specifically include fairness measures that do not focus on discrete classifications but instead allow continuous labels, such as a prediction of student scores.

2.2.1 Normalized Mutual Information

The normalized mutual information as a criterion for fairness was introduced by (Kamishima et al., 2012). Similar measures were also introduced by (Calmon et al., 2017; Williamson and Menon, 2019). This measure may be used to detect whether the training set is balanced between groups and is targeted to identify indirect discrimination. Since the definition of mutual information is very general, it is versatile (may be used for discrete as well as continuous data features and/or labels) and applicable in different variants (may be used during, before or after the application of an algorithm).

In the discrete case, the normalized mutual information consists of the mutual information

$$\mathrm{MI}(Y;G) = \sum_{(y,g)\in D} \mathbb{P}_{Y,G}(y,g) \log \frac{\mathbb{P}_{Y,G}(y,g)}{\mathbb{P}_Y(y)\mathbb{P}_G(g)} \qquad (3)$$

and a normalization term, leading to

$$\mathrm{NMI}(Y;G) = \mathrm{MI}(Y;G)/\sqrt{H(Y)H(G)} \qquad (4)$$

as the normalized mutual information, where $H(\cdot)$ is an entropy function.

In practice, the probabilities $\mathbb{P}$ in (3) are replaced by estimates $\hat{\mathbb{P}}$, e.g., the corresponding relative frequencies in the training data or the outcomes $p_\Theta(Y)$ in the test data.

Note that the mutual information $\mathrm{MI}(U;V)$ between two random variables U and V measures the dependency of these variables. Following the Group Fairness definition (Dwork et al., 2012), the label Y is hoped to be independent of the protected group G. Furthermore, the denominator in (4) serves as a normalization, which ensures $\mathrm{NMI} \in [0,1]$, where 0 corresponds to independent variables, while 1 indicates highest possible dependence. Mutual information may be replaced by f-divergence (Komiyama and Shimao, 2017) or the Hilbert-Schmidt criterion (Pérez-Suay et al., 2017) as different measures of distance in (4).

2.2.2 Underestimation Index

The underestimation index calculated the Hellinger distance between two probability distributions, namely the distributions of the outcomes of different groups (Kamishima et al., 2012). This measure is targeted at the third type of discrimination mentioned in Section 2.1, underestimation, and detects it as a property of the algorithm by measuring its uncertainty.

Let $\tilde{\mathbb{P}}_{Y,G}$ be an estimator for $\mathbb{P}_{Y,G}$, given by an algorithm, and let $\hat{\mathbb{P}}_{Y,G}$ be the distribution of the label and the protected group in the training set. Then the underestimation index is given by

$$\mathrm{UEI}(Y;G) = \sqrt{1 - \sum_{(y,g)\in D} \sqrt{\hat{\mathbb{P}}_{Y,G}(y,g)\tilde{\mathbb{P}}_{Y,G}(y,g)}}, \tag{5}$$

which is a reformulation of the Hellinger distance between $\hat{\mathbb{P}}_{Y,G}(y,g)$ and $\tilde{\mathbb{P}}_{Y,G}(y,g)$.

2.2.3 Outcome-based Fairness

Many measures of fairness, mostly those built on the Equalized Odds definition of fairness (Hardt et al., 2016), work with elements of confusion matrices to calculate differences between groups. These measures fall in the third group of measures of fairness, the outcome-based measures, and aim to detect indirect discrimination. Since Equalized Odds requires the group membership *given the outcome* to add no information to the prediction, various measures of fairness for binary classification were constructed, using true/false positives/negatives (Verma and Rubin, 2018). In the following, we will introduce Slicing Analysis, which generalizes previous methods and has been used in an educational context (Gardner et al., 2019).

Assume that we have a prediction p for binary classification and that the final step of the algorithm consists of thresholding. Let $X_0(p)$ be the random variable of the p-score of a negative instance (which may be below or above the threshold) and $X_1(p)$ denote the random variable of the p-score of a positive instances, then

$$\mathrm{AUC}(p;G) = \mathbb{P}(X_1(p) > X_0(p)|G) \tag{6}$$

$$= \int_0^1 \mathrm{TPR}_{(p|G)}\left(\mathrm{FPR}^{-1}_{(p|G)}(x)\right)\,\mathrm{d}x \tag{7}$$

denotes the *area under curve*, which may be used to investigate algorithms, where

$$\mathrm{TPR}_{(p|G)} : \mathbb{R} \to [0,1],$$
$$\mathrm{FPR}_{(p|G)} : \mathbb{R} \to [0,1],$$

map classification thresholds of p given a group G to their true/false positive rates. This leads us to the fairness measure of Slicing Analysis, which is defined as

$$\mathrm{ABROCA}(p) := \mathrm{AUC}(p;g_1) - \mathrm{AUC}(p;g_2). \tag{8}$$

Since the AUC value is an accuracy value, we strive for equal performance between groups, meaning that the ABROCA(p) value should be as close to 0 as possible.

One can also define

$$\mathrm{ROC}_{(p|G)}(x) = \mathrm{TPR}_{(p|G)}\left(\mathrm{FPR}^{-1}_{(p|G)}(x)\right), \tag{9}$$

which explains the name “ABROCA” as an abbreviation for the area between ROC curves. The area between these curves should be as small as possible.

2.3 Establishment of Fairness

Several authors split the existing research on establishing algorithmic fairness into three categories (Ruggieri et al., 2010; Williamson and Menon, 2019; Calmon et al., 2017; Zafar et al., 2015):

1. **Fairness by Manipulation of Data (preprocessing).** In (Zemel et al., 2013), Zemel and colleagues propose an algorithm which creates a representation of data where group information is obfuscated. Calmon and colleagues (Calmon et al., 2017) provide a probabilistic framework for the detection and removal of bias in data. Further nameworthy contributions were made by auditing data (Adler et al., 2018) or by reweighing/resampling data (Kamiran and Calders, 2012).
2. **In-process Fairness by Optimization (constraints).** By using constraints, several researchers have implemented algorithms which optimize fairness as part of the learning algorithm. Zafar and colleagues (Zafar et al., 2015) design a method for convex margin-based classifiers (such as logistic regression and SVM). In (Calders and Verwer, 2010), a naive Bayes approach is formulated, while (Kamishima et al., 2012) defines an algorithm on the basis of logistic regression and (Komiyama et al., 2018) optimizes using a least squares regression. Recently, fairness optimizations using risk measures have also been introduced (Williamson and Menon, 2019).
3. **Post-process Fairness.** Several authors propose post-process corrections (Feldman et al., 2015; Hardt et al., 2016) after using algorithmic classification. Calders and Verwer (Calders and Verwer, 2010) have offered a Bayesian approach for the post-process correction of classified data.

In the following, we will examine how to establish fairness in learning analytics, as the third of our leading questions, and compare different algorithms, which might be used in a learning analytics context. The pre- and post-process correction methods ensure balances (with respect to different metrics) in the training and outcome data. In the case of pre-process alterations, it is expected that these balances transfer to the classified data. It seems, therefore, that these corrections are mostly independent of the algorithms. Whether fairness according to a certain measure (for example the mean difference in the case of Calders (Calders and Verwer, 2010)) implies fairness according to a different measure, however, would go beyond the scope of this paper, as we try to restrict ourselves to possible applications in the field of learning analytics.

Since the comparison of methods, not datasets, lies in the focus of this paper, we focus on two in-process fairness-optimization methods which we picked based on the number of citations and their comparability. In the following, we will briefly introduce these algorithms.

2.3.1 Kamishima's Prejudice Remover

Kamishima and colleagues (Kamishima et al., 2012) approach was to include a regularization term to directly decrease prejudice within a logistic regression. To do so, they define a regularization term

$$R_{\mathrm{PR}}(D;\Theta) = \sum_{(x_i,g_i)\in D} \sum_{y\in\{0,1\}} \mathbb{M}[y|x_i,g_i,\Theta] \log \frac{\hat{\mathbb{P}}_Y(y|g_i)}{\hat{\mathbb{P}}_Y(y)},$$

where $\mathbb{M}$ is a logistic regression prediction model. This regularizaion term is closely related to the mutual information.

2.3.2 Zafar's Margin-based Classifier

The margin-based classifier of Zafar and colleagues is described in detail in (Zafar et al., 2015). Similar to Kamishima's Prejudice Remover (Kamishima et al., 2012), they bounded their classifier using a fairness criterion, which was an extension of the DI measure introduced in Section 2.1.

2.3.3 Trade-off between Fairness and Performance

Note that many authors identify a trade-off between fairness and performance of an algorithm (Menon and Williamson, 2018; Komiyama et al., 2018; Friedler et al., 2019). We would also like to verify this by using constraint-based algorithms and establishing their accuracies.

2.4 Performance Prediction in Learning Analytics

Understanding and improving students' performance plays an important role in producing work force and innovators in the labor market (Yadav et al., 2012; Chen et al., 2019; Shahiri et al., 2015). The university of Porto, for example, has prioritized the modeling of (un-)successful students in order to devise strategies to reduce failures and understand general trends in student performance (Strecht et al., 2015).

In order to accomodate different levels of academic performances in an institution, data mining methods are being used to mitigate failures and to better manage resources (Miguéis et al., 2018). Often, predictions are made early on, in order to have more leeway for interventions (Baneres et al., 2019).

Shahiri et al. (Shahiri et al., 2015) have conducted a literature review on students' performance prediction using data mining techniques and have summarized typical features used in such analyses. They have found that

1. the cumulative grade point average (CGPA) was most frequently used as the main attribute to predict student performance. According to the authors, the reason for this might be that the CGPA has a tangible value for future educational and career mobility (Shahiri et al., 2015).
2. the second most often used attributes were demographic and external assessments, extra-curricular activities, high school background and social interactions (Shahiri et al., 2015; Oladokun et al., 2008).

Several researchers use internal factors, such as tendencies to procrastinate (Michinov et al., 2011), persistence (Morris et al., 2005), engagement (Anderson et al., 2014), or other internal factors (Angeline, 2013) as a basis for the classification. Chen et al. (Chen et al., 2019) have conducted a feature analysis of demographic, internal and external features and have found that, among other things, gender and location were strong predictors for poor performance in online learning. Furthermore, Chen et al. (Chen et al., 2019) divide the online behavior of students into four categories: operational behaviors, cognitive behaviors, collaborative behaviors and problem-solving behaviors (Peng, 2013).

2.5 Ethics in Learning Analytics

Ethical guidelines for educational research have existed for decades (Cohen et al., 2002) and include the demands for informed consent, privacy, non-maleficence and human dignity among other things (Cohen et al., 2002). In recent years, with the evolution of learning analytics as an autonomous research area, several more taylored guidelines and frameworks have been published (Yun et al., 2019; Welsh and McKinney, 2015; Slade and Prinsloo, 2013; Drachsler and Greller, 2016; Sclater and Bailey, 2015) to guide researchers to use student data ethically.

Most of these guidelines focus on research practice, on consent and transparency or data ownership (Prinsloo and Slade, 2017; Drachsler and Greller, 2016; Ferguson et al., 2016; Sclater and Bailey, 2015; Sclater, 2016). Also, privacy and legal responsibilies in experimental setups are relevant (Sclater and Bailey, 2015; Ferguson et al., 2016; Drachsler and Greller, 2016).

Recently, researchers have argued that models and algorithms should be "sound and free from bias" (Slade and Boroowa, 2014; Sclater and Bailey, 2015). However, they do not go into detail as to how data or an algorithm can be balanced or even how to check for balance.

3 DATASETS AND METHODS

In this section, we will introduce the data used as well as the methods for the prediction of a course outcome.

3.1 OULAD Dataset

The OULAD dataset[2] is an open dataset published by the Open University (Kuzilek et al., 2017). The dataset contains anonymized student data from a virtual learning environment (VLE) for seven courses in the years 2013 and 2014. Furthermore, the OULAD dataset contains data from roughly 30,000 students of different gender, age and origin.

The first group of variables entails the demographic data of the students. This data contains typical demographic information which might be saved in a virtual learning environment, such as the gender, age band and highest education of a student. Also, whether or not a student has declared a disability is recorded. In a first descriptive analysis, we see that the distribution of the genders seems to be balanced, while the students with declared disabilities vs. the ones without declared disabilities are not balanced (see Figure 1).

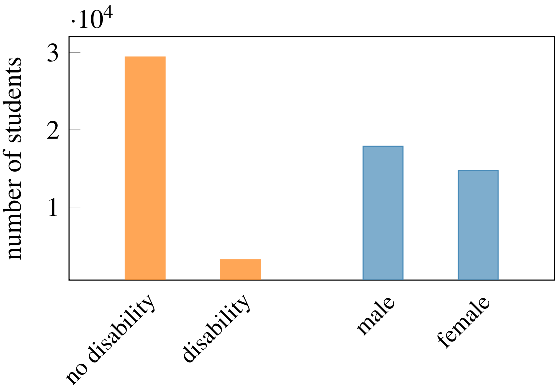

Figure 1: Histogram of the Features for Declared Disability and Gender.

Table 1: Variables and Virtual Learning Environment Data Used as Features in the Classification.

Variable	Explanation
`CMA`	count of computer-marked assignment submitted
`TMA`	count of tutor-marked assignment submitted
`login/day`	logins per day
`num_of_logins`	total number of logins
`forumng`	number of clicks for this resource page type
`glossary`	number of clicks for this resource page type
`homepage`	number of clicks for this resource page type
`resource`	number of clicks for this resource page type

The second group of features used were computed from course-specific data and are explained in detail in Table 1. First, a total of 25 features were generated from the original data. In order to select the most relevant features for the performance of the algorithm, mutual information was used, yielding the most informative variables.

3.2 Algorithms for Course-outcome Prediction

In the following, we will briefly introduce the algorithms used for the course-outcome prediction of courses in the OULAD dataset. The training and test data were split randomly for each execution of an algorithm, where the training data was chosen to have 20% of the samples and the training data to contain the remaining samples. The first group of algorithms contains three types of logistic regression:

- Kamishima's Prejudice Remover (KPR),
- Zafar's Margin-Based Classifier (ZMBC), and
- Classical Logistic Regression (LR),

where the second algorithm may be applied to any margin-based classifier. These algorithms, except for the classical logistic regression, were introduced in Section 2.3.

Further algorithms were picked by their usability for the task at hand, the prediction of the course outcome for students in a virtual learning environment:

- Naive Bayes (NB),
- Decision Tree Classifier (DT), and
- Multi-Layer Perceptron (MLP).

[2] The dataset is available under https://analyse.kmi.open.ac.uk/open_dataset.

Table 2: Different Methods of at-Risk Prediction for the OULAD Dataset and Their Fairness, When Comparing Gender/disability Groups. The Methods That Were Used without Sensitive Information Are Indicated with "ns" at the End. For Easy Comparability, the DI and NMI Values in the Training and Test Sets Were Also Added.

		Gender				Disability			
Methods	Acc	DI	NMI	UEI	ABROCA	DI	NMI	UEI	ABROCA
Training		1.05	0.0008			0.84	0.0035		
Test		1.06	0.0010			0.77	0.0077		
KPR	96.8	1.07	0.0017	0.0093	0.0011	0.87	0.0031	0.1945	0.0155
ZMBC	96.7	1.08	0.0020	0.0078	0.0173	0.80	0.0081	0.1954	0.0080
LR	96.7	1.08	0.0019	0.0081	0.0196	0.80	0.0081	0.1967	0.0108
NB	90.9	0.97	0.0002	0.0056	0.0322	0.67	0.0190	0.1840	0.0169
DT	96.2	1.07	0.0010	0.0006	0.0017	0.74	0.0089	0.1446	0.0018
MLP	88.7	1.08	0.0013	0.0101	0.0081	0.84	0.0053	0.2057	0.0150
LRns	95.9	0.97	0.0003	0.0046	0.0101	0.80	0.0074	0.1883	0.0115
NBns	94.9	0.93	0.0014	0.0010	0.0165	0.80	0.0057	0.1589	0.0153
DTns	94.3	1.07	0.0009	0.0003	0.0109	0.86	0.0020	0.1230	0.0399
MLPns	85.7	0.99	0.0000	0.0073	0.0085	0.80	0.0078	0.2028	0.0203

Each of these algorithms has been used (at times in different variations) in recent publications to predict the performance of students in learning environments (Rüdian et al., 2018; Yadav et al., 2012; Miguéis et al., 2018; Shahiri et al., 2015). Table 2 shows accuracy and fairness measures for the prediction of course success (passing) using those algorithms and comparing fairness between genders and students with or without declared disability.

In an attempt to test the best approach for the fairness of algorithms, we will execute each of the aforementioned algorithms in two variations: first while feeding them sensitive information (gender, declared disability) as features and subsequently while leaving them out. We will examine the impact this has on the accuracy as well as the fairness of these algorithms. The results in terms of accuracy and fairness of the algorithms without sensitive features are displayed in the lower part of Table 2.

4 RESULTS

In the present paper, we have developed and tested several machine learning algorithms (listed in Section 3.2) for the prediction of course outcomes in the OULAD dataset. Among the algorithms, there were also constraint-based algorithms which optimized fairness.

The fairness measures used to evaluate these algorithms were introduced in detail in Section 2.2. They include:

1. **Acc:** the accuracy of an algorithm, measured as the rate of all correct classifications
2. **DI:** disparate impact
3. **NMI:** normalized mutual information
4. **UEI:** underestimation index
5. **ABROCA:** differences between the area under curve

Overall, the accuracies of all algorithms (Table 2) were high enough to compare the student performance prediction to similar publications (Strecht et al., 2015). Table 2 compares the fairness of training and test data and of the different models with regard to gender. A DI value larger than one can be interpreted as a higher probability for male participants to pass the course in comparison to female participants. The disparate impact of 5–6 percentage points in favor of male participants from the training and test set is reproduced and slightly amplified by most algorithmic approaches.

While the gender data is rather balanced, students with declared disability are strongly underrepresented in the data. However, the bias against students with declared disability gets reproduced but not consistently amplified. Again, the models tend to lose accuracy when the sensitive attribute is not used, with the exception of the Naive Bayes classifier.

When comparing the fairness measures for different genders, most of the values of the different algorithms are quite similar, except for the Naive Bayes algorithm, which had the lowest NMI value *and* the highest ABROCA value.

Moreover, it is notable that the KPR algorithm has low NMI values, but otherwise, the constraint-based algorithms (KPR and ZMBC) compare with the other algorithms not only in accuracy but also in fairness.

It is apparent that neither the accuracy of the classification nor their fairness differed much when leaving out sensitive information.

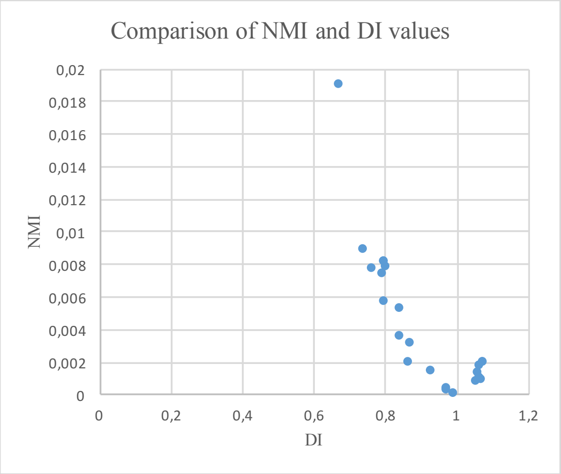

Figure 2: Comparison of NMI and DI Values for Different Methods of Course-Outcome Prediction on the OULAD Dataset.

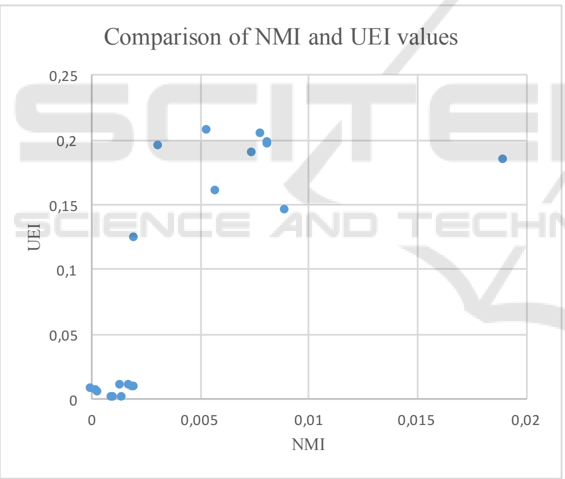

Figure 3: Comparison of NMI and UEI Values for Different Methods of Course-Outcome Prediction on the OULAD Dataset.

We can see in Figures 2 and 3 that the NMI value is negatively correlated to the DI value,

$$\text{Corr(NMI, DI)} = -0.8336, \tag{10}$$

$$p_{\text{NMI, DI}} = 7.8 \cdot 10^{-7}, \tag{11}$$

and that the NMI value is positively correlated to the UEI value,

$$\text{Corr(NMI, UEI)} = 0.7335, \tag{12}$$

$$p_{\text{DI, UEI}} = 3.5 \cdot 10^{-4}, \tag{13}$$

where $p_{\text{NMI, DI}}$ and $p_{\text{DI, UEI}}$ are the p-values to the corresponding t-tests. This supports the validity of the NMI value as an indicator for the shared information of a prediction outcome and group membership.

We can not support the feasibility of the ABROCA measure as it was found by (Gardner et al., 2019). This measure did not show significant values for imbalanced group sizes nor imbalances in the group values.

As noted in Section 2.3.3, we would have expected a trade-off between the performance (accuracy) of an algorithm and its fairness. This was not confirmed when comparing the values of KPR, ZMBC and regular LR. We can see that, while KPR has the smallest NMI values, the accuracy is the highest out of all three, though only by a thousandth.

In a natural approach to make the algorithms more independent of the sensitive attributes, gender and disability, we have also run the algorithms without feeding them these attributes as features into the algorithms. Overall, the fairness values slightly improved, though disparate impact could still be detected for students with declared disabilities (for LR and NB). Furthermore, the UEI values were similarly high for students with and without declared disabilities. Similar results have been found by (Kamishima et al., 2012; Zafar et al., 2015).

5 DISCUSSION

Though fairness in education is highly relevant (Slade and Boroowa, 2014; Sclater and Bailey, 2015), not many publications deal with fairness in learning analytics, specifically in virtual learning environments (Gardner et al., 2019; Riazy and Simbeck, 2019). In the following, we will discuss the results of the preceding section, especially in order to explore the influence of imbalances in training data to different algorithms in learning analytics. In summary, our results are as follows:

1. Depending on the balance of the training set, bias may be reproduced or amplified.
2. UEI and NMI values reliably detected imbalances.
3. The constraint-based algorithms, as well as leaving out sensitive information, slightly improved the fairness values.

5.1 Reproduction and Amplification of Bias

For sensitive groups that are represented in a balanced way in the dataset (in this case: gender), bias is reproduced by the models. For sensitive groups that are underrepresented in the dataset (in this case: disability), bias is sometimes amplified and sometimes decreased by the models. The effect of removing the sensitive datapoints is erratic as well.

5.2 Measuring Bias

As one would expect, the underestimation index (UEI) is generally rather low when comparing gender groups and rather high when comparing the groups of students who declared / did not declare a disability. Thus in our case, the UEI was correctly able to identify imbalances in the data. The NMI value correlated negatively with the DI value and positively with the UEI value. This supports its validity as a measure for informativity of group membership. The ABROCA measure, an initially promising fairness measure introduced in a learning analytics context in (Gardner et al., 2019), did not detect the disparate impact measured for students with declared disability.

5.3 Mitigating Bias

As expected, a loss in accuracy is associated with leaving out sensitive features. There was usually a loss of 1 to 3 percentage points in accuracy, except for the Naive Bayes algorithm, where the accuracy increased by 4 percentage points. The constraint-based algorithms, which optimized fairness, had slightly improved fairness values when comparing declared disability.

6 CONCLUSION AND OUTLOOK

All in all, by following our three leading questions, we have investigated possible ways of algorithmic discrimination in learning analytics, and considered ways to measure and mitigate them i.e., to establish fairness.

Since the prediction of student performance can lead to interventions, recommendations or adaptation of their learning environment, these decisions have to be tested for their validity.

In the present paper, we have examined the OULAD dataset, which contains real student data from a virtual learning environment, and found a great underrepresentation of students with declared disability. This underrepresentation – in some cases – lead to unfair classifications, meaning that students with declared disabilities were predicted to fail courses with a higher probability. This erratic behavior of the models, when a group is underrepresented, has yet to be included in guidelines and ethical codes, to lead and warn researchers when working with minorities. In order to test for imbalances in the data, we suggest to compute the UEI and NMI values. The DI measure, with the 80%-threshold, presents an easy tool for the determination of group differences. Its simplicity as a group comparison makes it valuable as a marker in order to find (un-)fair classifications. For performance prediction in learning analytics, we suggest a prior analysis of the data using at least the three values: DI, NMI and UEI.

In further research, we plan to investigate possibilities for the comparison of continuous values, which might be used in predictive tasks in learning analytics. Here, we would like to include other accuracy-based fairness measures, such as group comparisons of mean squared error values. Furthermore, an in-depth analysis is needed, in order to explain the different behavior of the algorithms, especially those of the outliers, such as the Naive Bayes algorithm.

ACKNOWLEDGEMENTS

This research has been funded by the Federal Ministry of Education and Research of Germany in the framework “Digitalisierung im Bildungsbereich” (project number 01JD1812A).

REFERENCES

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2014). Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*, pages 687–698. ACM.

Angeline, D. M. D. (2013). Association rule generation for student performance analysis using apriori algorithm. *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, 1(1):12–16.

Baneres, D., Rodriguez-Gonzalez, M. E., and Serra, M. (2019). An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies*.

Barocas, S. and Selbst, A. D. (2016). Big data’s disparate impact. *Calif. L. Rev.*, 104:671.

Brown, P. (2014). The civil rights act of 1964. *Wash. UL Rev.*, 92:527.

Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001.

Chen, Y., Zheng, Q., Ji, S., Tian, F., Zhu, H., and Liu, M. (2019). Identifying at-risk students based on the phased prediction model. *Knowledge and Information Systems*, pages 1–17.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148.

Cohen, L., Manion, L., and Morrison, K. (2002). *Research methods in education*. routledge.

Davila, A. and Bohara, A. K. (1994). Equal employment opportunity across states: The eeoc 1979-1989. *Public Choice*, 80(3/4):223–243.

Drachsler, H. and Greller, W. (2016). Privacy and analytics: it's a delicate issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 89–98. ACM.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Ebner, M. and Ebner, M. (2018). Learning analytics an schulen–hintergrund und beispiele. *Medienimpulse*, 56(1).

Ellis, E. and Watson, P. (2012). *EU anti-discrimination law*. OUP Oxford.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM.

Ferguson, R., Hoel, T., Scheffel, M., and Drachsler, H. (2016). Guest editorial: Ethics and privacy in learning analytics. SoLAR.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338. ACM.

Gardner, J., Brooks, C., and Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 225–234. ACM.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE.

Grandl, M., Taraghi, B., Ebner, M., Leitner, P., and Ebner, M. (2017). Learning analytics. *Handbuch E-Learning: Expertenwissen aus Wissenschaft und Praxis-Strategien*, pages 1–16.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer.

Komiyama, J. and Shimao, H. (2017). Two-stage algorithm for fairness-aware machine learning. *arXiv preprint arXiv:1710.04924*.

Komiyama, J., Takeda, A., Honda, J., and Shimao, H. (2018). Nonconvex optimization for regression with fairness constraints. In *International Conference on Machine Learning*, pages 2742–2751.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

Kuzilek, J., Hlosta, M., and Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific data*, 4:170171.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9.

Learning Analytics and Knowledge (2019). Proceedings of the 9th international conference on learning analytics & knowledge. New York, NY, USA. ACM.

Menon, A. K. and Williamson, R. C. (2018). The cost of fairness in binary classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118, New York, NY, USA. PMLR.

Michinov, N., Brunot, S., Le Bohec, O., Juhel, J., and Delaval, M. (2011). Procrastination, participation, and performance in online learning environments. *Computers & Education*, 56(1):243–252.

Miguéis, V. L., Freitas, A., Garcia, P. J., and Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, 115:36–51.

Morris, L. V., Finnegan, C., and Wu, S.-S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, 8(3):221–231.

Ochoa, X. and Merceron, A. (2018). Quantitative and qualitative analysis of the learning analytics and knowledge conference 2018. *Journal of Learning Analytics*, 5(3):154–166.

Oladokun, V., Adebanjo, A., and Charles-Owaba, O. (2008). Predicting students academic performance using artificial neural network: A case study of an engineering course.

O'Neill, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. *Nueva York, NY: Crown Publishing Group*.

Peng, W. (2013). Analysis and modeling of e-learning behavior.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. (2017). Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355. Springer.

Prinsloo, P. and Slade, S. (2017). Ethics and learning analytics: Charting the (un) charted. SOLAR.

Riazy, S. and Simbeck, K. (2019). Predictive algorithms in learning analytics and their fairness. In Pinkwart, N. and Konert, J., editors, *DELFI 2019*, pages 223–228, Bonn. Gesellschaft für Informatik e.V.

Rosenthal, R. and Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, 1(3):377–386.

Rüdian, S., Lui, Z., and Pinkwart, N. (2018). Comparison and prospect of two heaven approaches: Svm and ann for identifying students' learning performance. In *2018 Seventh International Conference of Educational Innovation through Technology (EITT)*, pages 156–161. IEEE.

Ruggieri, S., Pedreschi, D., and Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):9.

Sclater, N. (2016). Developing a code of practice for learning analytics. *Journal of Learning Analytics*, 3(1):16–42.

Sclater, N. and Bailey, P. (2015). Code of practice for learning analytics.

Shahiri, A. M., Husain, W., et al. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72:414–422.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400.

Slade, S. and Boroowa, A. (2014). Policy on ethical use of student data for learning analytics. *Milton Keynes: Open University UK*.

Slade, S. and Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10):1510–1529.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., and Abreu, R. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. *International Educational Data Mining Society*.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE.

Welsh, S. and McKinney, S. (2015). Clearing the fog: A learning analytics code of practice.

Williamson, R. C. and Menon, A. K. (2019). Fairness risk measures. *arXiv preprint arXiv:1901.08665*.

Yadav, S. K., Bharadwaj, B., and Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *arXiv preprint arXiv:1202.4815*.

Yun, H., Riazy, S., Fortenbacher, A., and Simbeck, K. (2019). Code of practice for sensor-based learning. In Pinkwart, N. and Konert, J., editors, *DELFI 2019*, pages 199–204, Bonn. Gesellschaft fuer Informatik e.V.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.