FairEd: A Systematic Fairness Analysis Approach Applied in a Higher Educational Context

Jonathan Vasquez Verdugo
jvasqu6@gmu.edu
George Mason University
Fairfax City, Virginia, USA
Universidad de Valparaiso
Valparaiso, Chile

Xavier Gitiaux
xgitiaux@gmu.edu
George Mason University
Fairfax City, Virginia, USA

Cesar Ortega
cesar@fen.uchile.cl
Universidad de Chile, Facultad de Economia y Negocios
Santiago, Chile

Huzefa Rangwala
rangwala@gmu.edu
George Mason University
Fairfax City, Virginia, USA

ABSTRACT

Higher education institutions increasingly rely on machine learning models. However, a growing body of evidence shows that these algorithms may not serve underprivileged communities well and at times discriminate against them. This is all the more concerning in education as negative outcomes have long-term implications. We propose a systematic process for framing, detecting, documenting, and reporting unfairness risks. The systematic approach's outcomes are merged into a framework named FairEd, which would help decision-makers to understand unfairness risks along the environmental and analytical fairness dimension. The tool allows to decide (i) whether the dataset contains risks of unfairness; (ii) how the models could perform along many fairness dimensions; (iii) whether potentially unfair outcomes can be mitigated without degrading performance. The systematic approach is applied to a Chilean University case study, where a predicting student dropout model is aimed to build. First, we capture the nuances of the Chilean context where unfairness emerges along income lines and demographic groups. Second, we highlight the benefit of reporting unfairness risks along a diverse set of metrics to shed light on potential discrimination. Third, we find that measuring the cost of fairness is an important quantity to report on when doing the model selection.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Machine learning approaches*; • **General and reference** → **Empirical studies**; **Metrics**; **Evaluation**; • **Applied computing** → *Computer-managed instruction.*

KEYWORDS

algorithmic fairness, student dropout, educational data mining

LAK22, March 21–25, 2022, Online, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9573-1/22/03.
https://doi.org/10.1145/3506860.3506902

ACM Reference Format:
Jonathan Vasquez Verdugo, Xavier Gitiaux, Cesar Ortega, and Huzefa Rangwala. 2022. FairEd: A Systematic Fairness Analysis Approach Applied in a Higher Educational Context. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3506860.3506902

1 INTRODUCTION

Data mining offers an increasing number of tools to understand factors that contribute to academic performance (e.g. [63]), identify students in needs of additional resources, make course recommendations (e.g. [54]) and predict graduation rates (e.g. [3]). However, a growing body of empirical evidence have raised concerns about the fairness, equity and transparency of algorithmic systems (e.g. [58], [12], [57], [29]). In education data mining the problem is all the more acute as the outcomes of predictive models might have long-term implications for students ([17], [62]) and at a macroeconomic level, for the growth of human capital ([65], [67]).

An emerging subfield of machine learning has developed methods to audit data mining algorithms for potentially unfair outcomes (e.g [1], [31], [50]); pre-process data to remove correlations between features and sensitive attributes (e.g. [49], [22], [32], [14], [69]); or, mitigate some reported biases ([2], [35], [68]). Recent research in education data mining evaluates the fairness of predictive models for graduation rates ([41]), dropout models ([29], [39], [40]) or college admissions. However, to the extent of our knowledge, there is no tool to systematically explore and document whether the developed mining algorithm in an educational setting can lead to future discriminatory outcomes. We believe that fairness in education mining will progress if the release of a data is systematically accompanied with a notice that highlights potential fairness pitfalls.

In this paper, we propose a systematic approach framed as a report card system – FairEd – to document fairness pitfalls of educational data mining algorithms. The objective is to provide a framework that allows to report the context of a dataset and evaluate its fairness risk measured along multiple fairness metrics and performance-fairness trade-offs. The report card systematically (i) stress tests a dataset before its use in a data mining pipeline with diverse and simple learning algorithms and reports their performances along multiple performance and fairness metrics; and, (ii) evaluates how much performance may need to be traded-off to mitigate potential source of unfairness. The rational for our report card is to ensemble the fairness and performance of simple models into a high level and standardized summary of the fairness risk presented by a dataset when used in the context of education data mining.

We apply FairEd to a dataset of students enrolled in three bachelor's programs of a business school in Chile between 2013 and 2019. We show that our framework identifies demographic groups that would be under-served by algorithms predicting student dropout in the context of Chilean higher education: students self-identified as female and students from public high schools. By exploring different class of predictive models, FairEd also documents that fair outcomes can be obtained by pre-processing mitigation strategies, but the resulting loss in predictive power might question whether a fair version of the dataset may be a useful input to a data mining pipeline.

2 RELATED WORK

2.1 Student Dropout

In United States, more than one third of the students do not complete their bachelor's degree within four years of under-graduate studies [27]. In Chile, first year retention was on average 74% for the 2019 cohort [19]. Dropouts have long-term implications for students leading to a negative reputation (and financial losses) for higher education institution [17, 62] and at the macroeconomic level, hindering the growth of human capital ([65, 67]).

A growing literature in education data mining has offered predictive models for student dropout (e.g. [20], [29], [39], [40]) that could serve as early-warning systems and help academic counselors identify at-risk students and allocate additional resource to mentor/tutor students before they drop out. This effort in education data mining is a natural extension of a stream of research (e.g [5, 6, 60, 61, 64, 66]) that has conceptualized dropout behavior in higher education and emphasized the role of family and social contexts and academic background to explain student dropout [9, 11, 13, 18, 47, 52, 55].

2.2 Fairness in Education Data Mining

In the recent years, a sub-field of machine learning has emerged to develop metrics and algorithms that prevent predictive models to replicate or exacerbate social biases (see [16] for a review). This fair machine learning effort has ramifications within the education data mining field ([43]), all the more as recent research shows that models predicting outcomes could differ by race or gender ([29], [39]).

However, current work limits the definition of fairness to one or very few metrics; and does not offer a standardized approach for data owners to explore whether the future use of their data as input of machine learning could be problematic from a fairness perspective. [44] demonstrate that not all definitions of fairness metrics are compatible. Different notions of fairness capture different types of discrimination ([21]) and biases. [37] illustrates the growing interest in learning analytic and educational data mining tools in Latin America. Also, they support further understanding of the ethical implications of educational data mining in the specific context of higher education in Latin America. We argue for an agnostic approach that systematically reports many metrics to explore how they correlate with each other.

2.3 Risks in data-powered Algorithms

There is a growing effort in the fair machine learning literature to standardize how we communicate the fairness pitfalls that the use of data may generate. [8] and [30] define frameworks to help stakeholders provide a context around the data. On the other hand, [51] propose to accompany a model's deployment with information related to its intended use, data description, metrics, and ethical considerations through a model card comprised of key components. The proposed work complements the previous approaches regarding the tool-kits on evaluating fairness in algorithms. In particular, we supplement the effort by proposing an agnostic-systematic process for detecting, documenting, and reporting whether the use of the data could potentially lead to unfair outcomes and whether existing mitigation strategies are effective against these potential outcomes. This risk audit and mitigation strategies are framed within the context of the stakeholder's objective (of using the ML-based solutions for decision making). Further, we implement our proposal within the educational domain by studying student dropout prediction in a Latin American higher institution.

3 BUILDING A SYSTEMATIC APPROACH FOR FAIRNESS RISK ALGORITHM EVALUATION

Our systematic approach comprises five independent modules. Each one contributes to the evaluation of risks of unfair over the model construction from essential perspective. The first two elements frame the context of the model, where stakeholders must participate in defining the purpose and use. Then, the later three modules are focused on performing analysis regarding potential fairness issues as well as, evaluation of mitigation approaches to reduce the negative effects.

ENVIRONMENT

- **Objective**. A concise and clear explanation of the objective would provide boundaries in the models construction, such as where and when the outputs of the model can be used and not. This would help to channel the model construction, as well as, avoid non-desired uses that would deal into undesired discrimination. This can happen in cases when a models are used for objectives that never were intended initially. For example, consider the case of an institution that initially build a dropout predictive algorithms whose outputs are initially intended to use as inputs for intervention strategies design. If a clear objective is not declared, there are risks that the models are also used for other goals, e.g. as part of the students admission process, so the discrimination risks arise. Furthermore, a right composition of the stakeholders should be leveraged, taking account that a large variety and number of members can interrupt the continuity in the development of the model but a very homogeneous group with few members decreases the likelihood of adoption and identification of fairness concerns regarding the model.

- **Data Context**. The definition of sensitive attributes are very context-dependent [43] and may require subject-matter expertise when deciding which characteristics and their metrics should be used. We suggest that the description of any dataset should be constructed as a triplet X, Y, A, where X should be defined as a $m-$ dimensional features vector (or variables/characteristics), Y as potential the potential outcome, i.e., the information that the model provide to reach the objective, and A as the sensitive attribute(s). Most of the time, A represents the membership to the groups to be protected of discrimination. Outcomes Y are also context-dependent. They represent the variables on which a decision maker will act upon. In the context of education data mining, examples for Y might be student dropout, admission decisions and degree completion. As we are building a report data card, we may not be able to anticipate all the use of this data. Therefore, it is recommended to use domain knowledge and stakeholder engagement to choose for X, A, and Y. Recall that even the definition of the measures for X, A, and Y could lead discrimination [28], the idea of using domain knowledge and stakeholder engagement, is to focus on the evaluation of unfairness on the variables of interest.

FAIRNESS ANALYSIS

- **Data Exploration**. The data exploration should identify signs of unfair sources through simpler analysis of relationship between the sensitive attributes A and the features X as well as the outcome Y. These relationships can be performed through correlation analysis, or by implementing a features importance analysis of a model where the sensitive attributes are intended to predict. Note that the output of this exploration would shed light what kind of mitigation approaches should be appropriate to reduce the unfairness.
- **Fairness Stress Tests**. The analysis on the relationship between A and X, Y is not enough, since it is important to measure the final effects on the unfair outcomes. Therefore, it should be perform an evaluation through the definition of clear metrics. The final set of metrics is context dependent and should be in line with the description provided in Objective and Data Context. From fairness perspective, several notions of fairness are discussed in literature and how to measure it in the dataset. For example, dataset can be stressed for detecting fairness issues through the construction of simpler classifiers functions $C : X \longrightarrow Y$ that predict the outcome Y from the data features X subject to criteria: (1) the computational cost to learn C should be limited, i.e., the objective of this step is not to construct a state-of-the-art classifier, but to evaluate whether outcomes from classifiers using the data might discriminate against some demographic groups; and (2) the set of classifiers C needs to be diverse enough so that it describes different aspects of the data manifold, including linear and non-linear geometry of the class boundaries, sparsity and redundancy of the feature space. Notice that the industry and literature provide a vast set of opensource toolkit for fairness analysis, e.g., AIF360 [7], scikit-learn [56], and fair-learn [10]
- **Mitigation**. Finally, a set of mitigation approaches should be evaluated. The outputs from the previous step would help to constraint what kind of mitigation are proper to reduce the risks of discrimination found in the dataset. From fairness point of view, there are three type of mitigations: (1) pre-processing techniques, which consist of a mapping $T : X \longrightarrow Z$ of the data features X into a transformed features Z where correlations with sensitive attributes have been removed, (2) in-processing mitigation strategies, which optimize a classifier's objective function (e.g. cross entropy) under a pre-specified fairness constraint, and (3) post-processing techniques consisting on modeling threshold – limit used on scores provided by one dimensional score function of the modes – as a random variable which a distribution depending on the value of the sensitive attribute [35]; these thresholds are adjusted so that decisions based on adjusted thresholds satisfy a pre-specified fairness constraint, while keeping these decisions as similar as possible as the ones obtained without fairness constraint.

Finally, after describing each elements, all the pieces –objective, data context, exploration, stress tests and mitigation – should be put in a succinct report. The report should summaries key findings and be complemented by charts and graphics. Notice that this card report also provides a visual guidelines, since it can be used for future development of the model, as well as, guide the developer in the model construction.

4 CASE STUDY: CHILEAN BACHELORS PROGRAMS

The systematic approach is tested in a Case Study of a Chilean College that has three bachelor programs. The description of each element previously discussed is defined below[1].

4.1 Objective

The college wants to build a data-driven algorithm that help to early identify students at high-risk of dropout after first year. The outcome of the algorithm is intended to be used as input for the design interventions to prevent early leaving of students. These intervention are benefits, such as financial aids and tutorials. It is stated a strict use of the resulting model, which claims that alternative uses are not allowed unless its new use is previously reviewed.

4.2 Data Context

The dataset contains a total of 4,983 students enrolled in three bachelor's program of a business school in Chile between 2013 to 2019. The collected features include

- **Information related to the admission Process**, e.g application preferences and application scores in the National Standard University Admission Tests (NSUAT).
- **Information related to past academic performances**, e.g. high school GPA in different academic fields or type of high school (public v.s. private).

[1] In the following repository you will find the code used in the experiments: https://github.com/jovasque156/FairEd-LAK22. Dataset are not shared for confidentiality reasons.

- **Information related to socio-economic status**, e.g income decile, gender or residential information.
- **Information related to college academic performance**, e.g GPA per semester and field, number of credits passed and failed per semester, and dropout status.

We define Y as equal to one if a student leaves the program in the third or fourth semester and equal to zero otherwise. We only use information available until the end of the second semester to create features related to measure a student's academic performance in college. Finally, we split our data into train/test, with 75% of the students being used for training and the rest of them for testing.

Preliminary data processing consist of imputation, encoding, and transformation. We impute missing values using the mode of the distribution for each categorical feature. We impute numerical features as a linear function of other features. Additionally, categorical variables were encoding in one-hot vectors. Moreover, all features were demeaned and normalized by their l_2 norm, where we compute mean and norm from the training dataset. The final dataset has 114 features in X.

Most of the fair machine learning research focus on three types of sensitive attributes([4]): race or ethnicity, nationality, and gender. However, in Chile, most segregation occurs along poverty lines [46], which excludes low-income individuals from climbing the social ladder [34]. This income segregation results into an education system where students from households with better socio-economic status attend private high schools, while students from poorer household go to public high schools. This difference between private and public schools affects admission in top-ranked universities and later, access to the best employment opportunity [23, 70]. Therefore, it seems important to monitor whether classifiers would discriminate against the student with lower socio-economic status as proxied by their attendance to a public high school.

Income segregation is exacerbated by gender inequality [70]. This is consistent with literature that identifies gender as a sensitive attribute (e.g [2, 15, 33, 35, 38, 42, 43, 45, 68]), including in education data mining (e.g. [29], [38]).

Therefore, the Chilean dataset includes three sensitive attributes (see 1: (i) *school*, which means that $A = 1$ if the student attended a public school; (ii) *gender*, which means that $A = 1$ if the student is self-identified as a female; (iii) *gender-school*, which means that $A = 1$ if the student is self-identified as a female or attended a public school).

Table 1: Definition of sensitive attributes for the Chilean dataset.

Scenario	Privileged $A = 0$	Unprivileged $A = 1$	Input Dataset
Gender	Male	Female	(X, A_{gender}, Y)
School	Not Public School	Public School	(X, A_{school}, Y)
Male + Private	Male *and* Not Public School	Female *or* Public School	$(X, A_{malepriv}, Y)$

4.3 Data Exploration

Figure 1 shows which features X correlates with the sensitive attributes A. These correlations are measured as the importance of features obtained from a random forest that predicts the sensitive attribute A. Variables that leak information about high school status include high school size, family income and the average scores in NSUAT of high school. This is consistent with the observation that private schools are smaller, their students richer and better tutored for the NSUAT test. NSUAT math scores correlates with gender.

Further exploration confirms that test scores correlates with the sensitive attribute, but are very noisy signal of college grades. The math-section scores of NSUAT displayed in Figure 2 shows that the unprivileged group has a higher mean and lower standard deviation. However, this difference is not observed on the academic-performance of first and second semesters. Similar observations are reported in [46]: students from private high school tend to get higher scores on standard admission tests but their performance is not different than their peers at the university.

4.4 Fairness Stress Tests

4.4.1 Fairness Notions. Even if the root causes of unfairness were well understood, it would be still challenging to define precisely what fairness means for a classifier C. The literature has focused on two types of notions, *statistical* and *individual.* Although statistical notions of fairness provides guarantees for an average representative of each demographic group ([21]), they are more operational: they can be consistently estimated from finite samples. For our data card, we follow [43] and focus on statistical notions of fairness.

Independence [43] defines fairness as independence between classifier's outcomes $C(X)$ and sensitive attributes A. Metrics in the *independence* category vary depending on how they measure independence between outcomes and group membership.

Demographic Parity (demPI): A classifier C satisfies $\alpha-$ demographic parity *DemPI* ($\alpha > 0$) [21] if and only if

$$|Pr(C(X) = 1|A = 1) - Pr(C(X) = 1|A = 0)| \leq \alpha. \tag{1}$$

Disparate Impact (dispIR): Another measure of independence, disparate impact *DispIR*, follows the recommendations of US Equal Employment Opportunity Commission ([24]) and imposes that a classifier C satisfies:

$$\frac{Pr(C(X) = 1|A = 1)}{Pr(C(X) = 1|A = 0)} > \alpha \tag{2}$$

where $\alpha > 0$ is a threshold. In United States, this α is set to 0.8, but in this paper, we measure unfairness as deviations of the ratio $Pr(C(X) = 1|A = 1)/Pr(C(X) = 1|A = 0)$ from 1.

Demographic parity and disparate impacts are strong definitions of fairness since they require statistical independence of $C(X)$ and A, unconditional on other features or outcomes. Moreover, [2] observe that this notion does not constrain the accuracy of the classifier to be similar across demographic groups.

Separation [43] relaxes demographic parity and disparate impact by only requiring statistical independence conditional on outcomes Y, i.e measures fairness within classes separated by their outcomes Y. The literature offers two metrics based on the concept of separation: *equalized odd* (eqODD) and *equalized opportunities* (eqOPP) [2, 35].

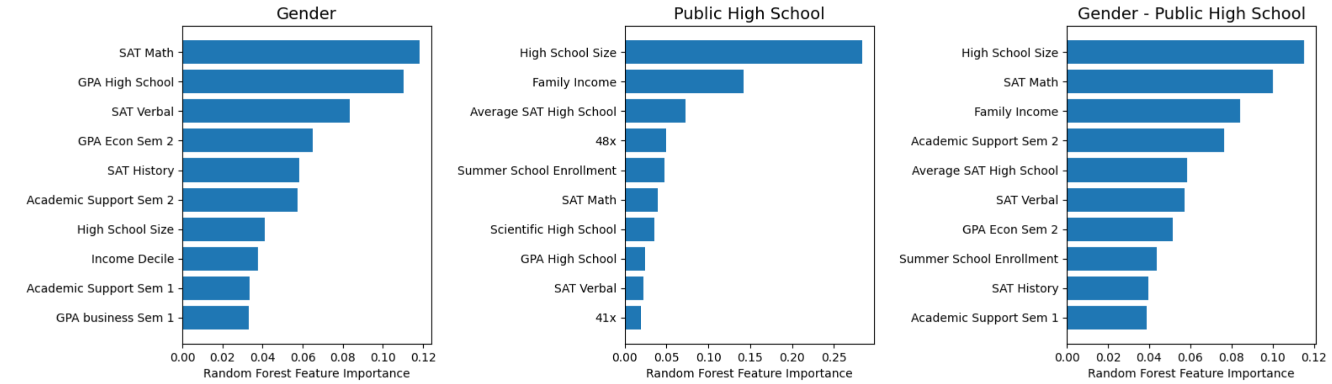
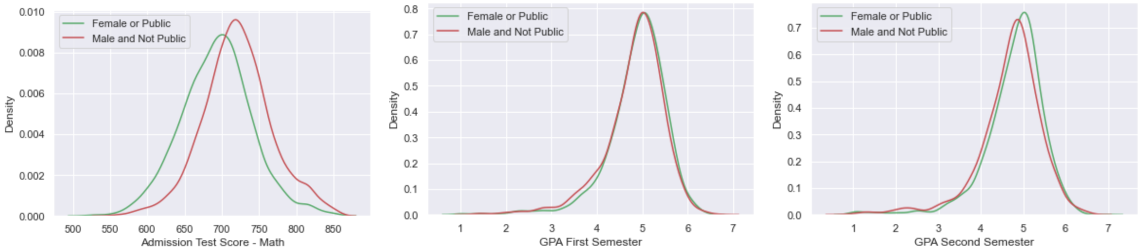


Figure 1: Feature importance obtained from a random forest classifier that predicts the sensitive attribute A from the data features X. The larger the importance of a given feature, the more impact it has on the output of the random forest classifier and the more information it leaks about the sensitive attribute A.



(a) (b) (c)

Figure 2: Score of math-section admission test (a), and GPA for first and second semester (b,c) between male from private high school and female from public high school.

Equalized Opportunity (eqOPP). A classifier C satisfies $\alpha-$ equalized opportunity ($\alpha > 0$) if and only if

$$|Pr(C(X) = 1|A = 1, Y = 1) - Pr(C(X) = 1|A = 0, Y = 1)| \leq \alpha. \quad (3)$$

Equalized Odd (eqODD) A classifier C satisfies $\alpha-$ equalized opportunity ($\alpha > 0$) if and only if for all y

$$|Pr(C(X) = 1|A = 1, Y = y) - Pr(C(X) = 1|A = 0, Y = y)| \leq \alpha. \quad (4)$$

In order to capture both cases $Y = 1$ and $Y = 0$ in a single metric, we follow [42] and measure equalized odd as

$$E_y KL(Pr(C|A = 1, Y = y)||Pr(C|A = 0, Y = y)), \quad (5)$$

where the Kullback-Leibler divergence KL measures a statistical distance between $Pr(C|A = 1, Y = y)$ and $Pr(C|A = 0, Y = y)$.

Note that equalized opportunity relaxes equalized odd since eqOPP constrains only the dependence of A and Y conditional on positive outcomes. eqOPP is applicable when $Y \in \{0, 1\}$ and the positive class $Y = 1$ is considered as a benefit [33].

Sufficiency. [43] introduces a notion of *sufficiency* (hereafter, *suffic*) that requires independence between sensitive attributes and true outcomes conditional on the model's predictions. A classifier C satisfies $\alpha-$ sufficiency ($\alpha > 0$) if and only if

$$|Pr(Y = 1|C = 1, A = 1) - Pr(Y = 1|C = 1, A = 0)| \leq \alpha. \quad (6)$$

Table 2 summarizes the fairness metrics used as we construct our FairEd data card. By systematically exploring how each stress test classifier in Table 3 performs along each of the fairness metrics in Table 2, our FairEd data card allows to evaluate whether the data at hand presents fairness risk when used as input to machine learning algorithms.

4.4.2 *Stressing Models.*

In order to evaluate whether the use of a data could lead to unfair outcomes, we use a set of classifiers or test functions $C : X \rightarrow Y$ that predict the outcome Y from the data features X. When choosing for the set of classifiers C, we consider two selection criteria. First, the computational cost to learn C should be limited: the objective of the data card is not to construct a state-of-the-art classifier, but to evaluate whether outcomes from classifiers using the data might discriminate against some demographic groups. Second, the set of classifiers C needs to be diverse enough so that it describes different aspects of the data manifold, including linear and non-linear geometry of the class boundaries, sparsity and redundancy of the feature space.

Table 2: List of Fairness Notions and Metrics.

Notion	**Metric**
Independence	demPI: $\lvert Pr(C(X) = 1 \mid A = 1) - Pr(C(X) = 1 \mid A = 0)\rvert \leq \alpha$ dispIR: $\frac{Pr(C(X)=1 \mid A=1)}{Pr(C(X)=1 \mid A=0)} > \alpha$
Separation	eqODD: $\lvert Pr(C(X) = 1 \mid A = 1, Y = y) - Pr(C(X) = 1 \mid A = 0, Y = y)\rvert$ eqOP: $\lvert Pr(C(X) = 1 \mid A = 1, Y = 1) - Pr(C(X) = 1 \mid A = 0, Y = 1)\rvert \leq \alpha.$
Sufficiency	suffic: $\lvert Pr(y = 1 \mid C(X) = 1, A = 1) - Pr(Y = 1 \mid C = 1, A = 0)\rvert \leq \alpha$

Therefore, we select five families of classifiers (see Table 3) that are commonly used in the education data mining literature [20, 26, 40, 53, 59, 67]. Hyperparameters are selected using a standard five-fold cross-validation and *F1- score* as performance metric. If the classes defined by Y are imbalanced, which is frequent in a education data mining setting, we re-weight the classes and favor *F1- score* as a performance metric over *Area Under the Receiver Operator Characteristic (AUC)* and *Accuracy.* The weights are compute as $n/(c \cdot m(y))$ where n is the number of samples, c the number of classes, and $m(y)$ the number of samples in each class, therefore, class with less samples weight more.

Table 3 enlists each model's parameters set for hyperparameter optimization procedure, where it was used a cross-validation of 5 folds and the *F1-score* as the performance to optimize. Anyway, *Area Under the Receiver Operator Characteristic-AUC* and *Accuracy* were computed for analysis purposes. Y shows imbalance, which is common on student dropout problem setting [25, 26, 40, 67]. Hence, *class weighting* approach was used for handling the imbalance.

Table 3: Classifiers and tuning parameters used to generate our stress test functions.

Model	Tuning Parameters
Support Vector Machine (svm)	$kernel$: {rbf, sigmoid} C: {0.01, 0.1, 1}
Logistic Regression (lr)	C: {0.01, 0.1, 1} $fit\_intercept$: {True, False} $solver$: {liblinear, lbfgs}
Random Forest (rf)	$n\_estimators$: {10, 50, 100} $criterion$: {gini, entropy} max_depth: {None, 2, 5, 10, 15}
Decision Tree (dt)	$splitter$: {best, random} $criterion$: {gini, entropy} max_depth: {None, 2, 5, 10, 15}
K-Nearest Neighbors (knn)	$weights$: {uniform, distance} $n\_neighbors$: {5, 10, 15, 20, 25, 30}

4.4.3 Fairness Stress Results. Table 4 shows the results of our stress test. Among all the models studied, random forest *rf* performs the best in terms of $F1$-score, while support vector machine *svm* performs the best in terms of AUC score. Note that given how imbalance the data is ($Y = 0$ for 90% of the students), accuracy is not a useful metric of performance and is not reported here.

By comparing classifier's performances across demographic groups, our systematic approach allows to capture two patterns. First, the predictive performance is generally better for students who were enrolled in a public high school: this is consistent with our observation in the data exploration that some features – test scores – are more noisy for students who were enrolled in a private high school. Moreover, predictive performance is also quite consistently better for students that self-identifies as male.

Secondly, this predictive performance gaps between demographic groups vary significantly across classifiers. For decision trees and nearest neighbor methods, the difference in F1 score between male and female students is 0.13 and 0.2, respectively. The logistic regression *lr* has a similar $F1$-score for male and female students. These results highlight the benefit of our systematic approach when evaluating differences between groups: models belonging to different class may use differently sensitive attributes A and their correlation with features X.

Figure 3 explores further how each classifier performs across the fairness metrics defined in Table 2. There is strong evidence that without fairness mitigation, the use of the Chilean dataset by common machine learning techniques would raise fairness concerns. For example, when sensitive attributes are defined by gender, our measure for disparate impact show that all models disproportionately estimate that student self-identified as female are $25 - 50\%$ less likely to dropout than students self-identified as male. When sensitive attributes are defined by high school status, students from a private high school are considered $20 - 25\%$ more likely to dropout across all models except nearest neighbors. Equalized opportunity shows that all models underestimate by 0.1 to 0.3 percentage point the ground truth dropout risk of female students relative to male students; and overestimate by 0.1 to 0.2 the ground truth dropout risk of students previously enrolled in a public high school.

Figure 3 supports further the use of many classifiers to understand the fairness risk associated with a data. Although the signs of the fairness metrics is quite consistent across models, there is significant variations in the magnitude of measurements.

4.5 Mitigation

In this case study we explore a wide range of mitigation strategies, including pre-processing, in-processing and post-processing techniques. The objective is not to design a new mitigation strategy,

Table 4: Performances of baseline classifiers (see Table 3) across demographic groups. Bold highlights represent for which group a classifier performs better, where groups are defined by gender, public high school or intersection of gender and public high school.

Classifiers	F1 score						
	All	**Gender**		**School**		**Gender + School**	
		Male	*Female*	*Private*	*Public*	*Female or Public*	*Male and Private*
rf	**0.54**	**0.56**	0.51	0.53	**0.587**	0.53	**0.56**
lr	0.44	0.44	0.44	0.43	**0.47**	0.43	**0.45**
svm	0.46	0.45	**0.47**	0.45	**0.48**	0.44	**0.47**
dt	0.46	**0.51**	0.38	**0.48**	0.40	**0.50**	0.41
knn	0.43	**0.53**	0.23	0.43	**0.44**	**0.51**	0.34
	AUC score						
	All	**Gender**		**School**		**Gender + School**	
		Male	*Female*	*Private*	*Public*	*Female or Public*	*Male-Private*
rf	0.72	**0.73**	0.69	0.71	**0.75**	0.71	**0.72**
lr	0.73	**0.74**	0.72	0.72	**0.78**	0.72	**0.75**
svm	**0.74**	0.74	0.74	0.72	**0.79**	0.72	**0.76**
dt	0.71	**0.74**	0.67	**0.72**	0.70	**0.73**	0.70
knn	0.65	**0.69**	0.56	0.65	0.65	**0.68**	0.61

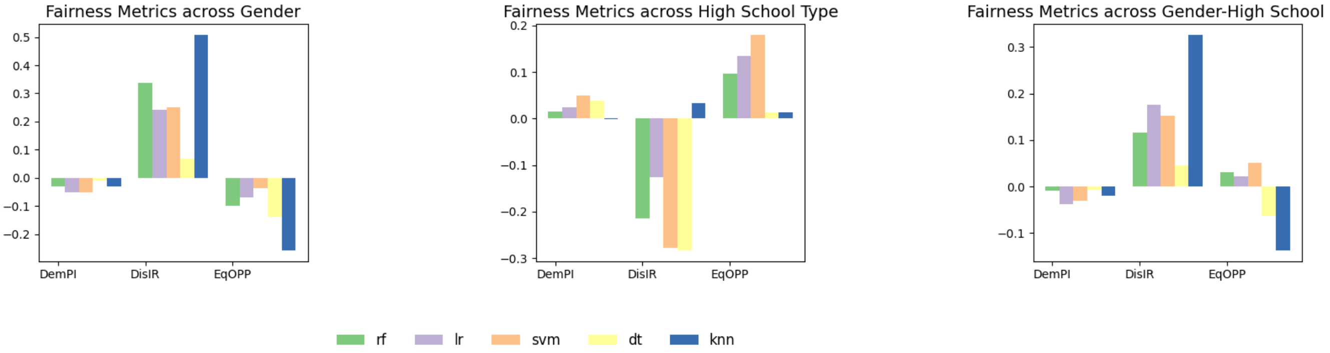


Figure 3: Stress tests for three fairness metrics across demographic groups. Negative values of $DemPY$ imply that a model predicts a lower risk of dropout for $A = 1$ (female, student from public high school); which can also be measured as a postive DispIP; Negative values of $EqOPP$ imply that a model underestimate the true dropout risk of students in the group $A = 1$.

but to investigate whether the current state-of-the-art approaches can alleviate the fairness pitfalls identified by our stress-tests. A negative answer should foster stakeholder engagement and serve as a warning against the use of the data as input to decision-making systems.

Pre-processing techniques consist of a mapping $T : X \longrightarrow Z$ of the data features X into a transformed features Z where correlations with sensitive attributes have been removed. Pre-processing approaches do not rely on any outcome of interest and thus, are particularly well suited to offer protection against *independence*-based definitions of fairness. We consider three types of pre-processing approaches: (i) *Unaware*; (ii) *Feature Selection*; (iii) *Fair representation learning*. The unaware approach involves training the classifiers C without using the sensitive attribute A. This technique forms a benchmark to evaluate whether removing the sensitive attribute from the data could effectively reduce the likelihood of unfair outcomes. However, in practice, because of complex social and historical biases, it is likely that the feature set (or a subset of it) X would encode the sensitive attribute A and thus, *Unaware* would not be an effective mitigation strategy. In Feature Selection, the goal is to identify $X' \subset X$ such that using X' is orthogonal to A. The FS procedure uses a backward feature selection strategy, where a regression model based on Ordinary Least Squares (OLS) is used to predict A and variables with p-values lower than 0.05 are dropped. The procedure is iteratively repeated until no variable have a p-value equal o lower than 0.05 when predicting A. Finally, Fair representation learning [22, 49] involves mapping the input features X into a representation Z with the condition that the learned representation

Table 5: Overall and fairness performances of best models and mitigation approaches per scenario. The bold font highlights the mitigation approach with the closest fairness metric to 0 within a scenario.

Scenario	Model	Mitigation	F1-score	AUC	demPI	dispIR	eqOPP	eqODD
Gender	*rf*	Unaware	0.525	**0.869**	-0.004	**-0.039**	0.093	0.068
		FS	0.508	0.786	-0.015	-0.148	-0.045	0.061
		adv	0.182	0.538	0.046	0.133	0.033	0.006
		β-VAE	0.161	0.545	-0.037	-0.182	**-0.023**	0.010
		Red-DemPI	0.540	0.711	-0.0240	0.200	-0.087	0.003
		Red-EqODD	**0.530**	0.701	**-0.021**	0.280	-0.030	**0.001**
		Post-DemPI	0.487	0.712	0.021	-0.260	-0.064	0.012
		Post-EqODD	0.510	0.712	-0.009	0.110	-0.031	**0.001**
School	*svm*	Unaware	**0.52**	**0.846**	-0.002	-0.014	0.067	0.086
		FS	0.511	0.834	0.045	0.327	0.154	0.095
		Adv	0.180	0.508	0.032	0.051	0.105	0.006
		β-VAE	0.200	0.514	**0.000**	**0.000**	**0.000**	**0.000**
		Red-DemPI	0.465	0.744	0.021	-0.151	0.206	0.020
		Red-EqODD	0.465	0.745	0.060	-0.351	0.104	0.002
		Post-DemPI	0.448	0.745	-0.002	0.010	-0.043	0.001
		Post-EqODD	0.412	0.745	0.002	-0.008	0.126	0.004
Male+Private	*rf*	Unaware	0.525	**0.869**	-0.004	-0.039	0.093	0.068
		FS	0.439	0.795	0.001	0.018	0.032	0.044
		Adv	0.186	0.534	0.003	0.011	-0.099	0.003
		β-VAE	0.140	0.473	-0.015	-0.058	0.017	**0.000**
		Red-DemPI	**0.526**	0.707	**0.000**	-0.007	0.056	0.0009
		Red-EqODD	**0.526**	0.710	0.010	-0.050	0.005	**0.000**
		Post-DemPI	0.506	0.710	**0.000**	**0.000**	0.070	0.001
		Post-eqODD	0.510	0.707	-0.004	0.005	**-0.002**	**0.000**

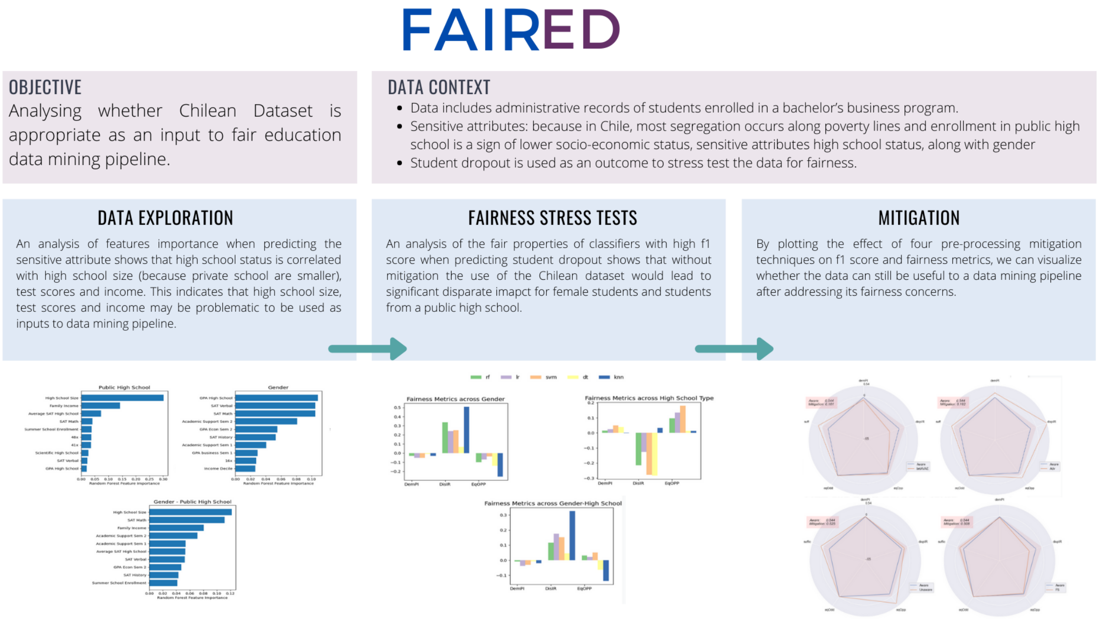
Z is independent of A. In this study, we obtain independence of Z and A with two methods: *Adversarial* or $beta-VAE$. *Adversarial* (Adv) ([49], [22]) minimizes the cross-entropy of an adversarial classifier that predicts A from Z. $beta-VAE$ ([36], [48]) controls the bit rate of the representation via a variational auto-encoder [36]).

In-Processing mitigation strategies optimize a classifier's objective function (e.g. cross entropy) under a pre-specified fairness constraint. [2] show how to reduce such constrained optimization problem to a cost-sensitive classification that can be solved by a two-player minmax game. We apply [2]'s reduction approach with either a constraint on demographic parity – *Red-DemPI* – or on equalized odds –*Red-EqODD*.

Post-Processing. Most classifiers we use to stress test the data generate a one-dimensional score function and then, decisions are made by thresholding the scores. Post-processing techniques consist of modeling the threshold as a random variable which a distribution depending on the value of the sensitive attribute [35]. These thresholds are adjusted so that decisions based on adjusted thresholds satisfy a pre-specified fairness constraint, while keeping these decisions as similar as possible as the ones obtained without fairness constraint. We apply [35]'s post-processing approach with either a constraint on demographic parity – *Post-DemPI* – or on equalized odds –*Post-EqODD*.

In Table 5, we look at the impact of mitigation approaches that attempts to offer fairness guarantees for demographic groups identified by gender (top panel); by enrollment in high school (middle panel); and by a combination of both (bottom panel). Table 5 shows how state-of-the-art mitigation strategies affect both the predictive power of our stress test classifiers and the performance gaps between demographic groups that we observe in Table 4 and Figure 3.

The table allows to answer three questions: (i) can the fairness issues raised in the previous sections be alleviated; (ii) which party along the data pipeline best positioned to mitigate fairness issues; and, (iii) what is the cost of fairness mitigation? First, in the context of the Chilean dataset, our analysis shows that the *unaware* strategy, albeit simple, is quite effective at alleviating some fairness concerns. It implies that from a data owner perspective, hiding the sensitive attributes – gender, high school status – is a good mitigation approach. None of the individual mitigation strategies outperform consistently across the three scenarios. The same is observable when we analyze the type of fairness mitigation approaches (pre, in, and post-processing). However, there is an exception for *School* case, where *β-VAE* addresses all the fairness issues raised by our stress test analysis. This is consistent with our remark in Section 3 that most of these mitigation strategies are targeted for some specific fairness metrics and that their fairness guarantees do not necessarily transfer to other metrics. The results show that the lack



FAIRED

OBJECTIVE
Analysing whether Chilean Dataset is appropriate as an input to fair education data mining pipeline.

DATA CONTEXT
- Data includes administrative records of students enrolled in a bachelor's business program.
- Sensitive attributes: because in Chile, most segregation occurs along poverty lines and enrollment in public high school is a sign of lower socio-economic status, sensitive attributes high school status, along with gender
- Student dropout is used as an outcome to stress test the data for fairness.

DATA EXPLORATION
An analysis of features importance when predicting the sensitive attribute shows that high school status is correlated with high school size (because private school are smaller), test scores and income. This indicates that high school size, test scores and income may be problematic to be used as inputs to data mining pipeline.

FAIRNESS STRESS TESTS
An analysis of the fair properties of classifiers with high f1 score when predicting student dropout shows that without mitigation the use of the Chilean dataset would lead to significant disparate imapct for female students and students from a public high school.

MITIGATION
By plotting the effect of four pre-processing mitigation techniques on f1 score and fairness metrics, we can visualize whether the data can still be useful to a data mining pipeline after addressing its fairness concerns.

Public High School
Gender
Gender - Public High School
Random Forest Feature Importance
rf
lr
svm
dt
knn
Fairness Metrics across Gender
Fairness Metrics across High School Type
Fairness Metrics across Gender-High School
DemPi
DispR
EqOPP

Figure 4: FairEd card design comprises objective, data context, data exploration, fairness stress tests, and mitigation. Each section is based on discussion Section 3.

of winning mitigation strategies illustrates the need for a systematic and context-dependent exploration of the data with different mitigation strategies, the prime takeaway of our study. Second, in-processing and post-processing techniques do not outperform pre-processing techniques. This is consistent with results from [42] who show that pre-processing could be an efficient mitigation approach. From a data owner's perspective, it means that it is possible to mitigate the fairness issues associated with the dataset before the distribution of the dataset to third parties. Lastly, Table 5 provides indications to the data owner on whether applying data mining algorithms on the dataset is still useful when fairness concerns have been addressed. On one hand, complete mitigation, as obtained by β*-VAE*, comes at the cost of very low predictive performances (F1 score down from 0.5 to 0.2). On the other hand, Table 5 makes clear that to maintain a F1 score above 0.5 will force some trade-off in terms of fairness. BY illustrating these trade-offs, Table 5 is a useful tool for data owners and stakeholders to decide whether they are willing to bear some fairness risk by allowing third parties to access the data and use it to train decision-making systems.

Finally, we build a report card and label it as FairEd. Figure 4 presents this FairEd card, which offers a summary of the fairness risks associated with the use of the Chilean dataset in an education data mining pipeline.

5 CONCLUSION

In this paper, we propose a framework –FairEd – to systematically explore whether using an education dataset could lead to unfair outcomes. The objective is to allow data owners and stakeholders to anticipate the potential fairness risks of allowing the data to be used as an input to complex data mining pipelines and encourage the participation of stakeholders around the design of the model. This participation framed the design into an algorithmic fairness perspective. The FairEd framework is tested in a Chilean University case study where a predictive model on dropouts is aimed to build. We provide the analysis and definitions performed regarding each element in the framework. The case study allows to show the benefits of FairEd and exemplifies the use of the proposed systematic approach. We believe that any algorithmic fairness must be necessarily framed by the objectives of the model construction and the data contextualization according to the stakeholders and experts participants in the definition of these elements. Their outcomes provide guidelines for the fairness analysis comprised of three elements. The data exploration provides clear signs of potential variables of unfairness risk sources from the fairness analysis domain. These signs would guide the data's stress testing, where many classifiers and many fairness metrics can be selected to analyze potential unfair outcomes. Finally, the third element of the fairness analysis domain allows identifying a proper mitigation approach by balancing the power of prediction and discrimination costs. We believe that FairEd can help education institutions make an informed decision when considering authorizing third parties to access the data and use it as an input to decision-making systems.

We highlight that our proposed framework is a technical tool to guide and generate stakeholder engagement around the ethics of learning analytics. Additionally, it does not provide a road map on how to initiate/foster this engagement or on how to alleviate potential ethical concerns, although it highlights the awarness of unfairness risks on the outcomes that a learning algorithm may generate.

A variety of roles play in the implementation of FairEd framework. Each one should play concrete responsibilities and relate to other parts differently. Although these roles are not defined in this paper, it is expected to identify and describe them for future research. In addition, there are some inquiries about the potential effects that this framework may generate. For example, the first two elements could eventually generate spaces that, rather than encouraging the use of machine learning models, could discourage it. The evaluation of these effects, as well as the mitigation approaches, are assessed for future researches.

ACKNOWLEDGMENTS

This study was supported by the National Agency for Research and Development (ANID - Agencia Nacional de Investigación y Desarrollo/Subdirección de Capital Humano), "Becas Chile" Doctoral Fellowship 2020 program; Grant No. 72210492 to Jonathan Patricio Vasquez Verdugo.

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54, 1 (2018), 95–122.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[3] Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. 2019. Mining University Registrar Records to Predict First-Year Undergraduate Attrition. *International Educational Data Mining Society* (2019).

[4] Ryan S. Baker and Aaron Hawn. 2021. lgorithmic Bias in Education. *EdArXiv* (2021).

[5] John P Bean. 1980. Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in higher education* 12, 2 (1980), 155–187.

[6] John P Bean and Barbara S Metzner. 1985. A conceptual model of nontraditional undergraduate student attrition. *Review of educational Research* 55, 4 (1985), 485–540.

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[8] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[9] Trudy H Bers and Kerry E Smith. 1991. Persistence of community college students: The influence of student intent and academic and social integration. *Research in higher Education* 32, 5 (1991), 539–556.

[10] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[11] John M Braxton, Ana V Shaw Sullivan, and Robert M Johnson. 1997. Appraising Tinto's theory of college student departure. *HIGHER EDUCATION-NEW YORK-AGATHON PRESS INCORPORATED-* 12 (1997), 107–164.

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[13] Alberto F Cabrera, Amaury Nora, and Maria B Castaneda. 1993. College persistence: Structural equations modeling test of an integrated model of student retention. *The journal of higher education* 64, 2 (1993), 123–139.

[14] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001.

[15] Hongyan Chang and Reza Shokri. 2020. On the Privacy Risks of Algorithmic Fairness. *arXiv preprint arXiv:2011.03731* (2020).

[16] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[17] Kristof Coussement, Minh Phan, Arno De Caigny, Dries F. Benoit, and Annelies Raes. 2020. Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems* 135 (2020), 113325.

[18] Deana B Davalos, Ernest L Chavez, and Robert J Guardiola. 1999. The effects of extracurricular activity, ethnic identification, and perception of school on student dropout rates. *Hispanic Journal of behavioral sciences* 21, 1 (1999), 61–77.

[19] Consejo Nacional de Educacion. [n. d.]. Retencion Primer Año. https://www.cned.cl/indices/retencion-primer-ano. Accessed: 2020-11-30.

[20] Dursun Delen. 2011. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice* 13, 1 (2011), 17–35.

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference.* 214–226.

[22] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[23] Gregory Elacqua. 2012. The impact of school choice and public policy on segregation: Evidence from Chile. *International Journal of Educational Development* 32, 3 (2012), 444–453.

[24] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* 259–268.

[25] Antonio Jesús Fernández-García, Roberto Rodríguez-Echeverría, Juan Carlos Preciado, José María Conejero Manzano, and Fernando Sánchez-Figueroa. 2020. Creating a Recommender System to Support Higher Education Students in the Subject Enrollment Decision. *IEEE Access* 8 (2020), 189069–189088.

[26] Josep Figueroa-Cañas and Teresa Sancho-Vinuesa. 2020. Early Prediction of Dropout and Final Exam Performance in an Online Statistics Course. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* 15, 2 (2020), 86–94.

[27] National Center for Education Statistics. 2020. Undergraduate retention and graduation rates. (2020).

[28] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[29] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge.* 225–234.

[30] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[31] Xavier Gitiaux and Huzefa Rangwala. 2019. mdfa: Multi-Differential Fairness Auditor for Black Box Classifiers.. In *IJCAI.* 5871–5879.

[32] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. 2019. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning.* 2357–2365.

[33] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, Vol. 1. 2.

[34] Lani Guinier. 2015. *The tyranny of the meritocracy: Democratizing higher education in America.* Beacon Press.

[35] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413* (2016).

[36] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. (2016).

[37] Isabel Hilliger, Margarita Ortiz-Rojas, Paola Pesántez-Cabrera, Eliana Scheihing, Yi-Shan Tsai, Pedro J Muñoz-Merino, Tom Broos, Alexander Whitelock-Wainwright, and Mar Pérez-Sanagustín. 2020. Identifying needs for learning analytics adoption in Latin American universities: A mixed-methods approach. *The Internet and Higher Education* 45 (2020), 100726.

[38] Qian Hu and Huzefa Rangwala. 2020. Metric-Free Individual Fairness with Cooperative Contextual Bandits. *arXiv preprint arXiv:2011.06738* (2020).

[39] Qian Hu and Huzefa Rangwala. 2020. 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).* 431–437.

[40] Huade Huo, Jiashan Cui, Sarah Hein, Zoe Padgett, Mark Ossolinski, Ruth Raim, and Jijun Zhang. 2020. Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach. *Journal of College Student Retention: Research, Theory & Practice* (2020), 1521025120963821.

[41] Stephen Hutt, Margo Gardner, Angela L Duckworth, and Sidney K D'Mello. 2019. Evaluating Fairness and Generalizability in Models Predicting On-Time Graduation from College Applications. *International Educational Data Mining*

Society (2019).

[42] Sajad Khodadadian, AmirEmad Ghassami, and Negar Kiyavash. 2021. Impact of Data Processing on Fairness in Supervised Learning. *arXiv preprint arXiv:2102.01867* (2021).

[43] René F Kizilcec and Hansol Lee. 2020. Algorithmic Fairness in Education. *arXiv preprint arXiv:2007.05443* (2020).

[44] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[45] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*. 4066–4076.

[46] David Kynaston and Francis Green. 2019. *Engines of privilege: Britain's private school problem*. Bloomsbury Publishing Plc.

[47] Dean R Lillard and Philip P DeCicca. 2001. Higher standards, more dropouts? Evidence within and across time. *Economics of Education Review* 20, 5 (2001), 459–473.

[48] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830* (2015).

[49] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*. PMLR, 3384–3393.

[50] Charles T Marx, Richard Lanas Phillips, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Disentangling Influence: Using disentangled representations to audit model predictions. *arXiv preprint arXiv:1906.08652* (2019).

[51] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[52] Paul A Murtaugh, Leslie D Burns, and Jill Schuster. 1999. Predicting the retention of university students. *Research in higher education* 40, 3 (1999), 355–371.

[53] Senthil Kumar Narayanasamy and Atilla Elçi. 2020. An effective prediction model for online course dropout rate. *International Journal of Distance Education Technologies (IJDET)* 18, 4 (2020), 94–110.

[54] Zachary A Pardos, Zihao Fan, and Weijie Jiang. 2019. Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User modeling and user-adapted interaction* 29, 2 (2019), 487–525.

[55] Angie Parker. 1999. A study of variables that predict dropout from distance education. *International journal of educational technology* 1, 2 (1999), 1–10.

[56] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[57] Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H Shah. 2019. Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 271–278.

[58] ProPublica. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (2016).

[59] Nor Samsiah Sani, Ahmad Fikri Mohamed Nafuri, Zulaiha Ali Othman, Mohd Zakree Ahmad Nazri, and Khairul Nadiyah Mohamad. 2020. Drop-Out Prediction in Higher Education Among B40 Students. *International Journal of Advanced Computer Science and Applications* (2020).

[60] William G Spady. 1970. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange* 1, 1 (1970), 64–85.

[61] William G Spady. 1971. Dropouts from higher education: Toward an empirical model. *Interchange* 2, 3 (1971), 38–62.

[62] Nate Sutter and Sharon Paulson. 2017. Predicting college students' intention to graduate: a test of the theory of planned behavior. *College Student Journal* 50, 3 (2017), 409–421.

[63] Mack Sweeney, Huzefa Rangwala, Jaime Lester, and Aditya Johri. 2016. Next-term student performance prediction: A recommender systems approach. *arXiv preprint arXiv:1604.01840* (2016).

[64] Vincent Tinto. 1975. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research* 45, 1 (1975), 89–125.

[65] Vincent Tinto. 2007. *Taking student retention seriously*. Syracuse University Syracuse, NY.

[66] Vincent Tinto and John Cullen. 1973. Dropout in Higher Education: A Review and Theoretical Synthesis of Recent Research. (1973).

[67] Jonathan Vásquez and Jaime Miranda. 2019. Student Desertion: What Is and How Can It Be Detected on Time? In *Data Science and Digital Business*. Springer, 263–283.

[68] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*. 229–239.

[69] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[70] Seth D Zimmerman. 2019. Elite colleges and upward mobility to top jobs and top incomes. *American Economic Review* 109, 1 (2019), 1–47.