Automating Procedurally Fair Feature Selection in Machine Learning

Clara Belitz
University of Illinois
Urbana–Champaign
Champaign, IL, USA
cbelitz2@illinois.edu

Lan Jiang
University of Illinois
Urbana–Champaign
Champaign, IL, USA
lanj3@illinois.edu

Nigel Bosch
University of Illinois
Urbana–Champaign
Champaign, IL, USA
pnb@illinois.edu

ABSTRACT

In recent years, machine learning has become more common in everyday applications. Consequently, numerous studies have explored issues of unfairness against specific groups or individuals in the context of these applications. Much of the previous work on unfairness in machine learning has focused on the fairness of outcomes rather than process. We propose a feature selection method inspired by fair process (procedural fairness) in addition to fair outcome. Specifically, we introduce the notion of *unfairness weight*, which indicates how heavily to weight unfairness versus accuracy when measuring the marginal benefit of adding a new feature to a model. Our goal is to maintain accuracy while reducing unfairness, as defined by six common statistical definitions. We show that this approach demonstrably decreases unfairness as the unfairness weight is increased, for most combinations of metrics and classifiers used. A small subset of all the combinations of datasets (4), unfairness metrics (6), and classifiers (3), however, demonstrated relatively low unfairness initially. For these specific combinations, neither unfairness nor accuracy were affected as unfairness weight changed, demonstrating that this method does not reduce accuracy unless there is also an equivalent decrease in unfairness. We also show that this approach selects unfair features and sensitive features for the model less frequently as the unfairness weight increases. As such, this procedure is an effective approach to constructing classifiers that both reduce unfairness and are less likely to include unfair features in the modeling process.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; **Machine learning**.

KEYWORDS

Feature selection, fairness, bias, machine learning

ACM Reference Format:
Clara Belitz, Lan Jiang, and Nigel Bosch. 2021. Automating Procedurally Fair Feature Selection in Machine Learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462585

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462585

1 INTRODUCTION

Approaches to improving algorithmic fairness in the context of machine learning applications have mainly focused on three categories of methods: pre-processing [9, 28], in-processing [17, 29, 46], and post-processing [23]. Each proposes an intervention at a specific stage of the machine learning process to achieve fairness, either before, during, or after the model training process. In addition, a fourth category of work involves ensuring that even earlier steps, like data collection and annotation, are fair [27]. Most of these methods focus on fair predictions and derive their assessment of fairness by measuring outcome alone. In this paper, however, we explore a machine learning method that modifies the fairness of the model building process by selecting which variables may be used when making decisions. This allows us to balance not only the compromise between accuracy and cost in the outcome, but also the fairness of the features used in the learning process.

The previously described approaches tend to focus specifically on the concept of fairness. In related work outside of machine learning, however, a distinction has been drawn between that of *fairness* and *unfairness*. For example, Cojuharenco and Patient [13] show that when presented with language around unfairness, people were more likely to consider inputs and processes rather than outcomes. Given our focus on inputs (features), we therefore use measurements of unfairness in this paper. This allows us to measure the reduction of unfairness as we improve processes, rather than outcome fairness alone.

Outcome fairness is also referred to as *distributive* fairness, which refers to the fairness of the outcomes of decision making, while *procedural* fairness refers to the fairness of the decision making processes that lead to those outcomes [22]. We can interpret the idea of fair process in this context to mean building the model itself in a way that incorporates concerns of fairness [22]. This leads us to ask questions of the model. For example: are protected features included? Protected features describe groups that must not be discriminated against in a formal or legal sense, as defined by each notion of (un)fairness [33]; examples of common protected features are race and gender. Other demographic categories may not be legally protected, but may still be questionable to use, such as whether a student lives in a rural area. The combination of protected and otherwise undesirable features will together be referred to as *sensitive* features in this paper. A second question we might ask is: are unfair features being included in decision making processes?

Unfair features are those where one group is likely to benefit more than another from their inclusion. Unfair features may be a proxy for protected features (e.g., ZIP code as a proxy for race), biased due to historical or social contexts (e.g., standardized college entrance exam scores [37]), or otherwise statistically skewed. Of course, a dataset may include more than one sensitive or unfair feature and it can be difficult to determine the full set of these features. As such, one previous approach has examined how humans rate the inclusion of sensitive features, in general and when given knowledge about how those features affect the accuracy and fairness of the classifier [22]. In this paper we take inspiration from Grgić-Hlača et al. [22] to move beyond distributive fairness, and examine how automating feature selection can contribute to a fair process for building classifiers. The value of our approach to practitioners is that our feature selection process can avoid sensitive features that may have been overlooked. Rather than declaring one or more features sensitive and thus off-limits, this approach generalizes the desired statistical measure of unfairness to ensure that each feature selected is more fair to use.

In sum, in this paper we define a straightforward process for building procedurally fair classifiers by incorporating both unfairness and accuracy considerations during feature selection. We investigate how this process affects both accuracy and unfairness outcomes in practice. As such, we explore how an automated feature selection process can incorporate elements of fairness in order to improve both process and outcomes. We also specifically investigate how a fairer feature selection process affects the inclusion of both sensitive features and unfair features in models. We explore how this approach works when applied to three commonly used machine learning classifiers, and how it performs in terms of both accuracy, as measured by AUC, and unfairness, as measured by six different statistical definitions (defined in the **Unfairness Definitions** section). We demonstrate that our approach generally reduced unfairness, with an inevitable reduction in accuracy [17, 18, 22], for three real-world datasets. Additionally, for simulated data, we demonstrate that our method chose both a sensitive feature and an unfair feature less frequently. This demonstrates that the method caused the classifier to be less likely to include sensitive and unfair features in the decision-making process, leading to a fairer model with regards to procedure as well as predictions.

1.1 Previous Conceptualizations of Fairness

The study of fairness in machine learning has seen a dramatic increase because unfairness is pervasive in algorithmic decision making [36]. As machine learning has proliferated, so have questions of its impacts [3, 11, 34]. One of the commonly studied impacts is that of fair usage and outcomes. Much of the work evaluating the fairness of machine learning models has focused on a variety of statistical definitions describing fair outcomes [20]. These quantitative definitions primarily arose from fairness literature based in educational testing, but have been based in legal definitions or strictly mathematical uses as well [25]. It has been proven that, except in highly specialized cases, different definitions of fairness cannot all be satisfied at the same time [11, 30]. In addition, which definition is most applicable is often context dependent and has different implications that depend on how outcomes are measured. For example, statistical parity is a measure of population-wide fairness, and may not account for discrimination against specific demographics [15]. Given this constraint, and the fact that fairness is a social concept [11], it is generally necessary for researchers and users to choose which definitions are appropriate to apply to a given classifier and dataset, based on the targeted fair prediction outcomes. In this paper, we thus demonstrate the effect of our proposed strategy on a variety of these fairness measures.

Our proposed strategy is based on altering the definition of accuracy to incorporate a conceptualization of unfairness during the feature selection process. Selecting which features to include in a model is crucial because adding more features may lead to worse predictions [44], in addition to increasing model complexity. When training a classifier, then, only the "best" features should be included in the model. How "best" is defined, however, has a strong effect on the resulting model [41]. Past research has tended to use "most accurate," as defined by a specific measure of error, as the indicator of best. However, measures of accuracy optimize prediction outcomes based on a particular dataset, but the data used can be unfair because it was created in an unfair world, it is missing information, or it is otherwise unrepresentative [3, 5]. Therefore, accuracy alone is not a neutral measure of classifier value. Merely excluding sensitive features (e.g., protected group status like race) is not, however, a perfect remedy for fixing unfairness. Often, removing a sensitive feature does not remove all correlated proxy features, and even removing all of the correlated features does not always provide an acceptable trade-off in accuracy and fairness [33]. In addition, previous work has shown that including seemingly "unfair" features can actually lead to better outcomes in terms of both fairness and accuracy, since the model can then account for past unfairness that created the initial dataset [10, 11, 30, 33]. Therefore, selecting appropriate features for a fair model is not a simple matter of applying a pre-processing rule to the data, and may be better accomplished by a more complex feature selection approach.

1.2 Our Study

We propose incorporating unfairness into feature selection as an alternative to optimizing for accuracy alone. We demonstrate an approach to model building with classification algorithms that aims to minimize unfairness while simultaneously maximizing accuracy during the feature selection step. Furthermore, we trained three classical machine learning algorithms—Gaussian naïve Bayes, logistic regression, and decision trees—in order to explore the generality of our approach.

We define the following two research questions (RQs) to explore this problem:

RQ1: Does our proposed method reduce unfairness? If so, according to which unfairness definitions? Addressing this question will determine whether the method affects model predictions as expected. In particular, we assess how accuracy, as measured by AUC, and unfairness, as defined in the **Unfairness Definitions** section, are affected by our approach. We look at this question overall as well as examining the impact across levels of the trade-off between accuracy and unfairness.

RQ2: How does the selection of an unfair feature and the sensitive feature affect accuracy and unfairness? We examine simulated data to understand when the unfair feature and sensitive feature are included during feature selection. We look at whether this varies with different definitions of unfairness and perform statistical tests to determine whether the sensitive group status and an unfair feature are selected less frequently as unfairness is weighted more heavily in the accuracy–unfairness trade-off.

Note that while we are measuring a single sensitive feature per dataset in this paper, the approach can be applied to multiple sensitive features—for example, by averaging unfairness metrics for each sensitive feature during feature selection.

1.3 Related Work

As discussed above, approaches to reducing unfairness in machine learning generally fall into three categories: pre-processing, in-processing, and post-processing. Kamiran and Calders [28] demonstrated four now-common approaches to pre-processing: 1) suppression, which entails removing the sensitive feature and its most correlated features from the data; 2) "massaging" the dataset, which involves changing a strategic selection of labels in the dataset in order to remove discrimination; and 3) "reweighing", which involves resampling instances in the dataset to make the data discrimination-free with regards to the sensitive feature. Using these approaches, they demonstrated that removing the sensitive feature from the dataset does not always result in the removal of the discrimination. Massaging and resampling were more effective, "leading to an effective decrease in discrimination with a minimal loss in accuracy." For example, they showed a decrease in discrimination from 16.48% to 3.32% for their decision tree, while accuracy only decreased from 86.05% to 84.3%. Similarly, Calmon et al. [9] use a distortion function based on an unfair feature in the initial data to create a transformed dataset. They showed that this transformed dataset led to fairer classification, though the reduction in unfairness came with an accuracy penalty. This reduction was due to the restrictions imposed on transformation. In general, restrictions in feature or data usage will tend to lead to a reduction in accuracy due to a loss of information [9, 22].

The second category of approaches are in-processing, referring to making the learning algorithm itself less unfair. Many papers have addressed this topic [8, 15, 17, 29, 45, 46]. Fish et al. [17] use two in-processing approaches. The first is a shifted decision boundary which finds the minimal-error decision boundary shift for the sensitive group that achieves statistical parity (equal predicted rate for all groups). The second is fair weak learning, which is specific to adapted boosting (a machine learning method) and replaces a standard boosting weak learner with one which tries to minimize a linear combination of error and unfairness. They demonstrated that using a shifted decision boundary allows a substantial reduction in bias before there is significant drop-off in accuracy for large enough datasets. Zafar et al. [46] proposed a new statistical definition of unfairness, and proceeded to train logistic regressors that were constrained by false positives. They were able to build a classifier that did not use sensitive features in decision making, and still achieved similar results to post-processing approaches like those of Hardt et al. [23].

The final category is post-processing, which makes the model decisions less biased after the model has been built. Hardt et al. [23] demonstrate an example of this; they designed a simple post-processing step that allowed them to avoid changing a complex training pipeline and avoid loss of information from the original data. They proposed post-processing as a last resort when better features, and more and better data, cannot be obtained. They proposed a new notion of non-discrimination, "obliviousness," based only on the joint distribution of the true target outcomes, the predicted target outcomes, and the sensitive feature. Obliviousness does not evaluate the features nor the functional form of the prediction algorithm. They used a case study of financial risk assessment to show that it was possible to maintain close to full profitability with some approaches, such as a race-blind one, but that others, such as equalized odds (equal bias and equal accuracy across all categories of the sensitive feature), are costlier. Their findings highlight the inherent trade-offs of fairness and accuracy, and show that different fairness standards affect accuracy and other measurements of success differently.

2 EXPERIMENTS AND METHODS

The approach in this paper builds on previous work, as described above, by incorporating constraints on unfairness during the feature selection step. In this section, we describe the method in detail, as well as the process by which we evaluated it. We also describe the datasets used for evaluation.

2.1 Feature Selection Approach

Our feature selection method is a form of *wrapper feature selection*. In wrapper feature selection, a machine learning model is trained with a given feature or set of features, and then a model evaluation metric (e.g., AUC) is used as the measure of how good that particular feature set is. We specifically utilized *forward* feature selection. Forward feature selection consists of training all one-feature models (i.e., one model per feature), then choosing whichever feature was best according to the model evaluation metric. The process then repeats with every possible two-feature model, using the best feature from the first round and pairing it with each remaining feature. This continues until reaching a stopping criterion: in our case, once all features had been added to the model. Finally, we selected the best set of features from among all those that we explored during the feature selection process.

Our method differs from typical forward feature selection in the model evaluation step. Typically, models (and thus feature sets) would be evaluated based on an accuracy metric, such as AUC [26], Cohen's kappa [12], or another metric [41]. We instead evaluated models based on a combination of accuracy and unfairness. Specifically, we maximized $AUC - weight * unfairness$, where $weight$ is a hyperparameter selected by the experimenter to balance the relationship between accuracy and unfairness, and $unfairness$ is a measure of inequality in the model's predictions. We explored six different unfairness definitions, described in the **Unfairness Definitions** section, which were all implemented such that they ranged from 0 (no unfairness) to 1 (maximal unfairness). Thus, features are selected in this method if they yielded a model that improved

accuracy while not adding a large amount of unfairness. For example, consider a situation where one feature has a correlation with the outcome variable in one demographic group but not another, whereas another feature has a weaker correlation that is consistent across groups. In this case, we expect the method will preferentially select the latter feature.

Note that AUC, used during forward feature selection, tends to yield models that predict positive and negative cases at rates that differ from the original data (the base rate), versus some measures like mean squared error and Cohen's kappa [41]. Matching predicted rates to base rates is related to some unfairness definitions (e.g., statistical parity), and thus there may be an interaction between choice of accuracy metric and unfairness metric. We leave this consideration to future work, though in principle our method works with whatever accuracy metric is suitable for a particular problem.

We incorporated our method into existing feature selection functionality in the *MLxtend* Python package [39], which in turn integrates with the *scikit-learn* package [38]. Our code for all experiments is publicly available[1]. We then evaluated the method via machine learning experiments implemented with *scikit-learn*, as described below.

2.2 Model Training

We explored three common classification algorithms: Gaussian naïve Bayes, logistic regression, and decision trees (specifically, *classification and regression trees*, or CART [7]). We selected these three as examples because they are well-known, widely-used, and vary considerably in how they make classification decisions. Naïve Bayes has no inherent feature selection capabilities, and is sensitive to the "curse of dimensionality", wherein unnecessary features negatively impact model predictions [19]. Hence, feature selection is a typical step in the training process for naïve Bayes models. Similarly, logistic regression models may suffer when unnecessary features are included, especially if those features are highly collinear [42]; hence, feature selection is common. Logistic regression often incorporates feature selection via L1 regularization [43], which effectively eliminates features from a model by setting their weight coefficients to 0. However, logistic regression may still benefit from our feature selection method in terms of reducing unfairness, since L1 regularization does not penalize unfairness. Finally, we trained decision trees, which are typically robust to the presence of unnecessary features [6]. Features that are uncorrelated with the outcome can be ignored in a decision tree, since the model may simply choose to make branching decisions based on other features. However, we expect that our method will reduce unfairness even if feature selection is not otherwise needed.

We trained models with 4-fold cross-validation, in which training data came from 75% of randomly-chosen individuals, and data from the remaining 25% of individuals were used to test the model accuracy and unfairness. We repeated the process 4 times so that each individual was in the test data exactly once. We repeated each experiment 100 times with different randomly-chosen cross-validation data partitions, then calculated means of all evaluation metrics so that results were not influenced by the random choice of training and testing datasets. We performed the feature selection method, described above, using nested 4-fold cross-validation within training data only. That is, we further split training data into subsets to train and evaluate the models built as part of feature selection, thereby avoiding overfitting the feature selection step based on testing set accuracy. Apart from feature selection, we did not tune any classifier hyperparameters, leaving them at the *scikit-learn* default settings. We expect that the key results explored in this study would be unaffected by hyperparameter tweaks made to classifiers, and leave such analysis to future work.

[1] https://github.com/pnb/fairfs/releases/tag/v0.1-aies

2.3 Unfairness Definitions

In this paper, we used six statistical unfairness definitions to measure unfairness in our experiments. Specifically, from among many possible definitions, we chose the set of definitions in Berk et al. [4].

As background for these definitions, let there exist a machine learning classifier that predicts a set of values, Y', based on both "legitimate" predictors of the outcome of interest and "sensitive" predictors, which may be legally protected, such as race, or otherwise questionable to use, such as ZIP code. Y', the outcome predicted by the classifier, is an estimate of Y, the true outcome of interest. Let TP = true positives, FN = false negatives, FP = false positives, and TN = true negatives, such that $TP + FN + FP + TN$ add up to the sample size. The category of "true" results are when Y' equals Y, while "false" results are where they differ; "positive" refers to values of Y or Y' that belong to a specific class (e.g., students with above-median grade in the Student Academics dataset described below), while "negative" refers to all other values.

Our fairness definitions are based on the relationships between measured outcomes and ground truth. We consider only example datasets with a binary sensitive group (e.g. Male or Female) and two outcome classes, but we have implemented these metrics for general multi-class, multi-group cases by taking the maximum unfairness across classes and groups.

- **Overall Accuracy Equality**: achieved by a classifier when overall procedure accuracy is the same for each category of the sensitive feature. That is, true positives + true negatives is the same proportion ($\frac{TP+TN}{TP+FN+FP+TN}$) for each group. We measured unfairness with this definition by calculating the absolute difference between proportion correct per group. This definition is imperfect in that it does not distinguish between accuracy for correctly identified positives and accuracy for correctly identified negatives. That is, overall accuracy equality assumes that true negatives are as desirable as true positives. However, no metric can account for every conceivable dimension of fairness, so we also evaluated other related metrics.
- **Statistical Parity**: achieved by a classifier when the marginal distributions of the predicted classes are the same for each category of the sensitive feature. That is, the proportion of all positives ($\frac{TP+FP}{TP+FN+FP+TN}$) and all negatives ($\frac{FN+TN}{TP+FN+FP+TN}$) are the same for all groups. This definition of statistical parity, sometimes called "demographic parity," also has flaws because it can lead to highly undesirable decisions [15]. It is not necessarily valuable to force all groups to

have equivalent statistical distributions if the groups don't require equal treatment (e.g., recommending interventions for students who don't need them just to keep the numbers equivalent). We calculated unfairness for this definition as the difference between the predicted rate (proportion of positive class predictions) across groups.

- **Conditional Procedure**: achieved by a classifier when conditional procedure accuracy is the same for each category of the sensitive feature. That is to say, true positives versus true positives and false negatives (actual positives; $\frac{TP}{TP+FN}$) is the same for all groups, and true negatives versus true negatives and false negatives (actual negatives; $\frac{TN}{FP+TN}$) is the same for all groups. In other words, when conditioning on the known outcome, is the approximating function equally accurate across sensitive group categories? This is the same as considering whether the false negative rate and the false positive rate, respectively, are the same for all groups. Thus, we measured unfairness with this definition as the absolute difference between recall scores across groups. Note that this measure has a special case called "equality of opportunity" that effectively is the same as conditional procedure accuracy equality, but only for the outcome class that is more desirable (e.g., positives for hiring), which may raise the question of which outcome class is more desirable. In addition, equality of opportunity is not always a useful measure, in that it cannot account for bias that was present at the time of data creation [33].
- **Conditional Use Accuracy Equality**: achieved by a classifier when conditional use accuracy is the same for each category of the sensitive feature. This definition is conditioning on the algorithms' predicted outcome, not the actual outcome. That is, true positives versus true and false positives (all predicted positives; $\frac{TP}{TP+FP}$) is the same for all groups, and true negatives versus true and false negatives (all predicted negatives; $\frac{TN}{FN+TN}$) is the same for all groups. We calculated unfairness according to this definition as the absolute difference in precision across groups.
- **Treatment Equality**: achieved by a classifier when the ratio of false negatives and false positives ($\frac{FP}{FN}$ or $\frac{FN}{FP}$) is the same for all categories of the sensitive feature. The term "treatment" is used to convey that such ratios can be a policy lever with which to achieve other kinds of fairness. This allows the modeler to decide how costly a wrong outcome is for a specific category of the sensitive feature (e.g., a false negative is worse than a false positive), which is the quality missing from overall accuracy equality. We measured unfairness by whichever ratio was larger; however, unlike the other unfairness measures, these ratios may exceed a value of 1. Thus, we applied a sigmoid function to transform values into the [0, 1] range like the other measures, since our feature selection method assumes the accuracy and unfairness measures have similar scale.
- **Total Average Equality**: defined as the mean of the previous five equality metrics. Note that it is impossible for all five of the preceding definitions to be satisfied at the same time. For example, conditional procedure equality (matching false positive and negative rates across groups) cannot be achieved alongside statistical parity (matching positive and negative predicted rates across groups) if the ground truth proportion of positive or negative instances differs across groups.

2.4 Datasets

We tested our feature selection method on four datasets. The first three datasets are all publicly available from the UCI Machine Learning Repository, while the fourth was generated by the authors. The **Student Performance** and **Student Academics** datasets were collected in educational contexts, where machine learning is valued because early prediction of student success can drive adaptations that improve students' learning [2, 16]. However, machine learning models derived from such data may be unfair [21, 35], including potential structural unfairness (e.g., predicting a student will succeed simply because of their affluence). The third dataset, **Adult**, is commonly used in machine learning literature, and as such is helpful for baseline comparisons. The fourth dataset is a simulated dataset, which we created to explicitly test whether the feature selection method would select a feature known to be unfair.

Student Performance: This dataset consists of person-level survey responses from students and administrative data provided by schools (including demographic information and students' grades) for two courses [14]. In this study, we analyzed data from a mathematics course with 395 students, predicting whether each student's final grade was above or below the median. The data include predictor variables like number of absences and attitudes toward school, some of which are examples of variables that may be predictive but which could be procedurally unfair to use for grade prediction (e.g., age, romantic relationship status). We extracted 33 features from the numeric and nominal information in this dataset. We treated students' home community size (rural versus not rural) as the sensitive feature for measuring unfairness, given previous research showing the potential for machine learning models to generalize poorly across these categories [35].

Student Academics: This dataset consists primarily of demographic information for students, including parental occupations, family size, gender (only male and female), and others for 131 students [24]. In total, we extracted 22 features. Final grades are also included, which we predicted in this study as a binary outcome variable (above or below the median). This dataset includes some non-demographic predictors, such as attendance records, but includes several demographic variables with potential to induce model unfairness if selected. As in the Student-Performance dataset, we treated rural/non-rural status as the sensitive feature for measuring unfairness.

Adult: This dataset consists of information from the 1994 U.S. Census database [31]. This dataset contains information for 48,842 American adults, predicting whether a person earns above or below $50K a year. In total, we extracted 22 features, after transforming categorical features to binary features and removing infrequently-occurring categories. The data include predictor variables like education level, marital status, and hours worked per week, but also include some demographic variables with potential to introduce unfairness if selected, such as race. In this study, we treated sex as

the sensitive feature for measuring unfairness, as has been done in past literature that uses this dataset [17, 28, 29].

Simulated Data: This dataset is artificially constructed, and contains 1,000 rows in total. It consists of four columns: group (the sensitive group membership status, 0 or 1), the binary outcome column, a "fair" feature (each group has a similar correlation between this feature and the outcome), and an "unfair" feature (where the correlation with outcome differs for group 0 and group 1). Sensitive group status consisted of 500 rows each for groups 0 and 1. We generated a normally-distributed random variable, normalized the values in the [0, 1] range, and generated the outcome column by rounding to 0 or 1. We then transformed the same underlying normally-distributed random variable to generate the fair and unfair features. When generating the fair feature, we randomly replaced 300 out of 1,000 variable's values with uncorrelated values. Conversely, when generating unfair feature, we randomly replaced 225 of the values from group 0 with uncorrelated values but only 75 of the values from group 1. Point-biserial correlations indicated the expected patterns; the overall correlation with outcome for the fair and unfair features were similar (r_{pb} = .328 and r_{pb} = .345, respectively), but the unfair feature's correlation differed by group (r_{pb} = .145 for group 0, r_{pb} = .554 for group 1) while the fair feature's correlation differed little (r_{pb} = .290 vs. r_{pb} = .367).

3 RESULTS

We focus results on the research questions outlined in **Introduction: Our Study**. In this section, we describe our results as well as the tests used to assess effectiveness of our method. We examine trends across datasets and classifiers, as well as the effect of unfairness weight on accuracy and unfairness. Most importantly, we demonstrate that our approach reduced unfairness for our test datasets. We conclude by looking at how unfairness weight affects the inclusion of specific features, and show that the sensitive feature and unfair feature were selected less frequently as the unfairness weight was increased.

3.1 Reduction in Unfairness (RQ1)

RQ1 asked whether the proposed method reduces unfairness. Overall, our method reduced unfairness in our datasets as measured by our six definitions. For the Student Performance dataset, for example, at unfairness weight 4 (i.e., where reducing unfairness is approximately 4 times as important as accuracy) averaging across all metrics, mean unfairness decreased 54% (0.215 to 0.123) for logistic regression, 55% (0.217 to 0.123) for Gaussian naïve Bayes, and 27% (0.158 to 0.121) for decision trees, when compared to unfairness weight 0. This aligned with our expectations that decision trees, which are robust to the presence of unnecessary features, would still benefit from feature selection to decrease unfairness.

3.1.1 Effect of Unfairness Weight. As we increased the unfairness weight, for the majority of unfairness metrics and classifiers, the unfairness of predictions significantly decreased (as expected). We measured the decrease statistically with Mann-Whitney U tests, given that unfairness values were ordinal but not normally-distributed. In total, we ran 90 experiments per dataset, consisting of each combination of 3 classifiers, 6 unfairness metrics, and 5 unfairness weights. Of these 90 experiments, 72 had a positive unfairness weight (i.e., at least some unfairness penalty during feature selection). For the Simulated data, 72% (52 out of 72) of the experiments yielded a significant decrease in unfairness relative to the previous (1 unit smaller) unfairness weight. Findings were similar for the Student Performance, Student Academics, and Adult datasets, which had significant decreases for 58%, 56%, and 82% of the experiments, respectively.

For each classifier, different unfairness metrics started off as more or less unfair at unfairness weight = 0. This was expected, since each metric measures a different aspect of unfairness that may be more or less present in a particular dataset. The relevance of particular metrics can be observed in the Adult dataset, shown in Figure 1. Unfairness was most evident in terms of Treatment Equality and Statistical Parity measures, but the unfairness was greatly reduced by our feature selection method without excessive decrease in accuracy. At unfairness weight 1 for logistic regression using Statistical Parity, unfairness decreased 60% (from 0.183 to 0.098) while accuracy only decreased 6% (from 0.759 to 0.716). This reduction in unfairness brought statistical parity to between 5% and 12% at unfairness weight 1 for all three models.

For most of the datasets, one or more metrics showed relatively low initial unfairness and maintained that regardless of unfairness weight. For example, overall accuracy equality had no statistical decrease, as measured by Mann-Whitney U tests, in unfairness for any classifier applied to the Student Academics dataset. We observed a similar resistance to the method across datasets and classifiers, though for different unfairness metrics (Figure 2). In general, unfairness metrics were not reduced by our feature selection method if unfairness was already relatively low according to that metric before applying the method (i.e., when unfairness weight = 0).

In general, the most fair and unfair metrics tended to be consistent across classifiers for a given dataset. For example, as can be seen in Figure 1, statistical parity showed substantial unfairness for the Adult dataset across all three classifiers. This is unsurprising, given that 30.38% of men, but only 10.93% of women, reported earnings of more than $50K in the dataset.

3.1.2 Unfairness vs. Accuracy Relationship. As Figure 2 shows, when the unfairness weight increased and unfairness decreased, so did accuracy. Accuracy decrease was expected, because as we constrained which features could be selected, there was less information for the model to learn from. Previous work in fair learning has demonstrated that some decrease in accuracy is unavoidable [17, 18, 22]. In cases where dimensionality is a large problem, reductions in accuracy may be alleviated because feature selection is essential [44], but our datasets do not have a particularly large number of features (e.g., more features than instances). Since we had a limited number of features to choose from—22 for Student Academics, 33 for Student Performance, 22 for Adult, and 3 for the Simulated Data—we anticipated a loss of accuracy when we constrained feature selection. We demonstrated the expected reduction in accuracy with our experimental results by showing that accuracy and unfairness were moderately correlated. For the 80% of metrics that were impacted by this approach, mean r = .545 with $p < .01$. As a demonstration of the correlation between accuracy and unfairness, we plotted three metric and classifier combinations

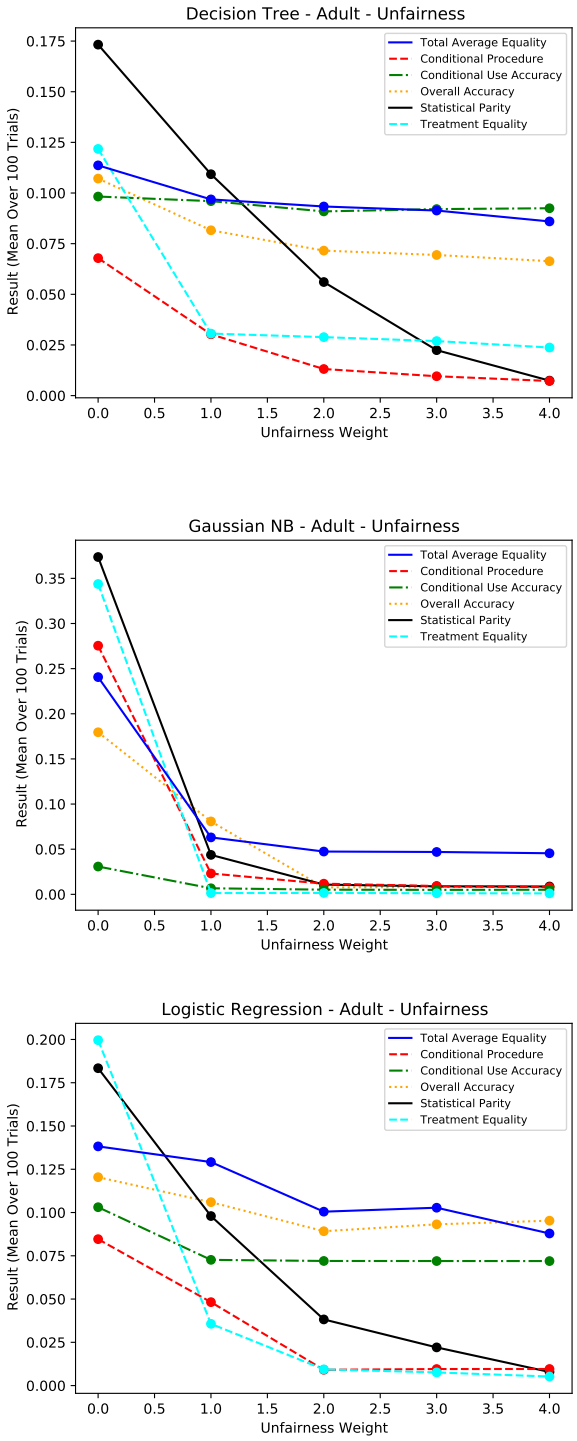


Figure 1: Three classifiers for the Adult dataset. These figures plot the mean of unfairness across 100 iterations using each unfairness metric.

for the Student Academics dataset, as seen in Figure 3. These combinations are a decision tree using conditional procedure accuracy (Figure 3a), logistic regression using total average equality (Figure 3c), and Gaussian naïve Bayes using treatment equality (Figure 3b).

For the remaining 20% of metrics, mean r = -.081. The set of metrics that did not have any correlation between accuracy and unfairness were the same as the ones that were most resistant to change using our method. We hypothesize that is because they started with low unfairness. For example, for the Simulated data, statistical parity had weak or no correlation between accuracy and unfairness for all classifiers. But, both accuracy and unfairness were unaffected by the change in unfairness weight (Figure 2).

For the Student Academics and Simulated data at unfairness weight 1, all three classifiers had a "best" or "worst" metric (or both). That is to say, a metric existed that had both the highest unfairness and lowest accuracy or lowest unfairness and highest accuracy. These metrics were the same as the ones that were the least or most unfair with the baseline classifier (unfairness weight of 0). In Figure 2, we can see that for logistic regression on the Simulated data, the best metric was statistical parity, while the worst was treatment equality. For decision trees on the Student Academics data, the best metric was overall accuracy equality, while the worst was treatment equality.

As unfairness weight was increased to 2 or more, however, the trend of a best or worst metric broke. At unfairness weights 3 and 4, some unfairness metric results were mixed, exhibiting the least unfairness and the worst accuracy—for example in the treatment equality results for the Simulated data, as seen in Figure 2. This could be attributed to an over-optimization for unfairness as the unfairness weight was increased, sacrificing accuracy in the process. In comparison, some metrics appeared to achieve results that balanced unfairness and accuracy at weights 2 and 3. At these weights some metrics were able to attain very similar accuracy to weight 0, with significantly less unfairness. For example, this decrease in unfairness occurred with logistic regression and the statistical parity metric for the Simulated data (Figure 2). Thus, results show that some amount of human intervention is still likely necessary to determine what unfairness weight is appropriate for a specific application, since the results will vary depending on the dataset and model.

We previously mentioned that some unfairness metrics remained unaffected by our feature selection approach for specific combinations of datasets and classifiers. The metrics that were resistant to change were resistant in terms of both accuracy and unfairness. Generally, we saw that the unfairness metric that changed the least was frequently the least unfair at the beginning. Moreover, when the method failed to reduce unfairness it also did not reduce accuracy. Thus, there was no cost to trying this approach. For example, conditional use accuracy as well as statistical parity for both logistic regression and Gaussian naïve Bayes were unaffected by changing the unfairness weight for the Simulated data (Figure 2). Accuracy was also unaffected, and there was no correlation for statistical parity between accuracy and unfairness. These specific metrics were also the least unfair from the beginning, ranking second and first, respectively, for the Simulated data. In other words, when a model was already fair and accurate, applying our method did not produce worse results.

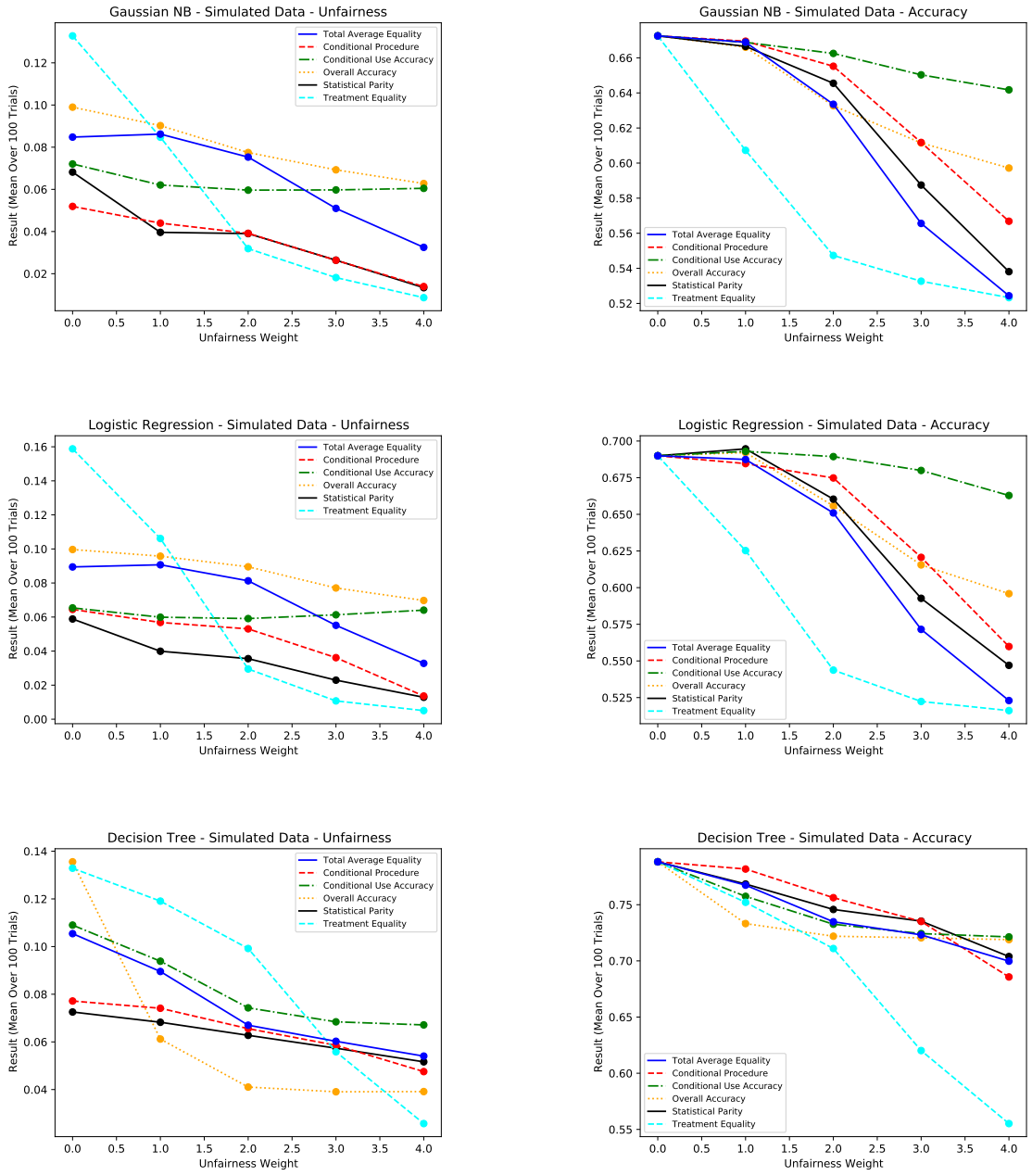
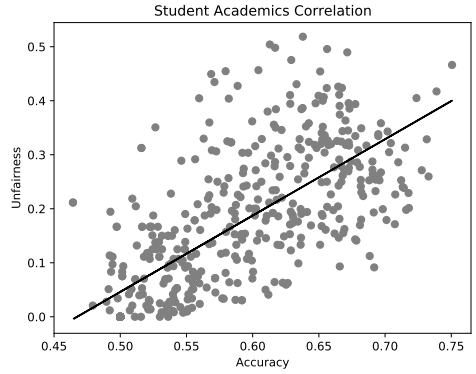


Figure 2: All classifiers for the Simulated data. These figures plot the mean of unfairness or accuracy (AUC) across 100 iterations using each unfairness metric.

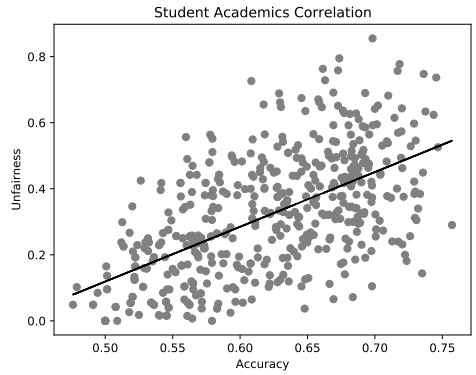
3.2 Inclusion of Sensitive and Unfair Features (RQ2)

RQ2 asked how the selection of the unfair feature and the sensitive feature (group status) affected outcomes. Using Mann-Whitney U tests, we found that both the unfair feature and the sensitive feature were selected significantly less frequently as the unfairness weight increased, for the Simulated data. We exclusively analyzed the Simulated data for this research question, since we knew exactly which features were fair and unfair, and in what way the unfair feature was unfair.

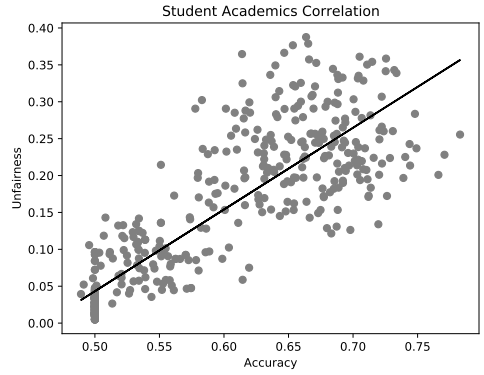
We measured whether the sensitive feature was included in the classifier less frequently for each increase (by 1) of the unfairness weight. For the sensitive feature, 70% (51 out of 72) of unfairness



(a) Decision Tree, Conditional Procedure Equality



(b) Gaussian NB, Treatment Equality



(c) Logistic Regression, Total Average Equality

Figure 3: Plots of correlation between accuracy and unfairness for three combinations of metrics and classifiers for the Student Academics dataset.

weight increases resulted in the sensitive feature being selected significantly less frequently ($p < .05$). This result was across all combinations of classifiers and metrics. The unfairness metrics that did not exhibit a statistically significant decrease in selection of the sensitive feature aligned with the ones that were also unaffected by the method in general. Worth noting, however, is that there were no metrics for which every single step was ineffective. For example, for Gaussian naïve Bayes, optimizing for statistical parity did not significantly affect the selection of the sensitive feature from unfairness weights 2 to 3 and 3 to 4, but did significantly change from steps 0 to 1 and 1 to 2. For metrics that start with low unfairness, increases in the unfairness weight may not effect change beyond a certain point. In our results, weighting unfairness at all (i.e., weight > 0) affected the selection of the sensitive feature. Beyond that, however, we may have already made whatever intervention was available to be made.

For the unfair feature, 64% (46 out of 72) of increases by 1 to the unfairness weight resulted in the unfair feature being selected significantly less frequently ($p < .05$). Similarly to the sensitive feature, selection of the unfair feature was less affected by our feature selection method for unfairness metrics that did not decrease significantly when the unfairness weight increased. In fact, there was one instance where no effect was observed; statistical parity with Gaussian naïve Bayes did not select the unfair feature with any less frequency at any unfairness weight. Statistical parity was, however, the metric that started out with low unfairness compared to other metrics, so this finding aligned with our other findings regarding resistant metrics.

4 LIMITATIONS AND FUTURE WORK

The experiments in this paper have a few limitations. First, two of the publicly available datasets we analyzed were similar education-related datasets, though education is not the only field to which our method applies. We chose these datasets as examples because education has a long history of looking at unfairness and bias (e.g., test unfairness), and many of our current formal definitions in machine learning echo those defined in educational contexts [25]. The Adult dataset provided a baseline to other machine learning work, but future work should explore our method in additional fairness-sensitive domains, such as healthcare.

Second, our datasets were relatively small; most notably, the Simulated data had only three features. It was intentionally constructed in this way to make the relationship between the selection of the unfair feature and the unfairness weight as clear as possible—however, in the future we will evaluate this approach on datasets with more features and thus a larger search space for feature selection.

Third, the method we described is based on wrapper feature selection, which integrated well with our approach and is appropriate for smaller datasets. Wrapper feature selection, however, is slow compared with other feature selection methods because of its computational complexity [40]. Furthermore, wrapper feature selection is not the only way to select features. As such, future work could apply our method to other feature selection methods, such as RELIEF-F [32] or model-based methods [1].

5 CONCLUSION

We devised an approach to fair feature selection, inspired by the general framework of fair model building. In particular, we introduced the notion of an unfairness weight in feature selection, which indicates how heavily to weight unfairness versus accuracy when measuring the marginal benefit of adding a new feature to a model.

Our goal was to maintain accuracy while reducing unfairness, as defined by six common statistical definitions.

We demonstrated that our method decreased unfairness. Moreover, our method does not reduce accuracy when fairness will not be improved. There is, however, an inevitable trade-off between improved fairness and accuracy, which we showed by measuring the correlation of accuracy and unfairness. By automating the selection of features to produce fairer predictions, we provide an approach that improves model fairness by eliminating decisions based on unfair information. Because this approach affects feature selection only, it can also be combined with other common fairness methods, such as pre-processing. As long as the desired definition of unfairness is known, it can be applied to this approach. Moreover, our method could easily be extended to other statistical unfairness definitions beyond the six that we tested. This procedure is an effective approach to constructing classifiers that both reduce unfairness and are less likely to include unfair features in the modeling process.

REFERENCES

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 10 (May 2010), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

[2] Ryan S. J. d Baker and Kalina Yacef. 2009. The state of educational data mining in 2009: a review and future visions. *JEDM | Journal of Educational Data Mining* 1, 1 (Oct. 2009), 3–17. https://doi.org/10.5281/zenodo.3554657

[3] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 671 (2016).

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in criminal justice risk assessments: the state of the art. *arXiv:1703.09207* (May 2017).

[5] danah boyd and Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication & Society* 15, 5 (June 2012), 662–679. https://doi.org/10.1080/1369118X.2012.678878

[6] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[7] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees.* CRC Press.

[8] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (Sept. 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[9] 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.

[10] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (July 2019), 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

[11] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (June 2017), 153–163. https://doi.org/10.1089/big.2016.0047

[12] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960. *Educational and Psychological Measurement* 20, 1 (April 1960), 37–46. https://doi.org/10.1177/001316446002000104

[13] Irina Cojuharenco and David Patient. 2013. Workplace fairness versus unfairness: Examining the differential salience of facets of organizational justice. *Journal of Occupational and Organizational Psychology* 86, 3 (2013), 371–393. https://doi.org/10.1111/joop.12023

[14] Paulo Cortez and Alice Silva. 2008. Using data mining to predict secondary school student performance. In *Proceedings of the 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008).* 5–12.

[15] 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 (ITCS '12).* Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[16] Christian Fischer, Zachary A. Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. 2020. Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education* 44, 1 (March 2020), 130–160. https://doi.org/10.3102/0091732X20903304

[17] Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. 2016. A Confidence-Based Approach for Balancing Fairness and Accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining.* https://doi.org/10.1137/1.9781611974348.17

[18] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* Association for Computing Machinery, New York, NY, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[19] Jerome H. Friedman. 1997. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery* 1, 1 (March 1997), 55–77. https://doi.org/10.1023/A:1009778005914

[20] Pratik Gajane and Mykola Pechenizkiy. 2018. On Formalizing Fairness in Prediction with Machine Learning. *arXiv:1710.03184* (May 2018). http://arxiv.org/abs/1710.03184 arXiv: 1710.03184.

[21] 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 (LAK19).* Association for Computing Machinery, New York, NY, USA, 225–234. https://doi.org/10.1145/3303772.3303791

[22] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

[23] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems.* 3323–3331. https://doi.org/10.5555/3157382.3157469

[24] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Ba-Alwib, and Ribata Najoua. 2018. Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science* 9, 2 (Feb. 2018), 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

[25] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600

[26] László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.* 245–251. https://doi.org/10.1109/ACII.2013.47

[27] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3351095.3372829

[28] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[29] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science, Vol. 7524)*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer, Berlin, Heidelberg, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

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

[31] Ron Kohavi. 1996. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* AAAI Press, 202–207.

[32] Igor Kononenko, Edvard Šimec, and Marko Robnik-Šikonja. 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence* 7, 1 (Jan. 1997), 39–55. https://doi.org/10.1023/A:1008280620621

[33] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. 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., 4066–4076.

[34] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2019. Delayed impact of fair machine learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.* International Joint Conferences on Artificial Intelligence Organization, Macao, China, 6196–6200. https://doi.org/10.24963/ijcai.2019/862

[35] Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan. 2014. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology* 45, 3 (May 2014), 487–501. https://doi.org/10.1111/bjet.12156

[36] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy* (1 ed.). Crown Publishers, New York, NY.

[37] Jason W. Osborne. 2001. Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology* 26, 3 (July 2001), 291–310. https://doi.org/doi:10.1006/ceps.2000.1052

[38] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830.

[39] Sebastian Raschka. 2018. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software* 3, 24 (April 2018), 638. https://doi.org/10.21105/joss.00638

[40] Payam Refaeilzadeh. 2007. On comparison of feature selection algorithms. In *Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07) Workshop Program*. 6.

[41] Debopam Sanyal, Nigel Bosch, and Luc Paquette. 2020. Feature selection metrics: Similarities, differences, and characteristics of the selected models. In *13th International Conference on Educational Data Mining (EDM 2020)*. 212–223.

[42] Robert L. Schaefer. 1986. Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation* 25, 1-2 (Aug. 1986), 75–91. https://doi.org/10.1080/00949658608810925

[43] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[44] G. V. Trunk. 1979. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 3 (July 1979), 306–307. https://doi.org/10.1109/TPAMI.1979.4766926 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[45] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*. PMLR, 6373–6382.

[46] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web* (April 2017), 1171–1180. https://doi.org/10.1145/3038912.3052660