That’s Not Fair:
Identifying and Reducing Bias in Machine Learning

Carleton College

Evan Allgood, Cayden Ehrlich, Steph Herbers,
Malcolm Mitchell, Madeline Prins, and Irene Sakson

Abstract

Machine learning algorithms are proliferating in many industries with the digitalization of numerous processes. One category of these algorithms is the classification algorithm, which predicts outcomes based on the attributes of individuals within a dataset for tasks such as bank loan requests and job applications. These algorithms have been used in applied settings with the aim of removing human biases and expediting classification processes. However, classification algorithms may also pick up the biases of the datasets on which they are trained. These biases can be especially problematic when they are based on sensitive attributes such as gender, age, income, and education level. In this paper, we discuss our implementation of several algorithms that reduce bias in classification problems, how to quantify bias and fairness, and what measures we can use to evaluate the fairness of these algorithms. We implemented the Feldman et al. (2015) Disparate Impact Detector, the Feldman et al. (2015) Repair, the Naive Bayes Classifier, the Calders and Verwer (2010) Modified Bayes Classifier, and the Calders and Verwer (2010) Two Bayes Classifier. Then, we analyze how well the algorithms improve the fairness of classifications for seven datasets on which we ran the algorithms. We find that the success of these algorithms is somewhat dependent on the original degree of discrimination present and the peculiarities of the dataset that we are working with, but we see a trend of improving fairness.

Introduction

We are surrounded by news of technology making decisions that are normally made by humans. Machine learning algorithms have the potential to reduce human bias by standardizing the way these decisions are made—but are algorithms always less biased than humans?

A number of studies have indicated that these algorithms may be more biased than people expect and these biases can have large and small impacts. For instance, ads for higher paying jobs on sites like LinkedIn are shown more frequently to men than women, which can affect women's ability to find and attain these jobs (Datta et al., 2015). On a smaller scale, Pokémon Go had fewer Pokéstops in predominantly Hispanic and Black neighborhoods, limiting residents' ability to enjoy the game (Juhász & Hochmair, 2017). These facts are not consistent with the goal of using algorithms that are less biased than humans. Instead, the algorithms may still be making decisions that are more contingent on the identities of individuals rather than on more relevant information. Specifically, we are concerned with reducing bias based on sensitive attributes such as gender, race, and age, because they have no correlation to how someone will perform in targeted applications.

In this study, we focused on bias in classification algorithms. Classification algorithms train a model to divide individuals into groups based on their characteristics. Examples of classification problems include choosing a subset of job candidates to hire or predicting whether individuals will default on their loans. If a classifier is trained on a biased dataset, the classifier may replicate these biases in its classifications. For example, if in a given dataset more men than women are hired for a job, the classifier may learn from the data that being male is correlated with getting hired, and thus predict that more men than women in the dataset will be hired. Even if a person could not tell by looking at the data that the data is biased, the classifier might pick up on unintended discrimination in the data and return classifications that reflect this implicit bias.

In this paper, we will discuss related work that inspired our research. Then, we will discuss the five algorithms we implemented to identify and reduce bias in classification problems. One algorithm takes a pre-processing approach (modifying the data to be more fair before the data are provided to the algorithm), while three others take an in-processing approach (modifying the algorithm to better account for fairness as it solves the classification problem). The final algorithm determines whether or not a dataset contains disparate impact. Finally, we will evaluate the performance of these algorithms on seven datasets using eight metrics based on different definitions of fairness.

Related Work

To get a deeper understanding of the problem our project is trying to solve, we investigate the depth of effect of algorithmic bias in a handful of industries and break down the semantic complexity of defining “fairness.”

Depth of Effect

Biases in classification algorithms have been found in seemingly trivial cases to life threatening situations. For example, Engin Bozdag’s 2013 article *Bias in algorithmic filtering and personalization* describes algorithms that were originally designed to tailor themselves to individuals through “personalization.” Companies like Google and Facebook create user profiles based on attributes such as name, gender, education, and country. However, this filtering process can be manipulated by third parties or the companies themselves (Bozdag, 2013).

Additionally, Raub M.’s paper *Bots, Bias, and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability In Hiring Practices* (2018) analyzes the bias in the increasingly common software being used to screen job applicants. This software was originally designed to help remove bias from the screening process by relying on trained machine learning models. However, the conscious decisions that people make when utilizing this software affect the fairness of the machine learning process. For example, the choice of data input, whether intentional or not, can cause discriminatory results, which could become a learned pattern in the machine learning model. Thus, machine learning algorithms have the potential power to worsen social inequalities already plaguing minority groups (Raub, 2018).

Stephanie K. Glaberson’s *Coding Over The Cracks: Predictive Analytics and Child Protection* (2019), presents an example of algorithmic bias in the child welfare system. Algorithms generate risk scores for children potentially in danger of maltreatment based on attributes like income, parental drug abuse history, and parental criminal history. These scores are used to assess whether intervention with families is necessary. This predictive classifier's high-pressure task can allow great harm if it misses warning signs of abuse or neglect and wrongfully classifies a case as not needing intervention (Glaberson, 2019).

Definitions of Fairness

There are many different suggested frameworks for measuring fairness. Several papers, including Yao and Huang's *New Fairness Metrics for Recommendation that Embrace Differences* (2017), propose using four metrics for measuring fairness: *value unfairness*, *absolute unfairness*, *underestimation unfairness*, and *overestimation fairness*. These fairness metrics are based less on human understanding of fairness and more on mathematical fairness. Yao and Huang's article also offers metric baselines, including simply removing "sensitive attributes" like gender or race, or enforcing demographic parity, which gives equal weight to people of all races or genders. Although this article does not provide one concise definition of fairness, it gives a handful of options of ways to measure different varieties of fairness (Yao & Huang, 2017).

Disparate Impact

Feldman et al.'s article *Certifying and removing disparate impact* (2015) proposes a pre-processing approach to improve the fairness of generic categorization algorithms. The Feldman et al. process aims to minimize "disparate impact" (2015), using two algorithms: one to detect disparate impact, and one to reduce it.

Disparate Impact

Disparate impact is the legal theory used to determine *unintended* discrimination. It occurs when a selection process appears neutral, but in reality has vastly different outcomes for different groups (Feldman et al., 2015). For a dataset with a particular sensitive attribute, the Disparate Impact Detector determines whether or not the sensitive attribute's value can be predicted from the other attributes. If it is consistently easy to determine what a particular row's sensitive attribute would be, then the dataset has disparate impact. Quantifying the disparate impact on the dataset will tell us whether or not the model we are building from that data is fair.

Feldman et al. Disparate Impact Detector

To determine whether or not a dataset has disparate impact, a classifier is run on all the columns of the dataset except the sensitive attribute column. We call its output $f(Y)$. Then, the algorithm calculates the balanced error rate (BER) using $f(Y)$, which is defined as the unweighted average class-conditioned error of f. The BER describes the likelihood of recovering the sensitive attribute values from the non-sensitive attributes.

$$BER((f(Y), X) = (P[f(Y) = 0|X = 1] + P[f(Y) = 1|X = 0]) / 2$$

The BER is optimized at a value of ½, which means that a predictor has a 50/50 chance of correctly determining which sensitive attribute is related to a particular Y value. A *high* BER indicates the sensitive attribute was difficult to predict based on the other attributes of the individual. In order to find the allowable BER threshold, the algorithm builds a confusion matrix such that one diagonal indicates a correct $f(Y)$ prediction and the other indicates an incorrect $f(Y)$ prediction. If the BER is above this threshold, the algorithm returns "no disparate impact," and if the BER is below this threshold, the algorithm returns "possible disparate impact." In the latter case, if we were to remove the sensitive attribute from the dataset, the classifier could predict the sensitive attribute value with high probability, and this allows for the possibility of disparate impact.

Mitigating Bias

Feldman et al. Repair

The Feldman et al. (2015) preprocessing algorithm provides a way to "repair" the dataset with respect to a particular sensitive attribute X. Repairing the dataset means that the algorithm aims to mitigate bias by altering values in the data itself. The algorithm iterates through the numerical columns in the data and repairs each one. For a given column, the algorithm finds the distributions (F_x) of values (Y) for different values of sensitive attributes (x) and creates a median distribution. The Repair preserves rank, which means that people in a certain percentile for Y in their sensitive attribute distribution F_x will end up in that same percentile in the median distribution.

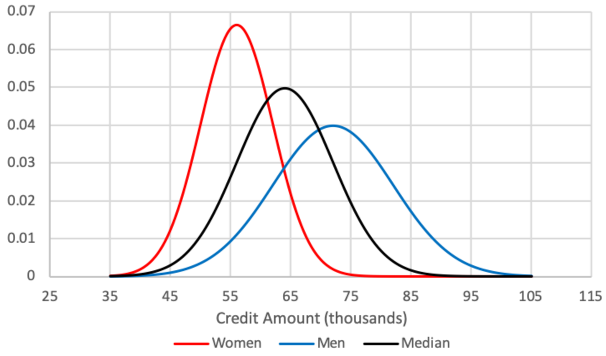


Figure 1: An example median distribution for the credit amount column of a dataset with sex as sensitive attribute.

To find the median distribution, the algorithm will identify the percentile for each person in that distribution. However, if a dataset has few people in one sensitive attribute category *x*, then the algorithm may not be as specific about the percentiles for people in F_x. To solve this, the Feldman et al. Repair algorithm (2015) divides each of the distributions (F_x) into groups called "buckets," based on the size of the smallest sensitive attribute group. If there are *n* members of the smallest sensitive attribute category, then the algorithm will divide each of the distributions F_x into *n* buckets and compute the medians across those *n* buckets (as opposed to the medians across *n* = 100 percentiles). The algorithm uses quantile bucketing, where each bucket has the same number of people to ensure that there are no empty buckets. The algorithm compares the bucket medians across sensitive attribute distributions. These values form the median distribution for this column of the data. Finally, the algorithm iterates through the rows of the data, determines which bucket the row's Y value is in (say, bucket *b*) for its sensitive attribute distribution F_x, and replaces the original Y with the median distribution's value for bucket *b*. The Feldman et al. Repair algorithm can repair any numerical attribute where the numbers correspond to a scale (for instance, test scores but not id numbers), so we ran the Repair on all columns in our datasets that met these two conditions.

Dividing Data

When running the following classification algorithms, we must avoid classifying the same data on which we have trained our models. Doing so would make it difficult to assess the effectiveness of our algorithms, because the model may pick up on eccentricities in the particular data we train it on, and thus be more highly successful in classifying those data than it would for any data drawn from the same distribution. To remedy this issue, we split the data into a training set and a test set. Eighty percent of the rows are randomly selected and added to the training set, and the other twenty percent of the rows are added to the test set. We train our classification models using the train set and then we classify the test set with the original classification column removed. This process allows us to have more confidence in the performance of our models.

Naive Bayes

Naive Bayes, one of the most common classification algorithms, makes what is referred to as the "independence assumption." It assumes that all attributes of an individual (e.g. gender or employment status) contribute equally and independently to their classification. In other words, the classifier ignores any possible correlation between the attributes of an individual, as depicted in Figure 2.

The algorithm itself is broken up into two sections: train and classify. To train, Naive Bayes populates a model with conditional probabilities of the relationship between every value of every attribute and the two possible classifications (e.g. $P(A_n|C_+)$; the probability an individual has attribute value A_n given they are positively classified). It attains these conditional probabilities by averaging the classifications of every individual in the training set of the data.

To classify, Naive Bayes iterates through every row of the test set and computes two overall probabilities, one for each possible classification, and one for each individual. To do so, it retrieves the conditional probability of each attribute for the classification and the probability of that classification from the model. Then, for each of the two classifications, the algorithm multiplies all of these probabilities together. Finally, the algorithm classifies the individual with whichever of these two classifications results in a higher overall probability.

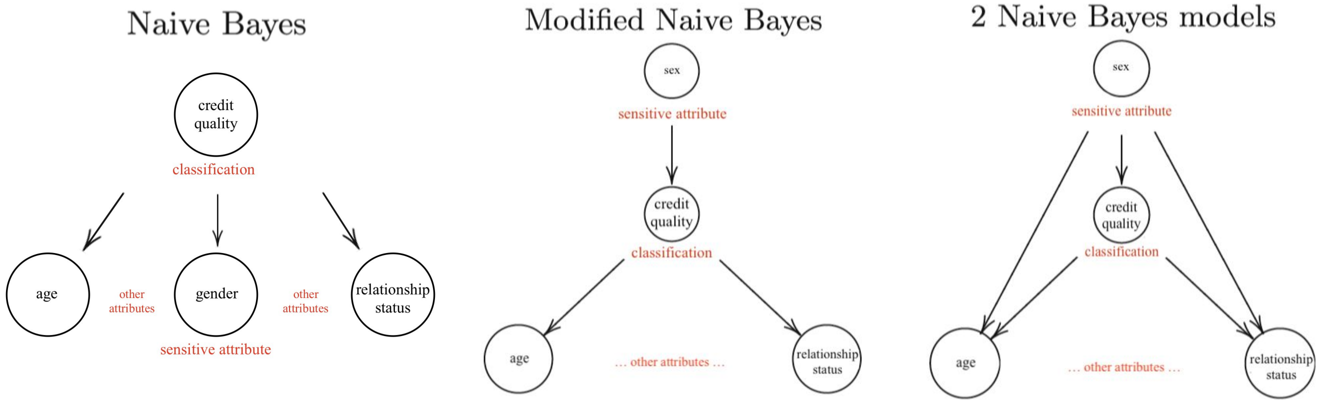


Figure 2: Diagrams of Naive Bayes, Modified Naive Bayes, and Two Bayes displaying the hierarchical differences between the Bayesian algorithms (Calders & Verwer, 2010; modified for paper).

Modified Bayes

Because the independence assumption can cause Naive Bayes to replicate the biases of a dataset, Calders and Verwer (2010) propose a modification to the algorithm. Modified Bayes follows a similar structure to Naive Bayes but additionally takes in the name of a sensitive attribute. The algorithm modifies the conditional probabilities associated with that attribute such that the two sensitive groups have similar likelihoods of being classified positively. A model trained by Modified Bayes will store the conditional probabilities of a sensitive attribute differently than a model trained by Naive Bayes. Rather than classifying based on the probability $P(C) \times P(S|C)$, we use $P(S) \times P(C|S)$. In doing so, we allow later modification of the impact of the sensitive attribute on the classification of an individual. For all other attributes, however, the conditional probabilities will be stored in the same manner.

The Modified Bayes algorithm first trains a model on the training set, just as in Naive Bayes. We classify the training set based on that model and calculate a discrimination score from these classifications. The discrimination score is defined as $P(C^+|S^+) - P(C^+|S^-)$, which essentially finds the difference between the probability that a privileged individual will be positively classified and the probability that an underprivileged individual will be positively classified. Then, as long as this score is greater than 0 (i.e. the privileged sensitive group has an advantage over the underprivileged group) we repeatedly modify the conditional probabilities of the sensitive attributes and reclassify the training set.

To start, we compute the counts for the four possible combinations of classifications (C) and sensitive attributes (S), which we refer to as C^+S^+, C^+S^-, C^-S^+, C^-S^-, where $+$ is a positive classification and — is a negative classification. We check how the number of positive classifications we have assigned compares to the number of actual positive classifications from the ground truth column, which results in two options. In the first case, the number of positive classifications we have assigned is less than the actual total number of positive classifications, which means that we may assign more positive classifications. Therefore, we slightly increase the number of positive classifications for the underprivileged group and slightly decrease the number of negative classifications for the underprivileged group. We do this by computing two new, artificial counts, defined by:

$$C^{+}S^{-}Count = C^{+}S^{-}Count + (weight \times C^{-}S^{+}Count)$$

$$C^{-}S^{-}Count = C^{-}S^{-}Count - (weight \times C^{-}S^{+}Count)$$

The weight of change refers to the degree by which we are changing the counts with each iteration of the Modified Bayes loop. A smaller weight of change allows us to get closer to perfect fairness with a discrimination score of exactly 0.0 but can perform very slowly due to the increased number of iterations needed. We used a weight of change of .01. We update the probabilities of the two values whose counts we have just artificially modified.

In the second case, the number of positive classifications we have assigned is greater than the actual number of positive classifications, meaning that we have given out too many positive classifications. We therefore will slightly increase the number of negative classifications given to the privileged group and decrease the number of positive classifications given to the privileged. We use the same weight of change value and a similar equation:

$$C^{-}S^{+}Count = C^{-}S^{+}Count + (weight \times C^{+}S^{-}Count)$$

$$C^{+}S^{+}Count = C^{+}S^{+}Count - (weight \times C^{+}S^{-}Count)$$

We update the probabilities the same way we do in the other case. Then we simply reclassify the train set, recomputing the classification for every instance, incorporating the newly updated probabilities into the product sum involved in the Bayesian classification, and recompute the discrimination score to see if we have to repeat the process. The structure of the modification drastically changes the number of positive classifications assigned.

Two Naive Bayes

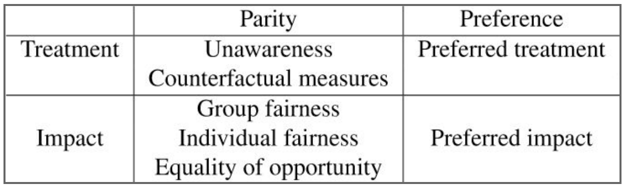
The second algorithm modification proposed by Calders and Verwer (2010), called Two Bayes, eliminates the correlation between the classification and the sensitive variable altogether. The simplest way to remove this correlation would be to train an algorithm on all attributes except the sensitive attribute. However, doing so results in a large loss in accuracy due to fewer attributes. Moreover, it is possible for a classification algorithm to be biased against a certain group even without training on the sensitive attribute. This phenomenon, called the red-lining effect, occurs because the distribution of the attribute may depend on sensitive attribute values.

Two Bayes does not remove the sensitive attribute itself, instead removing its ability to be used to classify an individual. First, we split the train set into two disjoint sets, one for each sensitive group. Next, we train a model on each of the two datasets in the same way we would in Naive Bayes. We will then have one model trained only on the privileged group and one trained only on the underprivileged group. Finally, we modify the two algorithms using the Modified Bayes algorithm. Since we have two models, the process by which we modify the probabilities is slightly different. We use $P(C)$ in each model in place of $P(C|S)$. This is because the classification probabilities are naturally conditional as the models are trained on the sensitive group-specific dataset. As in Modified Bayes, we iteratively tweak the probabilities of the two models until the difference between classifications satisfies our discrimination score requirement.

To classify our test set with the Two Bayes algorithm, we iterate row by row. We classify each individual using the model that matches their sensitive attribute value. For instance, if the individual is of the privileged sensitive group we use the model trained on the privileged set and vice versa. Classified with Two Bayes, we remove the correlation between the sensitive attribute and other attributes without sacrificing accuracy.

Evaluation Metrics

In order to evaluate how well these algorithmic interventions are performing, we decided upon five fairness metrics, and three accuracy metrics. We based our definitions of fairness on Gajane & Pechenizkiy's theoretical paper *On formalizing fairness in prediction with machine learning* (2017). The authors created a framework for defining their conceptions of fairness through intersections of parity or preference, and treatment or impact, as shown in Table 1. Gajane and Pechenizkiy (2017) define the question of parity or preference as "whether fairness means achieving parity or satisfying the preferences," and the question of treatment or impact as "whether fairness is to be maintained in treatment or impact (results)." In other words, we are interested in equality or equity, and during which steps of the process we should enforce it.



	Parity	Preference
Treatment	Unawareness Counterfactual measures	Preferred treatment
Impact	Group fairness Individual fairness Equality of opportunity	Preferred impact

Table 1: The categorization of different metrics along the axes of treatment vs. impact and parity vs preference (Gajane & Pechenizkiy, 2017).

We implemented five of Gajane and Pechenizkiy's (2017) definitions as fairness metrics: *counterfactual measures*, *preferred treatment*, *group fairness*, *individual fairness*, and *equality of opportunity*. We decided to omit *fairness through unawareness* since, as mentioned in the Disparate Impact Detector section, just because a dataset does not explicitly contain a particular sensitive attribute, that does not mean that a classifier is not capable of determining what the sensitive attribute was from other information in the dataset. Therefore, checking how the outcomes change when simply removing the sensitive attribute from the dataset did not feel to be effective for the goal of this project. We also omit *preferred impact* because this requires comparisons between models that can interact with data both as aggregated data and as data that has been split up by sensitive attribute, which this project does not have. Finally, we implemented a basic accuracy metric to investigate if our accuracy trends correlate with our fairness trends.

Counterfactual Measures

Counterfactual measures is a parity of treatment metric, in that its goal is to show whether or not a classifier is treating every individual in a dataset the same. It does this by looking at the original output after passing data through a classifier, and then creating a counterfactual dataset in which all data is the same except the sensitive attribute is swapped to its counterfactual label (for example, if “Male” and “Female” were our attribute labels, all “Male” would become “Female” and vice versa), and comparing their outcomes. Gajane and Pechenizkiy (2017) define this formally, as seen in Table 2. This says, in essence, that the probability that the outcome of a predictor $\mathcal{H}$ for a sensitive attribute label a in A should equal the probability of the same outcome of the same predictor $\mathcal{H}$ for the other sensitive attribute label a' in A, where Z is the set of remaining attributes.

We compare the two different sets of outcomes by treating one as the true classifications and the other as our model’s classification output. We then calculate the accuracy, which means, the higher the output of our counterfactual measures metric, the less impact that the sensitive attributes have on our classification process, and the fairer our classifier is.

Group Fairness

Group fairness is a parity of impact metric. Under its definition of fairness, a classifier is fair if all groups, regardless of sensitive attribute label, have the same probability of getting a positive outcome. This is a method which is independent of the ground truth— rather than looking at whether or not they’re getting true positive outcomes, it’s just investigating generic positive outcomes from the classifier. The output of this method is a set of two proportions of positive outcomes, one for each sensitive attribute group, and we want these proportions to be relatively equal.

Individual Fairness

Individual fairness is also a parity of impact metric in a very similar way to group fairness. Where group fairness seeks parity amongst the positive outcomes of groups, individual fairness wants similar people to be impacted similarly— that is, similar individuals should have similar outcomes.

For our project, we found similar individuals by using the Euclidean distance and seeing how far apart each individual was from each other individual. For categorical variables, we dummified the values (see Disparate Impact Detector), which allowed us to also use categorical information in our determination of similarity. We then z-scored all values to put everything on the same scale. All pairs of individuals who were in the bottom 10% of the distribution— that is, the closest 10% of pairs— were then considered "similar." The output of this metric was the proportion of these similar individuals who receive the same outcome. We want this proportion to be fairly high.

Equality of Opportunity

Equality of opportunity is a parity of impact metric, as well, in that we want people to have similar chances of getting a true positive outcome. This means that everyone, ideally, should have an equal chance of getting a positive outcome in the original dataset, and that the classifier is classifying people well.

For our project, the output of this metric is the p-value of whether or not people have similar chances of getting a true positive outcome in a statistically significant sense. In order to calculate this, we perform a chi-square test on the true positive and true negative counts. We use a chi-squared test because we wanted to find a way to standardize what "similar chances" meant. If they are similar enough, we will receive a small p-value. If the p-value is small enough, that means that a model has Equality of Opportunity. This metric, therefore, outputs a p-value and a Boolean value, and we want small values, and True as our output if our model is fair.

Preferred Treatment

Preferred treatment is a preference of treatment metric, which means our intent is to treat different groups differently within a model. This is the one metric that inherently compares submodels within a model, which means that we can only use this metric with our Two Bayes models— every other model always receives an output of True. A group-conditional model is said to satisfy preferred treatment if each group receives more benefit from their respective model than they would have received from any other predictor. Given that Two Bayes makes one model for each sensitive attribute, the hope is that the model tailored towards a particular sensitive attribute group would give that sensitive attribute group more benefit. The output of this metric is a Boolean that states whether or not the two submodels of Two Bayes give their specified sensitive attribute group the most benefit.

Metric	A model $\mathcal{H} : X \to Y$ is considered fair…
Counterfactual Measures	… given $Z = z$ and $A = a$, for all y and $a \neq a'$, iff $P\{\mathcal{H}_{A=a} = y \mid Z = z, A = a\} = P\{\mathcal{H}_{A=a'} = y \mid Z = z, A = a\}$
Group Fairness	… with bias ϵ with respect to groups $S, T \subseteq X$ and any subset of outcomes $O \subseteq A$ iff $\lvert P\{\mathcal{H}(x_i) \in O \mid x_i \in S\} - P\{\mathcal{H}(x_j) \in O \mid x_j \in T\} \leq \epsilon$

Individual Fairness	… iff $\mathcal{H}(x_i) \approx \mathcal{H}(x_j) \vert d(x_i, x_j) \approx 0$ where $d : X \times X \rightarrow R$ is a distance metric for individuals
Equality of Opportunity	… with respect to group S iff $P\{\mathcal{H}(x_i) = 1 \vert y_i = 1, x_i \in S\}$ $= P\{\mathcal{H}(x_j) = 1 \vert y_j = 1, x_j \in X \backslash S\}$
Preferred Treatment	… if $B_S(\mathcal{H}_S) \geq B_S(\mathcal{H}_T)$ for all $S, T \subset X$

Table 2: A table containing formal definitions of fairness metrics (Gajane & Pechenizkiy, 2017).

Accuracy

We first measured the accuracy of our outputs. While our main goal is to achieve fairness, accuracy is also a crucial aspect of categorization problems. Thus, we measured the true positive (TP) rate and the true negative (TN) rate, which also allows us to find the false positive (FP) and false negative (FN) rates easily, if we need them. The TP rate is the percentage of cases who are correctly given a positive output, “correctly” being as defined by the original dataset’s classification. The FN rate is the percentage of cases that are correctly given a negative output. Our accuracy metric, then, is $\frac{TP+TN}{TP+FP+TN+FN}$.

Data Sets

We chose seven different datasets to run our algorithms on. Of those seven, three were also used in the Friedler et al. paper (2019). For all of the datasets, we removed any rows that had empty values in them. Table 3 contains information about each dataset’s size, classification, and sensitive attribute. Table 4 contains a breakdown of the outcomes by sensitive attribute for each dataset.

Dataset	**Sample Size**	**Number of Attributes**	**Classification**	**Sensitive Attribute**
Ricci	118	6	Promotion or no promotion	Race
Portuguese	649	33	Passed or failed class	Sex
German	1,000	22	Good or bad credit quality	Sex
Jury	1,517	6	Not struck or struck	Race
Restaurants	22,168	14	Passed or failed inspection	Borough
Credit	29,655	24	Defaulted or did not default on credit card	Education level

Income	30,162	15	Income >= $50k or <$50k	Education level

Table 3: Information about all of the datasets we used in our project.

The first and smallest dataset we used was the Ricci dataset (which we refer to as "Ricci"). The Ricci dataset was used by Friedler et al. (2019). It comes from the *Ricci v. DeStafano* court case (2009), where a group of White and Hispanic firefighters sued the city of New Haven, Connecticut after they did not receive promotions despite passing the exam required for a promotion. The city invalidated the test results because no Black candidates were going to be considered for promotions based on the test scores. The Supreme Court ruled in favor of the prosecutors (*Ricci v. DeStefano*, 2009). Since our algorithms could only handle binary sensitive attributes, we edited the race column so that instead of having values of White, Black, and Hispanic, it only had White and Person of Color (POC) (Miao, 2010).

The second dataset we used was the student performance dataset (which we refer to as "Portuguese"). The dataset contains information about the demographics of different students in two Portuguese schools as well as their grades in the subject of Portuguese on a 20 point scale. Using a grade conversion chart (Dale, 2011), we made the grade column binary by counting any scores of 12 or higher as passing and 11 or lower as failing (Dua & Graff, 2017).

The German credit dataset (which we refer to as "German") is also used by Friedler et al. (2019). It contains information about the credit quality of German people, where credit quality is calculated by a cost matrix. The original dataset had sex and relationship status as a single column; in order to use sex as the sensitive attribute, we separated them into two columns (Dua & Graff, 2017).

The next largest dataset we used was the *Flowers v. Mississippi* dataset (which we refer to as "Jury"). This dataset is from the *Flowers v. Mississippi* (2019) court case, where Curtis Flowers, a Black man, was tried six times for the murder of four people. The prosecution continuously struck all Black prospective jurors, leading to multiple hung juries and mistrials. The Supreme Court ruled that this case falls under the precedent set by *Batson v. Kentucky*, and that the Mississippi court was wrong to affirm Curtis Flowers' death sentence (*Flowers v. Mississippi*, 2019). The dataset contains information about all of the prospective jurors in the different trials of Flowers' case (Craft et al., 2018).

The third largest dataset we used was a dataset of New York City restaurants (which we refer to as "Restaurants"). This dataset contains information on restaurants in New York City such as their name, location, and type of cuisine. We chose to use the borough of the restaurant as the sensitive attribute because of the potential correlation between the wealth of the borough and the grade the restaurants were given. We made this attribute binary by dividing boroughs into the categories "less wealthy" (Queens, Brooklyn, and The Bronx) and "more wealthy" (Manhattan and Staten Island) using census data (Bureau, n.d.). To make the classification binary, we changed the score column to be pass or fail so that a score of 14 or less was a pass and a score of more than 14 was a fail, as more points means that the restaurant has more violations. We chose 14 points based on the scale provided by the New York Department of Health and Mental Hygiene. We also removed columns such as the phone number and address of the restaurants because these are not columns that should impact whether or not a restaurant passes or fails, but not in the same way that a sensitive attribute shouldn't impact categorization, especially since each value is unique. Some restaurants appeared multiple times in the dataset because they had been inspected multiple times. In these cases, we kept the most recent inspection results (by date) for each restaurant and deleted previous inspections (Health & Hygiene, 2020).

The next dataset we used was a credit card clients dataset from Taiwan (which we refer to as “Credit”). The sensitive attribute we used for this dataset was education level, which we split into “high school or below” and “higher education” (Dua & Graff, 2017).

Our final and largest dataset was the adult income dataset (which we refer to as “Income”). This dataset contained information from the 1994 Census database. We chose to use education level as the sensitive attribute, which we once again made binary by splitting it into “high school or below” and “higher education” (Dua & Graff, 2017).

Dataset	Sensitive Attribute	Sensitive Attribute Values	Percent with positive outcome	Percent with negative outcome
Ricci	Race	White	34.7%	22.9%
		POC	12.7%	29.7%
Portuguese	Sex	Female	35.0%	24.0%
		Male	18.6%	22.3%
German	Sex	Female	20.0%	11.0%
		Male	50.6%	18.9%
Jury	Race	White	56.6%	8.1%
		Black	14.8%	20.4%
Restaurants	Borough	Less Wealthy	52.1%	4.5%
		More Wealthy	40.4%	3.0 %
Credit	Education level	High School or Below	4.2%	12.8%
		Higher Education	18.1%	54.9%
Income	Education Level	High School or Below	6.1%	38.9%
		Higher Education	18.8%	36.2%

Table 4: A breakdown of the outcomes by sensitive attribute in each of the original datasets.

Results

Disparate Impact

After running the Feldman et al. Disparate Impact Detector on all of the datasets, it returned "possible disparate impact" for all datasets except Jury and German datasets. The Jury felt counterintuitive because with columns like "race" and "same race," we had predicted it would conclude possible disparate since race was the sensitive attribute. However, disparate impact tries to describe discrimination that is not explicit, and because of this, it is not very intuitive and challenging to detect by looking at the dataset. Therefore, it is possible that datasets, that we believed to be biased towards a majority group, were deemed acceptable in terms of disparate impact. Disparate impact is just one way to indicate whether the data may lead to unfair results. It is an excellent legal way to have companies and organizations make positive changes to be less biased, but it is not the only solution. For this reason, we studied many more definitions of fairness to evaluate the outcomes of these datasets.

Naive Bayes — Feldman et al. Trends

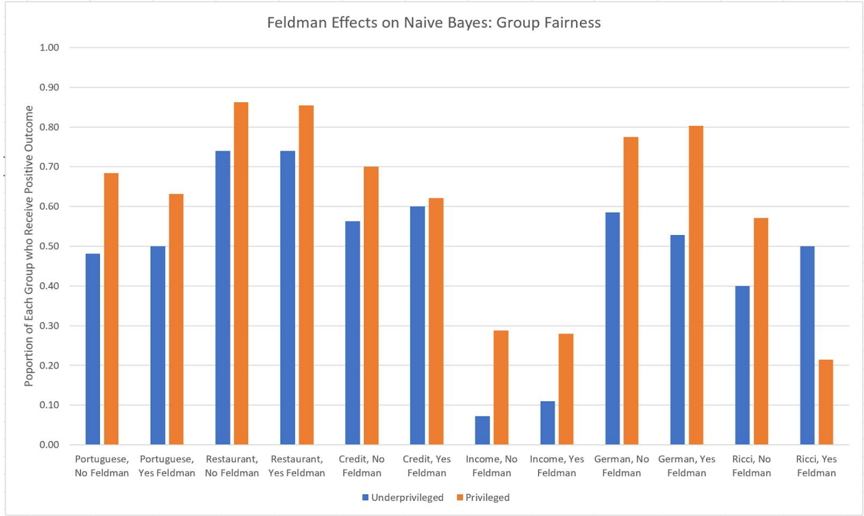


Figure 3: The impact on group fairness of pairing the Feldman et al. Repair with the Naive Bayes classifier.

In general, when looking at the Feldman et al. impact just within the Naive Bayes models, there do not appear to be many trends between metrics. The accuracy, true positive, and true negative rates do not appear to alter much between "No Feldman" and "Feldman" trials, with the accuracy only dropping by a maximum of 1.5%, and the true positive and true negative rates only dropping by a maximum of 2.5%. In terms of fairness, it appears that our equality of opportunity p-values often go down, but never changes our Boolean outcome. The counterfactual measures already start quite high, with all values in the 90s, and all except Portuguese and Income have a very small increase, on the order of 0.2% to 0.3%. Portuguese decreases by 2%

and Income by 4%. Given that preferred treatment is always true for Naive and Modified Bayes, all our preferred treatment outcomes are True. Our individual fairness measures already begin quite high, and appear to have no trend overall.

Group fairness, though, is the metric that most clearly measures the impact of Feldman et al. Given that we are trying to reduce the difference between the proportions of positive classifications each sensitive attribute group gets, and group fairness shows us those proportions, this metric shows us our clearest trends. Generally, for Naive Bayes, Feldman et al. improves the underprivileged scores and slightly decreases privileged scores, causing the proportions to be more similar with Feldman et al. than before, as seen in Figure 3. The Jury dataset is not included in this figure and in other Feldman et al. Repair algorithm comparisons because it did not have any numerical columns to repair. This can be most clearly seen in the Credit dataset, as their proportions grow closer by 10%. It appears that for Credit, Income, and Portuguese, Feldman et al. has strong positive impacts. For Ricci, Feldman et al. has strong negative impacts, as well as for German. For Restaurants, there is no significant change.

Modified Bayes — Feldman et al. Trends

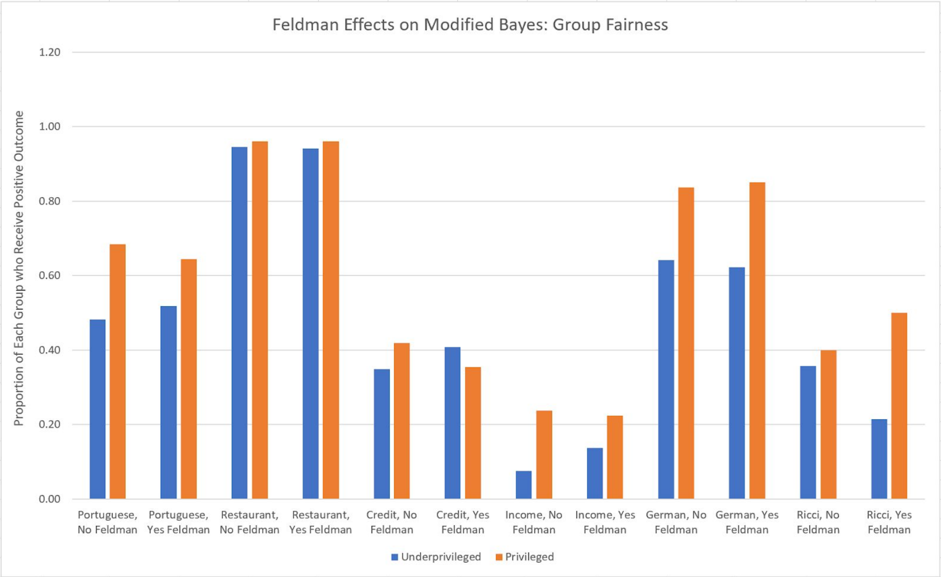


Figure 4: The impact on group fairness of pairing the Feldman et al. Repair with the Modified Bayes classifier.

Running the Feldman et al. Repair had little to no change in impact on accuracy across datasets. Similarly, running the Feldman et al. Repair had little to no change on true positive rates and true negative rates. All true positive rates changed by less than 2% across datasets (except for Ricci, which dropped by 9%) and all true negative rates changed by less than 6%. There was no trend across datasets in the equality of opportunity (EOO) outcomes after running the Feldman et al. Repair. While EOO increased by .29 for the Credit dataset (changing the

Boolean from True to False) and decreased by .5 for the Ricci dataset (where the Boolean was False in both cases), no other EOO changes were above .01, and the rest of the EOO Booleans were all True before and after the Feldman et al. Repair.

There was also no consistent trend in the counterfactual measures across datasets. However, most of the counterfactual measures were high (in the 90s) before and after the Feldman et al. Repair, with the one exception being the Jury dataset, where the counterfactual measures value was 15.5. Preferred treatment was True across all datasets. There were little to no changes across datasets in changes to individual fairness, with the largest change being a 6% decrease for the Credit dataset.

Finally, there were a number of different results with regards to group fairness for the underprivileged and privileged groups, as seen in Figure 4. Three of the underprivileged group fairnesses improved in the experiments with the Feldman et al. Repair (Portuguese, Credit, and Income datasets), with the highest rate increase being 6% for the Income dataset. The German and Restaurant datasets underprivileged group fairness changed very little (differences under 2%) while the Ricci dataset showed the largest change as underprivileged group fairness dropped 14%. The group fairness scores for the privileged group generally stayed the same or dropped slightly (under 5%) except for the Ricci dataset, where the privileged group fairness increased by 10%. Finally, most of the distances between group fairness scores decreased across datasets (with the largest decrease of 8% for the Portuguese dataset) except the Ricci dataset, which increased by 14%.

Two Bayes — Feldman et al. Trends

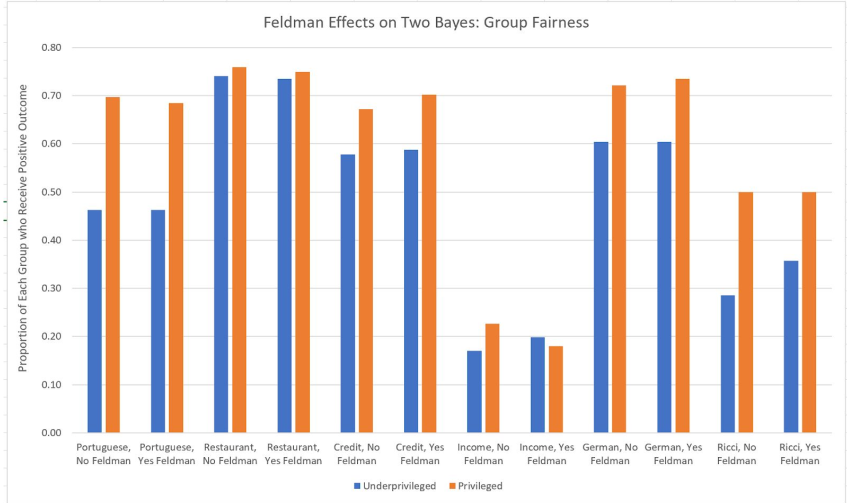


Figure 5: The impact on group fairness of pairing the Feldman et al. Repair with the Two Bayes classifier.

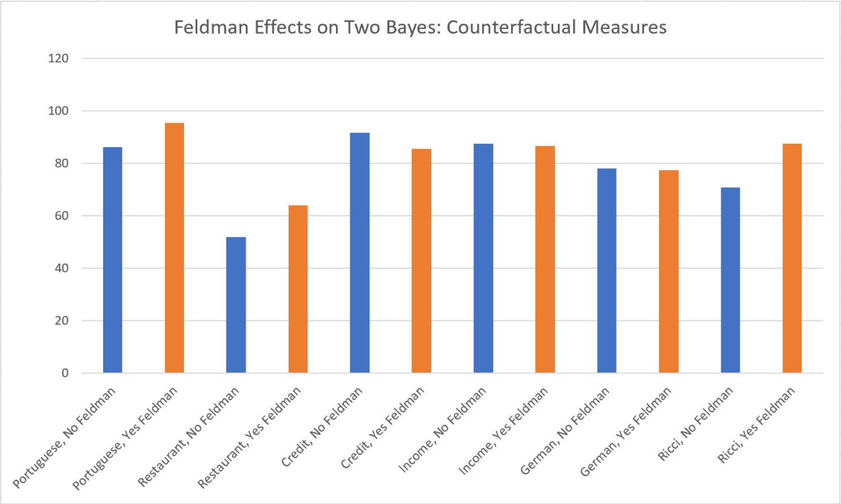


Figure 6: The impact on counterfactual measures of pairing the Feldman et al. Repair with the Two Bayes classifier.

Running Feldman et al. didn’t seem to have a consistent significant effect on the accuracy for Two Bayes. The only significant change was in the Ricci dataset, whose accuracy increased from 91.7% to 95.8% when Feldman et al. Repair was added. The true positive rates mostly stayed roughly the same except for the Income and Ricci datasets. The Income dataset’s true positive rate drops by 6% with Feldman et al. Repair, whereas the Ricci dataset’s increases by 9%. The true negative rates didn’t change much at all.

The EOO p-values don’t have a very clear trend with the addition of Feldman et al., but there were some relatively large changes in value. The Restaurants dataset’s p-value changes from 0.06 to 0.03, causing the Boolean value to go from True to False. The German dataset’s p-value decreases by .05, while Ricci’s increases by .15. In counterfactual measures, the Portuguese, Restaurants, and Ricci datasets all have a significant increase with Feldman et al. Repair, increasing by 9%, 13%, and 17% respectively. The Credit dataset’s counterfactual measures metric decreases by 6%. The values for the other datasets didn’t change significantly. The preferred treatment Boolean values were all true, regardless of running Feldman et al. Repair. Individual fairness didn’t change much with Feldman et al., except for on the German dataset, where the value increased by 3%.

Lastly, the group fairness values didn’t show many trends on Two Bayes with the addition of Feldman et al. The only dataset where there was a significant change in the group fairness for the underprivileged group is the Ricci dataset, where the value increases by 7%. The group fairness values for the privileged group also didn’t show much of a consistent trend. The only dataset that had a significant increase with Feldman et al. is the Credit dataset, which increased 3%. The only dataset with a significant decrease with Feldman et al. is the Income dataset, which decreased by 4%. The differences between the group fairness values of the different groups, however, mostly decrease. The Ricci dataset has the biggest decrease, with a decrease of 7%. The only two datasets with differences that don't decrease are the German and Credit datasets, which increase by 1.3% and 2% respectively.

Bayesian Trends

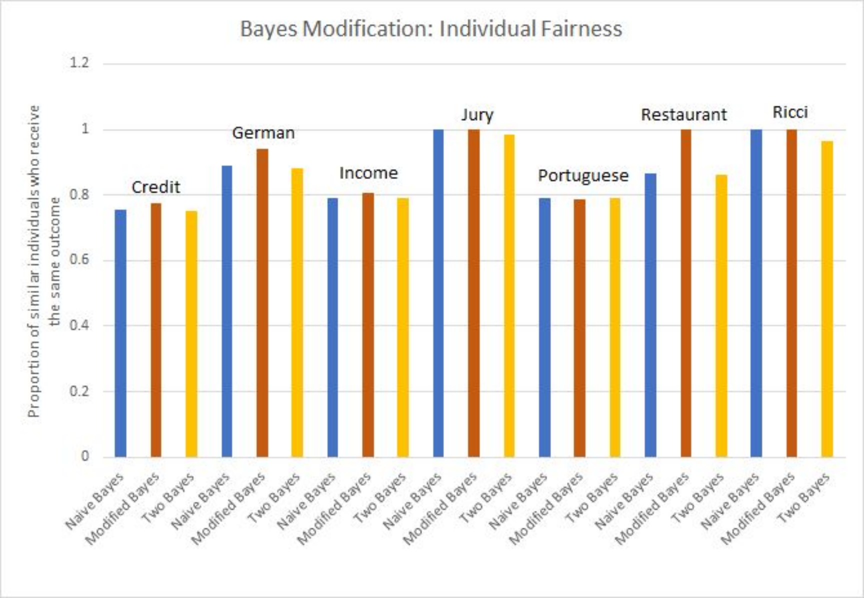


Figure 7: Individual fairness across different Bayes models (with the Feldman et al. Repair).

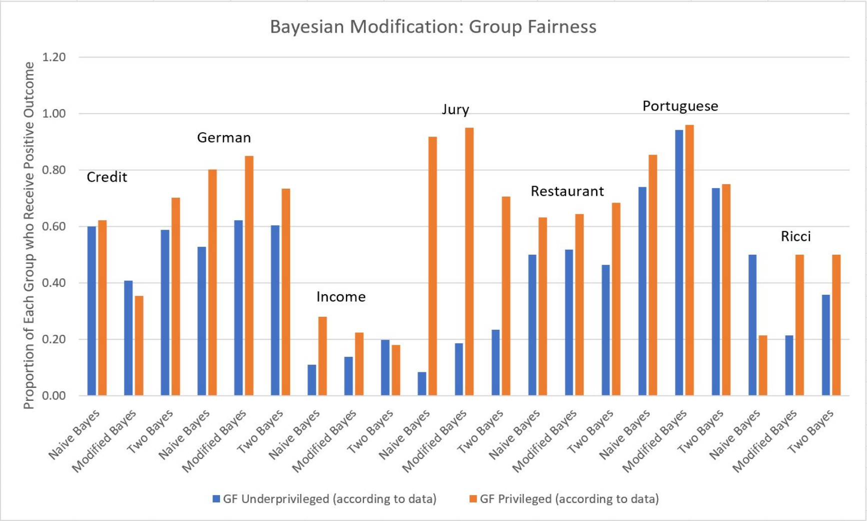


Figure 8: Group fairness across different Bayes models (with the Feldman et al. Repair).

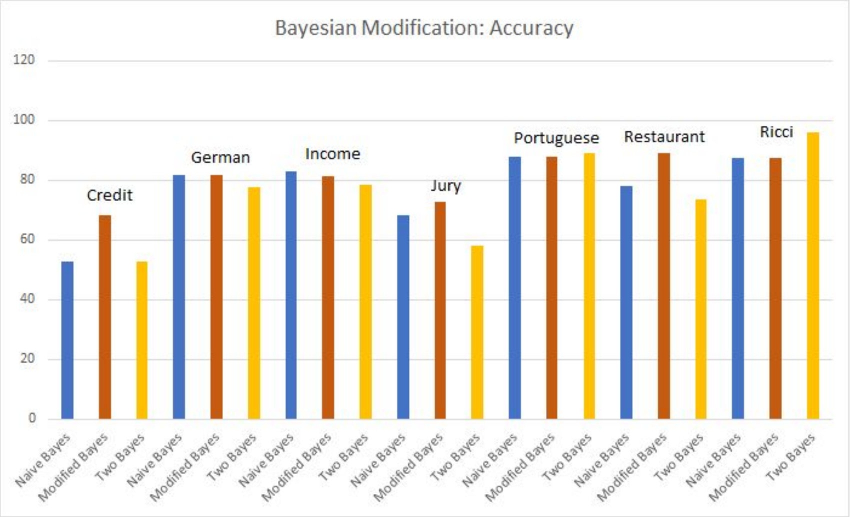


Figure 9: Accuracy across different Bayes models (with the Feldman et al. Repair).

In terms of accuracy, the variations of the Bayesian algorithms follow one of two trends. For several datasets, accuracy stays relatively similar regardless of the algorithm trained upon it. These datasets include Portuguese, Income, German, and Ricci. Accuracy varies by less than 4% regardless of Bayesian algorithm. Following the other trend are the Restaurant, Credit, and Jury datasets. For these datasets, accuracy remains low when trained with Naive Bayes, increases by 5 to 18% and then decreases again when trained with Two Bayes by about 14 to 18%.

The group fairness metric illustrates the effectiveness of Calders and Verwers' (2010) alternate Bayesian algorithms. For every dataset, the difference in group fairness of the two sensitive groups is extremely high when trained with Naive Bayes. In some cases, they differed by as much as 0.8. For most datasets, training with Modified Bayes significantly improves fairness, decreasing the gap between group fairnesses of the two sensitive groups by around 0.1. Finally, when trained with Two Bayes, the difference in group fairness decreased even more. In the Jury dataset, for example, the difference in group fairness between the privileged and underprivileged groups decreased 0.25.

For nearly all datasets, the individual fairness metric remains similar regardless of the Bayesian algorithm trained on it. Those metrics are all above 0.75 and vary by only 0.01 to 0.03. The only exceptions to this trend were the Restaurant and German datasets, in which the accuracy remained similar for the Naive and Two Bayes algorithms, but Modified Bayes boosted accuracy by more than 0.1. This may be an indication that these datasets are already fairer towards underprivileged groups.

Outliers and Oddities

Although these general trends persisted, there were a few notable outliers. For some of the datasets, we see the accuracies remain the same. For example, the Ricci accuracy from Naive Bayes with and without Feldman have almost identical accuracies. Additionally, we note an interesting result with the Portuguese dataset because the accuracy improves with the switch to Two Bayes, which is not in line with the general decrease in accuracy from Naive to Modified and Two Bayes. We also found some cases, markedly with the Jury and Credit sets, where the accuracies make an improvement from Naive Bayes to Modified Bayes, and then drop back down with Two Bayes.

Some of these instances may be a result of the eccentricities of particular datasets. For instance, only 24.9% of the people in the Income dataset make above $50,000 per year, so the classifier tends to predict that relatively few people will make above $50,000 per year. Relatedly, 92.5% of all the restaurants in the Restaurant dataset pass inspection, leading to the possibility that a classifier might pass nearly everyone in the dataset to achieve good fairness and relatively good accuracy (which we see with Modified Bayes' 5% true negative rates). The Ricci dataset also had many outlier values among our results and we attribute this to its unusually small size compared to the other datasets. Likewise, it is possible that if the number of occurrences for sensitive groups is greatly skewed in a dataset (e.g. if there were significantly more women than men) it could cause an unfair shift in the probabilities. Additionally, it is important to remember that the Naive Bayes classifier uses the independence assumption that can potentially lead to preliminary classifications that have worsened the fairness from the original dataset's classifications. Therefore, it is not necessarily the perfect baseline to compare results to. Furthermore, when we move from Naive Bayes to Modified Bayes, we switch to using the probability of the sensitive attribute $P(S) \times P(C|S)$ instead of the $P(C) \times P(S|C)$ that Naive Bayes uses. We then switch back to using $P(C) \times P(S|C)$ for Two Bayes, which we believe could be an origin of the shifting accuracies from Naive to Modified and Two Bayes.

One other unusual result was that the Feldman et al. (2015) Disparate Impact Detector found that the Jury dataset had no disparate impact, despite the fact that the Jury dataset had the lowest underprivileged group fairness and the highest difference between underprivileged and privileged group fairnesses. We are not entirely sure what to make of this finding, but perhaps future work could explore the details of how disparate impact results like this one might occur.

Discussion

The goal of the Feldman et al. Repair (2015) is to bring distributions of data that are skewed by a sensitive attribute closer together—in essence, making the two groups have more similar probabilities of getting the same outcomes. The intent of the Bayesian modifications is to tweak the probabilities that each sensitive attribute group will be assigned a positive classification. Given that group fairness is the metric that is investigating the proportions of each sensitive attribute group who receives a positive classification as their outcome, group fairness is the metric that we would expect to be most impacted by these interventions.

In line with these intuitions, the Feldman et al. (2015) Repair generally decreases the difference between the underprivileged group fairness and the privileged group fairness, especially when combined with Modified Bayes, bringing the difference between group fairness rates 7% closer on average. However, when combined with Two Bayes, the Feldman et al. (2015) Repair seems to make less of a difference in decreasing group fairness rates, only bringing them an average of 3% closer. Because Feldman et al. (2015) Repair and Two Bayes intend to accomplish similar goals, an explanation for this may be that the impact of fairness is less obvious when the two are combined.

The Bayesian modifications also bring the underprivileged and privileged group fairness rates closer together. Running a dataset through Modified Bayes brings the groups 4% closer, on average, than Naive Bayes. Running a dataset through Two Bayes brings the groups 11% closer, on average, than Modified Bayes— a total of 15% closer than Naive Bayes. This is a fairly significant increase.

When looking at the group fairness results, we realized that for some datasets, the underprivileged group was not the sensitive attribute that we expected. For instance, in the Portuguese dataset, more women than men in the test set for all three types of Bayes passed the Portuguese class. Similarly, in the Naive Bayes experiments, more people with high school education had good credit scores than people with college education. Our algorithms rely on the data to determine which group is the underprivileged group: the Feldman et al. (2015) Repair finds the median distribution between sensitive attribute groups, and thus implictly identifies the underprivileged group by identifying which distribution has lower scores overall. This means that the Feldman et al. (2015) Repair may even identify different sensitive attribute groups to be the underprivileged group for different numerical attributes based on the distributions. Modified Bayes and Two Bayes similarly rely on the data to determine the underprivileged sensitive attribute group. They do so by determining which group is the minority (by counting the number of members of each sensitive attribute group and determining the smaller group). The advantage of determining the underprivileged group from the data is that the algorithms are detecting systematic bias in the dataset itself rather than relying on human input, but this also means that ideas about which groups are systematically underprivileged in a society may not be taken into account by these algorithms.

In contrast to the group fairness findings, our results indicate that the Feldman et al. (2015) Repair , Calders and Verwer's (2010) Modified Bayes, and Calders and Verwer's (2010) Two Bayes algorithms do not significantly change fairness with regards to several of the fairness metrics we used. Across all datasets, equality of opportunity, counterfactual measures, preferred treatment, and individual fairness do not consistently improve with any combination of interventions, including the Feldman et al. Repair when combined with Two Bayes. When considering what these metrics are measuring, though, this makes sense. Given that Modified Bayes and Two Bayes both hold the number of true positives output constant, the equality of opportunity and counterfactual measures should not change very much. Preferred treatment also should not change, given that all versions of Bayes should have preferred treatment, and they do. Finally, individual fairness should not be impacted very much either, as none of the algorithms are changing the way that individuals relate to each other—Feldman preserves rank and none of the Bayes alter individuals relationships within groups.

Limitations and future work

We faced a number of limitations in the implementation of our project which could be alleviated by future work. For one, the classification algorithms we used can only be trained and run on a single sensitive attribute at a time. In reality, nearly every classification situation involves multiple sensitive attributes, for instance, the Income dataset had attributes for both education level and gender, both of which are arguably sensitive attributes, but we only trained our algorithms using education level as the sensitive attribute.

Similarly, several algorithms, including Modified Bayes and Two Bayes, can only be applied to binary sensitive attributes. This does not reflect the reality of the world, and in some cases can seriously hamper the fairness of our classifications. Due to this limitation, we were forced to turn attributes which were not originally binary into binary attributes, for example in the Ricci dataset, we converted race values from W, B, H (i.e. white, black, hispanic) into W, POC (i.e. white, person of color). The two Bayesian modification algorithms were made under the assumption that the sensitive attribute would be binary, so major rework would be required to apply them to non-binary sensitive attributes. To do so, however, would certainly improve the potential of our research.

Additionally, because the test-train split randomly divides our dataset into two groups, there is always a chance for an unbalanced split which might limit the effectiveness of our classification algorithms. Cross-validation and multiple trials could help to remedy this issue and improve the consistency of our results.

With additional time, our research would have benefited from further exploration of algorithms. Post-processing algorithms, for example, work to modify results after the classification algorithm has finished classifying. Such algorithms could have added another tool with which fairness might be improved beyond pre- and in-processing algorithms. Furthermore, we tuned our algorithms to work on the seven datasets we discussed in this paper, but they may not successfully run on any possible dataset. Unusual datasets (in terms of size or attribute distribution) may cause our algorithms to perform poorly or fail altogether. To make our findings more applicable, our algorithms must be able to run on any dataset.

In conclusion, there may not be an algorithmic fix that *perfectly* classifies individuals such that there is an even 50-50 split between privileged and underprivileged groups. However, we have shown that there are algorithms that do improve the fairness of classification distributions between and therefore these measures are valid and justifiable to pursue - any advancements toward equality should be seen as worthwhile.

Acknowledgments

We are thankful for everyone that has helped us throughout this project. We are particularly grateful for the assistance given by Anna Rafferty as her role as a stellar advisor. Thank you to Layla Oesper for support in the comps process. We offer special thanks to Mike Tie for software and hardware help during our testing and presenting stages of the project. The advice given by Kristin Partlo and Iris Jastram has been a great help in finding our datasets. We would also like to thank the UCI Machine Learning Repository for providing four out of our seven datasets. And finally, thank you to our friends and family for attending our comps presentation and general support throughout this process.

Works Cited

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227.

Bureau, U. S. C. (n.d.). *U.S. Census Bureau QuickFacts*. https://www.census.gov/quickfacts/fact/table/newyorkcountymanhattanboroughnewyork,bronxcountybronxboroughnewyork,queenscountyqueensboroughnewyork,kingscountybrooklynboroughnewyork,richmondcountystatenislandboroughnewyork,newyorkcitynewyork/INC110218#INC110218

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

Craft, W., Montgomery, D., Tungekar, R., & Yesko, P. (2018). *Jurors*. American Public Media. https://github.com/APM-Reports/jury-data/blob/master/jurors.csv

Dale, E. (2011). *Grade Conversion Chart*. Elizabeth Dale. http://plaza.ufl.edu/edale/Grade conversion chart.htm

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92–112.

Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

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

Flowers v. Mississippi (Vol. 139). (2019). Supreme Court.

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

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *ArXiv Preprint ArXiv:1710.03184*.

Glaberson, S. K. (2019). Coding Over the Cracks: Predictive Analytics and Child Protection. *Fordham Urb. LJ*, *46*, 307.

Health, N. Y. C. D. of, & Hygiene, M. (2020). *DOHMH New York City Restaurant Inspection Results*. New York City Department of Health and Mental Hygiene.

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Juhász, L., & Hochmair, H. H. (2017). Where to catch ‘em all?–a geographic analysis of Pokémon Go locations. *Geo-Spatial Information Science*, *20*(3), 241–251.

Miao, W. (2010). Did the results of promotion exams have a disparate impact on minorities? Using statistical evidence in Ricci v. DeStefano. *Journal of Statistics Education*, *18*(3).

Raub, M. (2018). Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. *Ark. L. Rev.*, *71*, 529.

Ricci v. DeStefano (Vol. 557). (2009). Supreme Court.

Yao, S., & Huang, B. (2017). New fairness metrics for recommendation that embrace differences. *ArXiv Preprint ArXiv:1706.09838*.