Measuring and Mitigating Bias for Tabular Datasets with Multiple Protected Attributes*

Manh Khoi Duong[1,*], Stefan Conrad[1]

[1]Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany

Abstract

Motivated by the recital (67) of the current corrigendum of the AI Act in the European Union, we propose and present measures and mitigation strategies for discrimination in tabular datasets. We specifically focus on datasets that contain multiple protected attributes, such as nationality, age, and sex. This makes measuring and mitigating bias more challenging, as many existing methods are designed for a single protected attribute. This paper comes with a twofold contribution: Firstly, new discrimination measures are introduced. These measures are categorized in our framework along with existing ones, guiding researchers and practitioners in choosing the right measure to assess the fairness of the underlying dataset. Secondly, a novel application of an existing bias mitigation method, `FairDo`, is presented. We show that this strategy can mitigate any type of discrimination, including intersectional discrimination, by transforming the dataset. By conducting experiments on real-world datasets (Adult, Bank, COMPAS), we demonstrate that de-biasing datasets with multiple protected attributes is possible. All transformed datasets show a reduction in discrimination, on average by 28%. Further, these datasets do not compromise any of the tested machine learning models' performances significantly compared to the original datasets. Conclusively, this study demonstrates the effectiveness of the mitigation strategy used and contributes to the ongoing discussion on the implementation of the European Union's AI Act.

Keywords

Machine Learning, Bias Mitigation, Intersectional Discrimination, Fairness, AI Act

1. Introduction

Discrimination in artificial intelligence (AI) applications is a growing concern since the adoption of the *AI Act* by the European Parliament on March 13, 2024 [1]. It still remains a significant challenge across numerous domains [2, 3, 4, 5]. To prevent biased outcomes, *pre-processing* methods are often used to mitigate biases in datasets before training machine learning models [6, 7, 8, 9]. The current corrigendum of the *AI Act* [1] emphasizes this in Recital (67):

> *"[...] The data sets should also have the appropriate statistical properties, including as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with specific attention to the mitigation of possible biases in the data sets [...]"*

Since datasets often consist of multiple protected attributes, pre-processing methods should be able to handle these cases. However, only a few works have addressed this issue [7, 10, 11, 12, 13] and de-biasing such datasets is still an ongoing research topic. In addition, there is no straightforward approach to managing multiple protected attributes, as shown in Figure 1.

Our paper mainly focuses on how to measure and mitigate discrimination in datasets where multiple protected attributes are present. In our first contribution, we provide a comprehensive categorization of discrimination measuring methods. Besides introducing new measures for some of these cases, we also categorize existing measures from the literature. Some of the listed measures specifically address *intersectional discrimination* and *non-binary groups*. The second contribution deals with bias mitigation. For this, we use our published pre-processing framework, `FairDo` [9], that is *fairness-agnostic*. The fairness-agnostic property makes it possible to define any discrimination measure that should be

AEQUITAS 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain

*Corresponding author.

manh.khoi.duong@hhu.de (M. K. Duong); stefan.conrad@hhu.de (S. Conrad)

0000-0002-4653-7685 (M. K. Duong); 0000-0003-2788-3854 (S. Conrad)

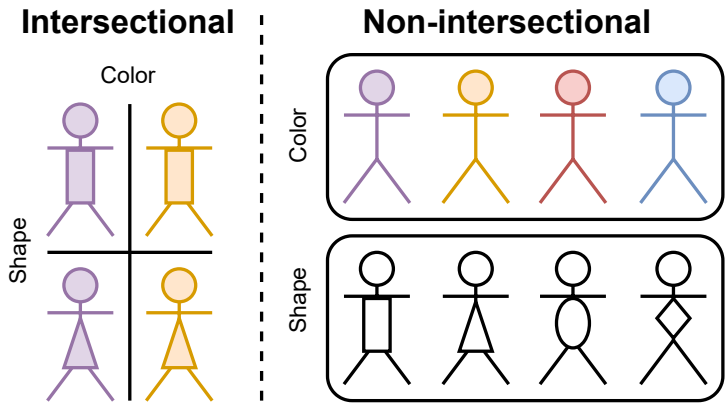

Figure 1: Stick figures can be differentiated by their color and shape. In intersectional discrimination, attributes are intersected, which leads to new subgroups. In non-intersectional, each attribute is treated independently, i.e., colors and shapes are not intersecting in this case.

minimized. By implementing the introduced measures, we can therefore mitigate biases for multiple protected attributes. Another advantage of `FairDo` is that it preserves data integrity and does not modify the features of individuals during the optimization process, unlike other methods [14, 3, 7].

We evaluated our methodology on popular tabular datasets with fairness concerns, such as Adult [15], Bank [16], and COMPAS [17]. We used different discrimination measures to evaluate the effectiveness of the bias mitigation process. Because a successful mitigation process does not guarantee that the outcomes of machine learning models are fair, we trained machine learning models on the transformed datasets and evaluated their predictions regarding fairness and performance. The code for the experiments can be found in the accompanying repository: https://github.com/mkduong-ai/fairdo/evaluation.

The results of the bias mitigation process as well as the performance of the machine learning models are promising. They indicate that achieving fairness in datasets with multiple protected attributes is possible, and `FairDo` is a proper framework for this task. Overall, our work contributes technical solutions for stakeholders to enhance the fairness of datasets and machine learning models, aiming for compliance with the *AI Act* [1].

2. Preliminaries

To handle multiple protected attributes, we define $\mathcal{Z} = \{Z_1, \ldots, Z_p\}$ as a set of protected attributes. It can represent the set of sociodemographic features such as age, gender, and ethnicity. These factors may make individuals vulnerable to discrimination. Each protected attribute $Z_k \in \mathcal{Z}$ is formally a *discrete random variable* that can take on values from the sample space g_k. In this context, we refer g_k to groups that describe distinct social categories of a protected attribute. For example, let Z_k represent gender; then g_k is a set containing the genders male, female, and non-binary. To avoid limitations to a particular group fairness notion, we introduce a generalized notation based on the works of Žliobaitė [2], Duong and Conrad [9] in the following.

Definition 2.1 (Treatment)**.** *Let E_1, E_2 be events and Z_k be a random variable that can take on values from g_k, then we call the conditional probability*

$$P(E_1 \mid E_2, Z_k = i)$$

treatment, *where $i \in g_k$. E_1 describes some favorable outcome, such as getting accepted for a job, while E_2 often represents some additional information about the individual, such as their qualifications.*

Definition 2.2 (Fairness Criteria)**.** *With the definition of treatment, we can define fairness criteria that demand equal treatment for different groups. Let $P(E_1 \mid E_2, Z_k = i)$ and $P(E_1 \mid E_2, Z_k = j)$ be*

treatments, then we call the following equation:

$$P(E_1 \mid E_2, Z_k = i) = P(E_1 \mid E_2, Z_k = j)$$

a fairness criterion*, for all* $i, j \in g_k$.

Definition 2.2 allows us to define various group fairness criteria, including *statistical parity* [18], *predictive parity* [3], *equality of opportunity* [19], etc. They all demand some sort of equal outcome for different groups and can be defined by configuring the events E_1, E_2. For instance, statistical parity [18] requires that two different groups have an equal probability of receiving a favorable outcome $(Y = 1)$.

Example 2.1 (Statistical Parity [18])**.** *To define statistical parity for the attribute* Z_k *using our notation, we set* $E_1 := (Y = 1)$ *and* $E_2 := \Omega$*. By setting* E_2 *to the sample space* Ω*, we compare the probabilities of the event* $Y = 1$ *across different groups without conditioning on any additional event:*

$$\begin{aligned} & P(Y = 1 \mid \Omega, Z_k = i) = P(Y = 1 \mid \Omega, Z_k = j) \\ \iff\ & P(Y = 1 \mid Z_k = i) = P(Y = 1 \mid Z_k = j), \end{aligned}$$

where $i, j \in g$ *represent different groups.*

In real-world applications, achieving equal probabilities for certain outcomes is not always possible. Due to variations in sample sizes in the groups, it is common to yield unequal treatments, even when they are similar. Thus, existing literature [2] uses the absolute difference to quantify the strength of discrimination.

Definition 2.3 (Disparity)**.** *Let* $P(E_1 \mid E_2, Z_k = i)$ *and* $P(E_1 \mid E_2, Z_k = j)$ *be two treatments, then we refer to*

$$\delta_{Z_k}(i, j, E_1, E_2) = |P(E_1 \mid E_2, Z_k = i) - P(E_1 \mid E_2, Z_k = j)|$$

as the disparity*, for all* $i, j \in g_k$*. Trivially,* δ_{Z_k} *is commutative regarding* i, j*. In practice, it prevents reverse discrimination due to the absolute value.*

Definition 2.4 (Discrimination)**.** *We use* $\psi \colon \mathbb{D} \to \mathbb{R}$ *to denote some discrimination measure that quantifies the discrimination inherent in any dataset* $\mathcal{D} \in \mathbb{D}$*. A dataset* $\mathcal{D}$ *consists of features, protected attributes, and labels for each individual. The explicit form of* ψ *depends on the cases introduced in Section 3.*

3. Measuring Discrimination for Multiple Attributes

We found that numerous scenarios arise when dealing with multiple protected attributes. We categorize these scenarios based on the number of groups, denoted as $|g|$, and the number of protected attributes, denoted as $|\mathcal{Z}|$. By going through all cases, we present possible approaches from the literature as well as our own suggestions to measure discrimination.

3.1. Single Protected Attribute $(|\mathcal{Z}| = 1)$

In the case of having only one protected attribute, i.e., $|\mathcal{Z}| = |\{Z_1\}| = 1$, we distinguish between cases by the number of available groups $|g|$ in the dataset. We categorize the cases by $|g| = 0, 1, 2$, and $|g| > 2$.

3.1.1. No Groups ($|g| = 0$)

When there are no groups, the measurement of discrimination is impossible if no assumptions are being made. Discrimination can be assessed through proxy variables [20]; however, this approach can be imprecise and may introduce new biases. This case is equivalent to having no protected attribute, i.e., $|\mathcal{Z}| = 0$.

3.1.2. Single Group ($|g| = 1$)

Similarly to the case of having no groups, discrimination cannot be measured when having only one group. For this, we propose practices where prior information can be incorporated:

1. *No discrimination*: As no difference towards any other group can be measured, returning a discrimination score of 0 is one viable option.

$$\psi(\mathcal{D}) = 0. \tag{1}$$

2. *Difference to optimal treatment*: Another way is to return the absolute difference of the group's outcome to the optimal treatment. For example, group i has an 80% chance of receiving the favorable treatment. Ideally, having a 100% chance would represent the optimal scenario. Therefore, the discrimination score is 20% in this case. It is given by:

$$\psi(\mathcal{D}) = |P(E_1 \mid E_2, Z_1 = i) - 1|. \tag{2}$$

3. *Difference to expected treatment*: We can use the expected treatment as a reference point. For example, we know that a company has a 50% acceptance rate for job applications. Now a machine learning classifier is trained to predict whether an applicant will be accepted and the model's predictions result in a 60% acceptance rate for group i. Hence, the model is positively biased towards group i by 10%. This can be formulated as:

$$\psi(\mathcal{D}) = |P(E_1 \mid E_2, Z_1 = i) - p_{\text{expect.}}|, \tag{3}$$

   where $p_{\text{expect.}}$ is the expected treatment. It can describe the average treatment across all groups [21] or some other prior information that is not included in the dataset.

3.1.3. Binary Groups ($|g| = 2$)

Without using any prior information, we can calculate the discrimination score by taking the absolute difference between the treatments of the two groups, as advised by Žliobaitė [2]. The discrimination measure ψ is then simply given by the disparity as mentioned in Definition 2.3.

3.1.4. Non-binary Groups ($|g| > 2$)

While the case for binary attributes is straightforward, it becomes non-trivial for non-binary attributes that arise naturally in real-world data. We can fall back to $|g| = 2$ by calculating the absolute difference between every distinct group $i, j \in g$. Because the discrimination between i and j is the same as between j and i, only $\binom{|g|}{2}$ pairs need to be compared and we use an aggregation function $\text{agg}^{(1)}$ to report the differences [2]. Lum et al. [22] refers to measures that aggregate or summarize discrimination scores as *meta-metrics*. The aggregate can be the *sum* or *maximum* function, depending on the use case. The result for a single protected attribute Z_k with two or more groups can be computed as follows:

$$\psi(\mathcal{D}) = \underset{i,j \in g_k, i<j}{\text{agg}^{(1)}} \delta_{Z_k}(i, j, E_1, E_2), \tag{4}$$

where δ_{Z_k} is the disparity as defined in Definition 2.3 and $i < j$ ensures that each pair is considered only once (assuming label-encoded groups). According to Žliobaitė [2] and her personal discussions with legal experts, she advocates using the maximum function, i.e.,

$$\psi(\mathcal{D}) = \max_{i,j \in g_k, i<j} \delta_{Z_k}(i, j, E_1, E_2) \tag{5}$$

$$= \max_{i \in g_k} P(E_1 \mid E_2, Z_k = i) - \min_{j \in g_k} P(E_1 \mid E_2, Z_k = j). \tag{6}$$

Equation (5) describes the maximum discrimination obtainable between two groups. An alternative and equivalent formulation is given in Equation (6) [7]. The latter is computationally more efficient as it requires $\mathcal{O}(2|g|)$ operations compared to $\mathcal{O}(|g|^2)$ operations for the former.

A more general approach to measuring discrimination is to calculate some form of *correlation coefficient* between the protected attribute and the outcome. The correlation coefficient can be calculated using Pearson's correlation [23], Spearman or Kendall's rank correlation [24, 25]. The discrimination measure can then be defined as the absolute value of the correlation coefficient:

$$\psi(\mathcal{D}) = |\text{Corr}(E_1, Z_k)|. \tag{7}$$

This approach can be applied to any number of groups. `Fairlearn` provides a pre-processing method that removes the correlation between the protected attribute and the outcome by transforming the data [7]. However, the given approach violates data integrity constraints as categorical attributes are transformed into continuous values. Moreover, zero correlation does not imply independence between two variables.

3.2. Multiple Protected Attributes ($|\mathcal{Z}| > 1$)

There are several ways to measure discrimination for multiple protected attributes ($|\mathcal{Z}| > 1$). Based on the works of Kearns et al. [21], Yang et al. [11] and Kang et al. [13], we categorize them into two approaches: *intersectional* and *non-intersectional* (see Figure 1). Intersectional approaches consider the intersection of identities. The overlapping of such identities forms *subgroups* [21]. Non-intersectional approaches treat each protected attribute independently [11].

3.2.1. Intersectional Discrimination

The central idea of intersectionality is that individuals experience overlapping forms of oppression or privilege based on the combination of multiple social categories they belong to. In the following, we will introduce definitions to formulate intersectional discrimination, which is based on the work of Kearns et al. [21].

Definition 3.1 (Subgroup [21])**.** *Let $\mathcal{Z} = \{Z_1, \ldots, Z_p\}$ be a set of discrete random variables representing protected attributes that can take on values from corresponding groups $g_1, \ldots, g_p$. A* subgroup *i is defined as $i = (i_1, \ldots, i_p) \in g_1 \times \ldots \times g_p$. In other words, a* subgroup *encompasses multiple groups from different protected attributes.*

Definition 3.2 (Subgroup Treatment)**.** *Let i be a subgroup as defined in Definition 3.1 and let $\mathcal{Z} = \{Z_1, \ldots, Z_p\}$ be a set of discrete random variables.* Subgroup treatment *is then defined as:*

$$P(E_1 \mid E_2, Z_1 = i_1, \ldots, Z_p = i_p).$$

Definition 3.3 (Subgroup Disparity)**.** *Let $\mathcal{Z} = \{Z_1, \ldots, Z_p\}$ be a set of discrete random variables. Let $i, j \in g_1 \times \ldots \times g_p$ be two subgroups with $i = (i_1, \ldots, i_p)$ and $j = (j_1, \ldots, j_p)$. The disparity between two subgroups is denoted as $\hat{\delta}_{\mathcal{Z}}$ and is given by:*

$$\hat{\delta}_{\mathcal{Z}}(i, j, E_1, E_2) = |P(E_1 \mid E_2, Z_1 = i_1, \ldots, Z_p = i_p) - P(E_1 \mid E_2, Z_1 = j_1, \ldots, Z_p = j_p)|.$$

Similarly to Equation (4), we can calculate the discrimination score for multiple protected attributes by aggregating disparities across all subgroups. A subgroup can be treated like a normal group. According to Definition 3.1, there are theoretically at least 2^p subgroups, where p is the number of protected attributes. However, not all subgroups may be available in the dataset. For unavailable subgroups, the disparity cannot be calculated as the corresponding treatment is undefined.

Let us denote the set of available subgroups as $G_{\text{avail}} \subseteq g_1 \times \ldots \times g_k$. To finally capture the discrepancies across all available subgroup pairs, an aggregation function $\text{agg}^{(1)}$ is applied to the subgroup disparities $\hat{\delta}_{\mathcal{Z}}$:

$$\psi_{\text{intersect}}(\mathcal{D}) = \underset{i,j \in G_{\text{avail}}}{\text{agg}^{(1)}} \hat{\delta}_{\mathcal{Z}}(i, j, E_1, E_2). \tag{8}$$

Table 1
Example dataset of individuals receiving a favorable ($Y = 1$) or unfavorable ($Y = 0$) outcome. The dataset shows four individuals with their respective age group and sex.

Individual	Age	Sex	Outcome (Y)
1	Old	Male	1
2	Old	Female	0
3	Young	Male	0
4	Young	Female	1

Equation (8) represents the aggregated discrimination between all available subgroups in the dataset. When using the maximum function as the aggregator, the calculations are equivalent to Equation (5) and Equation (6). The only difference is that the conditionals are now subgroups instead of groups:

$$\begin{aligned}\psi_{\text{intersect}}(\mathcal{D}) &= \max_{i,j \in G_{\text{avail}}} \hat{\delta}_{Z_k}(i, j, E_1, E_2) \\ &= \max_{i \in G_{\text{avail}}} P(E_1 \mid E_2, Z_1 = i_1, \ldots, Z_p = i_p) - \min_{j \in G_{\text{avail}}} P(E_1 \mid E_2, Z_1 = j_1, \ldots, Z_p = j_p).\end{aligned} \tag{9}$$

Kang et al. [13] also dealt with intersectional discrimination in their work by introducing a multivariate random variable Z where each dimension represents a protected attribute. Their fairness objective is to minimize the mutual information between the outcome and the multivariate random variable. By minimizing the mutual information, the outcome is independent of the protected attributes, which is a desirable property for fairness [14, 26]. In this context, zero mutual information implies the absence of intersectional discrimination [13]. However, this approach relies on expensive techniques to approximate the mutual information. Using our notation, their formulation can be written as [13]:

$$\psi_{\text{MI}}(\mathcal{D}) = \text{MI}(E_1, Z), \tag{10}$$

where MI denotes the mutual information.

3.2.2. Non-intersectional Discrimination

The problem with measuring discrimination for intersectional groups is that it has an upward bias when using meta-metrics [22]. This is because the number of subgroups grows exponentially with the number of protected attributes. This leads to many subgroups where the number of samples in each subgroup is possibly small, resulting in larger noise in the treatment estimates [22].

Besides intersectional groups, Yang et al. [11] listed a non-intersectional definition of groups, called *independent groups*. Building on the definition of *independent groups*, we propose an appropriate approach to measure discrimination for this type of groups. It is more suitable when dealing with a large number of subgroups or when intersectional discrimination is not deemed important. Our non-intersectional approach treats each protected attribute independently and aggregates the discrimination scores across all protected attributes. For this, a second aggregate function with $\text{agg}^{(2)}$ is introduced, yielding the following equation:

$$\psi_{\text{indep}}(\mathcal{D}) = \underset{Z_k \in \mathcal{Z}}{\text{agg}^{(2)}} \left\{ \underset{i,j \in g_k, i<j}{\text{agg}^{(1)}} \delta_{Z_k}(i, j, E_1, E_2) \right\}. \tag{11}$$

The first-level aggregator $\text{agg}^{(1)}$ aggregates disparities within a protected attribute, considering unique pairs of groups i and j. The second-level aggregator $\text{agg}^{(2)}$ then combines the results across all protected attributes. By applying both operators, we obtain a discrimination measure that captures disparities between groups across multiple attributes.

3.2.3. Example

Let us consider a dataset with two protected attributes, age and sex (see Table 1). The set of protected attributes is $\mathcal{Z} = \{Z_1, Z_2\} = \{\text{Age}, \text{Sex}\}$ and the set of available subgroups in the dataset is $G_{\text{avail}} = \{\text{Old}, \text{Young}\} \times \{\text{Male}, \text{Female}\}$. We measure discrimination using *statistical disparity*. For simplicity, all aggregation functions are set to the maximum function. The *intersectional approach* yields the following discrimination score:

$$\begin{aligned}\psi_{\text{intersect}}(\mathcal{D}) &= \max_{i,j \in G_{\text{avail}}} \hat{\delta}_{\mathcal{Z}}(i, j, (Y = 1), \Omega) \\ &= \max_{i,j \in G_{\text{avail}}} \hat{\delta}_{\{\text{Age, Sex}\}}(i, j, (Y = 1), \Omega) \\ &= \max_{i \in G_{\text{avail}}} P(Y = 1 \mid Z_1 = i_1, Z_2 = i_2) - \min_{j \in G_{\text{avail}}} P(Y = 1 \mid Z_1 = j_1, Z_2 = j_2) \\ &= |P(Y = 1 \mid \text{Age} = \text{Old}, \text{Sex} = \text{Male}) - P(Y = 1 \mid \text{Age} = \text{Young}, \text{Sex} = \text{Male})| = 1,\end{aligned} \tag{12}$$

while the discrimination score for the *non-intersectional approach* is given by:

$$\begin{aligned}\psi_{\text{indep}}(\mathcal{D}) &= \max_{Z_k \in Z} \left\{ \max_{i,j \in g_k, i<j} \delta_{Z_k}(i, j, (Y = 1), \Omega) \right\} \\ &= \max \left\{ \delta_{\text{Age}}(\text{Old}, \text{Young}, (Y = 1), \Omega), \delta_{\text{Sex}}(\text{Male}, \text{Female}, (Y = 1), \Omega) \right\} \\ &= \max\{|0.5 - 0.5|, |0.5 - 0.5|\} = \max\{0, 0\} = 0.\end{aligned} \tag{13}$$

The non-intersectional approach yields a discrimination score of 0 because the disparities for both protected attributes are 0. This is quite different from the intersectional approach, which reports a discrimination score of 1. As seen, the results can differ depending on the approach.

4. Experiments

Our experimentation follows a pipeline consisting of *data pre-processing*, *bias mitigation*, *model training*, and *evaluation*. To mitigate bias in tabular datasets with multiple protected attributes, we used the sampling method, `FairDo` [9], that constructs fair datasets by selectively sampling data points. The method is very flexible and only requires the user to define the discrimination measure that should be minimized. In our case, we are interested in a dataset that has minimal bias across multiple protected attributes. The experiments revolve around the following research questions:

- **RQ1** Is it possible to yield a fair dataset with `FairDo`, where bias for multiple protected attributes is reduced?
- **RQ2** Are machine learning models trained on fair datasets more fair in their predictions than those trained on original datasets?

4.1. Experimental Setup

Datasets and Pre-processing The tabular datasets employed in our experiments include the Adult [15], Bank [16], and COMPAS [17] datasets. They are known for their use in fairness research and contain multiple protected attributes. We pre-processed the datasets by applying one-hot encoding to categorical variables and label encoding to protected attributes. Table 2 shows important characteristics of the datasets after pre-processing.

Each dataset was divided into training and testing sets using an 80/20 split, respectively. We ensured that the split was stratified (if possible) based on protected attributes to maintain representativeness across different groups in both sets.

Table 2
Overview of Datasets

Dataset	Samples	Feats.	Label	Protected Attributes	Description
Adult [15]	32 561	21	Income	**Race**: White, Black, Asian-Pacific-Islander, American-Indian-Eskimo, Other **Sex**: Male, Female	Indicates individuals earning over $50,000 annually
Bank [16]	41 188	50	Term deposit subscription	**Job**: Admin, Blue-Collar, Technician, Services, Management, Retired, Entrepreneur, Self-Employed, Housemaid, Unemployed, Student, Unknown **Marital Status**: Divorced, Married, Single, Unknown	Shows whether the client has subscribed to a term deposit.
COMPAS [17]	7 214	13	2-year recidivism	**Race**: African-American, Caucasian, Hispanic, Other, Asian, Native American **Sex**: Male, Female **Age Category**: <25, 25-45, >45	Displays individuals that were rearrested for a new crime within 2 years after initial arrest.

Bias Mitigation Applying the bias mitigation method `FairDo` [9] to the datasets can be regarded as a pre-processing step, too. This is because the method simply returns a dataset that is fair with respect to the given discrimination measure. `FairDo` [9] offers a variety of options to mitigate bias, and we chose the *undersampling* method that removes samples. In this option, the optimization objective is stated as [9]:

$$\min_{\mathcal{D}_{\text{fair}} \subseteq \mathcal{D}} \quad \psi(\mathcal{D}_{\text{fair}}), \tag{14}$$

where $\mathcal{D}$ is the training set of Adult, Bank, or COMPAS, and ψ is the fairness objective function. We experimented with both $\psi_{\text{intersect}}$ and ψ_{indep} as objectives functions. Bias mitigation is only applied to the training set and the testing set remains unchanged. `FairDo` internally uses genetic algorithms to select a subset of the training set that minimizes the objective function. We used the same settings and operators as provided in the package and only adjusted the population size (200) and the number of generations (400).

Model Training We utilized the `scikit-learn` library [27] to train various machine learning classifiers, namely *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Random Forest* (RF), and *Artificial Neural Network* (ANN). These classifiers were trained on both the original and fair datasets. Classifiers trained on the original datasets serve as a baseline for comparison. We used the default hyperparameters given by `scikit-learn` package for each classifier.

Evaluation Metrics We evaluated the models' predictions on fairness and performance using the test set. For fairness, we assessed $\psi_{\text{intersect}}$ and ψ_{indep}. For the classifiers' performances, we report the *area under the receiver operating characteristic curve* (AUROC) [28], where higher values indicate better performances. Because removing data points can compromise the overall quality of the data, we also report the number of subgroups before and after bias mitigation to check for representativeness.

Trials For each dataset and discrimination measure combination, the bias mitigation process was repeated 10 times. The results were averaged over the trials to obtain a more robust evaluation.

4.2. Results

Fair Dataset Generation Table 3 shows the average discrimination before and after mitigating bias in the training sets. On all datasets, discrimination was reduced after applying `FairDo`. Without

Table 3
Average discrimination and number of subgroups before and after pre-processing the training sets with FairDo.

Dataset	Metric	Disc. Before	**Disc. After**	Subgroups Before	Subgroups After
Adult	ψ_{indep}	20%	13%	10	10
	$\psi_{\text{intersect}}$	31%	16%	10	10
Bank	ψ_{indep}	24%	5%	48	48
	$\psi_{\text{intersect}}$	33%	15%	48	46.2
COMPAS	ψ_{indep}	30%	5%	34	34
	$\psi_{\text{intersect}}$	100%	17%	34	28.8

considering group intersections, discrimination was reduced by 7%, 19%, and 25% for Adult, Bank, and COMPAS, respectively. When considering intersectionality, the discrimination was reduced by 15%, 18%, and 83%. Hence, discrimination was reduced by 28% on average across all datasets, thus answering **RQ1** positively. When comparing the discrimination scores, it can be observed that the intersectional discrimination scores are generally higher. This is because in the intersectional setting, more subgroups are considered, which potentially leads to larger differences between them [21].

We also report the number of subgroups before and after bias mitigation to assess the impact of the undersampling method on the dataset. The removal of subgroups can only be observed in the intersectional setting. In the COMPAS dataset 5.2 out of 34 subgroups were removed on average, indicating the largest amount of subgroups removed across all datasets. While the Bank dataset consists of 48 subgroups, only 1.8 subgroups were removed on average. Because the COMPAS dataset's initial intersectional discrimination score is 100%, removing more subgroups seems inevitable to reduce bias.

Model Performance and Fairness Figure 2 shows the results of the classifiers' performances on the test set. The classifiers' performances are displayed on the y-axis, while the discrimination values are shown on the x-axis. We note that the axes do not share the same scale across the subfigures for analytical purposes.

Classifiers trained on fair datasets did not suffer a significant decline in performance compared to those trained on original datasets. In all cases, only a slight decrease of 1%-3% in performance can be noted. This indicates that the bias mitigation process does not compromise the dataset's fidelity and, therefore, the classifiers' performances. Regarding discrimination, a significant reduction is evident. The x-axis scales are much larger than the y-axis scales, suggesting that changes in discrimination are larger than changes in performance. For example, the RF classifier trained on the Bank dataset (Figure 2g) shows a decrease in intersectional discrimination from 38% to 15%, while the performance only decreases by 2%. Similar results can be observed for the other classifiers and datasets as well, successfully addressing **RQ2**. The results suggest that FairDo can be reliably used to mitigate bias in tabular datasets for various measures that consider multiple protected attributes. Still, we advise users to carefully perform similar analyses when applying the method to their datasets.

5. Discussion

The results of our experiments show that the presented measures detect discrimination in datasets with multiple protected attributes differently. When using the intersectional discrimination measure, more groups are identified and compared to each other. While subgroups are not ignored by this measure, measuring higher discrimination scores by random chance becomes more likely [21, 22]. In contrast, treating each protected attribute separately prevents this issue but may lead to overlooking discrimination. The choice of measure is up to the stakeholders and depends on the context of the dataset and the regulations that apply to the AI system. We generally recommend using the intersectional discrimination measure if the number of individuals in each subgroup is large enough to draw statistically significant conclusions. Otherwise, treating each protected attribute separately is more suitable.

By using the mitigation strategy FairDo [9], the resulting datasets in the experiments have improved

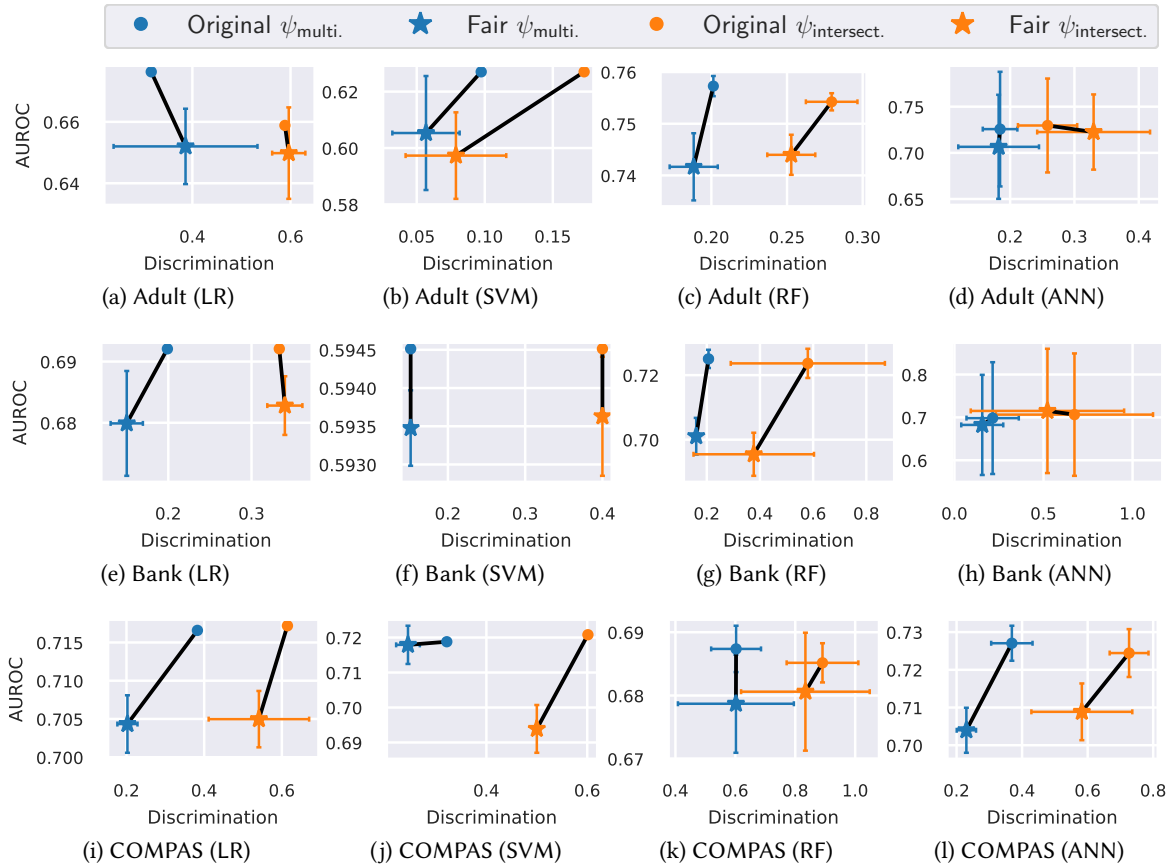

Figure 2: Results on the test set. The x-axis represents the discrimination values (legend indicates used measure) and the y-axis represents the classifiers' performances. We compare the pre-processed (fair) data with the original data. The points/stars represent averages, and the error bars display the standard deviations of the AUROC and discrimination values over 10 trials.

statistical properties regarding fairness. Whether intersectionality was considered or not, reducing discrimination in datasets was possible. At the current state, the AI Act [1] does not explicitly mention *intersectional discrimination* nor how to deal with multiple protected attributes generally. While recital (67) states that datasets *"should [...] have the appropriate statistical properties"*, it does not specify what these properties are. Hence, our work serves as an initial guideline for what these properties could be and how to achieve them in practice.

6. Conclusion

Datasets often come with multiple protected attributes, which makes measuring and mitigating discrimination more challenging. Most existing studies only deal with a single protected attribute, and works that consider multiple protected attributes often focus on intersectionality. In opposition to this, we proposed a new non-intersectional measure that treats each protected attribute separately. This is more suitable when the number of subgroups is too large or the number of individuals in each subgroup is small. We used both intersectional and non-intersectional measures as objectives and applied the `FairDo` framework to mitigate discrimination in multiple datasets. The experiments show that discrimination was reduced in all datasets and on average by 28%. Machine learning models trained on the bias-mitigated datasets also improved their fairness while maintaining performance compared to models trained on the original datasets.

References

[1] European Commission, Artificial Intelligence Act, Corrigendum, 19 April 2024, Available online: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf, 2024. Accessed: 17 May 2024.

[2] I. Žliobaitė, Measuring discrimination in algorithmic decision making, Data Mining and Knowledge Discovery 31 (2017) 1060–1089.

[3] M. B. Zafar, I. Valera, M. Gomez Rodriguez, K. P. Gummadi, Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment, in: Proceedings of the 26th International Conference on World Wide Web, 2017. doi:10.1145/3038912.3052660.

[4] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 797–806.

[5] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019.

[6] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[7] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in AI, Technical Report MSR-TR-2020-32, Microsoft, 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[8] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A reductions approach to fair classification, in: International Conference on Machine Learning, PMLR, 2018, pp. 60–69.

[9] M. K. Duong, S. Conrad, Towards fairness and privacy: A novel data pre-processing optimization framework for non-binary protected attributes, in: D. Benavides-Prado, S. Erfani, P. Fournier-Viger, Y. L. Boo, Y. S. Koh (Eds.), Data Science and Machine Learning, Springer Nature Singapore, Singapore, 2024, pp. 105–120.

[10] J. R. Foulds, R. Islam, K. N. Keya, S. Pan, Bayesian Modeling of Intersectional Fairness: The Variance of Bias, 2020, pp. 424–432. doi:10.1137/1.9781611976236.48.

[11] F. Yang, M. Cisse, S. Koyejo, Fairness with overlapping groups, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[12] L. E. Celis, L. Huang, V. Keswani, N. K. Vishnoi, Fair classification with noisy protected attributes: A framework with provable guarantees, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 1349–1361.

[13] J. Kang, T. Xie, X. Wu, R. Maciejewski, H. Tong, Infofair: Information-theoretic intersectional fairness, 2022 IEEE International Conference on Big Data (Big Data) (2021) 1455–1464.

[14] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International conference on machine learning, PMLR, 2013, pp. 325–333.

[15] R. Kohavi, Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid, KDD'96, AAAI Press, 1996, p. 202–207.

[16] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decision Support Systems 62 (2014) 22–31.

[17] J. Larson, J. Angwin, S. Mattu, L. Kirchner, Machine bias, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[18] T. Calders, F. Kamiran, M. Pechenizkiy, Building classifiers with independency constraints, in: 2009 IEEE International Conference on Data Mining Workshops, 2009, pp. 13–18. doi:10.1109/ICDMW.2009.83.

[19] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).

[20] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in

machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.
[21] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 2564–2572.
[22] K. Lum, Y. Zhang, A. Bower, De-biasing "bias" measurement, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 379–389. doi:10.1145/3531146.3533105.
[23] K. Pearson, Notes on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242.
[24] C. Spearman, The proof and measurement of association between two things, American Journal of Psychology 15 (1904) 72–101.
[25] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.
[26] A. Ghassami, S. Khodadadian, N. Kiyavash, Fairness in supervised learning: An information theoretic approach, in: 2018 IEEE International Symposium on Information Theory (ISIT), IEEE Press, 2018, p. 176–180. doi:10.1109/ISIT.2018.8437807.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[28] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861–874. doi:10.1016/j.patrec.2005.10.010.