Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students

Qian Hu
Department of Computer Science
George Mason University
Fairfax, Virginia
qhu3@gmu.edu

Huzefa Rangwala
Department of Computer Science
George Mason University
Fairfax, Virginia
rangwala@cs.gmu.edu

ABSTRACT

Over the past decade, machine learning has become an integral part of educational technologies. With more and more applications such as students' performance prediction, course recommendation, dropout prediction and knowledge tracing relying upon machine learning models, there is increasing evidence and concerns about bias and unfairness of these models. Unfair models can lead to inequitable outcomes for some groups of students and negatively impact their learning. We show by real-world examples that educational data has embedded bias that leads to biased student modeling, which urges the development of fairness formalizations and fair algorithms for educational applications. Several formalizations of fairness have been proposed that can be classified into two types: (i) group fairness and (ii) individual fairness. Group fairness guarantees that groups are treated fairly as a whole, which might not be fair to some individuals. Thus individual fairness has been proposed to make sure fairness is achieved on individual level. In this work, we focus on developing an individually fair model for identifying students at-risk of underperforming. We propose a model which is based on the idea that the prediction for a student (identifying at-risk students) should not be influenced by his/her sensitive attributes. The proposed model is shown to effectively remove bias from these predictions and hence, making them useful in aiding all students.

Keywords

Fairness, at-risk students detection, decision making, student modeling

1. INTRODUCTION

Educational data mining (EDM) approaches seek to analyze student-related data with the objective of improving learning outcomes for students. Many machine learning methods have been proposed for student modeling and forecasting. However, in the past few years, concerns have emerged about the fairness of machine learning models. An investigation by ProPublica has found that a machine learning tool COMPAS used to predict risk of recidivism exhibits alarming bias against African-American defendants. It shows that the false positive rate of African-American defendants is twice as their white counterparts (45% vs. 23%) [1]. Buolamwini et al. [3] observed imbalanced gender and skin type distributions in facial recognition datasets. Their study shows that facial recognition algorithms are more likely to misclassify darker-skinned females with error rates up to 34.7%, while the maximum error rate for light-skinned males is 0.8%. In health care, an algorithm used to guide health decisions found that African-American patients assigned the same level of risk are sicker than white patients [24].

Qian Hu and Huzefa Rangwala "Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 431 - 437

In the domain of EDM, unfairness has also been observed. In academic performance prediction systems, social indicators have been found to predict low-performance of male students more accurately than that of female students [29]. A study by Doroudi et al. [7] showed that although personalized models were more equitable than treating all students the same, they were still not fair when relying on inaccurate models and the inequities could cascade as the amount of content increases.

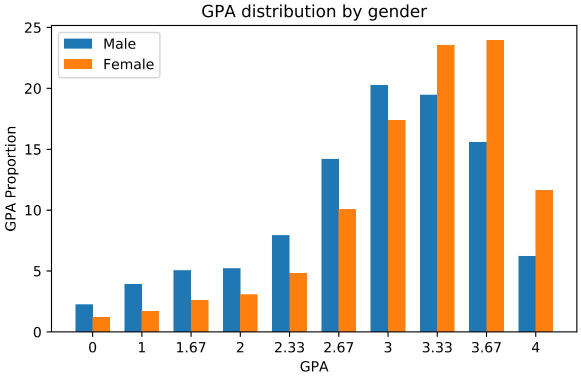


Figure 1: GPA distribution by gender.

Machine learning models learn from data. If bias is recorded in data, models trained on the biased data can also be biased [3]. Bias is also observed in educational data. Figures 1 and 2 show the average GPA of students by gender and race at George Mason University over a period of ten years. The GPA of a student is his/her accumulative GPA as of the last term before graduation. In Figure 1, average GPA of male

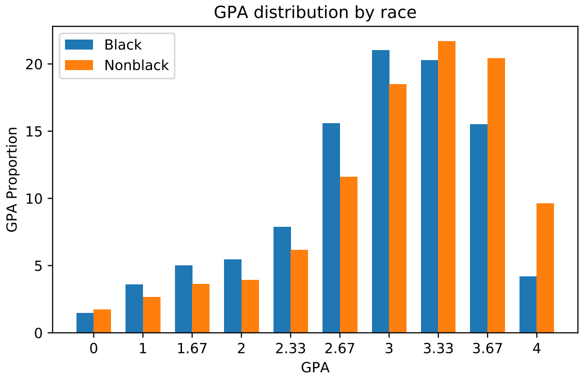


Figure 2: GPA distribution by race.

students is skewed towards lower GPAs, while average GPA of female students is skewed towards higher GPAs. The average GPA of overall female students is 3.15 which is higher than that of male students 2.86. Figure 2 shows the average GPA of African-American and non-African-American students. From the figure, we can observe that average GPA of African-American students leans towards left while that of non-African-American students leans towards right. The data shows that the average GPA of African-American students is 2.86, while it is 3.03 for non-African-American students.

Biased data can lead to biased machine learning models which can be harmful to minority groups. For example, models predicting a group of students to be at-risk or underperforming can discourage them and undermine their learning outcomes. To resolve the harmful results brought about by inequity of machine learning, there are critical needs to develop fair machine learning algorithms.

In this work, we build a fair machine learning model based on metric free individual fairness. Metric free individual fairness assumes that an individual's qualification should not be changed if his/her sensitive attribute is changed [19]. In this paper, without loss of generality we assume there are two sensitive attributes. The proposed model is composed of two classifiers. Each classifier corresponds to a sensitive group. The classifier corresponding to the individual's sensitive attribute predicts the individual's probability of being positive, while the probability of the other classifier indicates the individual's probability of being positive if his/her sensitive attribute is changed. According to the definition of metric free individual fairness, the two probability distributions should be nearly identical. The proximity of the two probability distributions is treated as fairness. The closer the two distributions, the fairer the prediction is. In addition to fairness, we also care about the accuracy of the classifier. Therefore, the overall objective we seek to optimize is the accuracy of the classifier corresponding to the individual and the proximity of the distributions of the two classifiers.

The proposed model is evaluated on datasets collected from George Mason University and the task is detecting at-risk students. The experimental results show the efficacy of the proposed model at mitigating bias. Although, the overall data shows that female and non-African-American students have higher overall performance, we observe that the bias is different for different courses. Specifically, in some courses female students belong to disadvantaged group, while in other courses male students are in disadvantaged group. This observation is useful for future work on developing fair machine learning models in educational setting.

The rest of the paper is organized as following. Section 2 discusses related work on EDM and fairness. The following section introduce preliminary on the definition of individual fairness. In Section 4, we propose our fair model for at-risk students detection. Datasets and experimental protocol is described in Section 5. Section 6 presents experimental results and analysis. The last section concludes the paper and discusses future work.

2. RELATED WORK

In this work, we focus on mitigating bias in classification tasks. We first describe related works in EDM that rely on classification. Then we describe the formalizations of fairness. Lastly, we talk about proposed methods for fair machine learning.

2.1 Classification Problems in EDM

In educational data mining, there are many tasks that can be formulated as a classification problem and several prior works have been proposed in this area such as affect detection [30], dropout prediction [4], graduation prediction [20], at-risk student detection [17, 28], knowledge tracing [31], etc.

Affect detection is the task of classifying a student's affective states such as boredom, confusion, delight, concentration and frustration by using sensor [26] and sensor-free [2] data. Vinayak et al. [15] proposed to predict student dropout using a Naive-Bayes classifier. Ojha et al. [25] proposed SVMs, Gaussian Processes and Deep Boltzmann Machines for student's graduation prediction using factors such as pre-university preparation. A set of human-interpretable features have been engineered by Polyzou et al. [28] for at-risk student detection. All these tasks can be formulated as a classification problem. However, all these works did not consider the potential bias and discrimination of the models. In this work, we try to build a general method that can be used for different kinds of tasks. To test the proposed method, we focus on the task of identifying at-risk students.

2.2 Fairness Formalizations

Over the years, different formalizations of fairness have been proposed that focus on different aspects. For example, statistical parity [11] requires that the probability of being predicted as positive across all the groups should be nearly the same. Equal odds imposes the constraint that the true positive rate should be the same for all the groups [14]. Equal opportunity requires a qualified individual should be predicted as qualified regardless of his/her sensitive attribute [14]. Another type of fairness formalization focuses more on individual level. The notion of individual fairness proposed by Cynthia et al. [8] assumes that similar individuals should be treated similarly. However, the requirement of a problem-specific similarity metric limits its adoption [5]. Hu et al. [19] proposed metric free individual fairness based on the assumption that the prediction outcome of an individual should be not be influenced by the individual's sensitive attribute. The elimination of similarity metric makes implementation of metric free individual fairness easier.

2.3 Fair Machine Learning Algorithms

Several algorithms have been proposed to achieve individual fairness. Based on John Rawls' notion of fair equality of opportunity, Joseph et al. [21] proposed an individual fairness notion that a worse individual should never be favored over a better one. The unfairness comes from the prediction's dependence on sensitive attribute. To remove the dependence, Zemel et al. [32] proposed learning a fair representation which does not contain sensitive information. The representation is a cluster of embedding vectors. Following the idea of learning fair representation, Edwards [9] proposed to remove sensitive information from the learned representation by using adversarial learning. The input feature vectors are mapped to an embedding vector by an encoder. An adversary tries to predict the sensitive attribute from the representation. The encoder and the adversary plays a minimax game to remove sensitive information. The fair representation learning algorithms achieve individual fairness by first learning a representation and then training a classifier based on the learned representation. Our proposed model directly puts fairness constraints on the predictions.

3. PRELIMINARIES

In this section, we discuss the formalization of individual fairness.

3.1 Individual Fairness

Cynthia et al. [8] introduces the concept of individual fairness, which is based on the idea that similar individuals should be treated similarly. This definition requires a similarity metric measuring the similarity between two individuals. Given two individuals x_i and x_j, a classifier H is individually fair if the difference of the predictions between the individuals are upper bounded by their dissimilarity. The definition is as following

$$D(H(x_i), H(x_j)) < d(x_i, x_j) \tag{1}$$

where D is the distance measure between the outputs of the classifier and d is the distance metric between the two individuals. The drawback of this definition is that a similarity metric is required. A similarity metric guaranteeing fairness is problem specific and requires strong assumptions, which obstructs its adoption [5].

3.2 Metric Free Individual Fairness

Hu et al. [19] proposed metric free individual fairness based on the idea that the qualification of an individual should not be influenced by his/her sensitive attribute. Thus, changing an individual's sensitive attribute should not change the prediction of a classifier. The definition of metric free individual fairness is following

$$D(P(Y|x_i, S = s_i), P(Y|x_i, S \neq s_i)) < \epsilon \tag{2}$$

where s_i is the sensitive attribute of individual i, D is the distance measure of the predictions, ϵ is an arbitrarily small positive number. This definition eliminates the requirement of a similarity measure between individuals. In this work, we develop a fair model based on this definition.

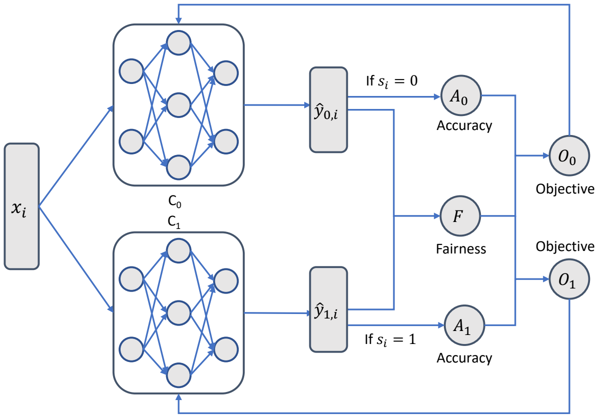


Figure 3: The architecture of the proposed model. The model consists of two classifiers C_0 and C_1 corresponding to sensitive attribute 0 and 1. An input vector x_i is fed into the two classifiers and the outputs are used to compute accuracy and fairness score. Note that if the sensitive attribute s_i is 0, accuracy A_0 and fairness F are combined to compute objective O_0 and only classifier C_0 is updated; otherwise, A_1 and fairness F are combined to form objective O_1 and classifier C_1 is updated.

4. METHODS

4.1 Problem Statement

In this work, we focus on the task of identifying at-risk students. Given a student i with $((x_i, s_i), y_i)$, $x_i \in \mathbb{R}^P$ encodes the student's grades in courses taken prior to the target course; $s_i \in \{0, 1\}$ is the student's sensitive attribute such as gender or race; $y_i \in \{0, 1\}$ is the ground truth label indicating whether a student is at-risk (1) or not (0). We focus on a binary sensitive attribute, though our method can be easily extend to scenarios where the sensitive attribute is n-ary. We want to build a classifier to predict if a student will underperform in a future target course. The classifier needs to satisfy two constraints: 1) make predictions as accurate as possible and 2) the output of the classifier is individually fair as specified by Equation 2.

The model is trained in a course-specific manner, namely, we train a model for each target course. Given a target course, we extract all the students who have taken it. The courses these students have taken prior to the target course are extracted as prior courses. The students' grades in the prior courses are extracted to form a matrix X and the students' grades in the target course are Y. Students' sensitive attributes are denoted as S. We train a course-specific model on (X, Y) to predict whether students who have not taken the target course will fail it or not. Note that sensitive attributes S are not used as features.

4.2 Proposed Algorithm

In this section, we present the proposed model, multiple cooperative classifier model (MCCM). Figure 3 shows the architecture of the proposed model. The model is composed of two classifiers, each of which corresponds to a sensitive attribute, e.g., male or female. Given an individual $((x_i, s_i), y_i)$, the feature vector x_i is fed into the two classifiers. The output of the classifier corresponding to s_i is the individual's probability of being positive, while the output of the classifier corresponding to $1 - s_i$ is the individual's probability of being positive if his/her sensitive attribute is changed. Based on the assumption of metric free individual fairness, to be fair the difference between the outputs of the two classifiers should be ignorable. In this work, the difference is the KL-divergence of the two outputs. In addition to fairness, we also care about the accuracy of the classifier. Therefore, for student i, the objective function we seek to optimize is as following

$$L_i = -y_i \log \hat{p}_{s_i,i} - (1-y_i)\log(1-\hat{p}_{s_i,i}) + \lambda \mathrm{KL}(\hat{p}_{s_i,i}, \hat{p}_{1-s_i,i}) \tag{3}$$

where λ is a hyperparameter trading off between accuracy and fairness, $\hat{p}_{s_i,i}$ is the probability of being positive predicted by classifier s_i and $\hat{p}_{1-s_i,i}$ is the probability predicted by classifier $1 - s_i$. Note that, for L_i only the classifier corresponding to s_i is updated. The classifiers are feed-forward neural networks with two hidden layers. The activation function is chosen to be ReLU [12]. Dropout [16] is used to prevent overfitting.

Algorithm 1: Multiple Cooperative Classifier Model

```
Input  : Data $D = \{((x_i, s_i), y_i)\}_{i=1}^N$, learning rate $\alpha$,
         $\lambda$, number of iterations $T$, classifier $C_0$ and
         $C_1$.
Initialize parameters $\{\theta_0^0, \theta_1^0\}$
for $t = 1, ..., T$ do
    Sample example $((x_i, s_i), y_i)$ from $D$
    Feed $x_i$ into classifier $C_{s_i}$ and $C_{1-s_i}$
    Compute the loss $L_i$ according to equation 3
    $\theta_{s_i}^{t+1} = \theta_{s_i}^t + \alpha \frac{\partial L_i}{\partial \theta_{s_i}^t}$
return $\{\theta_0^T, \theta_1^T\}$
```

5. EXPERIMENTAL PROTOCOL

5.1 Datasets

To evaluate the proposed model, we collect ten-year data at George Mason University from Fall 2009 to Fall 2019. We choose top five majors including Biology (BIOL), Civil Engineering (CEIE), Computer Science (CS), Electrical Engineering (ECE) and Psychology (PSYC). We only choose a course if there are at least 300 students who have taken it. We use a student's grade in prior courses to predict whether a student is at-risk of failing a target course. While preprocessing the data, we exclude courses that are not relevant to a major such as elective courses. Table 1 shows statistics of the data. From the table, we can see clear gender difference for different majors. Female students tend to choose Biology and Psychology majors, while male students are more prone to engineering majors such as Civil Engineering, Computer Science and Electrical Engineering. Overall, the proportion of African-American students is relatively small, especially for Civil Engineering and Computer Science.

We build course specific models, namely, for a target course we train a classifier to predict whether a student will fail that course in the future. We define as at-risk student if the student's grade is lower than 3.0. Given a target course, the data related to that course is split into 75%, 15%, 15% for training, validation and testing, respectively.

5.2 Baselines

As in this work we focus on individual fairness, we compare our proposed model with several individually fair algorithms.

*5.2.1 Logistic Regression (***LR***)*

This baseline does not have a fairness constraint. It directly predicts if a student is at-risk or not. The input is a feature vector encoding a student's grades in prior courses. The output is the student's probability of failing the target course.

*5.2.2 Rawlsian Fairness (***Rawlsian***)*

The concept of Rawlsian fairness is that a worse candidate should never be favored over a better one. Joseph et al. [21] proposed an individually fair algorithm utilizing a contextual bandits as building block to implement Rawlsian fairness.

*5.2.3 Learning Fair Representation (***LFR***)*

The unfairness of a prediction comes from the correlation of the output with the sensitive attribute. Zemel et al. [32] proposed to remove the correlation by learning an intermediate representation and train a classifier on it.

*5.2.4 Adversarial Learned Fair Representation (***ALFR***)*

Edwards et al. [9] propose to remove sensitive information from representation by adversarial learning. An encoder maps the original feature vector to a latent embedding vector, from which an adversary tries to predict the sensitive attribute. While the adversary tries to predict the sensitive attribute, the encoder seeks to generate a representation that prevent the encoder from predicting it.

5.3 Evaluation Metrics

To evaluate if the proposed algorithm satisfy the accuracy and fairness constraints, we utilize three evaluation metrics **accuracy**, **discrimination** and **consistency**.

The **accuracy** metric assesses the predictive accuracy of the model, defined as following

$$\mathrm{acc} = \frac{\sum_{i=1}^N \mathbb{1}(y_i = \hat{y}_i)}{N} \tag{4}$$

where N is the number of examples, $\hat{y}_i$ is the prediction and $\hat{y}$ is the ground truth label.

Discrimination measures the difference between the groups' rate of being predicted as positive, mathematically expressed as following

$$\mathrm{discri} = |\frac{\sum_{i=1}^N \mathbb{1}(s_i = 0) * \hat{y}_i}{\sum_{i=1}^N \mathbb{1}(s_i = 0)} - \frac{\sum_{i=1}^N \mathbb{1}(s_i = 1) * \hat{y}_i}{\sum_{i=1}^N \mathbb{1}(s_i = 1)}| \tag{5}$$

Consistency compares the predicted results of an individual with his/her k-nearest neighbors. If the predicted results

Table 1: Dataset Statistics

Major	#S	#C	#G	#M	#F	#AA	#NAA
BIOL	6,127	16	124,716	1,927(31.45%)	4,200(68.55%)	759(12.39%)	5,368(87.61%)
CEIE	450	7	23,708	338(75.11%)	112(24.89%)	27(6.00%)	423(94.00%)
CS	2,430	11	90,819	1,942(79.92%)	488(20.08%)	157(6.46%)	2,273(93.54%)
ECE	671	10	65,396	575(85.69%)	96(14.31%)	66(9.84%)	605(90.16%)
PSYC	5,110	17	84,504	1,200(23.48%)	3,910(76.52%)	694(13.58%)	4,416(86.42%)

#S total number of students, #C number of courses for prediction, #G total number of grades
#M number of male students, #F number of female students, #AA number of African-American students
#NNA number of non-African-American students.

is close to the results of the neighbors, consistency is high and the algorithm is fair. Consistency is defined as following

$$\text{consist} = 1 - \sum_{i=1}^{N} \frac{\sum_{n=1}^{K} |\hat{y}_i - \sum_{j \in \text{kNN}(x_i)} \hat{y}_j|}{K} \tag{6}$$

where $\mathbf{kNN}(x_i)$ is the k-nearest neighbors of individual i.

We use Gower similarity [13] to measure the similarity between individuals. Gower similarity is defined as

$$\text{Gower}(i, j) = \frac{\sum_{k=1}^{N} w_k S_{ijk}}{\sum_{k=1}^{N} w_k} \tag{7}$$

where N is the number of features and w_k is the weight of the k-th variable, in this paper the weights are set to one; S_{ijk} is the contribution by the k-th variable. If the k-th variable is continuous, S_{ijk} is defined as

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k} \tag{8}$$

where x_{ik} is the value of k-th feature of i and r_k is the range of values for the k-th variable. If the k-th variable is categorical, S_{ijk} is 1 if $x_{ik} = x_{jk}$ or 0, otherwise.

6. EXPERIMENTAL RESULTS

6.1 Results and Analysis

We train a classifier for each course in a major to predict if a student will fail that course. The predictions are evaluated by using accuracy, discrimination and consistency. The results are averaged across the courses in a major. Table 2 shows the experimental results with gender as sensitive attribute. From the table, we can see that the proposed model **MCCM** achieves the best performance in mitigating bias in terms of discrimination. It is able to achieve both group fairness and individual fairness, although, it is designed for achieving individual fairness. The reason is that group and individual fairness are highly correlated so that achieving one helps achieving the other.

The predictions from **LR** model is highly biased as there is no fairness constraint imposed on it, but it performs well with respect to predicting accuracy. On average, the discrimination of **LR** is 7.3%. Other methods achieve fairness at the cost of accuracy. It is interesting to see that **Rawlsian** is not able to remove bias and in some cases it leads to even more unfair predictions. **Rawlsian** is based on the idea that a worse candidate should never be favored over a better one, which is implemented by interval chaining that is a weak fairness constraint. We can also observe from the table that different majors have different level of bias, e.g., Psychology has the least bias while Computer Science has the highest bias with respect to the predictions of **LR**. The experimental results with race as sensitive attribute is shown in Table 3. The results are similar to those with gender as sensitive attribute.

6.2 Fine-grained analysis of the bias

To have a fine-grained view of the bias, we look at the data and predictions at the course level. In this section, we analyze the bias embedded in the data and predictions from **LR** and the proposed model **MCCM**. Figure 4 shows the fine-grained results with gender as sensitive attribute. For Figure 4, the data bias is that the proportion of at-risk female students subtracts the proportion of at-risk male students. Positive bias means female students are more likely to be predicted as at-risk; otherwise male students are more likely to be predicted as at-risk. For the predictions from the models, the bias is the female students' average probability of being predicted as at-risk students subtract that of male students.

First of all, as stated in Section 1, the overall data such as overall GPA by gender shows that male is minority groups. However, when looking at the course level, different courses have different minority groups. Figure 4 shows that in some courses male students are less likely to be at-risk. This insights can be used to inform future fairness work in educational data mining that a course specific model is desirable, considering that different courses have different minority groups. From the figures, we can also observe that data and machine learning models might have different bias direction. For example, in Figure 4(a), for course C0 the data bias is against male while **LR** and **MCCM** is against female. In addition, data bias does not necessarily lead to predictive bias. For example in Figure 4, all the courses show data bias. However, a no-fairness-constraint classifier, e.g., logistic regression has fair predictions in many courses.

7. CONCLUSION AND FUTURE WORK

The concerns about bias and discrimination of machine learning models are rising with the increasing of their adoption. In educational setting, we observe bias from a real-world dataset and machine learning models without fairness constraints exhibit non-ignorable biased predictions. Machine learning models are intended to aid students with their learning. However, unfair treatment of students can undermine their learning and graduation. To mitigate discrimination in educational data mining, in this paper, we proposed a fair machine learning model satisfying metric free individual

Table 2: Experimental results with gender as sensitive attribute.

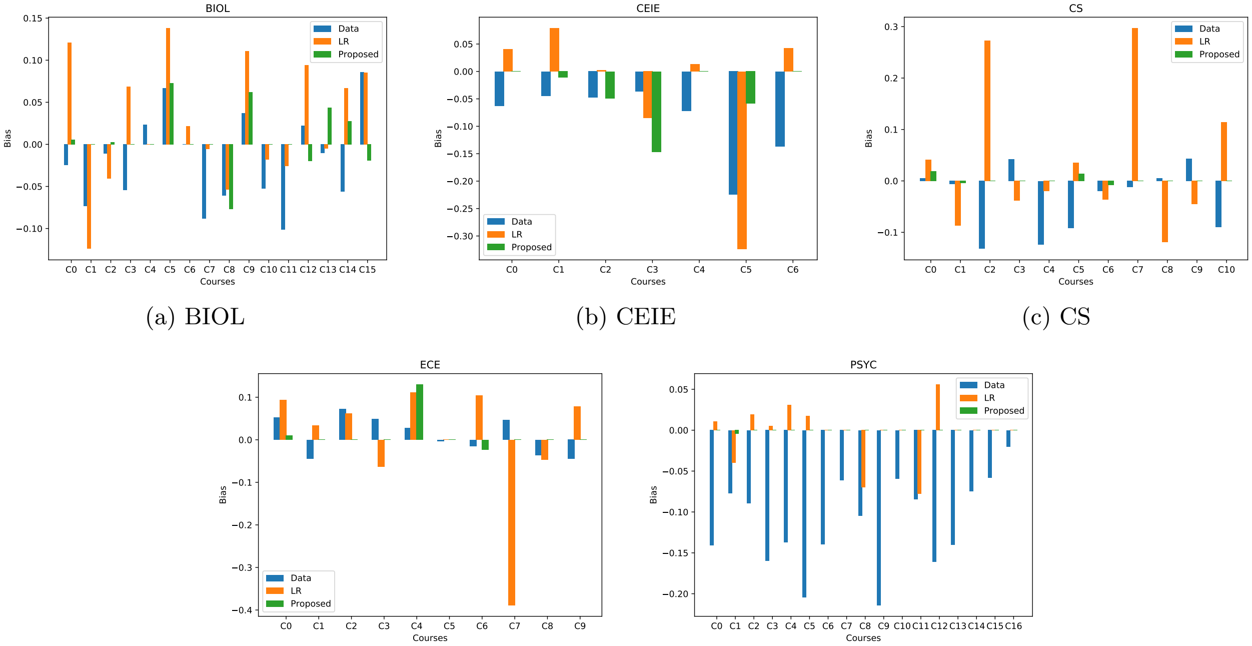
Method	BIOL acc(↑)|discri(↓)|consist(↑)	CEIE acc(↑)|discri(↓)|consist(↑)	CS acc(↑)|discri(↓)|consist(↑)	ECE acc(↑)|discri(↓)|consist(↑)	PSYC acc(↑)|discri(↓)|consist(↑)
LR	0.7662|0.0613|0.8152	0.6761|0.0837|0.7451	0.6628|0.1007|0.7569	0.7545|0.0980|0.7655	0.7769|0.0192|0.9578
Rawlsian	0.5889|0.0807|0.8120	0.6250|0.0866|0.7052	0.5582|0.0913|0.8301	0.6660|0.1498|0.7036	0.7559|0.0960|0.9396
LFR	0.6470|0.0369|0.9691	0.6983|0.0518|0.9631	0.6004|0.0228|0.9463	0.7389|0.0273|0.9912	0.7898|0.0248|0.9865
ALFR	0.6802|0.0202|0.9675	0.7062|0.0240|0.9855	0.6124|0.0134|0.9821	0.7465|0.0114|0.9783	0.7903|0.0125|0.9878
MCCM	0.6774|0.0163|0.9401	0.6415|0.0165|0.9823	0.6180|0.0038|0.9562	0.7394|0.0061|0.9717	0.7868|0.0023|0.9958

acc = accuracy, discri = discrimination, consist = consistency.
↑ means higher is better; ↓ menas lower is better.

Table 3: Experimental results with race as sensitive attribute.

Method	BIOL acc(↑)|discri(↓)|consist(↑)	CEIE acc(↑)|discri(↓)|consist(↑)	CS acc(↑)|discri(↓)|consist(↑)	ECE acc(↑)|discri(↓)|consist(↑)	PSYC acc(↑)|discri(↓)|consist(↑)
LR	0.7662|0.1004|0.8152	0.6761|0.1411|0.7451	0.6628|0.1085|0.7569	0.7545|0.1238|0.7655	0.7769|0.0276|0.9578
Rawlsian	0.5854|0.1129|0.7870	0.5849|0.3658|0.7349	0.5561|0.1857|0.8007	0.6999|0.1446|0.7416	0.7608|0.0776|0.9570
LFR	0.6202|0.0569|0.9051	0.7099|0.1722|0.9701	0.6107|0.0599|0.9897	0.7441|0.0800|0.9852	0.7874|0.0172|0.9933
ALFR	0.6850|0.0505|0.9504	0.7274|0.0862|0.9688	0.6129|0.0086|0.9715	0.7435|0.0384|0.9887	0.7898|0.0156|0.9882
MCCM	0.6563|0.0198|0.9340	0.7138|0.0114|0.9828	0.5895|0.0303|0.9968	0.7133|0.0013|0.9986	0.7857|0.0021|0.9974

acc = accuracy, discri = discrimination, consist = consistency.
↑ means higher is better; ↓ menas lower is better.



(a) BIOL

(b) CEIE

(c) CS

(d) ECE

(e) PSYC

Figure 4: Bias of different courses with gender as sensitive attribute.

fairness. We evaluate the model's performance on removing unfairness on datasets collected from an anonymous University. The results show the efficacy of the model on removing bias. Compared to other domains, educational data mining has its own characteristics. For example, in our dataset, when looking at university level, male and African-American students are biased against. However, at course level, different courses have different bias direction. This insights inform that future work on fairness in educational data mining should design course-specific models. In this work, we treat gender and race separately in terms of removing bias. In the future, we want to build models that treat gender and race as sensitive attributes simultaneously.

8. REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*.

[2] A. F. Botelho, R. S. Baker, and N. T. Heffernan. Improving sensor-free affect detection using deep learning. In *International Conference on Artificial Intelligence in Education*, pages 40–51. Springer, 2017.

[3] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[4] Y. Chen, A. Johri, and H. Rangwala. Running out of stem: a comparative study across stem majors of

college students at-risk of dropping out early. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 270–279, 2018.

[5] A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018.

[6] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[7] S. Doroudi and E. Brunskill. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 335–339, 2019.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. *CoRR*, abs/1104.3913, 2011.

[9] H. Edwards and A. Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[10] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala. Predicting student performance using personalized analytics. *Computer*, 49(4):61–69, 2016.

[11] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact, 2014.

[12] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[13] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[14] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[15] V. Hegde and P. Prageeth. Higher education student dropout prediction and analysis through educational data mining. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 694–699. IEEE, 2018.

[16] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[17] Q. Hu and H. Rangwala. Course-specific markovian models for grade prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 29–41. Springer, 2018.

[18] Q. Hu and H. Rangwala. Academic performance estimation with attention-based graph convolutional networks. *arXiv preprint arXiv:2001.00632*, 2019.

[19] Q. Hu and H. Rangwala. Cooperative contextual bandits for metric-free individual fairness. 2020.

[20] S. Hutt, M. Gardener, D. Kamentz, A. L. Duckworth, and S. K. D'Mello. Prospectively predicting 4-year college graduation from student applications. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 280–289, 2018.

[21] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 1(2), 2016.

[22] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[23] R. Morsomme and S. V. Alferez. Content-based course recommender system for liberal arts education. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, volume 748, page 753. ERIC, 2019.

[24] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[25] T. Ojha, G. L. Heileman, M. Martinez-Ramon, and A. Slim. Prediction of graduation delay based on student performance. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3454–3460. IEEE, 2017.

[26] L. Paquette, J. Rowe, R. Baker, B. Mott, J. Lester, J. DeFalco, K. Brawner, R. Sottilare, and V. Georgoulas. Sensor-free or sensor-full: A comparison of data modalities in multi-channel affect detection. *International Educational Data Mining Society*, 2016.

[27] A. Polyzou, N. Athanasios, and G. Karypis. Scholars walk: A markov chain framework for course recommendation. In *Proceedings of the 12th International Conference on Educational Data Mining*, pages 396–401, 2019.

[28] A. Polyzou and G. Karypis. Feature extraction for classifying students based on their academic performance. *International Educational Data Mining Society*, 2018.

[29] P. Sapiezynski, V. Kassarnig, and C. Wilson. Academic performance prediction in a gender-imbalanced environment. 2017.

[30] T.-Y. Yang, R. S. Baker, C. Studer, N. Heffernan, and A. S. Lan. Active learning for student affect detection. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, pages 208–217. ERIC, 2019.

[31] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*, pages 171–180. Springer, 2013.

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