A Principled Approach for Data Bias Mitigation

Bruno Scarone[1], Alfredo Viola[2], Renée J. Miller[3], Ricardo Baeza-Yates[4,5]
[1]Khoury College of Computer Sciences, Northeastern University, Boston, USA
[2]Casa de Investigadores Científicos La Comarca, La Floresta, Uruguay
[3]Cheriton School of Computer Science, University of Waterloo, Ontario, Canada
[4] WASP Professor, KTH Royal Institute of Technology, Sweden
[5] Department of Engineering, Universitat Pompeu Fabra, Barcelona, Spain
scarone.b@northeastern.edu, alfredo.viola@gmail.com, rjmiller@uwaterloo.ca, rbaeza@acm.org

Abstract

The widespread use of machine learning and data-driven algorithms for decision making has been steadily increasing over many years. *Bias* in the data can adversely affect this decision-making. We present a new mitigation strategy to address data bias. Our methods are explainable and come with mathematical guarantees of correctness. They can take advantage of new work on table discovery to find new tuples that can be added to a dataset to create real datasets that are unbiased or less biased. Our framework covers data with non-binary labels and with multiple sensitive attributes. Hence, we are able to measure and mitigate bias that does not appear over a single attribute (or feature), but only intersectionally, when considering a combination of attributes. We evaluate our techniques on publicly available datasets and provide a theoretical analysis of our results, highlighting novel insights into data bias.

1 Introduction

In this work, we seek to take advantage of new advances in table discovery (specifically, table discovery in data lakes (Fan et al. 2023)) to consider the problem of bias mitigation. Given a biased dataset, can we modify it to be unbiased or less biased? We do this in the context of real data where we do not want to simply modify the data to make it unbiased (for example, by changing the values associated with a tuple in a protected group). Our mitigation strategies use real data rather than synthetically altering data. We also do this in a context where we want to use the data for real data analysis. Hence, just removing tuples until we get a less biased subset that matches our fairness goal may not yield sufficient data to perform an analysis (such as training a machine learning (ML) model). In some cases, we may need to use table discovery to find new (real) data to meet our goals. An important contribution of our work is to help a data scientist explore the space of possible mitigation solutions that make the data less biased.

There are two main purposes when measuring data bias. On the one hand, we may be interested in quantifying the bias of a dataset as an assessment of its quality, given that bias provides information about the data's representativeness[1] or completeness (a well-studied data quality dimension (Batini and Scannapieco 2016)). But arguably, the fundamental practical application of such a metric is, upon detection of a significant bias, to obtain an unbiased (or less biased) dataset. We refer to algorithms that use data discovery to add tuples and/or use other table transformations (including tuple deletion or modification) to obtain a less biased dataset as bias mitigation algorithms. An underlying principle in bias mitigation is to perform a minimal change to a dataset that is sufficient to achieve a fairness goal.

1.1 Motivating Example

A major challenge in data fairness is to ensure that the dataset used for analysis has an appropriate representation of relevant demographic groups (Nargesian, Asudeh, and Jagadish 2021). This is because insufficiently representative training data has been repeatedly shown to be extremely problematic in a wide range of ML application domains (Hort et al. 2024; Pagano et al. 2023).

Consider a US Bank that decides to build an ML model for default loan prediction, i.e., predicting the probability that a person pays back a given loan. Since the bank does not have enough internal quality data, they decide to use the Adult Dataset (Becker and Kohavi 1996) for this purpose, a widely used dataset containing demographic information (14 attributes) from several thousand individuals, including an attribute indicating if a person's annual income exceeds $\$50,000$ or not. The bank decides to grant loans to people whose annual income exceeds $\$50,000$. We will call these tuples positively labeled or positive. In what follows, we consider the binary gender attribute (with values Male or Female) of this dataset. Tuples with the value Female form a *protected* group and we call them protected tuples.

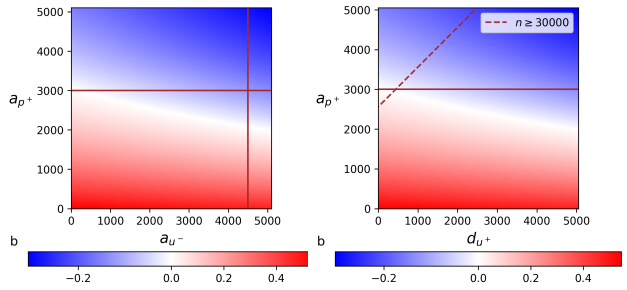
Upon inspection, the analysts realize that for the analysis they want to perform (predicting likelihood of paying back a loan), the data is biased against women: there are $1,179$ out of the $10,771$ tuples representing women that have a yearly income greater than $\$50,000$ (10.9%), while this fraction is $6,662/21,790$ for men (30.6%). The data scientists are worried that if they use this version of the data, the resulting model may grant fewer loans to women in a discriminatory way (in cases where they could indeed have paid the

[1]This can be referred to as an ideal scenario of reference.

loan back), so they aim at constructing a new version of the dataset that is group fair w.r.t. gender.

Since the total number of tuples is $n = 32,561$ and there are $p = 10,771$ women, we can modify the dataset so that $3,296$ of the women are positive (and hence at parity with the men) or we can lower the number of positive men to $2,375$, (so men are at parity with the women). Preprocessing techniques like feature normalization are common practice in ML. However, it is not acceptable for data scientists and domain experts to arbitrarily change attribute values from the tuples of the dataset. In our running example, it does not make sense to change the address, race or gender of a person. There are also cases where it is unreasonable to significantly change the salary value of an individual leaving the loan decision unchanged. Thus in many analyses, the only realistic operations are tuple additions (with new real data) and deletions. This turns even the fairness estimation task into a more challenging task as parameters like the number of tuples are no longer constant. Returning to our example, an additional requirement of the ML engineers is to have $n \geq 30,000$ for the learned model to have reasonable performance, hence only deleting tuples is not an option.

After analyzing the available data sources (open source data lakes as well as data available through data brokers) the team determines that they can find at most $3,000$ positive protected tuples and $4,500$ negative unprotected tuples. Using our approach, the data scientists are able to simulate different scenarios to determine, given the number of added and deleted tuples of both types, how far different mitigation strategies (different tuple additions and deletions) end up from creating a fair dataset. The results are shown in Figure 1, where $b = 0$ indicates that the resulting dataset is group fair (white region in Figure 1).



(a) Adding positive protected and unprotected tuples.

(b) Adding positive protected and deleting negative unprotected tuples.

Figure 1: Adult Dataset - Policies using two operations.

In Figure 1a, we model adding positive protected tuples (a_p^+, depicted on y-axis) and adding negative unprotected tuples (a_u^-, depicted on x-axis); while on Figure 1b we instead consider deleting positive unprotected d_u^+ for the second variable. In these figures, white indicates group fairness. Red indicates a bias against women (as in the original dataset) and blue is a bias against men. Here, the horizontal line $a_p^+ = 3,000$ represents the constraint that there are at most $3,000$ new positive protected tuples that can be added. The vertical line $a_u^- = 4,500$ represents the same for negative unprotected tuples. The dotted diagonal line in Figure 1b represents the restriction of needing at least $30,000$ tuples in the resulting dataset. Thus, in Figure 1a, the feasible region (datasets we can construct under the problem's restrictions) is the large lower left quadrant that includes the origin, while for Figure 1b it is the small triangular shape including the point $(0, 3000)$. Interestingly, Figure 1a shows how adding negative unprotected tuples (a_u^-, men who earn less than 50K) reduces the number of positive protected tuples required (a_{p^+}), but in a non-obvious way. For example, by adding $4,500$ negative unprotected tuples, we reduce the number of protected positive tuples needed to $2,077$. Figure 1b shows that by deleting a few of the positive unprotected tuples, we can add fewer than the available $3,000$ new positive protected tuples, but the space of options is much more limited. In this paper, we introduce mathematically sound methods to perform this process in an efficient and interpretable way.[2]

1.2 Contributions

The contributions of our work are summarized as follows:

- We introduce Uniform Bias (UB), the first intersectional, multi-label bias measure that can be computed directly from a given dataset T. Our measure is interpretable and can model bias in datasets with multiple classes (multi-label) and with multiple protected attributes.
- In contrast to the hundreds of papers surveyed by Hort et al. (Hort et al. 2024), we formally define bias for multi-class problems with multiple non-binary protected attributes.
- We present a new mitigation strategy to address data bias. Our strategies are interpretable and explainable. They can take advantage of new work on table discovery to create datasets that meet a fairness goal. Our mitigation algorithm guarantees a label frequency preservation property for the protected groups, meaning that the ratios of tuples with a given label (*e.g.,* people who are given a loan) is the same after bias mitigation is applied, which we argue is relevant for practical applications.
- We show how UB solves existing issues of anti employment discrimination rules used by the US Office of Federal Contract Compliance Programs.
- We evaluate our techniques on real datasets (recommended by a recent comprehensive survey (Hort et al. 2024)) and show that our mitigation strategies can be used in practice to produce unbiased training data that do not significantly lower the accuracy of ML models, and in some cases improves their performance.

The rest of this paper is organized as follows. In Section 2 we detail related work. In Section 3 we present the notation. In Section 4 we introduce Uniform Bias our new bias measure. Section 5 motivates the significance of Uniform Bias and how it can be used to solve open problems identified in the literature. In Section 6 we extend our approach to

[2]The analytical solution of this example is included in the full version of this work (Scarone et al. 2024).

be intersectional and to consider non-binary labels. Section 7 uses these ideas for bias mitigation in this new context. Section 8 extends our techniques in ways that are relevant in practice, while in Section 9 we evaluate how our mitigation affects the performance of ML models. We end with our conclusions and future work in Section 10.

2 Related Work

Nowadays, the concept of *bias* is prevalent in computer science literature. However, although the seminal paper on this topic is from 1996 (Friedman and Nissenbaum 1996), ten years ago this was not the case. Bias is used in heterogeneous settings, where the underlying meaning depends on the context. In some of these contexts, such as statistics, the definition is rigorous, while in others (*e.g.,* data mining, machine learning, web search and recommender systems) it is either handled informally or without a consensus within the broader community. This is in part due to the complexity and interdisciplinary nature of the problem, as noted by Zliobaitè (Žliobaitė 2017). The common idea across all these notions is that of a *systematic deviation from a predefined reference value*. A closely related concept is that of fairness, defined as the absence of (negative) discrimination.

Measuring Bias In the context of data mining and ML, where one of the main goals is to design fair models for the task at hand, researchers have proposed a variety of heuristic measures (Mehrabi et al. 2021) with the objective of quantifying bias, and thus being able to design new algorithms that would optimize these measures (Žliobaitė 2017). These measures can be classified into individual and group/statistical measures.[3] Individual fairness (introduced by Dwork et al. (Dwork et al. 2012)) centers on the idea that similar individuals should be treated similarly.[4] On the other hand, group fairness focuses on treating different demographic groups equally. The heterogeneous unprincipled nature of the measures and evaluation approaches makes it difficult to compare results (as recent examples see Osoba et al. (Osoba et al. 2024) and Lum et al. (Lum, Zhang, and Bower 2022)), as well as to establish guidelines for practitioners and policymakers. Verma et al. (Verma and Rubin 2018) collects several individual and group measures and computes each of them on a case-study intuitively explaining why the same case can be considered fair according to some definitions and unfair according to others. Yeh et al. (Yeh et al. 2024) focuses on two common families of statistical measures (ratio and difference based), theoretically analyzing their relationship and providing empirical results to establish initial guidelines for which can be used in different contexts. With the goal of providing a unifying view, Zliobaitè (Žliobaitė 2017) surveys and categorizes various statistical measures, experimentally analyzes them, and recommends which ones to use in different contexts. Its foundational notions consist of precisely defining the condition for a dataset to be *unbiased*, as well as a (randomized) bias addition algorithm. It is important to note that in this context, statistical measures are used as a tool to indirectly quantify the effect of added bias on tables.

Intersectionality Crenshaw introduced the term intersectionality (Crenshaw 2013), highlighting that the discrimination experienced by Black women is greater than the sum of racism and sexism individually and thus establishing that any bias analysis that does not take the intersection of sensitive groups into account cannot sufficiently address the particular manner in which individuals experience discrimination. In the context of data fairness, this translates into the fact that it is not enough to verify fairness for sensitive attributes independently. In fact, we can construct datasets that are group fair w.r.t. to gender and race, but not w.r.t. their intersections. This was also shown in the context of fair classification by Kearns et al. (Kearns et al. 2019). Wang et al. (Wang, Ramaswamy, and Russakovsky 2022) remark that research in fair ML has historically considered a single binary demographic attribute, and they study how to address intersectionality from a practical view in a ML pipeline.

Bias Mitigation Methods In a recent survey, Hort et al. (Hort et al. 2024) provide a comprehensive analysis of 341 publications concerning bias mitigation methods for ML classifiers. The authors identified three types of bias mitigation methods: Pre-processing, where mitigation is applied to the training data to prevent it from reaching ML models; In-processing, where mitigation is done while training the models and Post-processing, where it is done on trained models. Our method falls into the pre-processing category. The authors remark that there are almost twice as many publications with in-processing methods than pre-processing and as Salimi, Howe, and Suciu (Salimi, Howe, and Suciu 2020) highlight, one of the advantages of these methods is that they can be used in conjunction with any ML model.

Hort et al. remark that consolidating a common set of metrics is still an open challenge. Another related problem they highlight is to make sure that the metrics used are representative for the problem at hand and they state that future work should focus on multi-class problems, and non-binary sensitive attributes, which was mentioned by only 15 out of the 341 publications they considered. Out of these 15 publications, only one (Alabdulmohsin, Schrouff, and Koyejo 2022) can deal with both non-binary sensitive attributes and multi-class predictions, which is what we do with our UB (Uniform Bias) measure and our mitigation algorithm.

Another challenge identified by Hort et al. is to include trade-offs when dealing with accuracy and/or multiple fairness metrics. We evaluate our model on real datasets and show that our mitigation strategies can be used to produce training data that yield better performing ML models. In particular, we use the first two and seventh most widely used datasets according to Table 9 in (Hort et al. 2024).

Providing fairness guarantees is considered a relevant avenue of future work, stating that allowing for interpretable and explainable methods can aid in this regard. We provide mathematical guarantees in the most general case for both the interpretation of UB and our mitigation algorithm. Additionally, our techniques allow a user to specify acceptable or allowed levels of bias (*e.g.,* determined by domain experts)

[3] Sometimes the definition only specifies the fairness/unbiased condition without quantifying the deviation w.r.t. this value.

[4] This includes Counterfactual or Causal based fairness (Salimi, Howe, and Suciu 2020; Plecko and Bareinboim 2022).

in a natural way, which is also highlighted as one of the challenges by Hort et al. The survey also notes the importance of having continuous implementations of bias mitigation methods in real-world scenarios. Our methods are simple and can be naturally implemented to continuously monitor the bias level of data and compute mitigated versions.

Our Approach We present a new intersectional data bias measure that can directly be computed from the target table and argue why it improves the state-of-the-art. Namely, UB is the first measure that can handle an arbitrary number of multi-valued sensitive attributes and a multi-valued label in a simple way. The same ideas used to derive UB can serve as the basis for explainable and mathematically guaranteed bias mitigation strategies that can be computed for any dataset. We show how our bias mitigation produces new versions of well known datasets that can be used to improve the performance of ML models.

3 Preliminaries

As is common, we denote variables (*i.e.,* , dataset attributes) by uppercase letters, X, Y, Z; their values with lowercase letters, x, y, z; and denote vectors (or sets) of variables (values) using boldface (**X** or **x**). The domain of variable X is $Dom(X)$, the domain of a vector of variables is $Dom(\mathbf{X}) = \prod_{X \in \mathbf{X}} Dom(X)$. To simplify notation in formulas, when we use a value **x** we mean $|\mathbf{x}|$, which is the number of tuples from the dataset with value **x** and we omit the attribute **X** when it is clear from context.

Bias Notation We start by considering the classical setting of algorithmic fairness where we have a classification task (*e.g.,* hiring men and women candidates), where tuples have a target (class) attribute Y with a binary domain ($y \in \{0, 1\}$). In the literature, the values of Y are often called *binary target labels.* In addition, tuples have a binary *sensitive or protected attribute* S defining a protected (or unprivileged) group $s = 1$ and an unprotected (privileged) group $s = 0$. This simplified scenario has been extensively studied in the fair machine learning literature (Žliobaitė 2017; Wang et al. 2023; Hort et al. 2024). For this reason, Table 1 includes the notation for this binary scenario.

However, our work naturally extends to multi-class problems ($dom(Y)$ can be any set of values) and to account for multiple sensitive attributes (**S**) with non-binary values ($dom(\mathbf{S})$ can be any set of values). Let $|\mathbf{S}| = m$, then $\mathbf{s} \in (dom(S_1) \cup \{\epsilon\}) \times ... \times (dom(S_m) \cup \{\epsilon\})$ denotes a possible protected group. Here, ϵ represents all values in a given domain. For example, for attributes gender = {male, female, non-binary}, and age = {young, middle-aged, retired}, one possible protected group (non-binary, retired) represents all tuples that are both non-binary and retired, while another possible protected group (male, ϵ) represents all men independent of age. We drop the parenthesis and the ϵ when the context is clear, *i.e.,* use "male" instead of (male, ϵ) to represent the group of males. Table 1 also defines notation for this more general multi-class, non-binary case. All quantities refer to a given dataset T. Note that our measure, given a sensitive group s (or **s** in the non-binary case) and a class label y, can be computed only based on the table T, that is, our bias measure $\text{UB}(T, \mathbf{s}, y)$ as defined in Section 4. This is also the case for other preexisting measures we reference later and that are included in Table 1 (IR, OR and MD).

Variable	Definition
n	Total number (#) of tuples ($\|T\|$)
n^+	For binary problems, # positive tuples, $\{t \in T : t[Y] = 1\}$
n^y	For multi-class problems, # tuples with a given class label y: $\{t \in T : t[Y] = y\}$
p	For binary problems, # tuples in the protected group, $\{t \in T : t[S] = 1\}$ (resp., for unprotected group, u)
p^+	For binary problems, # positive protected tuples, $\{t \in T : t[Y] = 1 \wedge t[S] = 1\}$ (respectively, u^+)
P	Protected ratio, p/n (resp., $U = u/n$)
$f_{p,+}$	p^+/p, (respectively, for u)
f_y	n^y/n
$f_{\mathbf{s},y}$	Ratio of # tuples with protected attributes **s** and class label y over all tuples in the protected group **s**: $\|\mathbf{s}y\|/\|\mathbf{s}\|$
$\lfloor i \rceil$	Closest integer to i, $\lfloor i + 0.5 \rfloor$
IR	$\frac{f_{p,+}}{f_{u,+}} = \frac{p^+ \cdot (n-p)}{p \cdot (y^+ - p^+)}$
OR	$\frac{1-f_{p,+}}{f_{p,+}} \cdot \frac{f_{u,+}}{1-f_{u,+}} = \frac{u^+/u^-}{p^+/p^-}$
MD	$f_{u,+} - f_{p,+}$

Table 1: Notation used in this paper.

4 Uniform Bias

In this section, we present an example to motivate and introduce our new bias measure. Consider the setting where a company hires $n^+ = 200$ new employees from a set of $n = 600$ people, where $p = 150$ applicants are women (protected) and $u = 450$ are men (unprotected), thus gender is the protected attribute (S).

Suppose that after this selection process concludes, we have access to the corresponding tabular data showing the distribution of accepted and rejected candidates. There are three possible scenarios of interest, whose summary statistics are shown in Table 3:[5]

- T_0: The proportion of hired women ($f_{p,+}$) and men ($f_{u,+}$) are equal and coincide with the fraction of positive tuples of the population (f_+). Here, $f_{p,+} = f_{u,+} = f_+$, and we say T_0 is **unbiased**, following what is common practice in the algorithmic fairness community (Mehrabi et al. 2021);
- T_1: The proportion of hired women ($f_{p,+}$) is **lower** than that of men ($f_{u,+}$). Here, $f_{p,+} < f_+ < f_{u,+}$, thus T_1 exhibits a **negative bias** against women (protected group).
- T_2: The proportion of hired women ($f_{p,+}$) is **higher** than that of men ($f_{u,+}$). Here, $f_{u,+} < f_+ < f_{p,+}$, and so T_2

[5]Note that if we only consider the proportions (f, f_p and f_u), then we can associate each row in Table 3 to the set of tables whose parameters satisfy the proportions, which makes the analysis much more general (*e.g.,* independent of n).

Dataset	n	n^+	p	p^+	$p^+(0)$	UB	IR	OR	MD
Adult (Becker and Kohavi 1996)	32561	7841	10771	1179	2594	0.55	0.36	3.58	0.2

Table 2: Summary statistics of the Adult dataset.

	p^+	u^+	$f_{p,+}$	$f_{u,+}$	UB	IR	MD
T_0	50	150	.33	.33	0	1	0
T_1	40	160	.27	.36	.2	.75	.09
T_2	60	110	.4	.24	$-.2$	1.64	$-.16$

Table 3: Summary statistics and measures for three classes of tables. For all rows we have: $n = 600$, $y^+ = 200$, $p = 150$, $p^+(0) = 50$, $u^+(0) = 150$ and $f_+ = 1/3$.

exhibits a bias against men (unprotected group) or equivalently a **positive bias** in favor of women.

Now we introduce ideas to quantify bias from the viewpoint of a specific protected group, using the notation defined in Table 1. Consider the case of T_1. Table T_0 gives the ideal number of hires, which we call $p^+(0) = 50$. Ideally we would have wanted for $p^+(0) = 50$ women to get hired, but only $p^+ = 40$ were. Our goal is to quantify this bias, which is clearly related to the difference or deviation of 10 from the unbiased state T_0. We start by observing that this difference is 20% of the target quantity $p^+(0) = 50$, *i.e.,* 20% less women are being hired than desired. Thus, if we take b (the bias) to be this percentage we get $p^+ = p^+(0) - b \cdot p^+(0)$ with $b = 0.2$. If $b = 0$ we have unbiased data and if $b = 1$ negative bias against women is maximized. As stated before, representing these ideas in terms of ratios will be useful. Thus, we divide both sides of the last equation by p and noting that when $b = 0$, then $f_{p,+} = f_+$ (unbiased condition), we derive the following expression $f_{p,+} = (1-b) \cdot f_+$. Solving for b, we have a measure of the percentage of missing elements (20% for T_1). These ideas are the building blocks for our proposed formal bias definition.

Definition 4.1. *Given any table T, its* Uniform Bias *(UB) w.r.t. the group $s \in \{p, u\}$ and label $y \in \{+, -\}$ is given by*

$$UB(T, s, y) = b_{s,y} = 1 - \frac{f_{s,y}}{f_y} .$$

Observe that the right side can be directly computed based on the data, that it is linear w.r.t. $f_{s,y}$ and if $f_{s,y}$ is proportionally greater than (or less than) f_y by the same about, the bias will be the same but negated (for example, 20% or -20%). Additionally, note that by writing the definition in terms of relative table quantities, all tables with the same values of $f_{s,y}$ and f_y will result in the same UB, regardless of their size. For brevity, we use the symbol b when the context (T, s, y) is clear.

Note that, as discussed before, when $b_{p,+} = 0$ we have $f_{p,+} = f_+$, when $b_{p,+} > 0$ we have $f_{p,+} < f_+$ and when $b_{p,+} < 0$ we have $f_{p,+} > f_+$.

We say UB is uniform because it is not context dependent and the measure will be the same for any dataset T with equal relative parameters ($f_{p,+}$ and f_+). Finally, as remarked in the literature (Verma and Rubin 2018), in practice one does not expect a fair dataset to have a bias that is exactly zero. In this sense, one can determine an admissible range of bias values determined by domain experts (*e.g.,* $|b| \leq .1$) to declare a table to be sufficiently unbiased or fair.

5 Comparing UB to Existing Measures

The work by Gastwirth (Gastwirth 2021) centers around rules issued by the US Office of Federal Contract Compliance Programs[6] (OFCCP) in November 2020 to resolve employment discrimination issues. In this context, Gastwirth presents an in-depth analysis of how the agency will use and evaluate statistical evidence in its monitoring of government contractors' compliance with equal employment laws. The rules state that the agency will ordinarily use the impact ratio as its measure of practical significance and uses what is known as the "fourth-fifths rule" (used since 1970) to detect violations. As explained by Oswald et al. (Oswald, Dunleavy, and Shaw 2016), the four-fifths rule is violated when the selection rate of one applicant group (*e.g.,* Hispanic) is less than 80% of the selection rate for the group with the highest rate (*e.g.,* White).

Gastwirth observes that while the rules develop the agency's classification system in terms of the impact ratio (IR, Table 1), they also allow it to use other measures such as the odds ratio. In terms of practical significance measures[7] (including both the impact and odds ratio), there is an extensive literature (*e.g.,* (Gastwirth 2021; Oswald, Dunleavy, and Shaw 2016) and the references therein) analyzing these measures and illustrating their flaws in the context of determining disparate impact. However, these works do not present a systematic approach to study them (*e.g.,* to precisely characterize when they do not work and why) nor to solve the identified problems.

In order to show the limitation of IR and the mean difference (MD, Table 1) in not being able to distinguish datasets that are significantly different in terms of disparate impact, we construct two summary statistics (Tables 4 and 5) that have a fix IR and MD, but for which the values of UB (denoted by b) vary significantly. We take IR and MD as representatives of ratio based and difference based measures.

In these examples, we assume that an absolute value of UB lower than 10% denotes an acceptable disparity rate (unacceptable rates are highlighted in **bold** in the tables). For Table 4, IR (0.8) indicates that disparate impact is unlikely (per the four-fifths rule, (Oswald, Dunleavy, and Shaw 2016; Gastwirth 2021)). Meanwhile, the IR (0.25) shown in Table 5 is considered to strongly suggest this kind of discrimination (as presented in (Oswald, Dunleavy, and Shaw 2016; Gastwirth 2021)). Nevertheless, in both tables there

[6]It oversees employment practices and promotes efforts to diversify the work forces of government contractors in the US.

[7]Practical significance is given more attention in the rules issued by the OFCCP than in the original proposal (Gastwirth 2021).

p^+	p	u^+	u	f_+	$p^+(0)$	b
396	990	5	10	.401	396.99	.0025
388	970	15	30	.403	390.91	.0074
360	900	50	100	.410	369.00	.0244
320	800	100	200	.420	336.00	.0476
232	520	210	420	.442	229.84	.0950
160	400	300	600	.460	184.00	**.1304**
40	100	450	900	.490	49.00	**.1836**
12	30	485	970	.497	14.91	**.1951**
4	10	495	990	.499	04.99	**.1983**

Table 4: Summary statistics showing constant IR and MD, while $b \in [0.25\%, 19.83\%]$. For all rows we have: $n = 1000$, $f_p = .4$, $f_u = .5$, $IR = .8$ and $MD = .1$.

p^+	p	u^+	u	f_+	$p^+(0)$	b
199	995	4	5	.203	201.985	.0148
194	970	24	30	.218	211.46	.0826
180	900	80	100	.260	234.0	**.2308**
100	500	400	500	.500	250.0	**.6000**
40	200	640	800	.680	136.0	**.7059**
20	100	720	920	.740	74.0	**.7297**
10	50	760	950	.770	38.5	**.7403**
1	5	796	995	.797	3.9850	**.7491**

Table 5: Summary statistics showing constant IR and MD, while $b \in [1.48\%, 74.91\%]$. For all rows we have: $n = 1000$, $f_p = .2$, $f_u = .8$, $IR = .25$ and $MD = .6$.

are cases that, according to UB, either present strong discrimination or barely any discrimination at all. In our view, it is essential to provide a rigorous explanation of this.

For constructing these tables we start by fixing $f_{p,+}$ and $f_{u,+}$, which in turn determine the values of IR and MD. For a fix n, by varying p, the quantities p^+, u^+ and f are determined for each row. Notice that when p is large (resp. for u), most of the table is filled according to $f_{p,+}$ (resp. for $f_{u,+}$). As a consequence, when p is large, p^+ is close to the optimum $p^+(0)$ and so b is small. On the other hand, when p decreases (u increases), since $f_{u,+} > f_{p,+}$, p^+ gets further away from $p^+(0)$ and thus b increases. Recall we define the protected ratio $P = p/n$. Given that $IR(b) = \frac{1-b}{1+b \cdot P/U}$ (Table 1), it is easy to see that when $P \to 1$, $b \to 0$, and when $P \to 0$ then $b \to 1 - IR$.

The key observation is that IR and MD are not sufficiently descriptive, since they do not distinguish what is happening in each row. In fact, the b values in these rows present completely different scenarios in terms of the disparate impact they describe. Precisely, in Table 4, b varies in the range $[0.25\%, 19.83\%]$, while in Table 5 this range becomes $[1.48\%, 74.91\%]$. Given the interpretation of b as p^+ being a $1 - b$ fraction from $p^+(0)$, it is crucial to develop new disparate impact measures that are able to detect these significant disparities across the rows. Our measure UB achieves this goal.

Since $f_{p,+}$ and $f_{u,+}$ are constant, this variation is due to f_+ that ranges from $f_{p,+}$ when $P \to 1$ to $f_{u,+}$ when $P \to 0$. Neither IR nor MD are sensitive to f_+, while UB is. While these flaws in IR and MD (as well as for other measures) have been extensively pointed out in the literature, to the best of our knowledge, this is the first time a systematic explanation of the flaws has been presented. Based on this explanation, we believe UB to be a more suitable measure to use for determining disparate impact when analyzing data.

f_+	$f_{p,+}$	$f_{u,+}$			
$MD = .02$			IR	OR	$b_{p,+}$
.5	.482	.502	.960	.923	.036
.1	.082	.102	.804	.786	**.18**
$MD = .05$			IR	OR	b
.5	.455	.505	.901	.818	.09
.5	.055	105	**.542**	**.496**	**.45**
$MD = .10$			IR	OR	$b_{p,+}$
.5	.41	.51	.804	**.668**	**.18**
.1	.01	.11	**.091**	**.082**	**.9**

Table 6: Example data used by (Oswald, Dunleavy, and Shaw 2016).

To further illustrate the practical impact of UB we present now a solution to the contradictory judgments identified by Oswald *et al.* in Table 5.3 of (Oswald, Dunleavy, and Shaw 2016), arising from the use of IR, OR and MD. For instance, when looking at the case in Table 6 where $MD = .2$ and $f_+ = .1$, although neither IR nor OR indicate an adverse (bias) impacting the data, UB (18%) does. Furthermore, when $MD = .10$ and $f_+ = .5$, IR and OR even disagree on their judgments. Meanwhile, UB provides additional evidence supporting the claim made using OR. Moreover, the key observation is that $b_{p,+} = 18\%$ can be interpreted as the fraction of tuples away from the optimum $p^+(0)$. As far as we are aware, none of the measures presented in the literature give a similar quantitative explainable interpretation. This example provides further evidence that UB is a good metric to be used in this context.

6 Intersectionality and Multi-class Problems

In this section, we extend our approach to consider multi-valued labels, as well as to be intersectional, *i.e.,* to consider multiple sensitive groups and their intersections in the analysis. This approach is critical when assessing fairness in real world applications (Hort et al. 2024).

Example 6.1. *The need to consider non-binary labels comes very naturally in practice; as a paradigmatic example, we will consider the COMPAS dataset (Angwin et al. 2016) containing records for US criminal offenders and a score of their likelihood to reoffend (recidivism). The scores are given using three labels (low, medium, or high), so it would be ideal to capture this with our bias measure as well. The same goes for having multiple (potentially non-binary) sensitive attributes, in this case we consider two binary sensitive attributes gender and race, taking values in* {*men, women*} *and* {*white, non-white*} *respectively. We denote men with* m*, women with* w*, white tuples with* c *("caucasian") and non-white ones with* o *(for "others"). The data is summarized in Table 7.*

	Label	o	c	Total
m	**L**	19489	12202	31691
	M	7143	2862	10005
	H	4510	1273	5783
	Tot.	mo: 31142	mc: 16337	m: 47479
w	**L**	5637	4159	9796
	M	1589	894	2483
	H	665	375	1040
	Tot.	wo: 7891	wc: 5428	w: 13319
Total	**L**	25126	16361	41487
	M	8732	3756	12488
	H	5175	1648	6823
	Tot.	o: 39033	c: 21765	n: 60798

Table 7: Summary statistics (values) for initial version of the COMPAS dataset with ternary label.

One of the main benefits of our measure is that extending it to non-binary labels and multiple sensitive attributes is natural, as we can see in Definition 6.1.

Definition 6.1 (Uniform Bias, multi attributes, general label)**.** *Given a dataset T with sensitive attributes $\boldsymbol{S}$ we say that the Universal Bias of group $\boldsymbol{s} \in Dom(\boldsymbol{S})$ w.r.t. label $y \in \{y_1, \ldots, y_k\}$ is*

$$b_{\boldsymbol{s},y} = 1 - \frac{f_{\boldsymbol{s},y}}{f_y} \ .$$

Note how Definition 6.1 naturally leads to the generalized version of an unbiased dataset, given in Definition 6.2.

Definition 6.2. *In the setting of Definition 6.1, we say that T is unbiased w.r.t. $\boldsymbol{S}$ and label $y \in \{y_1, \ldots, y_k\}$ if for every $\boldsymbol{s} \in Dom(\boldsymbol{S})$ we have $f_{\boldsymbol{s},y} = f_y$.*

Remark. *Recall our notational convention explained in Section 3 on the use of the value ϵ for a sensitive attribute.*

Example 6.2. *In our COMPAS example, we have eight possible protected groups: four that only consider one attribute $\{m, w, c, o\}$, and four binary $\{mc, mo, wc, wo\}$. For each group and label, the data is unbiased if the frequency of the group having that label is equal to the frequency of the entire population having that label.*

Note that as before having $b_{\mathbf{s},y} = 0$ for all groups $\mathbf{s}$ and label values is equivalent to the unbiased condition stated in Definition 6.2. An important feature of UB is that we can analyze the bias of each group (and label) separately. To illustrate the usefulness of our techniques, we will use Uniform Bias to analyze the biases in the COMPAS dataset.

Example 6.3. *The values and biases of the COMPAS dataset are shown in Table 7 and Table 8 respectively.*

*This dataset is known to have multiple disparities in the treatment of different sensitive groups. This is the case for men and women, where the ratio of women with a low risk score is substantially greater than the ratio of men (*i.e., *$f_{w,L} > f_{m,L}$). It is also the case that $f_{w,M} < f_{m,M}$ and $f_{w,H} < f_{m,H}$. This case can be numerically quantified with Uniform Bias, $b_{m,L} = .022 > 0$ (bias against men) and $b_{w,L} = -.078 < 0$ (bias in favor of women). There is also a noticeable difference when the race attribute is considered: the magnitude of the bias of non-white people with a high score is half that of the one among white people ($b_{o,H} = -.181$ and $b_{c,H} = .325$). Note how our measure naturally captures what happens in terms of the frequencies: the rate of people with a high score among the non-white population is double the rate among white people (using the values in Table 7, $f_{c,H} = .076$ and $f_{o,H} = .133$). We observe a difference in terms of the gender attribute as well, namely $b_{w,H} = .304$ and $b_{m,H} = -.085$ ($f_{w,H} = .078$ and $f_{m,H} = .122$).*

	Label	o	c	Total
m	**L**	0.083	-0.095	0.022
	M	-0.117	0.147	-0.026
	H	-0.290	0.306	-0.085
w	**L**	-0.047	-0.123	-0.078
	M	0.020	0.198	0.092
	H	0.249	0.384	0.304
Total	**L**	0.057	-0.102	0
	M	-0.089	0.160	0
	H	-0.181	0.325	0

Table 8: Summary statistics (biases) for initial version of the COMPAS dataset with ternary label.

There is a consensus that the difference in the race attribute constitutes actual discrimination (Hort et al. 2024). This may not be the case for the difference found for the gender attribute (not reported for this dataset). This case could either be a true social phenomenon that is not to be corrected or discrimination. Within our framework, this can be done using external sources to, for example, determine that in a certain population $b_{w,H}$ may not be zero. In this example, sociologists may decide a value of 0.304 is a reasonable societal norm and therefore this dataset is not biased for this class. This is the type of decision an expert in the domain should make. Our mathematical methods are designed as a tool to help experts make these decisions. The mitigation algorithms we define next can be used to create a dataset with zero bias for a specific group or with a bias that is *a priori* set by experts. We will extend our techniques to be able to model these alternatives in Section 8.

7 General Data Bias Mitigation

We introduce our general mitigation algorithm, prove its correctness and use it to mitigate the COMPAS dataset.

Let T be a dataset with sensitive attributes $\mathbf{S}$ and label Y with values $y \in \{y_1, \ldots, y_k\}$. In the context of bias mitigation, we want to add or delete tuples to reduce the data bias as much as possible. We note that one can always consider tuple deletions as a preprocessing step, *i.e.,* first delete some of the tuples and then measure bias and determine the new tuples that should be added for mitigation. Thus, we will only consider tuple additions in this context. Nevertheless, we can consider both additions and deletions simultaneously with our framework. We denote the number of tuples to be

added from group $\mathbf{s}$ with label value y by $\Delta\mathbf{s}y$. The key idea for the algorithm comes from using Definition 6.2, since we want to determine $\Delta\mathbf{s}y$ such that this definition holds, we can write this as

$$f^{\text{new}}_{\mathbf{s},y} = \frac{\mathbf{s}y + \Delta\mathbf{s}y}{\mathbf{s} + \Delta\mathbf{s}} = f_y \Leftrightarrow \frac{\mathbf{s}y + \Delta\mathbf{s}y}{f_y} = \mathbf{s} + \Delta\mathbf{s} \quad (1)$$

where $f^{\text{new}}_{\mathbf{s},y}$ is the new value of $f_{\mathbf{s},y}$. This condition must hold for every group $\mathbf{s}$ and label value y, so we have a system of equations, where the unknowns are the $\Delta\mathbf{s}y$. The label frequency preservation that we introduce later is a direct consequence of the equation on the left. Then since the equality on the right in Equation (1) holds for every y, we can write

$$\frac{\mathbf{s}y + \Delta\mathbf{s}y}{f_y} = \frac{\mathbf{s}y_i + \Delta\mathbf{s}y_i}{f_{y_i}} \quad (2)$$

for every group $\mathbf{s}$ and label $y \in \{y_1, \ldots, y_k\}$. That is, we equate the equations of group $\mathbf{s}$ and label y with the one corresponding to the same group and y_i. Since when $y = y_i$ the equation is trivial, $\Delta\mathbf{s}y_i$ becomes a free variable of the system and we have precisely one free variable per group, making the system of equations undetermined (the system has fewer equations than unknowns). Solving for the unknown we are looking for we get, $\Delta\mathbf{s}y = -\mathbf{s}y + \frac{y}{y_i}(\mathbf{s}y_i + \Delta\mathbf{s}y_i)$. This may yield non-integral solutions, which do not make sense in our problem (our variables represent number of tuples). Thus, we need to find the exact integral solution or good approximations of it. The solution is integral when, in the second term, the second factor is a multiple of y_i.When $\mathbf{s}y_i \neq 0$ the smallest positive solution occurs when $k = 1$, *i.e.,* $\Delta\mathbf{s}y_i = y_i - \mathbf{s}y_i$. When considering approximate solutions, natural options include using the floor, closest, or ceiling operator on the second term. Note that this would allow us to choose **any** integral value of $\Delta\mathbf{s}y_i$, but we may introduce an error when doing so. We use the floor operator, since we want to minimize the total number of additions, leading to Theorem 7.1.[8]

Theorem 7.1. *Given a dataset T with sensitive attributes $\boldsymbol{S}$, label Y with values $y \in \{y_1, \ldots, y_k\}$ and index $i = \arg\max_j \frac{\boldsymbol{s}y_j}{y_j}$. A general approximate solution for the system given by Equation (2) for every group $\boldsymbol{s}$ and label value y is as follows*

$$\Delta\boldsymbol{s}y = -\boldsymbol{s}y + \left\lfloor \frac{y}{y_i}(\boldsymbol{s}y_i + \Delta\boldsymbol{s}y_i) \right\rfloor$$

for integral $\Delta sy_i \geq 0$. Once values are given to $\Delta\boldsymbol{s}y_i$ (free variables of the system, one per group), this solution determines the number of tuples that need to be added to each sensitive group to produce a mitigated dataset T_m that is unbiased. Specifically, T_m will contain $\Delta\boldsymbol{s}y$ new tuples from group $\boldsymbol{s}$ with label y.

Example 7.1. *For COMPAS, one such solution is shown in Table 9. Using our formulas we can check that this new version is unbiased for all sensitive groups. Note how in order to mitigate the gender bias, we needed to add more men than women ($\Delta m = 10,592$ and $\Delta w = 1,037$). However, when disaggregating the groups, we see that the majority of added men have a low risk score ($\Delta mL = 7,935$) and all added women have medium or high scores ($\Delta wM = 466$ and $\Delta wH = 571$). When also considering race, we see that all but 1 added white men have medium or high risk scores ($\Delta mcM = 811$ and $\Delta mcH = 734$) and that the majority of added non-white males have a low risk score ($\Delta moL = 7,934$). This showcases the importance of analyzing tuple additions for the different intersectional groups.*

[8] All proofs are included in the full version (Scarone et al. 2024).

	Label	o	c	Total
m	**L**	27422	12202	39624
	M	8254	3672	11926
	H	4510	2006	6516
	Tot.	mo: 40186	mc: 17880	m: 58066
w	**L**	5637	4159	9796
	M	1696	1251	2947
	H	927	683	1610
	Tot.	wo: 8260	wc: 6093	w: 14353
Total	**L**	33059	16361	49420
	M	9950	4923	14873
	H	5437	2689	8126
	Tot.	o: 48446	c: 23973	n: 72419

Table 9: Summary statistics (values) for a mitigated version of the COMPAS dataset without deletions. Note that this table is unbiased for all eight groups, *i.e.,* $f_{\mathbf{s},L} = f_L = .682$, $f_{\mathbf{s},M} = f_M = .205$ and $f_{\mathbf{s},H} = f_H = .112$.

A solution for the equations in Theorem 7.1 determines a number of tuples of each group to make T unbiased. But what is the minimal solution, *i.e.,* a solution with the least number of additions? Corollary 7.1.1 gives the answer.

Corollary 7.1.1. *In the same setting as Theorem 7.1, the least number of tuple additions for every group $\boldsymbol{s}$ and label value y for making T unbiased is given by*

$$\Delta\boldsymbol{s}y + \boldsymbol{s}y = \left\lfloor y\frac{\boldsymbol{s}y_i}{y_i} \right\rfloor \Leftrightarrow \boldsymbol{s}^{mit} = \frac{\boldsymbol{s}y_i}{y_i}n + C,\ 0 \leq C < k$$

where $\boldsymbol{s}^{mit}$ is the total number of tuples belonging to group $\boldsymbol{s}$ in the mitigated dataset and n is the total number of tuples in the initial version. Note how C is a small constant bounded by the number of labels k.

Example 7.2. *In fact, Table 9, is a minimal solution. When mitigating this dataset with our minimal solution, we have for $\boldsymbol{s} = mo$, $y_i = H$ and for the rest of the groups (mc, wo and wc) $y_i = L$. Note how $wo^{mit} = \frac{woL}{L}n = \lfloor (5637/41487) \cdot 60798 \rfloor = 8260$.*

Now we present bounds for the error of the floor approximate solution given in Theorem 7.1.

Theorem 7.2. *The output f^{new}_{sy} of our approximation algorithm satisfies the following bounds*

$$\left(1 - \frac{1}{y\frac{\boldsymbol{s}y_i}{y_i}}\right) \cdot f_y \leq f^{new}_{sy} \leq \left(1 - \frac{k}{n\frac{\boldsymbol{s}y_i}{y_i}}\right) \cdot f_y$$

where k is the number of distinct labels. Note that with the exact solution we get $f_{sy}^{new} = f_y$.

Since typically in data lakes or enterprise database tables y and n tend to be large numbers (*e.g.,* at least 1,000 or 10,000) and k is small, both constants are close to 1 and thus these bounds are tight. In our full version (Scarone et al. 2024), we verify the correctness of our bounds for the COMPAS dataset and show that the errors of the floor solution have at least an order of magnitude of 10^{-5}, which showcases the quality of our approximation. Another surprising feature of our approximate solution is the required number of tuples, the exact solution uses $182,394$ tuple additions, while the floor approximation only uses $11,621$ (6.37%).

7.1 Label Frequency Preservation Property

There are different statistical properties that a data scientist could want to preserve when altering the data in the mitigation process. An intuitive property that is easily motivated by the COMPAS example is the preservation of the label frequencies, hence this is the one we guarantee with our algorithm and prove in Theorem 7.3. Note that this property provides a mathematical guarantee for the mitigated data to be used in the same context as the original data. Most ML algorithms and statistical methods rely on assumptions about the input distribution, so it is of key importance to preserve the representativeness of the data as much as possible for these downstream tasks.

Theorem 7.3. *The mitigation algorithm given by Theorem 7.1 preserves the label frequencies among the total population (f_y values) from the original dataset.*

Example 7.3. *Recall the values (Table 7) for the COMPAS dataset. We have $f_L = 41487/60798 \approx .682$,, $f_M \approx .205$ and $f_H \approx .112$. Note that in the mitigated version (Table 9) these frequencies are preserved (*e.g., *$f_L = 49420/72419 \approx .682$).*

8 Generalizing the Unbiased Condition

Now we extend our framework to permit the unbiased condition to be non-zero. As mentioned in Section 6, while it is common to define a zero bias condition, this is not always the desired unbiased goal in practice. The reason is simple: a domain expert may consider that a given bias, for example the difference between women and men in Example 6.3, represents a social behavior that does not constitute discrimination. As a consequence, a mitigation algorithm should take this fact into consideration. We therefore generalize our unbiased conditions to allow a data scientist to specify a number $K_{\mathbf{s},y} > 0$ for each group $\mathbf{s}$ and label value y, which is the desired ratio of the success rate of the protected group to the success rate of the population:

$$\frac{f_{\mathbf{s},y}}{f_y} = K_{\mathbf{s},y}$$

To recover the original setting (where bias of 0 is the goal), we can set $K_{\mathbf{s},y} = 1$. For the selected values of $K_{\mathbf{s},y}$, the following identity holds $\sum K_{\mathbf{s},y} f_y = 1$.

Example 8.1. *In Table 10, we show the summary statistics of a mitigated version of the COMPAS dataset with gender and race protected attributes with values $\mathbf{s} = \langle g, r \rangle$ and the following K values: $K_{gr,y} = \frac{f_{g,y}}{f_y}$, chosen such that*

$$f_{gr,y} = \frac{gry + \Delta gry}{gr + \Delta gr} = f_{g,y}$$

*meaning that after mitigation is applied the proportion of any gender group is the same as prior to mitigation. Note that if $g = \epsilon$ (*i.e., *we only consider race), then $K_{gr,y} = f_y / f_y = 1$ as before.*

	Label	o	c	Total
m	**L**	24714	12202	36916
	M	7802	3852	11654
	H	4510	2226	6736
	Tot.	mo: 37026	mc: 18280	m: 55306
w	**L**	6268	4159	10427
	M	1588	1054	2642
	H	665	441	1106
	Tot.	wo: 8521	wc: 5654	w: 14175
Total	**L**	30982	16361	47343
	M	9390	4906	14296
	H	5175	2667	7842
	Tot.	o: 45547	c: 23934	n: 69481

Table 10: Summary statistics (values) of a mitigated version of the COMPAS dataset with ternary labels using constants $K_{\mathbf{s},y} \neq 1$.

In the full version (Scarone et al. 2024), we incorporate costs (or weights) on tuples of different groups and budgets into the mitigation algorithm.

9 ML Based Evaluation

In this section, we train six ML models on the mitigated and biased versions of three datasets to see how our data bias mitigation strategies affect ML performance.

Methodology In order to simulate the setting in which we have a biased dataset T and we want to collect external data (*e.g.,* from an open data lake) to mitigate a bias T has, we partition T into two sets uniformly at random: an "initial sample" of size x_0 and the remaining portion that we consider to be the external available data. Similarly to Salimi et al. (Salimi, Howe, and Suciu 2020), we then generate mitigated versions of the data set using different mitigation strategies if possible. We distinguish between two types of mitigation strategies: a strategy that preserves the initial label frequencies of the data and one that does not. Then, for each mitigated dataset, we generate a uniform sample of the same initial size. We train a set of ML models on both samples and evaluate their accuracy, precision and recall. Training is done using a random 80% portion of the dataset for training and the remaining 20% for evaluation.

Datasets & Models used We use three widely known datasets for our experiments: the Adult dataset (Becker and Kohavi 1996) introduced in Section 1, the Default dataset

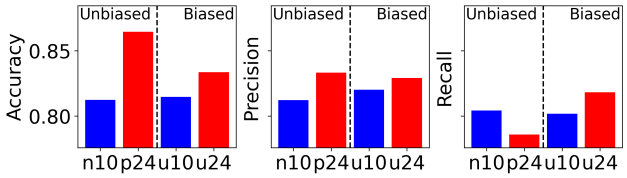


Figure 2: Adult dataset - GBDT ML model.

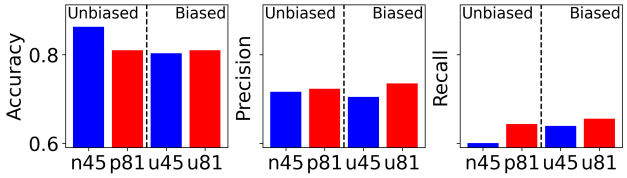


Figure 3: Default dataset - Extra Trees ML model.

(Yeh 2016) containing information about default payments of credit card clients in Taiwan and the COMPAS dataset (Angwin et al. 2016) introduced in Section 6. We apply the standard preprocessing techniques: we normalize numeric features and use one-hot encoding for categorical features.[9] In terms of ML models, we use Random Forest, Gradient Boost Decision Tress (GBDT), Extra Trees, Ada Boost, Multilayer Perceptron (MLP) and Logistic Regression. We use the open source ML library Scikit-learn[10] to implement our learning algorithms.

Experimental results We report results for one (different) model for each dataset, as the other experiments exhibit similar results. We segment the charts into two regions, the left containing the results for the unbiased versions of the dataset and the right the biased samples. We use the label "uX" for uniform samples of the data of size $X\%$, the label "nX" for a mitigated sample that does not preserve the initial label frequencies of the data and "pX" for the samples that do. For the Adult dataset (Fig. 2), the accuracy when using a mitigated dataset that preserves the label frequencies is significantly higher. Precision increases slightly and Recall decreases slightly. For the Default dataset (Fig. 3), similar accuracy is observed for all measures, only the accuracy for the mitigated version that does not preserve label frequencies is a bit higher. Surprisingly good results are observed for the COMPAS dataset (Fig. 4), where we consider ternary labels (instead of making them binary as is usually done in previous work (Hort et al. 2024)), using a mitigation that preserves the initial label frequencies of the data. In this case, the performance of the models is always consistently better than that of a uniform sample for all measures.

For the binary label case, when comparing our results to those reported by Salimi et al. (Salimi, Howe, and Suciu 2020) we can eliminate **all** the data bias without significant performance losses. What is more surprising is that for the case of ternary labels, we even improve the performance of all ML models when mitigating the bias. Note that although

[9]Source code is at https://github.com/bscarone/data-bias-pub.

[10]https://scikit-learn.org/stable/, accessed: 2024-12-10.

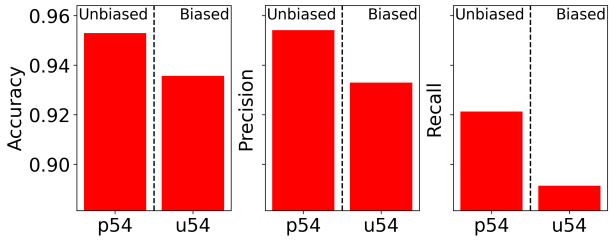


Figure 4: COMPAS dataset - Ada Boost ML model.

Salimi et al. introduce a bias mitigation technique, they measure bias using a slightly modified conditional version of the OR (Odds Ratio), one of the existing measures introduced in Section 3. Only an intuitive justification for this choice is given and their repair algorithm does not provide direct control over bias. We have argued why UB is a better measure than the OR in the binary setting in Section 5. Importantly, there is no version of OR for the general setting (multiple non-binary sensitive attributes and label values), while our UB directly translates to this setting (Section 6). We note that we can completely eliminate the bias in our dataset, *i.e.,* we reach $UB(T, \mathbf{s}, y) = 0$ for every group $\mathbf{s}$ and label value y, which in the binary case corresponds to $OR = 1$.

10 Conclusions and Future Work

We present Uniform Bias, the first interpretable and intersectional multi-label bias measure that can be computed directly and efficiently from a given dataset. To exemplify this we show how Uniform Bias solves issues in the anti employment discrimination rules currently being used by the US Department of Labor. The same ideas used to derive UB can serve as the basis for an explainable and mathematically guaranteed bias mitigation algorithm that can be computed for any dataset. Our algorithm preserves the label frequencies, a mathematical guarantee for the mitigated data to be used in the same context as the original data. We evaluate our techniques on widely used real datasets and show that our mitigation strategies can be used in practice to produce training data that yield better performing ML models.

In the future, we want to explore the relation between the bias of each group and the ML performance for the group. We plan to explore extensions of table union search (Fan et al. 2023) that incorporate the number of tuples from a group or with a particular label in the retrieved tables as part of the ranking criteria. This will allow us to implement scalable versions of our mitigation strategy over real data lakes. With our techniques, one can also obtain mitigated datasets that do not preserve the label frequencies. We would like to explore contexts where a solution without this property would be better than not mitigating the bias at all. Changing the free variables in our mitigation algorithm gives rise to different strategies. We are working to understand how this choice affects mitigation. We also want to consider the context in which we have missing values in our data, where imputation techniques may be useful.

Acknowledgments

We acknowledge the support of the NSF (IIS-2325632) and of the Canada Excellence Research Chairs (CERC) program. Nous remercions le Chaires d'excellence en recherche du Canada (CERC) de son soutien. We thank the anonymous reviewers for their insightful comments and suggestions.

References

Alabdulmohsin, I. M.; Schrouff, J.; and Koyejo, S. 2022. A reduction to binary approach for debiasing multiclass datasets. *Advances in Neural Information Processing Systems*, 35: 2480–2493.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. How We Analyzed the COMPAS Recidivism Algorithm.

Batini, C.; and Scannapieco, M. 2016. *Data and Information Quality*. Data-Centric Systems and Applications. Cham, Switzerland: Springer International Publishing, 1 edition.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Crenshaw, K. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, 23–51. Routledge.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: Association for Computing Machinery.

Fan, G.; Wang, J.; Li, Y.; and Miller, R. J. 2023. Table Discovery in Data Lakes: State-of-the-art and Future Directions. In *Companion of the 2023 International Conference on Management of Data*, SIGMOD '23, 69–75. New York, NY, USA: Association for Computing Machinery.

Friedman, B.; and Nissenbaum, H. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3): 330–347.

Gastwirth, J. L. 2021. A summary of the statistical aspects of the procedures for resolving potential employment discrimination recently issued by the Office of Federal Contract Compliance along with a commentary. *Law, Probability and Risk*, 20(2): 89–112.

Hort, M.; Chen, Z.; Zhang, J. M.; Harman, M.; and Sarro, F. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM J. Responsib. Comput.*, 1(2).

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2019. An Empirical Study of Rich Subgroup Fairness for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 100–109. New York, NY, USA: Association for Computing Machinery.

Lum, K.; Zhang, Y.; and Bower, A. 2022. De-biasing "bias" measurement. In *ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 379–389. New York, NY, USA: Association for Computing Machinery.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Nargesian, F.; Asudeh, A.; and Jagadish, H. V. 2021. Tailoring data source distributions for fairness-aware data integration. *Proc. VLDB Endow.*, 14(11): 2519–2532.

Osoba, O. O.; Badrinarayanan, S.; Cheng, M.; Rogers, R.; Jain, S.; Tandra, R.; and Pillai, N. 2024. Responsible AI update: Testing how we measure bias in the U.S.

Oswald, F. L.; Dunleavy, E. M.; and Shaw, A. 2016. Measuring practical significance in adverse impact analysis. In *Adverse Impact Analysis*, 112–132. Routledge.

Pagano, T. P.; Loureiro, R. B.; Lisboa, F. V.; Peixoto, R. M.; Guimarães, G. A.; Cruz, G. O.; Araujo, M. M.; Santos, L. L.; Cruz, M. A.; Oliveira, E. L.; et al. 2023. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1): 15.

Plecko, D.; and Bareinboim, E. 2022. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*.

Salimi, B.; Howe, B.; and Suciu, D. 2020. Database Repair Meets Algorithmic Fairness. *SIGMOD Rec.*, 49(1): 34–41.

Scarone, B.; Viola, A.; Miller, R. J.; and Baeza-Yates, R. 2024. A Principled Approach for Data Bias Mitigation. *arXiv preprint arXiv:2405.12312*.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, 1–7. New York, NY, USA: Association for Computing Machinery.

Wang, A.; Kapoor, S.; Barocas, S.; and Narayanan, A. 2023. Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *ACM J. Responsib. Comput.* Just Accepted.

Wang, A.; Ramaswamy, V. V.; and Russakovsky, O. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 336–349. New York, NY, USA: Association for Computing Machinery.

Yeh, I.-C. 2016. Default of credit card clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.

Yeh, M.-H.; Metevier, B.; Hoag, A.; and Thomas, P. 2024. Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 518–528. New York, NY, USA: Association for Computing Machinery.

Žliobaitė, I. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4): 1060–1089.