Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02555-7

RESEARCH

# Subgroup fairness based on shared counterfactuals

**Alejandro Kuratomi**[1] · **Zed Lee**[1] · **Panayiotis Tsaparas**[2] · **Evaggelia Pitoura**[2] · **Tony Lindgren**[1] · **Guilherme Dinis Junior**[1] · **Panagiotis Papapetrou**[1]

Received: 30 January 2025 / Revised: 18 July 2025 / Accepted: 21 July 2025

## Abstract

CounterFair is a group counterfactual search algorithm that detects and minimizes biases among sensitive groups and identifies relevant subgroups inside these sensitive groups based on shared counterfactual instances. We investigate the latter capability, analyzing the found subgroups from the perspective of fairness based on counterfactual reasoning, in order to evaluate whether they present different biases with respect to each other and to the sensitive feature groups they belong to. We perform these measurements on the subgroups extracted by CounterFair over six binary classification datasets, providing figures and their respective analysis on the presence of bias.

**Keywords** Counterfactual · Fairness · Explainability · Bias · Subgroups

E. Pitoura and T. Lindgren have contributed equally to this work.

✉ Alejandro Kuratomi
alejandro.kuratomi@dsv.su.se

Zed Lee
zed@hyundai.com

Panayiotis Tsaparas
tsap@uoi.gr

Evaggelia Pitoura
pitoura@uoi.gr

Tony Lindgren
tony@dsv.su.se

Guilherme Dinis Junior
guilherme@dsv.su.se

Panagiotis Papapetrou
panagiotis@dsv.su.se

[1] Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden

[2] Department of Computer Science and Engineering, University of Ioannina and Archimedes/Athena RC, Athens, Greece

Published online: 18 August 2025

## 1 Introduction

Explainable machine learning has attracted significant attention over the recent years with the main objective of increasing the trust in the predictions of opaque but well-performing models. Among several explainability techniques, counterfactual explanations help domain experts and users understand opaque machine learning (ML) models by exploring 'what-if' scenarios for individual instances [1]. Given a trained classifier that maps input instances to class labels, a *counterfactual* (CF) explanation can highlight the relevant feature value changes for an instance of interest that would result in an alternative class label prediction [1–3]. A CF is, therefore, also known as a *recourse* [4], as it suggests actions to improve the situation of a given instance [1, 2, 4–6]. For example, considering a loan application, CF explanations may highlight the feature changes that an individual should comply to (e.g., marital status, habits, education, occupation) for obtaining a positive answer on the application, or for changing their wealth status from low-wealth to a high wealth [4, 5]. Usually, the closest point with the desired label is selected as a CF for the given instance [1], since it results in a reduction in the feature changes that need to be applied for reaching the desired label.

In the presence of sensitive features, such as gender, race, or age, the suggested changes may, however, hide biases across sensitive groups. These biases, if left undetected or unattended, could lead to unfair and harmful outcomes. Hence, it becomes critical to identify and assess model biases or rithmic fairness through the recommendations suggested by CFs. *Counterfactual fairness* refers to the criterion according to which a model's predictions remain unchanged if counterfactuals for a given example are generated by applying changes to protected attributes, while keeping all other attributes the same [6–9]. For example, consider two commonly used public datasets: Adult[1] and Compas.[2] For Adult, the class label indicates whether a person earns over $50K/year or not, whereas for Compas it indicates whether a person is a recidivist (a person recommitting crimes). Figure 1 illustrates the average difficulty in achieving the desired state (i.e., high wealth or no recidivism) in the form of a measure called *burden* [4, 5]. Burden is the distance between an instance and its closest CF, and the figure shows its aggregated value per sensitive group. We observe a higher aggregated burden for females than for males in wealth prediction (Fig. 1a), implying that it is harder for females to achieve higher wealth. Moreover, we observe that it is harder for males than for females to not be recidivists (Fig. 1b). Similarly, it is harder for non-white people to achieve higher wealth, and harder for African-Americans to not be recidivists.

Although the generation of individual CFs provides personalized and actionable recommendations, biases can be introduced between sensitive feature groups, such as those observed in Fig. 1. There is no straightforward mitigation of these biases when generating individual CFs, since each individual CF ignores other feature groups. This is a problem when trying to explain biased models that are already trained. For example, if a group of 10 males and a group of 100 females receive a loan rejection, and the average recommendation for males is to increase their salary by $10K, whereas for females the recommended increase is by $25K, the model is biased not only in its predictions but also in the CF recommendations. Hence, model retraining is required to address these biases, so that these groups receive similar recommendations in relation to their gender or any other sensitive feature. To solve this, we consider not an individual but a group-based CF generation approach that permits the assessment of fairness across sensitive groups.

[1] https://archive.ics.uci.edu/dataset/2/adult.

[2] https://www.kaggle.com/datasets/danofer/compass.

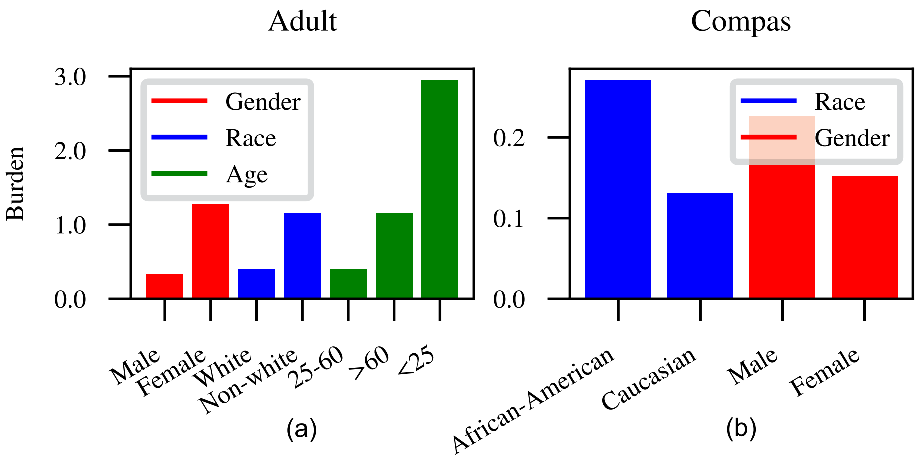


**Fig. 1** Burden for sensitive groups ($x$-axis) belonging to different features (colors), showing biases on Adult and Compas

Existing group CF algorithms obtain CFs through rule mining and *effectiveness* maximization [7, 10], which refers to the ratio of instances in the group that can apply changes to their own features to reach the feature values of the group CF. The main approach followed by these algorithms is to first mine the frequent subgroups of the undesired class label and, then, mine the frequent subgroups of the desired class. Nonetheless, these methods suffer from two main drawbacks: (1) the problem of fair CF generation is not addressed and (2) the relevant subgroups are not selected through the CF generation process. As a result, the solution space is limited and the generated CFs have a high burden and low effectiveness. Existing attempts to address these drawbacks are constrained to linear or decision-tree based models [8] and by the inability to output fair CFs among sensitive groups. We propose a method that can address these two drawbacks while also providing fairness assessment.

Recently, CounterFair [11] has been proposed to address these deficiencies. The main characteristic of CounterFair is that it produced model-agnostic CFs for assessing group CF fairness. Moreover, it can prioritize either burden minimization, subgroup identification, or fair recourse generation, leading to either more granular, group-oriented, or fairness-oriented CFs. Specifically, CounterFair detects and mitigates this burden measure as the bias measure across different subgroups, where certain demographic subgroups may experience a disproportionate negative disadvantage in terms of desired outcomes. This measure of bias estimates the aggregated effort each subgroup of people should do in order to be classified in the right, desired class label.

The extensions of CounterFair [11] proposed in this paper are as follows:

- **Bias detection in subgroups identified by CounterFair:** We extend CounterFair [11] by defining the concept of *subgroups* lying at the intersection of the different sensitive feature groups in order to assess the presence of biases in these particular sets of instances. We show that the presence of biases in these intersections is dataset-dependent. Furthermore, we provide a large set of figures indicating the presence of biases and perform an extensive analysis of the subgroup and intersectional fairness identified inside said relevant subgroups.
- **Evaluation:** We extend the experimental evaluation carried out in [11] to characterize the identified subgroups in terms of their relevance, and whether they present higher,

comparable, or lower biases with respect to each other, and with respect to their respective sensitive feature groups. Additionally, we provide insights on to how these subgroups may be subject to potentially biasing scenarios with respect to other features analyzed in their respective subgroups.

## 2 Related work

CF explanations are usually obtained for single instances [1, 2, 4, 12]. Different application scenarios are considered, including recommender system explanations for individual user-item combinations [13]. A few recent studies focused on group CFs [7, 8, 10]. There are several ways to generate group CFs. One way is to jointly generate a CF for each instance, to, for example, make their distribution similar to that of the dataset (*one-to-one* way) [8, 14]. Another way is to get several CFs for a single instance to maximize recourse diversity (*many-for-one* way) [15]. Finally, getting a CF for a group to characterize its instances (*one-for-many* way) [8].

Finding optimal group CFs is analogous to optimally locating facilities, such as hospitals or production plants [16]. This is known as the location analysis problem solved using mathematical programming (MP) [8]. CF algorithms, like the Actionable Recourse [12], use MP but are constrained to linear classifiers due to the difficulty in formulating the nonlinearities of highly accurate ML models. To preserve the formulation linearity and convexity, as well as the solutions optimality, one may apply a graph-oriented approach [17].

Other group CF approaches exist: Kavouras et al. [7] and Rawal and Lakkaraju [10] developed Fairness Aware Counterfactuals for Subgroups (FACTS) and Actionable Recourse Summaries (AReS), respectively, to generate group CFs through rule mining. The rules are in the form of a predicate and an action, e.g., *if gender == female then salary* $\geq$ *\$80K*. FACTS first finds subgroups from the undesired class label instances and sets of actions for these subgroups from the desired class using FP-growth. Then, for these subgroups, the algorithm finds their intersection, i.e., the subgroups or feature-value combinations that are common across them. These common subgroups are then used to find, from the space of actions, a set of valid, effective actions that have the same cost for the individuals on each subgroup. FACTS then uses a set of measures to establish whether there is a bias given the found CFs among the different sensitive groups. AReS is similar to FACTS but it extracts both predicates and actions from the training dataset. The recourse rules are then selected using an optimization procedure that aims to maximize the *correctness* (the fraction of instances for which the recourse rules effectively create a CF), the coverage (the amount of instances for which the "if" conditions apply) and the interpretability (the amount of recourse rules, their length and the number of subgroups). Particularly, these two methods also allow the user to identify subgroups of interest inside the sensitive groups (e.g., in the females sensitive group, those from the EU who are divorced) via the identified and stored predicates for the CF rules. Kavouras et al. [7], and Rawal and Lakkaraju [10] also discuss quality measures to assess the group CFs and the algorithmic fairness based on the CFs.

There are many quality measures to assess CF explanations [1, 4, 18]. These measures include proximity and likelihood (how likely is the given CF regarding the dataset [14]) among others. Likewise, there are many algorithmic fairness measures. These may benefit from CF reasoning. For example, predictive equality (the false-negative ratio of a sensitive group), which assesses biases jointly using the prediction and ground truth labels across groups, may miss the potential bias detection capability through the CF recommendations provided by the

CFs [19]. Sharma et al. [5] propose CF burden as a fairness proxy, while Kuratomi et al. [6] weighted the sensitive group burden with its predictive equality, combining accuracy-based and CF fairness. Group CFs quality may also be measured through effectiveness [7].

## 3 Preliminaries

Let $\mathcal{X}$ be a heterogeneous feature space with binary, categorical, ordinal, and continuous features. A dataset $\mathcal{D}$ is a collection of $n$ pairs of $(X, y)$ where $X$ is a data sample (i.e., instantiation) of $\mathcal{X}$ and $y$ is its corresponding binary class label $y \in \{\text{“}-\text{”}, \text{“}+\text{”}\}$. $\mathcal{D}$ is divided into a training and a test set, denoted as $\mathcal{D}_{\text{Train}}$ and $\mathcal{D}_{\text{Test}}$, respectively.

Moreover, let $\mathcal{S} \subseteq \mathcal{X}$ be a set of sensitive features in $\mathcal{X}$, such as *sex* or *race*. Each sensitive feature $s \in \mathcal{S}$ may be used to define different *sensitive groups* of data samples. A sensitive group of feature $s$ is denoted as $s_k$, where $k$ defines a condition on that feature, which is denoted by function $\text{cond}(\cdot)$. If $s$ satisfies condition $k$ then $\text{cond}(s, k)\ ==$ ‘true’. For example, sensitive feature *sex* may be used to define two sensitive groups, i.e., $s_{\text{female}}$ and $s_{\text{male}}$, corresponding to data samples for which *sex* == ‘female’ and *sex* == ‘male’, respectively. Given a classifier $f(\cdot)$, we define the set of false-negative test instances in sensitive group $s_k$ as:

$$\mathcal{D}^{s_k}_{\text{Test FN}} = \{(X, \text{“}+\text{”}) | f(X) = \text{“}-\text{”}, \text{cond}(s, k), X \in \mathcal{D}_{\text{Test}}\}$$

Additionally, we introduce the property of *feasibility*. A CF is *feasible* with respect to an instance of interest if it complies with the properties of *mutability*, *directionality*, and *plausibility*. A CF complies with the mutability property if only mutable features are changed from its corresponding instance in $\mathcal{D}_{\text{Test FN}}$. In the same manner, it complies with directionality if the features are changed only in possible directions, e.g., age or education cannot decrease. Finally, plausibility indicates that the CF feature values have all physically possible values.

The feasibility property, composed of the properties of mutability, directionality and plausibility, is essential not only to the counterfactual generation process, but also to the evaluation of fairness based on the proposed changes by the counterfactual instance. It is essential for the counterfactual point itself because these explanations should be *reachable* by the instances of interest, i.e., the people receiving the explanations (recommendations indicating a change of gender, or a lowering of age should not be given to people). With respect to fairness, the measurement of burden, that is, how difficult it is for people to reach their corresponding counterfactuals, is not relevant if the proposed recommendations inside these counterfactuals are not feasible or achievable from the perspective of the people of interest. The properties of mutability, directionality and plausibility of each feature are entered manually into the data processing pipeline, as per the recommendations in [2, 5]. As an example, education may change, but only in the upward direction, as does age, but gender or race cannot change.

For each instance $X_i \in \mathcal{D}^{s_k}_{\text{Test FN}}$ we use a CF generator to get its CF, $X'_i$. Let us now define the set of possible CFs for the instances in $\mathcal{D}_{\text{Test FN}}$ as $\mathcal{Q}$ and the function $F(\mathcal{D}_{\text{Test FN}}, \text{Q})$ as an indicator of the instances in $\mathcal{Q}$ that comply with the feasibility condition with respect to the instances in $\mathcal{D}_{\text{Test FN}}$. We now introduce the general problem formulation.

**Problem 1** (Bias detection, mitigation and Identification of relevant subgroups) *Given a classifier $f(\cdot)$ and the set of false-negative test instances $\mathcal{D}_{\text{Test FN}} = \bigcup_{s_k} \mathcal{D}^{s_k}_{\text{Test FN}}$, we want* *to obtain the set of CFs $\mathcal{D}'$ as follows:*

$$\mathcal{D}' = \arg\min_{\text{Q}}\{w_1 C_{\text{burden}}(Q) + w_2 C_{\text{fair}}(Q) + w_3 C_{\text{groups}}(Q) \mid F(\mathcal{D}_{\text{Test FN}}, \text{Q})\}, \tag{1}$$

*where $w_1$, $w_2$ and $w_3$ are the weights for the costs associated with: (1) the aggregated CFs burden ($C_{\text{burden}}$), (2) bias mitigation or fairness ($C_{\text{fair}}$) and (3) the number of relevant subgroups identified ($C_{\text{groups}}$), respectively.*

The formulation in problem (1) enables a flexible cost function definition to allow for the extraction of CFs that optimize different objectives. When prioritizing $C_{\text{burden}}$, burden is minimized, and when aggregated by sensitive groups, this measure elicits the biases among sensitive groups, indicating which groups require a higher effort to achieve the desired label. When prioritizing $C_{\text{fair}}$, the differences in burden across sensitive groups is reduced. This provides fair CF recommendations across different groups, since these would have similar application difficulty, as measured by burden. When prioritizing $C_{\text{groups}}$, a set of CFs that is minimal in size is obtained, which forces the CFs to be *shared* among the instances of interest, i.e., each of the false-negative instances will be subgrouped together with other instances, based on the shared CF, generating group CFs and identifying subgroups of interest simultaneously. We now explain CounterFair, the instantiation of the cost functions, $C_{\text{burden}}$, $C_{\text{fair}}$ and $C_{\text{groups}}$ used, and how CounterFair solves problem (1).

## 4 CounterFair

CounterFair is an MP-based algorithm that attains feasible and optimal group CFs in terms of a given cost function. This cost function is adaptable and can be defined in different ways. In our case, we define so that burden (leading to bias detection), burden differences (leading to bias mitigation) or the number of different CFs (identifying relevant subgroups) are minimized. We first provide an outline of the main steps of CounterFair, the instantiation of the cost function for CounterFair, and finally its MP formulation.

### 4.1 Outline

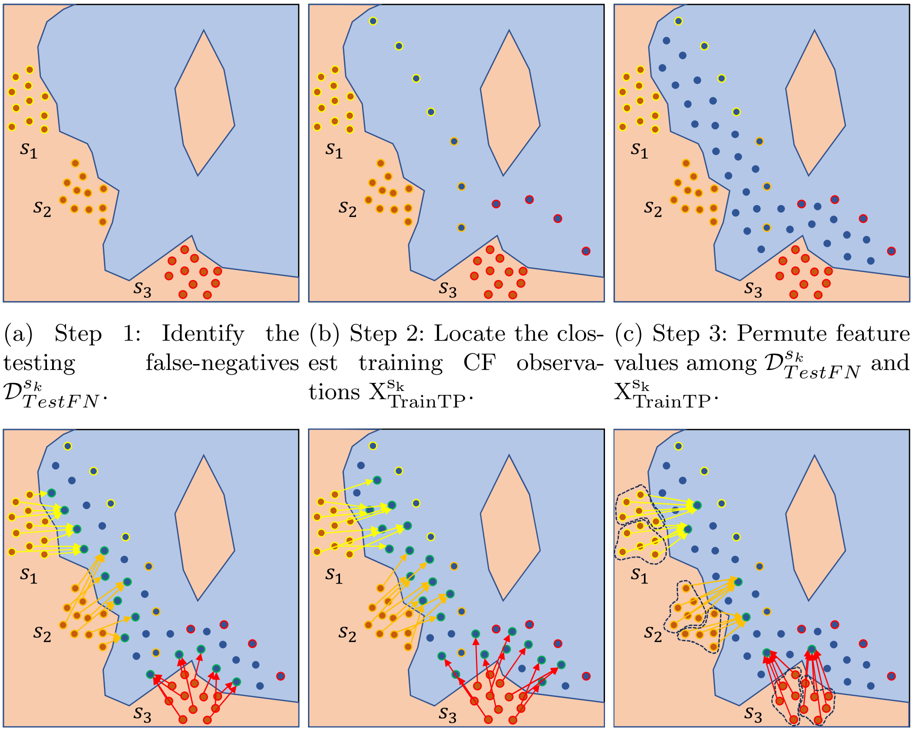
CounterFair creates a set of *points* from which it selects an optimal set of CFs given a cost function, following four steps: (1) obtain the sets of false-negative test instances $\mathcal{D}^{s_k}_{\text{Test FN}}$ and true-positive training instances $\mathcal{D}^{s_k}_{\text{Train TP}}$ per sensitive group; (2) obtain the nearest neighbor training CF for each instance in $\mathcal{D}^{s_k}_{\text{Test FN}}$ from the instances in $\mathcal{D}^{s_k}_{\text{Train TP}}$, which are then stored in set $\text{CF}^{s_k}_{\text{Train}}$; (3) find all the combinations of the feature values between each $X \in \mathcal{D}^{s_k}_{\text{Test FN}}$ and every CF in $\text{CF}^{s_k}_{\text{Train}}$ to generate a cloud of points, $\mathcal{P}$, which are the potential CFs; (4) solve the MP to select the best CFs simultaneously for every $X \in \mathcal{D}^{s_k}_{\text{Test FN}}$ using $\mathcal{P}$.

The steps of CounterFair are detailed in Algorithm 1 and depicted in Fig. 2. In step 1 (Fig. 2a), a ML model separates the undesired class (orange-shaded region) from the desired class (blue-shaded region). The orange points in the undesired region represent the false negatives. Each sensitive group $s_1$, $s_2$, and $s_3$ is outlined in yellow, orange and red, respectively.

In step 2, for each $X \in \mathcal{D}^{s_k}_{\text{Test FN}}$, the *Nearest Neighbor* (NN) is used to find the closest training CF observation from $\mathcal{D}^{s_k}_{\text{Train TP}}$ (blue-colored points in Fig. 2b) and store it in $\text{CF}^{s_k}_{\text{Train}}$. The set $\text{CF}^{s_k}_{\text{Train}}$ is filtered using: (1) the closest percentage $\Omega$ of CFs from the centroid of the

**Algorithm 1:** CounterFair

```
input  : $D$, $f$, $s$, $\vec{mut}$ (mutability vector), $\vec{dir}$ (directionality vector), $\vec{pla}$ (plausibility vector), $\Omega$
         (closest training percentage).
output: $CF_{CounterFair}$
$\mathcal{D}^{s_k}_{TestFN}, \mathcal{D}^{s_k}_{TrainTP} \leftarrow$ TestFNTrainTP($D, f, s$)
$CF^{s_k}_{Train} \leftarrow$ NN($\mathcal{D}^{s_k}_{TestFN}, \mathcal{D}^{s_k}_{TrainTP}, \Omega$)
$\mathcal{P}, C_{X_in}, F_{X_in} \leftarrow$ points($\mathcal{D}^{s_k}_{TestFN}, CF^{s_k}_{Train}, f, \vec{mut}, \vec{dir}, \vec{pla}$)
$CF^{X_in}_{CounterFair} \leftarrow$ solveMP($\mathcal{P}, C_{X_in}, F_{X_in}$)
return $CF_{CounterFair}$
```



(a) Step 1: Identify the testing false-negatives $\mathcal{D}^{s_k}_{TestFN}$.

(b) Step 2: Locate the closest training CF observations $\mathrm{X}^{\mathrm{s_k}}_{\mathrm{TrainTP}}$.

(c) Step 3: Permute feature values among $\mathcal{D}^{s_k}_{TestFN}$ and $\mathrm{X}^{\mathrm{s_k}}_{\mathrm{TrainTP}}$.

(d) Step 4.1: Solve the MP with a higher weight on CFs burden.

(e) Step 4.2: Solve the MP minimizing burden variance.

(f) Step 4.3: Solve the MP with a higher weight on group formation.

**Fig. 2** 2-Dimensional example of CounterFair. Steps 2d, 2e and 2f are obtained by prioritizing $C_{\text{burden}}$ (minimizing burden), $C_{\text{fair}}$ (minimizing burden differences) and $C_{\text{groups}}$ (minimizing the set of CFs), respectively

set of false-negatives instances of each $s_k$, and (2) a critical distance $d$ to the false-negative instances. The $\Omega$ value depends on the dataset but is usually 100%, i.e., all the training CFs in $\mathrm{CF}^{s_k}_{\text{Train}}$ are considered. The distance $d$ is the maximum of the distance averages of each sensitive group.

In step 3 (Fig. 2c), the cloud of blue points $\mathcal{P}$ is generated using all the possible combinations of the feature values between each $X_i \in \mathcal{D}^{s_k}_{\text{Test FN}}$ and the CF observations in $\mathrm{CF}^{s_k}_{\text{Train}}$. All the continuous features are discretized using an equal frequency binning. The generated points are stored in $\mathcal{P}$, if they: (1) are feasible CFs with respect to the instance from which they are generated, (2) lie inside the critical distance $d$ with respect to this instance. Finally, for each $X_i \in \mathcal{D}^{s_k}_{\text{Test FN}}$ and $n \in \mathcal{P}$, two parameters are calculated: (1) a cost parameter $C_{X_i n}$, representing the cost of using point $n$ as the CF for $X_i$ and (2) a feasibility parameter $F_{X_i n}$ indicating whether point $n$ is a feasible CF for $X_i$.

In step 4, the MP is solved in three separate ways, shown in Fig. 2d–f, by: (1) minimizing the aggregated burden, (2) minimizing the burden differences and (3) minimizing the number of distinct CFs. Figure 2d illustrates 15 unique CFs, outlined in green, presenting the lowest sensitive group-aggregated burden (the length of the arrows is minimized). Figure 2e shows CFs having similar distances from their respective instances of interest, minimizing the burden differences among sensitive groups (the difference in length of the arrows is minimized). In Fig. 2f there are six distinct, shared CFs. There are two unique subgroups per sensitive group (enclosed in the dashed lines), which are the relevant subgroups.

## 4.2 Cost function instantiation

To solve problem (1), we instantiate the cost functions $C_{\text{burden}}$, $C_{\text{fair}}$ and $C_{\text{groups}}$ and describe the MP formulation.

### 4.2.1 $C_{\text{burden}}$

In order to define $C_{\text{burden}}$, we hereby introduce the measure of accuracy weighted burden (AWB).

*Accuracy Weighted Burden:* the AWB measure, introduced in [6], is the product of predictive equality (the false-negative ratio) and the average burden per sensitive group. The AWB measure uses a distance function $\mathrm{d}(X_i, X'_i)$. This distance is the burden incurred by instance $X_i$ in trying to attain the feature values of $X'_i$, and it is a combination of the $L1$-norm and the $L0$-norm. Equation (2) indicates how to calculate the AWB measure.

$$\text{AWB}^{s_k} = \frac{\sum_{X_i \in \mathcal{D}^{s_k}_{\text{Test FN}}} \mathrm{d}(X_i, X'_i)}{|\{(X, y) \in \mathcal{D}_{\text{Test}} | \text{cond}(s, k), y = \text{“}+\text{”}\}|}, \tag{2}$$

where the denominator is the amount of true positives in the sensitive group $s_k$. Equation (2) indicates that a higher number of false negatives, or a higher distance between each instance and its CF in the $s_k$ sensitive group, make the $\text{AWB}^{s_k}$ burden higher. Then, we define $C_{\text{burden}}$ as the total AWB:

$$C_{\text{burden}} = \sum_{s_k} \text{AWB}^{s_k} \tag{3}$$

### 4.2.2 $C_{\text{fair}}$

We may define the cost associated with the presence of biases by estimating the differences of the burden among the sensitive groups. To do this, we define $\text{AWB}_{\min} = \min \text{AWB}^{s_k}$ and $\text{AWB}_{\max} = \max \text{AWB}^{s_k}$ as the minimum and maximum burden, respectively. $C_{\text{fair}}$ is then defined as the absolute difference between these two terms:

$$C_{\text{fair}} = \text{AWB}_{\max} - \text{AWB}_{\min} \tag{4}$$

### 4.2.3 $C_{\text{groups}}$

To define the cost associated with the number of distinct CFs, we define a variable and a set: $l_n, \forall n \in \mathcal{P}$ as a variable that indicates whether a point $n \in \mathcal{P}$ is selected as a CF for any of the instances $X_i \in \mathcal{D}_{\text{Test FN}}$ and $\mathcal{I} = \mathcal{D}_{\text{Test FN}} = \bigcup_{s_k} \mathcal{D}^{s_k}_{\text{Test FN}}$ as the set of all false-negative instances. Therefore, the cost $C_{\text{groups}}$ is defined as:

$$C_{\text{groups}} = \frac{\sum_{n \in \mathcal{P}} l_n}{|\mathcal{I}|} \tag{5}$$

In the worst-case scenario, every instance $X_i \in \mathcal{I}$ will have its own unique CF, making the cost $C_{\text{groups}} = 1$. We now continue with the MP formulation of the CounterFair algorithm and show how it solves problem (1).

## 4.3 CounterFair MP formulation

To solve problem (1), we split the implementation of CounterFair into two: one main implementation focusing on minimizing $C_{\text{burden}}$ and $C_{\text{groups}}$, and another focusing on minimizing $C_{\text{fair}}$, the latter requiring additional variables and constraints over the main implementation. In the main implementation, the MP uses only integer decision variables, making the MP an integer program. In the second implementation, a set of continuous decision variables must be added to the main implementation, which makes the MP a mixed integer linear program. We proceed to present the main formulation and then describe the added variables and constraints for the second.

We define the set of binary decision variables $p_{X_{\text{in}}}, \forall X_i \in \mathcal{I},\ n \in \mathcal{P}$. These variables indicate whether the point $n$ is selected as a CF for the instance $X_i$. In order to relate the $C_{\text{burden}}$ cost with the decision variable, we introduce the parameter $\text{AWB}^{s_k}_{X_{\text{in}}}$, which is the added burden when selecting point $n$ for the instance $X_i$ as a CF, i.e., when $p_{X_{\text{in}}} = 1$:

$$\text{AWB}^{s_k}_{X_{\text{in}}} = \frac{\text{d}(X_i, n)}{|\{(X, y) \in \mathcal{D}_{\text{Test}} | \text{cond}(s, k), y = \text{“}+\text{”}\}|}, \tag{6}$$

for all $X_i \in \mathcal{I}$ and for all $n \in \mathcal{P}$. Then, multiplying $\text{AWB}^{s_k}_{X_{\text{in}}}$ with the decision variable $p_{X_{\text{in}}}$:

$$\text{AWB}^{s_k} = \sum_{X_i \in \mathcal{I}} \sum_{n \in \mathcal{P}} \text{AWB}^{s_k}_{X_{\text{in}}} \cdot p_{X_{\text{in}}}, \tag{7}$$

and then $C_{\text{burden}}$ can be rewritten as:

$$C_{\text{burden}} = \sum_{s_k} \text{AWB}^{s_k} = \sum_{s_k} \sum_{X_i \in \mathcal{I}} \sum_{n \in \mathcal{P}} \text{AWB}^{s_k}_{X_{\text{in}}} \cdot p_{X_{\text{in}}} \tag{8}$$

For the $C_{\text{groups}}$ term, we use the binary decision variables $l_n, \forall n \in \mathcal{P}$, and the $C_{\text{groups}}$ cost remains as defined in (5). We define the objective function for the main implementation as:

$$Z_1 = \alpha C_{\text{burden}} + (1 - \alpha) C_{\text{groups}}, \tag{9}$$

where the weight $\alpha \in [0, 1]$. When weight $\alpha \approx 0$, $C_{\text{groups}}$ is prioritized and the number of distinct CFs will be minimized. This will force the *sharing* of CFs among similarly distanced instances, automatically identifying relevant subgroups via these shared CFs. Equivalently, when $\alpha \approx 1$, the total aggregated burden will be minimized, optimizing the recommendations found for each instance. In this scenario, the models biases will be observed as a higher relative aggregated burden for some of the sensitive groups.

A subgroup identified by CounterFair is the set of instances that are assigned to the same counterfactual and is mathematically defined as follows:

$$G_j = \{X_i \in \mathcal{I} : p_{X_{\text{in}}} = 1\}, \forall n \in \mathcal{P} \tag{10}$$

where $G_j$ is the $j$th subgroup found in the dataset, where $j \in \mathbb{Z}^+$.

We now define the block $\mathcal{R}$ of constraints as follows:

$\mathcal{R}$:

$$p_{X_{\text{in}}} \leq F_{X_{\text{in}}},\ \forall i \in \mathcal{I}, n \in \mathcal{P}, \tag{11}$$

$$\sum_{n \in \mathcal{P}} p_{X_{\text{in}}} = 1, \forall X_i \in \mathcal{I}, \tag{12}$$

$$p_{X_{\text{in}}} \leq l_n\ \forall X_i \in \mathcal{I}, \forall n \in \mathcal{P}, \tag{13}$$

$$p_{X_{\text{in}}}, l_n \in \{0, 1\}\ \forall X_i \in \mathcal{I}, \forall n \in \mathcal{P}, \tag{14}$$

Constraint (11) guarantees that the selected points $n$ are feasible for their instances $X_i$, while constraint (12) forces the selection of a single point $n$ per instance $X_i$. Finally, constraint (13) requires all $p_{X_{\text{in}}}$ variables to be less than or equal to the limiter $l_n$, with constraints (14) forcing $p_{X_{\text{in}}}$ and $l_n$ to be binary. Finally, the main MP formulation is:

$$\min Z_1,\ \ \text{subject to}\ \mathcal{R} \tag{15}$$

We now describe the second implementation, which aims at mitigating the biases among sensitive groups. To do this, we take the previously defined variables: $\text{AWB}_{\text{max}}$ and $\text{AWB}_{\text{min}}$, and add them as continuous decision variables. Then we add the following set of constraints to the block $\mathcal{R}$:

$$\text{AWB}_{\text{min}} \leq \sum_{X_i \in \mathcal{I}} \sum_{n \in \mathcal{P}} \text{AWB}^{s_k}_{X_{\text{in}}} \cdot p_{X_{\text{in}}} \leq \text{AWB}_{\text{max}}, \forall s_k, \tag{16}$$

which bounds the aggregated burden per sensitive feature. Then, we define the cost function as:

$$Z_2 = \text{AWB}_{\text{max}} - \text{AWB}_{\text{min}}, \tag{17}$$

which matches Eq. (4). When minimizing $Z_2$ the difference between the maximum and minimum burden is reduced down to zero. Given that $\text{AWB}_{\text{max}}$ and $\text{AWB}_{\text{min}}$ work as bounds on the burden, it then forces the selection of CFs that have equal burden across sensitive groups and thereby effectively decreasing the biases obtained from the recommendations.

By now, it is possible to conceive other cost functions and potential adaptations to the algorithm to achieve other objectives through the CounterFair MP formulation. Besides the objectives of bias detection, mitigation and subgroup identification, we consider CF effectiveness as a measure to optimize for in group CF generation. We now continue with the formulation to maximize group CF effectiveness.

As previously mentioned, CF effectiveness is the ratio of total instances that have a feasible CF. We define parameter $e_n$, $\forall n \in \mathcal{P}$, as the effectiveness associated with point $n$ with respect to the instances of interest:

$$e_n = \frac{\sum_{X_i \in \mathcal{I}} F(X_i, n)}{|\mathcal{I}|}, \forall n \in \mathcal{P}, \tag{18}$$

where $F(X_i, n)$ is the feasibility indicator function between the instance $X_i$ and the point $n$. This parameter, which can be estimated for every point $n \in \mathcal{P}$, indicates the ratio of the instances in the set of false negatives that can reach the $n$ CF point. Then, the cost function associated with effectiveness in the MP is defined as:

$$Z_3 = C_{\text{eff}} = -\sum_{X_i \in \mathcal{I}} \sum_{n \in \mathcal{P}} e_n \cdot p_{X_{\text{in}}}. \tag{19}$$

We now proceed to discuss the complexity of CounterFair.

*Complexity:* The most complex step of the CounterFair algorithm is step 4: the solution of the MP. In general, an integer program formulation, which is harder to solve than a mixed integer linear program formulation, is classified as a NP-complete problem [20–22] with a complexity determined by the number of rows (constraints) and columns (variables). Based on the block $\mathcal{R}$ of constraints, the number of constraints is $3(|\mathcal{I}| \cdot |\mathcal{P}|) + |\mathcal{I}| + |\mathcal{P}|$, while the number of variables, $v$, is $|\mathcal{I}| \cdot |\mathcal{P}| + |\mathcal{P}|$. The solution is usually obtained through a branch-and-bound approach, which requires the solution of a relaxed linear program at every node of the branch-and-bound tree. Each linear program solution is estimated to have a complexity of $\mathcal{O}(v^{2.5})$ [23]. The minimum number of nodes in a branch-and-bound tree is determined by $2^{\lfloor \frac{v}{2c} \rfloor}$, where $c$ is the maximum number of variables in any given constraint [24]. In this case, since $c = |\mathcal{P}|$ (see constraint 12) then $2^{\lfloor \frac{v}{2c} \rfloor} = 2^{\lfloor \frac{|\mathcal{I}| \cdot |\mathcal{P}| + |\mathcal{P}|}{2|\mathcal{P}|} \rfloor} = 2^{\lfloor \frac{|\mathcal{I}|+1}{2} \rfloor}$. Therefore, the total complexity of CounterFair is $\mathcal{O}(2^{\lfloor \frac{|\mathcal{I}|+1}{2} \rfloor}(|\mathcal{I}| \cdot |\mathcal{P}| + |\mathcal{P}|)^{2.5})$.

## 5 Empirical evaluation

In this section, we illustrate the experimental setup by describing the CF evaluation measures, the datasets and the classification performance achieved. We then evaluate CounterFair and compare it with AReS [10] and FACTS [7].

### 5.1 Experimental setup

We compare the aggregated AWB values of the sensitive groups using Eq. (7) and the number of subgroups obtained by summing the limiter variable $l_n$ for each of the sensitive groups. We define the set of points in the cloud of points $\mathcal{P}$ that belong to sensitive group $s_k$ as $\mathcal{P}^{s_k}$. Then, we define $L^{s_k} = \sum_{n \in \mathcal{P}^{s_k}} l_n$ as the number of distinct points selected as CFs for $s_k$. We calculate the effectiveness of the CFs per sensitive group, which is defined as $E^{s_k} = \frac{|\{X_i \in \mathcal{D}^{s_k}_{\text{Test FN}} | F(X_i, X'_i)\}|}{|\{X_i \in \mathcal{D}^{s_k}_{\text{Test FN}}\}|}$.

Moreover, we used six binary classification datasets. These datasets cover different application domains and sensitive groups, focusing mainly on gender, age, and race [25]. All datasets have been preprocessed and stored in our GitHub repository.[3] The preprocessing is based on Karimi et al. [2] and Le Quy et al. [25]. Further details about the CounterFair parameters and the features for each dataset may be found in the repository.

Additionally, we trained a random forest (RF) and a multi-layer perceptron (MLP) classifier and tuned their parameters using a 70%/30% train/test split and a grid search tuning on the training set. We used the F1 score as our classification metric. Details on performance, computing unit used and structure of the classifiers are provided in the repository.

[3] https://github.com/alku7660/CounterFair.

Finally, we benchmark CounterFair on three scenarios, one for each of the described cost functions in Sect. 4.3: (1) with cost function $Z_1$ and three different values of $\alpha$: $\alpha = [0.1, 0.5, 1.0]$ for all datasets. and obtain the $\text{AWB}^{s_k}$ and $L^{s_k}$ scores for each sensitive group $s_k$; (2) with cost function $Z_2$, to reduce the biases based on the aggregated burden; (3) with cost function $Z_3$ to showcase the adaptability of CounterFair and its performance when optimizing for effectiveness, as a group CF measure that has been previously prioritized by other methods. We compare the performance in terms of burden and effectiveness with AReS and FACTS.

### 5.2 Results

We present here the results of the experiments carried out with respect to 5 elements: (1) burden minimization for bias detection, (2) minimization of burden differences among sensitive groups for bias mitigation and fair recommendations, (3) impact of minimizing differences of burden and differences of distance on the group CF recommendations, (4) minimization of distinct CFs for relevant subgroup identification and (5) comparison of CounterFair with AReS and FACTS in burden, effectiveness and run times.

#### 5.2.1 Burden minimization

In the first experiment, we ran CounterFair to minimize cost function $Z_1$ with $\alpha = [0.1, 0.5, 1.0]$, i.e., starting with a low weight of 0.1 for the $C_{\text{burden}}$ cost and a high weight of 0.9 for the $C_{\text{groups}}$ cost, and ending with a high weight of 1.0 for $C_{\text{burden}}$ and 0.0 for $C_{\text{groups}}$. Figure 3 shows the aggregated burden per sensitive group as the bars when $\alpha$ increases for each dataset. The burden decreases as $\alpha$ increases for all the datasets, and shows the differences in burden among different sensitive groups.

The differences in burden across sensitive groups are best evidenced when using the highest $\alpha = 1.0$ because the nearest and easiest CF is selected for each instance. Since the aggregated burden for each sensitive group is calculated through Eq. (7), a higher number of false-negative instances (the number in parenthesis in the legend of each plot) would normally portray a higher burden for a given group. However, this is not always the case. For example, prioritizing AWB ($\alpha = 1.0$) in the German dataset does not show the same relative AWB behavior among genders as in the other two values for $\alpha$ and, even though there are less than half as many females as males, females present a higher AWB.

#### 5.2.2 Minimization of burden differences among sensitive groups

in this experiment we ran CounterFair to minimize cost function $Z_2$, i.e., the differences in burden among the CFs. Figure 3 shows the result of this experiment in the last set of bars on each plot, over the *Fair* *x*-axis label. Note that the obtained CFs show an equal burden as measured by $\text{AWB}^{s_k}$ among the sensitive groups for each dataset, effectively eliminating the biases in burden and producing group CF recommendations that are fair across these groups.

#### 5.2.3 Impact of minimizing the differences of burden among groups ($\text{AWB}^{s_k}$)

we illustrate the impact of the minimization of burden differences in the CounterFair CF recommendations for the false-negative instances in the German dataset and compare it to the minimization of distance differences, i.e., not considering the false negative ratio of the

**Table 1** Instances and their CFs obtained through CounterFair when mitigating biases across sensitive groups in the German dataset

| | Sex | Single | Unemployed | Purpose | Rate | Housing | Age | Credit | Duration |
|---|---|---|---|---|---|---|---|---|---|
| $X1$ | M | No | No | Elec | 3 | Rent | 22 | 1331 | 1.2 |
| $X1'^{\text{AWB}}_{\text{Fair}}$ | M | No | No | Elec | 3 | Rent | 22 | 1845 | 4.5 |
| $X1'^{\text{dist.}}_{\text{Fair}}$ | M | No | No | Elec | 3 | Rent | 22 | 1871 | 4.5 |
| $X2$ | F | No | No | Car | 4 | Owns | 34 | 1842 | 3.6 |
| $X2'^{\text{AWB}}_{\text{Fair}}$ | F | No | No | Car | 4 | Owns | 34 | 866 | 15 |
| $X2'^{\text{dist.}}_{\text{Fair}}$ | F | No | No | Car | 1 | Owns | 34 | 1490 | 15 |

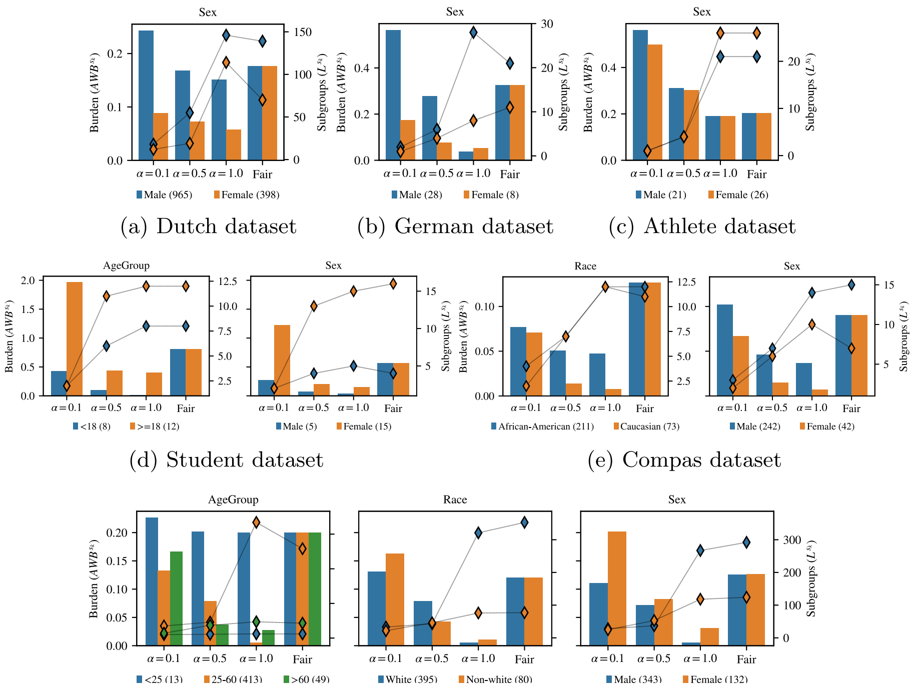
The recommended changes are larger for the female since the model is biased in accuracy favoring females

model. We randomly pick a male and a female from the $\mathcal{D}_{\text{Test FN}}$ set. We then run CounterFair on this dataset minimizing instead the differences in distance cost $\sum_{\mathcal{D}_{\text{Test FN}}} \text{d}(X_i, X_i')$ and extract the CFs for the selected male and female. The CFs (minimizing AWB differences and distance differences) are shown in Table 1. In the CFs minimizing AWB differences, the credit change is larger for the female than for the male, compensating for the higher amount of false-negative males (there are 8 false-negative females and 28 false-negative males). When minimizing for distance differences, there is no consideration of the false-negative ratio, so there is no compensation for the bias in accuracy, and the credit change is smaller for the females, although the rate also decreased.

We highlight the importance of having both the bias detection-oriented CF generation first, as this would allow the users to detect potential sensitive feature biases present in the trained model, and then, if required, generate fair CFs by minimizing the burden in deployed models. This is important, since only running CounterFair for bias mitigation (although beneficial for the fairness in the recommendations to users) might hide the models algorithmic biases. We recommend using CounterFair in the following manner: first identify the biases and then obtain the bias-mitigating CFs if needed, so that the recommendations are fair across groups.

### 5.2.4 Minimization of distinct CFs

We now show the structure of the relevant subgroups identified when minimizing cost function $Z_1$ with $\alpha = 0.1$. Figure 3 illustrates the number of relevant CFs and subgroups found for each sensitive group with the diamonds plotted using the secondary $y$-axis. Note that, as burden is increasingly prioritized, the number of subgroups increases. Figures 4 and 5 show the details of the relevant subgroups identified in the six studied datasets. As an example, we discuss now the Compas dataset. In the Compas dataset (Fig. 4c) the red subgroup shows Caucasian males with at least a felony and 15 priors, older than 45. The other subgroups are characterized by priors around 6, ages between 25 and 45. These are obtained by aggregating the instances that share the same group CF, as output by CounterFair. In the case of the Compas dataset, the identified subgroups may provide further data to analyze the people that are being misclassified as recidivists and why. For example, in the case of the red group (Caucasian males with a felony, almost 15 priors, older than 45), only 22 false-negative instances were found, while for the dark green (Caucasian males with a felony, around 6 priors, between 25–45 years of age) there were 210, indicating an almost 10 times larger false-negative ratio for the latter. This is interesting, since there are more false negatives in



(a) Dutch dataset (b) German dataset (c) Athlete dataset

(d) Student dataset (e) Compas dataset

(f) Adult dataset

**Fig. 3** Aggregated burden, identified subgroups and fair CFs burden output. Each plot has four points in the $x$-axis: three for the $\alpha$ values of 0.1, 0.5 and 1.0 using cost function $Z_1$, and one using cost function $Z_2$. The bars show the burden (on the left $y$-axis), while the diamonds the number of distinct subgroups for each sensitive group (on the right $y$-axis). The legends indicate the sensitive groups and their number of false negatives in parenthesis

the younger age group, even as there are less prior counts of crimes committed, indicating that the model might have an age bias. We now focus on all these identified subgroups.

We investigate these subgroups by performing two analysis: (1) *inter-subgroup analysis*: whether the identified subgroups present lower, comparable or higher biases with respect to other subgroups, and (2) *subgroup-group analysis*: whether the identified subgroups present lower, comparable or higher biases with regards to the sensitive feature groups that compose them. Explicitly, in the *subgroup-group analysis*, we compare the average distances of a given subgroup (the main component of the AWB measure) to their shared counterfactual, for example, of a subgroup of African-American Females, with respect to the average distances to the counterfactuals of only Females, and only African-Americans, separately. The hypothesis behind the *subgroup-group analysis* is: whenever a model presents biases against a given sensitive feature group (such as Female or African-American) it may present a similar or higher bias level against a subgroup lying in the intersection of both (African-American Females). The following section (Sect. 5.3), focuses on the study of subgroups identified for the Adult, Compas and Student datasets, which present at least 2 different sensitive features, allowing for subgroups formed by the intersection of their different sensitive groups. Section 5.3 will first focus on the evaluation of the sizes of each subgroup, then, based on the most relevant subgroups based on their size, we will proceed to perform the *inter-subgroup analysis* followed by the *subgroup-group analysis*.

**Table 2** The top ten largest subgroups in the Adult dataset

| $G$ | Sex | Race | Age | Size |
|---|---|---|---|---|
| 2 | Male | White | 25–60 | 60 |
| 8 | Male | White | 25–60 | 126 |
| 9 | Male | White | 25–60 | 81 |
| 13 | Male | White | 25–60 | 82 |
| 18 | Male | White | >60 | 52 |
| 20 | Male | White | 25–60 | 45 |
| 23 | Male | White | 25–60 | 49 |
| 28 | Male | Non-White | 25–60 | 50 |
| 51 | Female | White | 25–60 | 60 |
| 59 | Female | White | 25–60 | 108 |

### 5.3 Subgroups

Some of the datasets, as observed in Figs. 4 and 5, present a relatively high number of identified subgroups, while others a smaller set. The Adult dataset has the highest number of identified subgroups. This is due to the fact that it has the highest number of sensitive features, 3 in total: Sex, Race and Age. The Sex feature has two categories (Male and Female), the Race feature has two categories (White and Non-white) and the Age feature has three categories (Less than 25, 25–60 and Greater than 60). This leads to a set of 12 different combinations of sensitive subgroups.

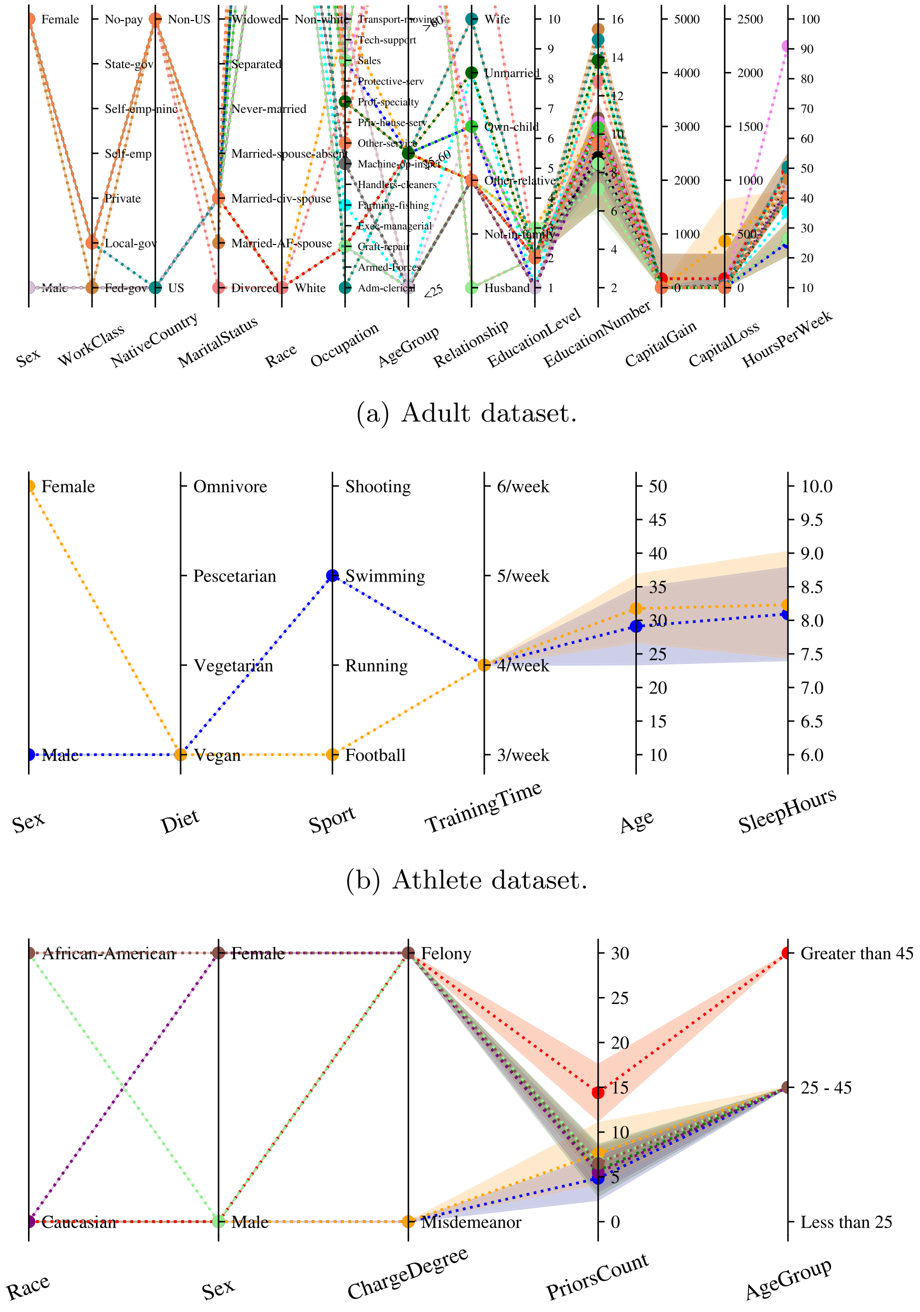
However, the CounterFair algorithm, when minimizing the cost function $Z_1$ with $\alpha = 0.1$, may obtain more than one counterfactual point for each of these potential combinations of sensitive groups, since these points may be located arbitrarily in space, leading to the formation of different subgroups inside these groups (for example, White Males under 25 that work less than 40 h per week, and White Males under 25 that work more than 80 h per week). Therefore, it is reasonable to observe the behavior in Fig. 4a where several smaller groupings are identified for a single sensitive subgroup. For the Adult dataset, a total of 84 subgroups are identified. Out of these subgroups, the top ten subgroups based on size are selected for the *inter-subgroup* and the *subgroup-group* analyses. These top ten groups are shown in Table 2.

In the Compas dataset, there are two sensitive features, namely Race and Sex, each having two categories: African-American and Caucasian, and Male and Female, respectively. Given that the subgroups are smaller, their sizes can be more easily observed in Fig. 6, while Table 3 shows the details of these identified subgroups.

Since there are only 7 different subgroups available, we perform the *subgroup-group* and the *inter-subgroup* analyses on all the subgroups.

Finally, in the student dataset, there are two sensitive features, namely Age and Sex, each having two categories: Less than 18 and Greater than or equal to 18, and Male and Female, respectively. The subgroups are observed in Fig. 7, while Table 4 shows the details of these identified subgroups.

We now proceed to present and discuss the results of the *inter-subgroup* and the *subgroup-group* analyzes.

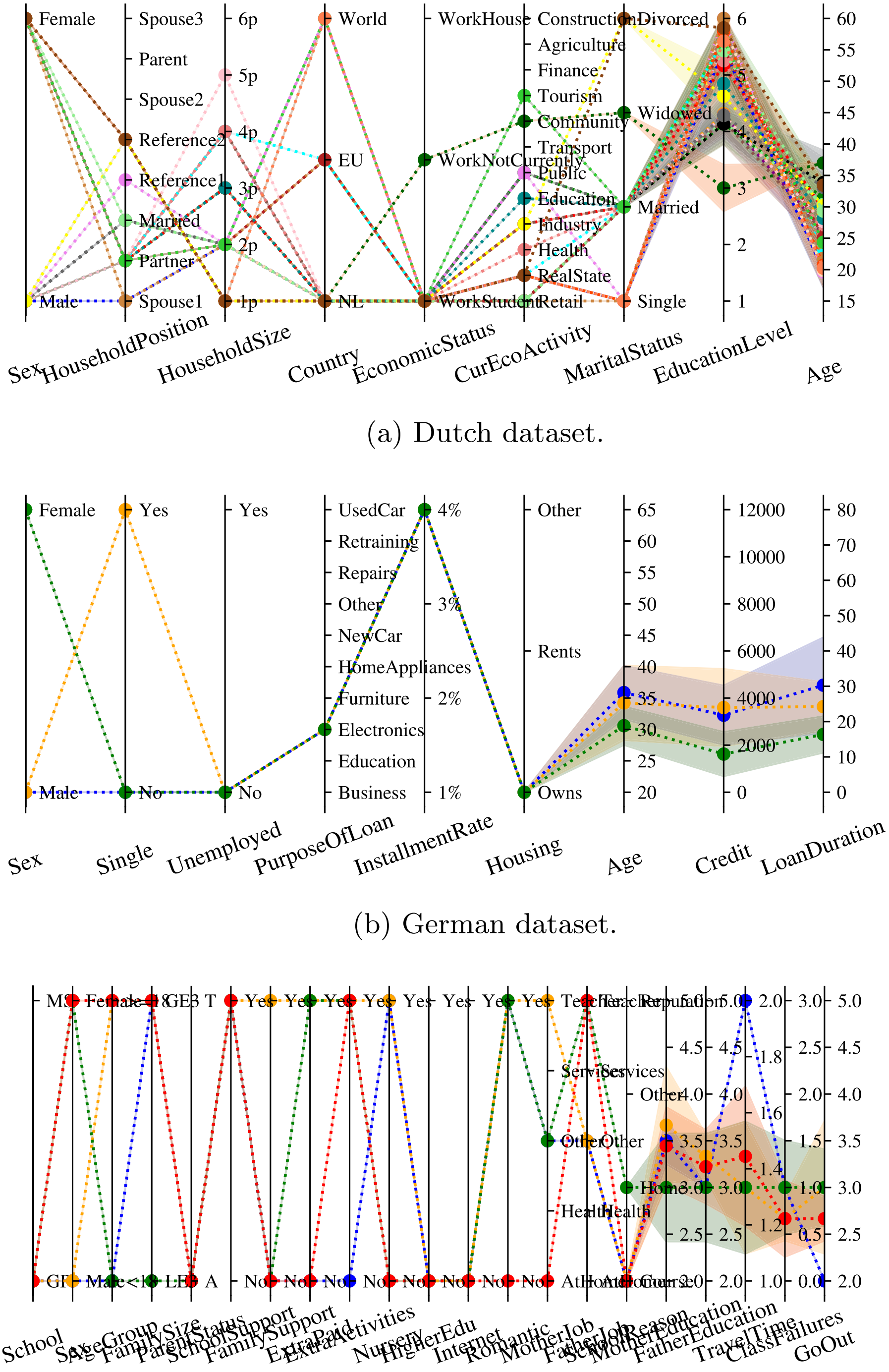


(a) Adult dataset.

(b) Athlete dataset.

(c) Compas dataset.

**Fig. 4** Subgroup details for the Adult, Athlete and Compas datasets, running CounterFair with $\alpha = 0.1$. The shaded regions have a width equal to one standard deviation of the features values of each subgroup



(a) Dutch dataset.

(b) German dataset.

(c) Student dataset.

**Fig. 5** Subgroup details for the Dutch, German and Student datasets, running CounterFair with $\alpha = 0.1$. The shaded regions have a width equal to one standard deviation of the features values of each subgroup

**Fig. 6** Pie chart indicating the relative importance of each subgroup identified in the Compas dataset

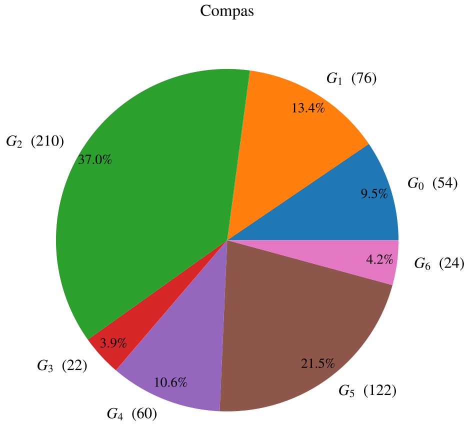


**Table 3** Different subgroups identified for the Compas dataset

| $G$ | Race | Sex | Size |
|---|---|---|---|
| 0 | African-American | Male | 54 |
| 1 | African-American | Male | 76 |
| 2 | African-American | Male | 210 |
| 3 | African-American | Male | 22 |
| 4 | African-American | Female | 60 |
| 5 | Caucasian | Male | 122 |
| 6 | Caucasian | Female | 24 |

**Fig. 7** Pie chart indicating the relative importance of each subgroup identified in the Student dataset

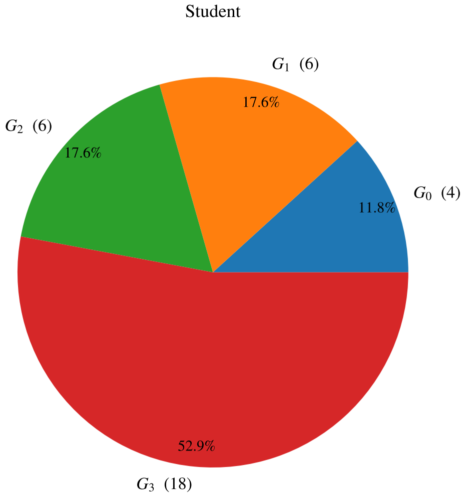


**Table 4** Different subgroups identified for the Student dataset

| $G$ | Sex | Age | Size |
|---|---|---|---|
| 0 | Male | $< 18$ | 4 |
| 1 | Male | $\geq 18$ | 6 |
| 2 | Female | $< 18$ | 6 |
| 3 | Female | $\geq 18$ | 18 |

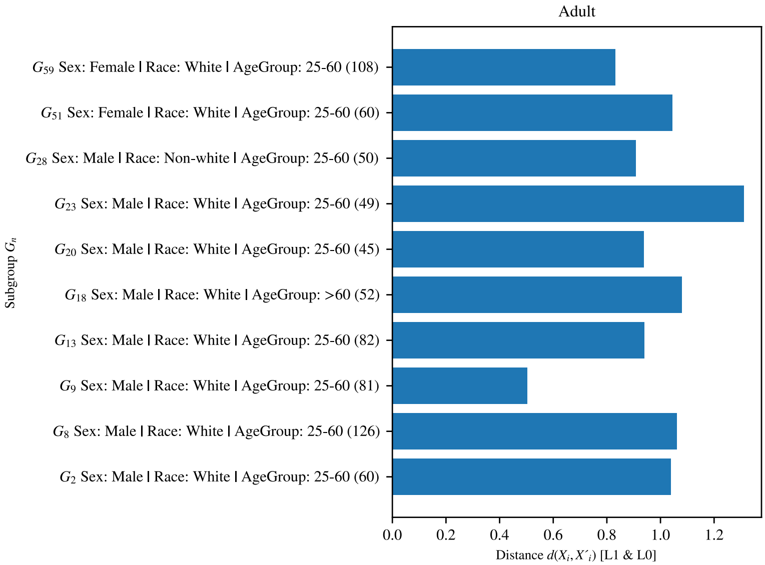


**Fig. 8** Bar plot showing the $L1$ and $L0$-norm for the top ten subgroups of the Adult dataset

### 5.3.1 Inter-subgroup analysis

In the Adult dataset the average distance between each of the top ten subgroups and their corresponding counterfactual is observed in Fig. 8. It is important to recall that the higher the distance, the higher the burden, and therefore the higher the biases *against* that specific subgroup.

The first observation from Fig. 8 is that six of the top ten subgroups by size (namely $G_2$, $G_8$, $G_9$, $G_{13}$, $G_{20}$ and $G_{23}$) correspond to White Males between the ages of 25 and 60. Additionally, the amount of instances compiled in these subgroups is 443, which is 62% of the 713 instances collected in the top ten subgroups. Note as well that there does not seem to be a bias in favor of these six subgroups based on the distances from their instances to their corresponding counterfactuals and the largest average distance corresponds to the subgroup $G_{23}$. The second largest distance corresponds to the subgroup $G_{18}$ of White Males older than 60, while the third largest to $G_8$, the largest subgroup with 126 instances of White Males between the ages of 25 and 60. Based on this information, there might be biases related to the increased number of White Males in the ages of 25 to 60, evidenced in the larger amount of instances in these subgroups. Based solely on the average distances to the counterfactuals, it is not evident that there might be larger, comparable or smaller biases across the different subgroups for the Adult dataset.

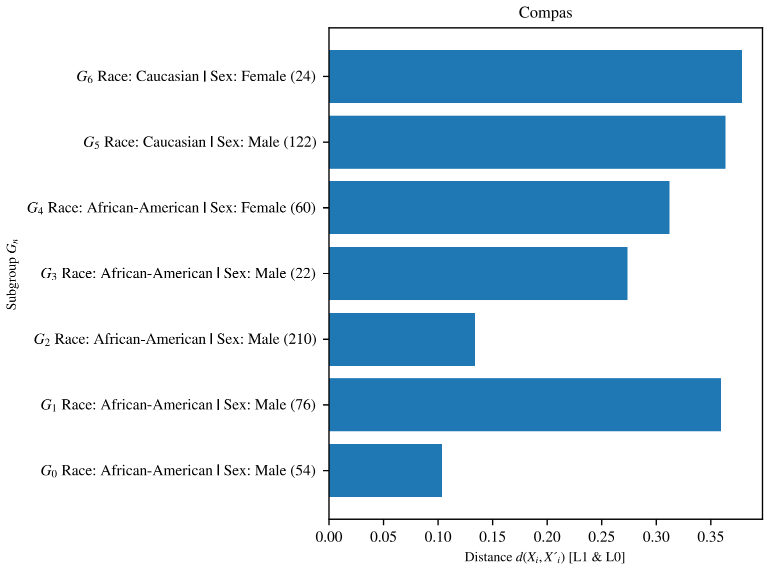


**Fig. 9** Bar plot showing the $L1$ and $L0$-norm for the Compas dataset

In the Compas dataset, the average distance between each of the subgroups and their corresponding counterfactual is observed in Fig. 9.

The largest distance is observed for the $G_6$ subgroup, corresponding to a set of Caucasian Female individuals. However, this group is small in size when compared to the largest subgroups of Males. Out of the 7 subgroups, 4 correspond to African-American Males (namely $G_0$, $G_1$, $G_2$ and $G_3$), adding up to 362 out of 568 false negatives in the dataset (64%) and 5 subgroups correspond to Males. The subgroup with the second highest average distance is $G_5$ corresponding to a set of 122 Caucasian Males. The third largest corresponds to $G_1$, a subgroup of 76 African-American Males. Similarly to the Adult dataset, the average distances between the subgroups and their corresponding counterfactuals does not seem necessarily related to the levels of bias, more than the amount of false negatives does.

In the Student dataset, the average distance between each of the subgroups and their corresponding counterfactual is observed in Fig. 9.

In the student dataset (see Fig. 10) there is a clear distinction between the 4 subgroups shown: the ones having Females have a much higher distance or bias against them. Subgroups $G_3$ and $G_2$ have almost double the average distance than subgroups $G_0$ and $G_1$, indicating a much higher bias for the Females than for the Males. However, note there is not such a clear distinction between the subgroups across the Age groups, with only a slight difference between the $G_1$ and $G_0$ subgroups of Males, where the younger ones in $G_0$ seem to have a higher average distance to their counterfactuals than the older Male students subgroup $G_1$. We now proceed to perform the *subgroup-group* analysis in the Adult, Compas and Student datasets.

### 5.3.2 Subgroup-group analysis

For this analysis, the average distances of the subgroups are compared to the average distances of the sensitive feature groups that compose the subgroups. As previously analyzed, the top

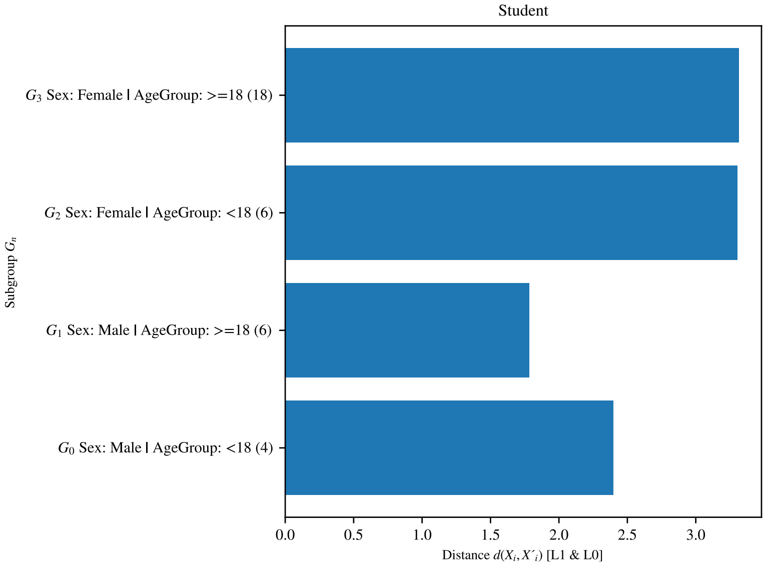


**Fig. 10** Bar plot showing the $L1$ and $L0$-norm for the Student dataset

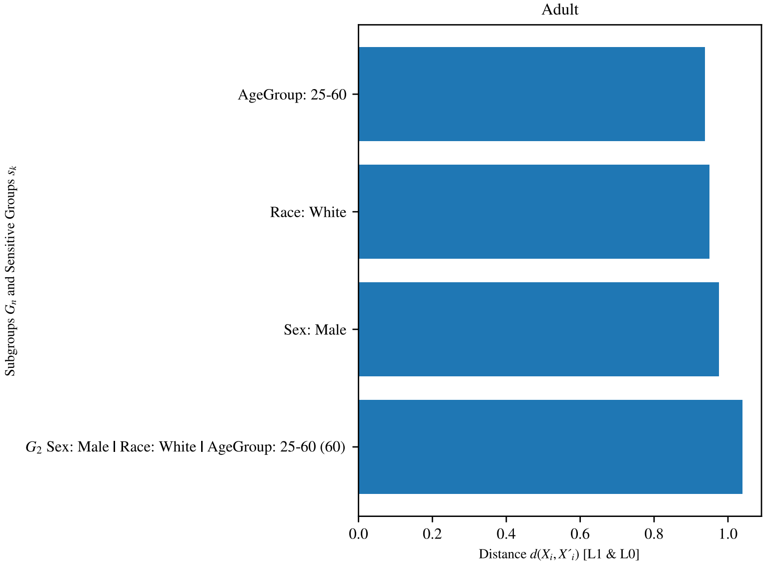


**Fig. 11** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_2$ of the Adult dataset with respect to the forming sensitive feature groups

ten subgroups of the Adult dataset are studied. For each of the subgroups, a plot and analysis is done and shown. For six of the subgroups, the sensitive groups to compare to are the same (White Males between the ages of 25 and 60).

Figure 11 indicates a slightly higher bias in terms of the average distance of the subgroup $G_2$, which is White Males between the ages of 25 and 60, with respect to the sensitive feature groups Male, Whites and 25–60 Age group, respectively. Figure 12 indicates a slightly higher

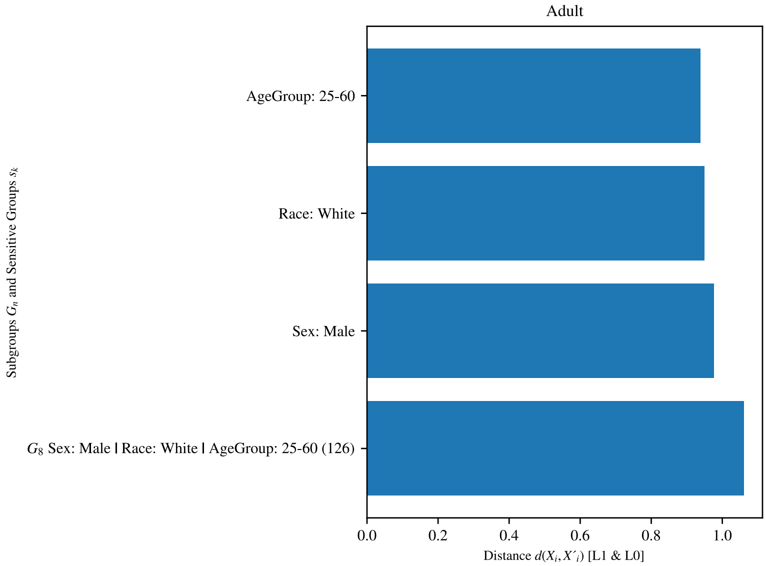


**Fig. 12** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_8$ of the Adult dataset with respect to the forming sensitive feature groups

bias in terms of the average distance of the subgroup $G_8$, with respect to the sensitive feature groups Male, Whites and 25–60 Age group, respectively. In this case, Fig. 13 indicates a lower bias in terms of the average distance of the subgroup $G_9$ with respect to the sensitive feature groups Male, Whites and 25–60 Age group, respectively. This was expected, according to the observations in Fig. 8, where the subgroup $G_9$ is shown to have the lowest average distance. Figure 14 indicates a comparable level of bias in terms of the average distance of the subgroup $G_{13}$, with respect to the sensitive feature groups Male, Whites and 25–60 Age group, respectively. Figure 15 indicates a higher level of bias in terms of the average distance of the subgroup $G_{18}$, with respect to the sensitive feature groups Males, Whites and above 60 Age group, respectively. So far, this indicates that the subgroups at the intersection of the sensitive groups may be, in general, showing higher biases than the forming sensitive features in most of the cases. In Fig. 16, the level of bias of the subgroup $G_{20}$ is relatively similar between the subgroup and the forming sensitive groups, as is the case of subgroup $G_{13}$. In Fig. 17, the level of bias of the subgroup $G_{23}$ is considerably higher than that of the forming sensitive groups, being almost 30%. In Fig. 18, the level of bias of the subgroup $G_{28}$ is relatively similar to that of the forming sensitive groups. In Fig. 19, the level of bias of the subgroup $G_{51}$ is considerably higher than that of the forming sensitive groups, specifically that of the Females, being 25% higher. In Fig. 20, the level of bias of the subgroup $G_{59}$ is relatively similar, and somewhat lower than that of the forming sensitive groups, specifically that of the White, and between 25 and 60.

The *Subgroup-group* analysis for the Compas dataset is done using the following figures. In Fig. 21, the level of bias of the subgroup $G_0$ is considerably lower than that of the forming sensitive groups (African-Americans and Males). In Fig. 22, the level of bias of the subgroup $G_1$ is considerably higher than that of the forming sensitive groups (African-Americans and Males). In Fig. 23, the level of bias of the subgroup $G_2$ is considerably lower than that of the forming sensitive groups (African-Americans and Males). In Fig. 24, the level of bias of

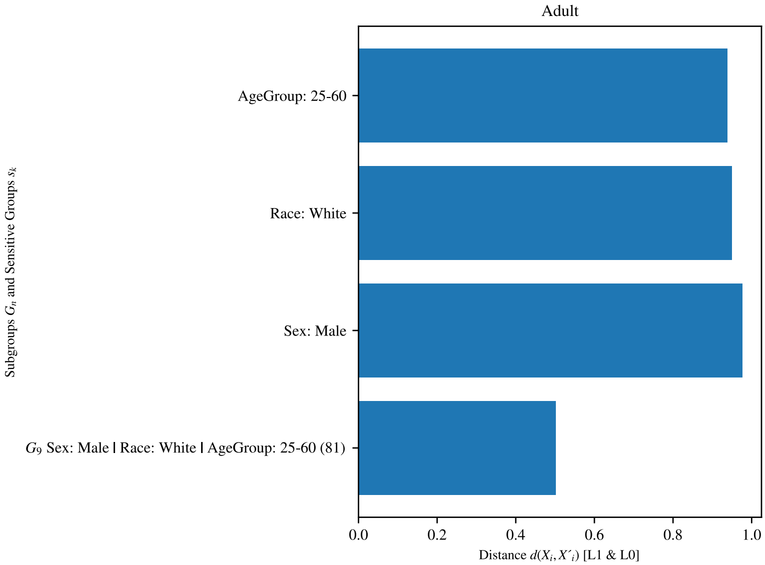


**Fig. 13** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_9$ of the Adult dataset with respect to the forming sensitive feature groups

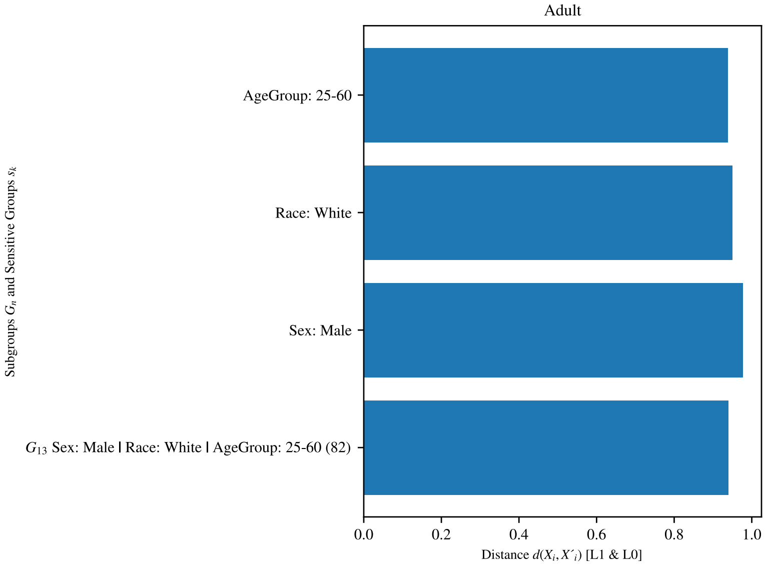


**Fig. 14** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_{13}$ of the Adult dataset with respect to the forming sensitive feature groups

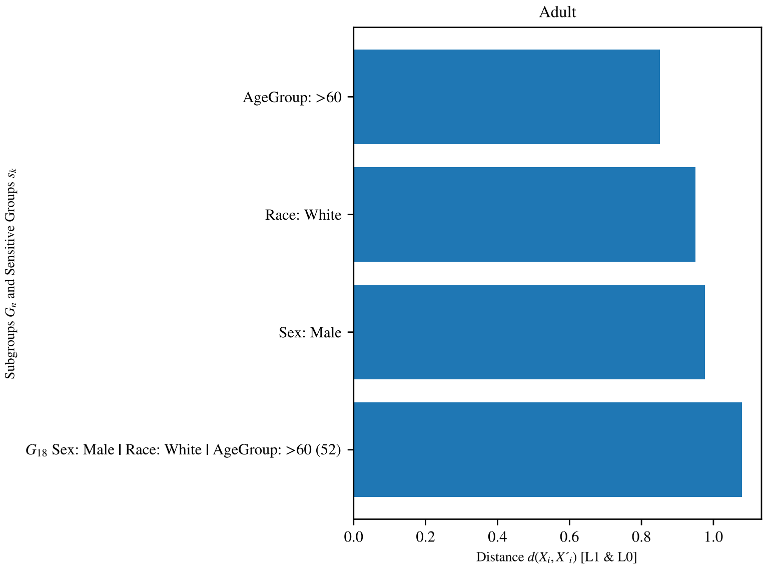


**Fig. 15** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_{18}$ of the Adult dataset with respect to the forming sensitive feature groups

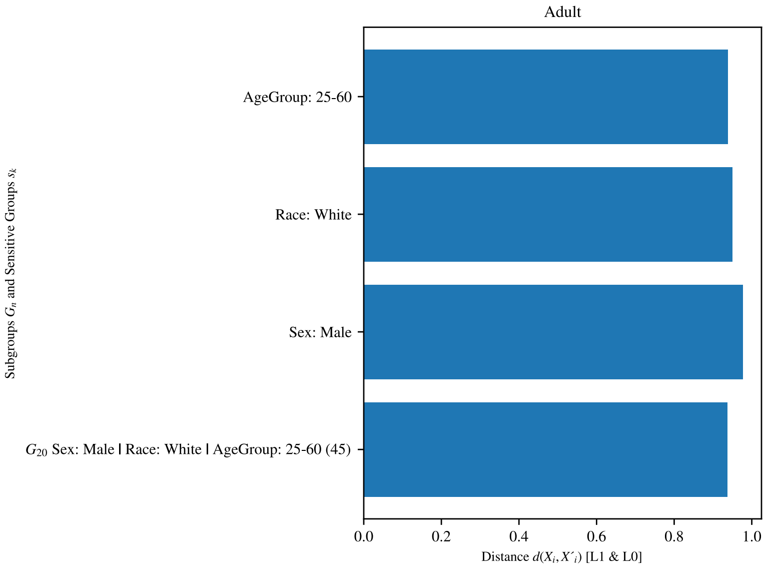


**Fig. 16** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_{20}$ of the Adult dataset with respect to the forming sensitive feature groups

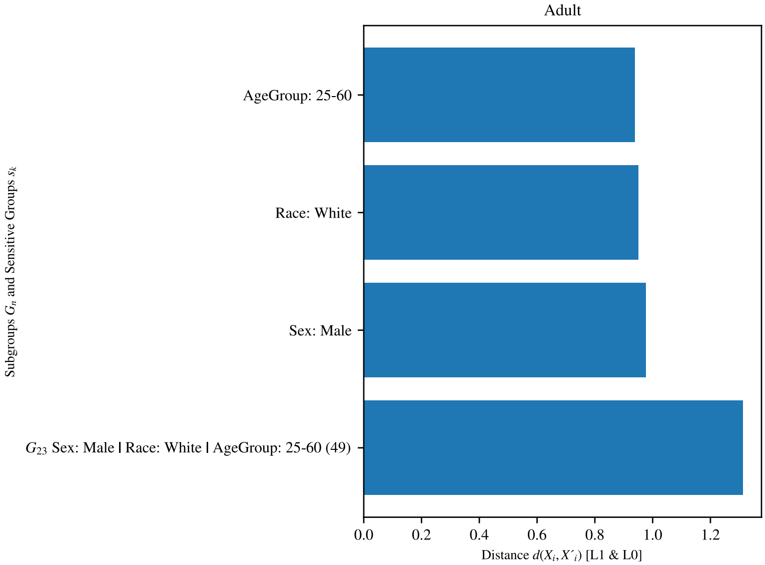


**Fig. 17** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_{23}$ of the Adult dataset with respect to the forming sensitive feature groups

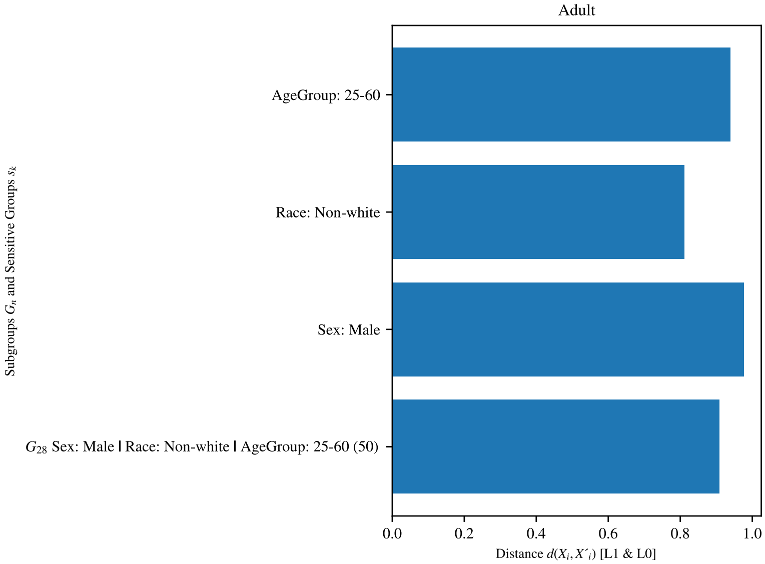


**Fig. 18** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_{28}$ of the Adult dataset with respect to the forming sensitive feature groups

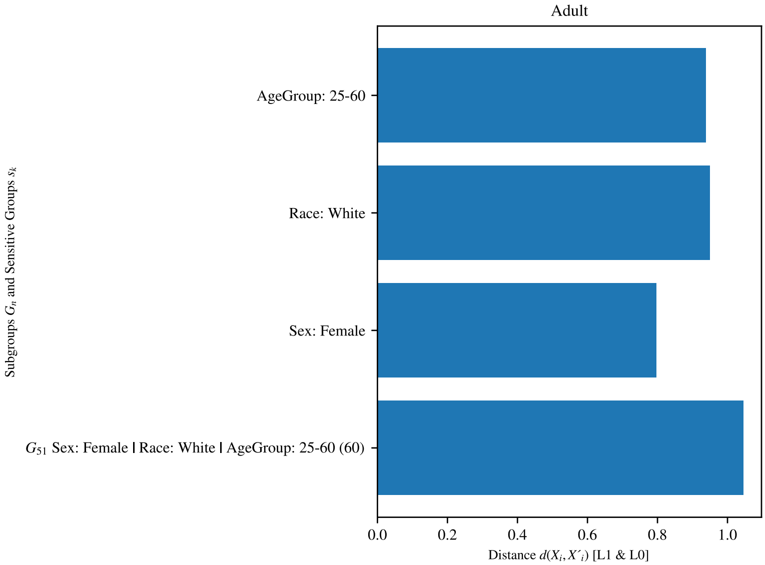


**Fig. 19** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_{51}$ of the Adult dataset with respect to the forming sensitive feature groups

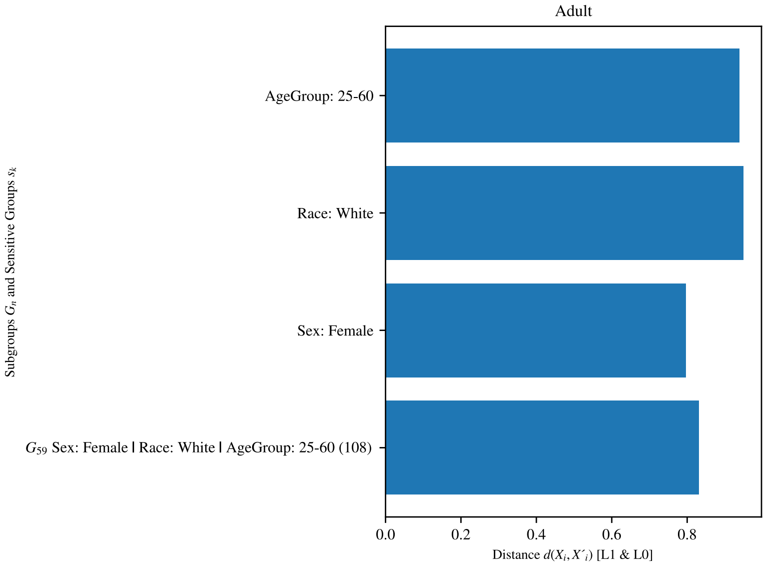


**Fig. 20** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_{59}$ of the Adult dataset with respect to the forming sensitive feature groups

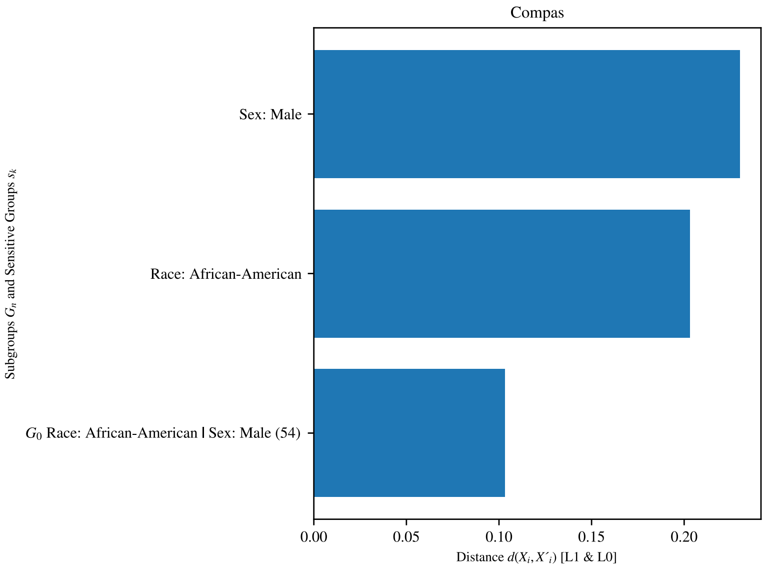


**Fig. 21** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_0$ of the Compas dataset with respect to the forming sensitive feature groups

the subgroup $G_3$ is considerably higher than that of the forming sensitive groups (African-Americans and Males). Out of 4 subgroups analyzed so far with the African-American Male groupings, 2 show lower biases and 2 show higher biases with respect to the forming sensitive subgroups. In Fig. 25, the level of bias of the subgroup $G_4$ is considerably higher than that of the African-Americans, but not of the Females. This indicates that being African-American and Female may be less favored off than being African-American and Male. In Fig. 26, the level of bias of the subgroup $G_5$ is high and indicates that more than being Male, the Race is affecting considerably, since the level of bias relates highly to the Caucasian Race. In Fig. 27, the level of bias of the subgroup $G_6$ is relatively high with respect to the forming sensitive groups.

In general, in the Compas dataset, there is no clear indication of higher or lower biases in the subgroups found, as was found in the Adult dataset.

In the student dataset, the following figures show the comparison between the subgroups and the forming sensitive groups. In Fig. 28, the level of bias of the subgroup $G_0$, which is Male and younger than 18, is relatively high with respect to the Male group, but considerably lower than that of the younger than 18 group. In Fig. 29, the level of bias of the subgroup $G_1$, which is Male and older than 18, is lower with respect to both the Male group and the older-than-18 group. In Fig. 30, the level of bias of the subgroup $G_2$, which is Female and younger-than-18, is similar the Female group, but higher than that of the younger-than-18 group. In Fig. 30, the level of bias of the subgroup $G_2$, which is Female and younger-than-18, is similar to the one of the Female group, but higher than that of the younger-than-18 group. In Fig. 31, the level of bias of the subgroup $G_3$, which is Female and older-than-18, is similar to the one of the Female group, but higher than that of the older-than-18 group. The last two groups probably indicate that the important variable here is more the age than the gender. As mentioned before, it seems the comparison among subgroupings and forming sensitive groups in the student dataset is not as clear as it was for this dataset in the comparison across

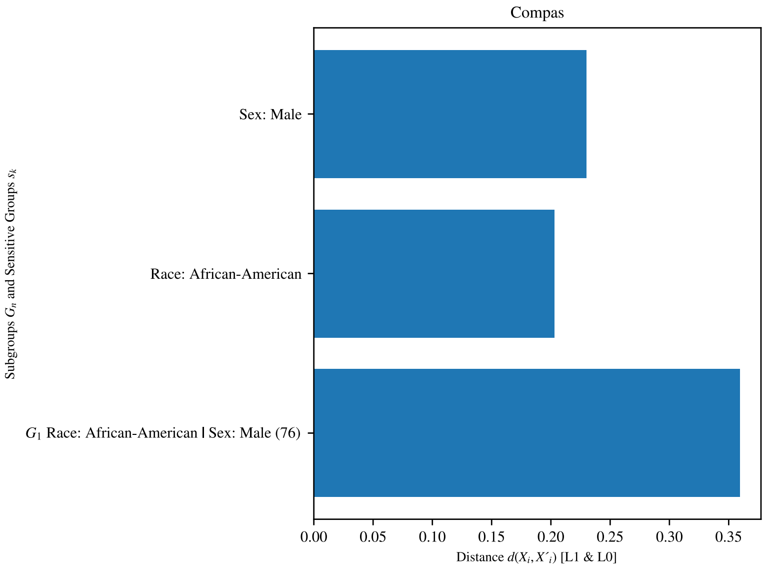


**Fig. 22** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_1$ of the Compas dataset with respect to the forming sensitive feature groups

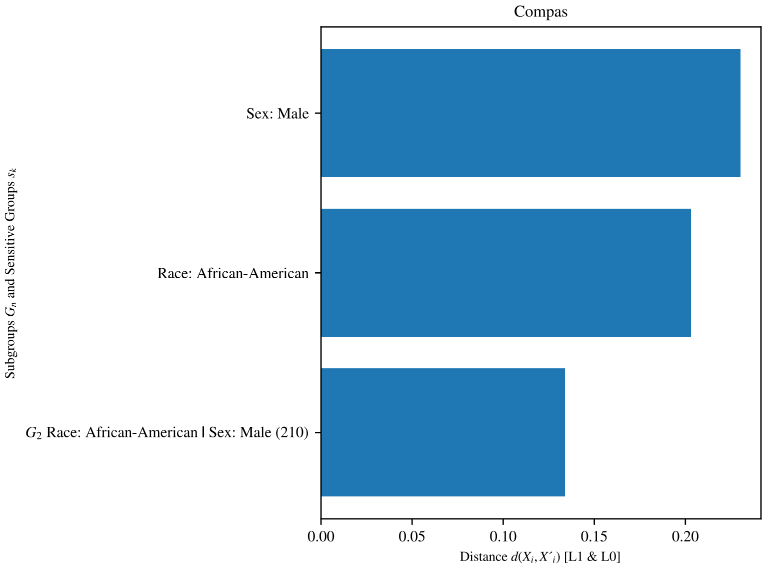


**Fig. 23** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_2$ of the Compas dataset with respect to the forming sensitive feature groups

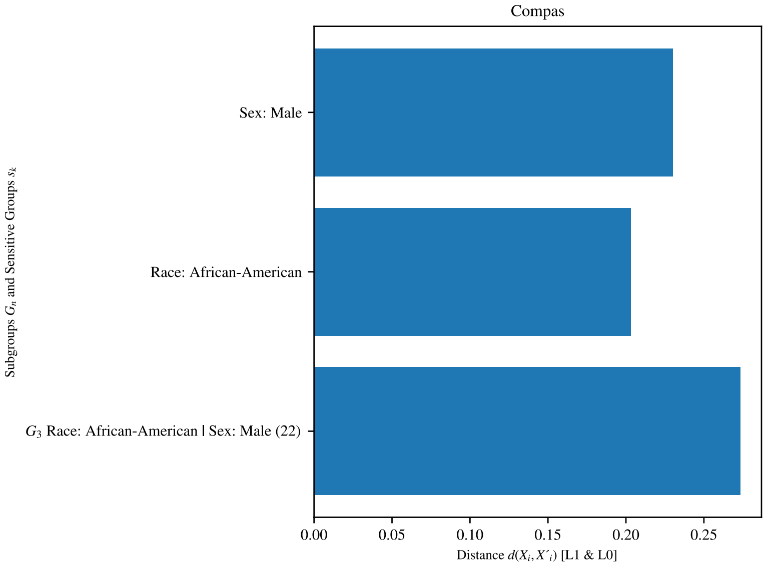


**Fig. 24** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_3$ of the Compas dataset with respect to the forming sensitive feature groups

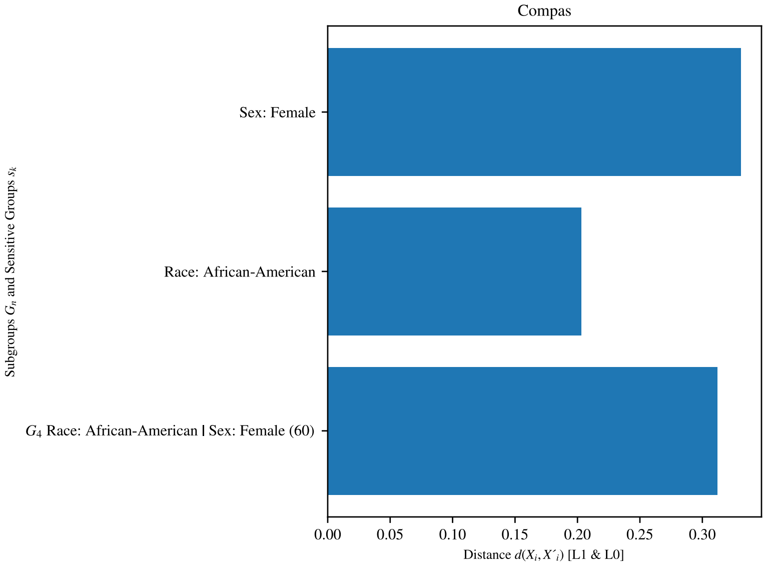


**Fig. 25** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_4$ of the Compas dataset with respect to the forming sensitive feature groups

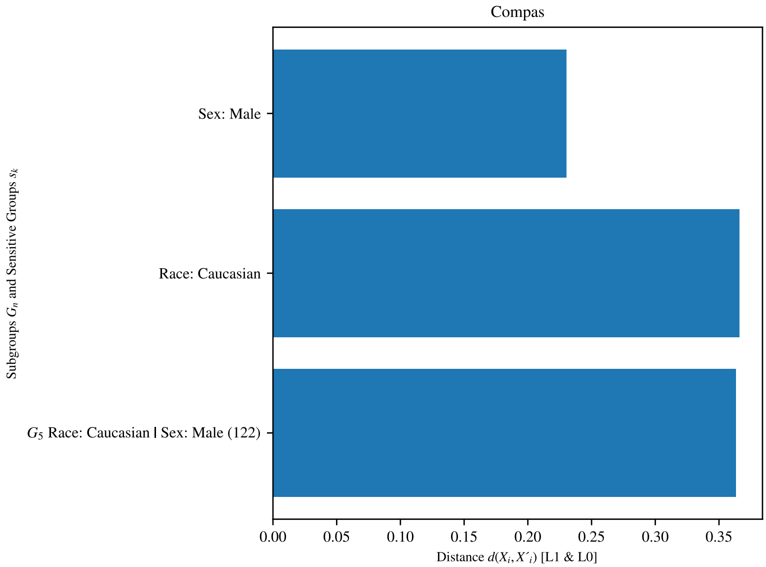


**Fig. 26** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_5$ of the Compas dataset with respect to the forming sensitive feature groups

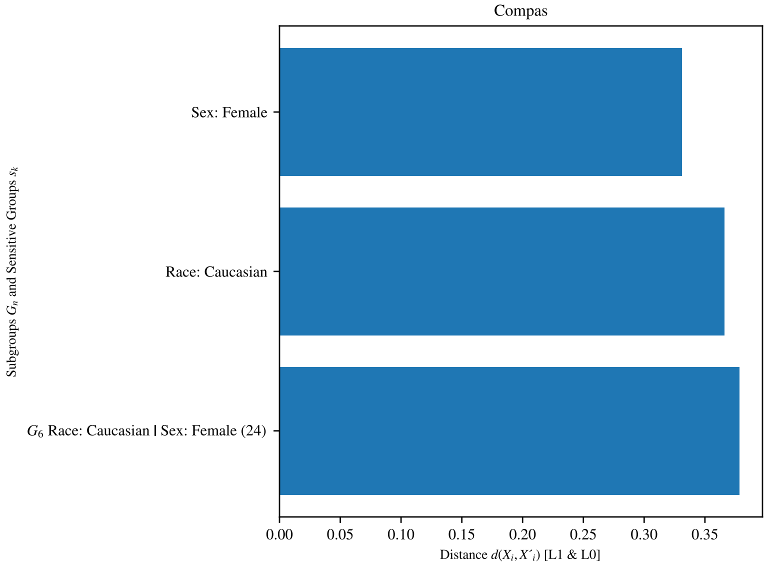


**Fig. 27** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_6$ of the Compas dataset with respect to the forming sensitive feature groups

different subgroups, however, specifically for the Females, it is perhaps possible to see that the gender is leading the most bias when compared to the Age group.

The analysis from the perspective of fairness into the subgroups found may shed light into the aspects of counterfactual intersectional fairness, where the intersections of these groups can be found. Specifically for the cases in which the subgroup has a larger distance or burden from their corresponding group counterfactuals, a further analysis can be done. In the Adult dataset, for the $G_2$ and $G_8$ subgroups, which are White Males between 25 and 60 years of age, these subgroups have a higher distance or burden when compared to the individual groups of Whites, Males and people between 25 and 60 alone. The general reason, which can cover all kinds of intersectional bias issues (not only this model and dataset), is that there is a smaller representation of the subgroup of White Males between 25 and 60 years of age, than there is of only Whites, only Males, or only people between 25 and 60. Looking more into detail, some interesting information is revealed that could shed light into what is driving these higher intersectional group biases.

Firstly, the Adult $G_8$ subgroup happened to also have all of its members married and with significant capital disadvantages (having 0 average capital gains and an average of 15 in capital losses). This indicates a very particular behavior for this group, because out of all the 60 white males inside it, none is single or divorced, and none of them has reported positive capital gains. This may be causing the model to learn a further or more clear decision boundary for this specific group of people.

A similar behavior can be observed for the Adult $G_2$ subgroup. This group curiously has all of its member divorced, and all have a significant capital gain (average of 160). Another thing to note is that they have one of the highest education levels, averaging 9.8 (10 is the top at College Professor level). Our intuition tells us that, with high levels of education and higher capital gains, these people should be closer to the desired (correct) class labels (remember these are false negatives). However, the model may be highly biased to consider divorce as a highly important feature to decrease the estimated level of wealth of these individuals, and therefore creates a potentially further decision boundary that contributes to these higher biases.

Furthermore, the Compas $G_1$ subgroup also shows a high subgroup bias. In this subgroup, the 76 people showed an average of 7.3 prior felony counts. For Compas subgroup $G_2$, which had an actual lower burden compared to that of $G_1$ with respect to their individual demographic groups, the number of prior felony counts was lower, with an average of 5.8. Again, these features may be causing a slightly different decision boundary trained around them that could be based on these features and leading to a higher difficulty for these false negatives to reach their corresponding group counterfactuals on the desired label space.

### 5.3.3 Comparison of CounterFair with AReS and FACTS with respect to burden, effectiveness and run time

We ran the experiments of AReS and FACTS with two considerations: (1) following the authors recommended support threshold of 1% [7, 10] and (2) limiting the execution time to maximum 1 week per dataset. However, the threshold had to be modified to run within the time limit, but at most to 10% (beyond this point, the performance significantly degrades). Figure 32 shows the AWB, effectiveness and run times of CounterFair, AReS and FACTS. CounterFair mostly outperforms AReS and FACTS in burden and effectiveness. AReS and FACTS beat CounterFair in AWB in the Dutch dataset Females (FACTS beats it on Males and Females). AReS also beats CounterFair in Males and Females in AWB in the Athlete dataset. CounterFair significantly beats them in effectiveness in all cases. For AReS this

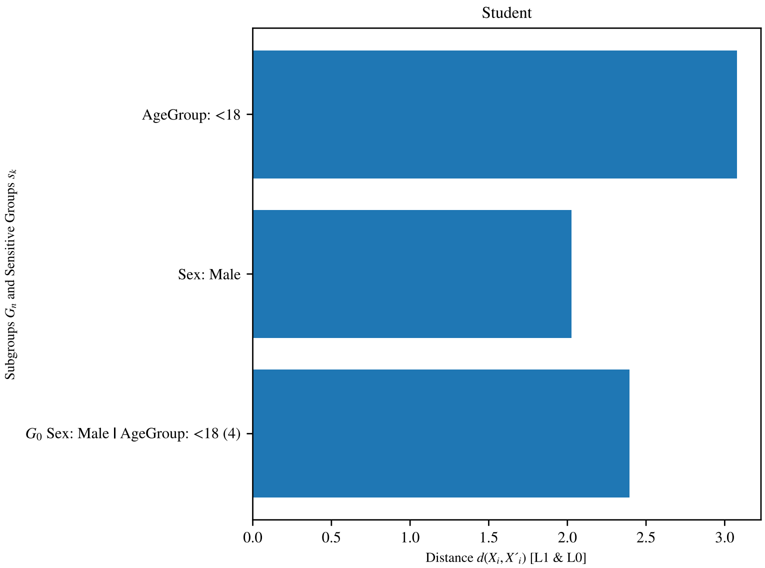


**Fig. 28** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_0$ of the Student dataset with respect to the forming sensitive feature groups

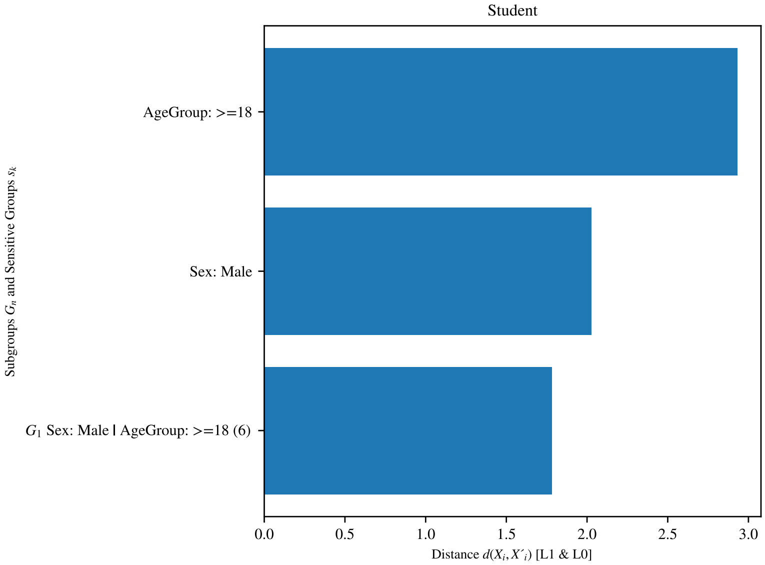


**Fig. 29** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_1$ of the Student dataset with respect to the forming sensitive feature groups

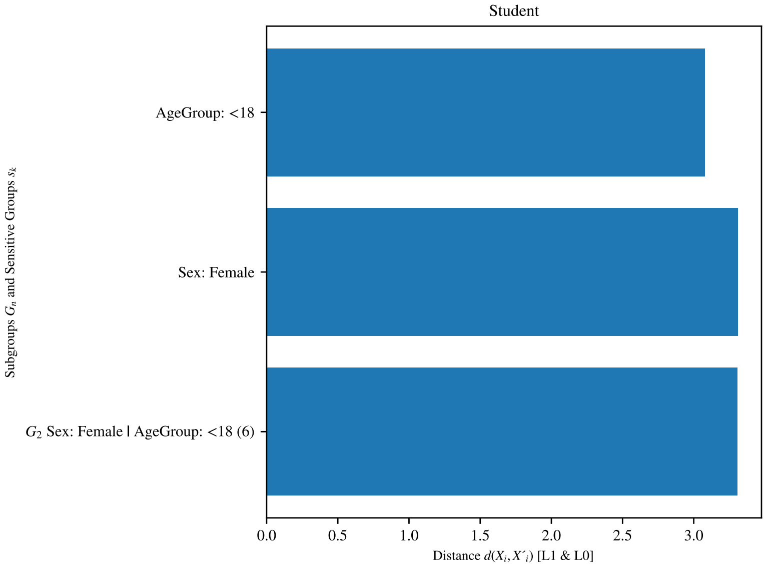


**Fig. 30** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_2$ of the Student dataset with respect to the forming sensitive feature groups

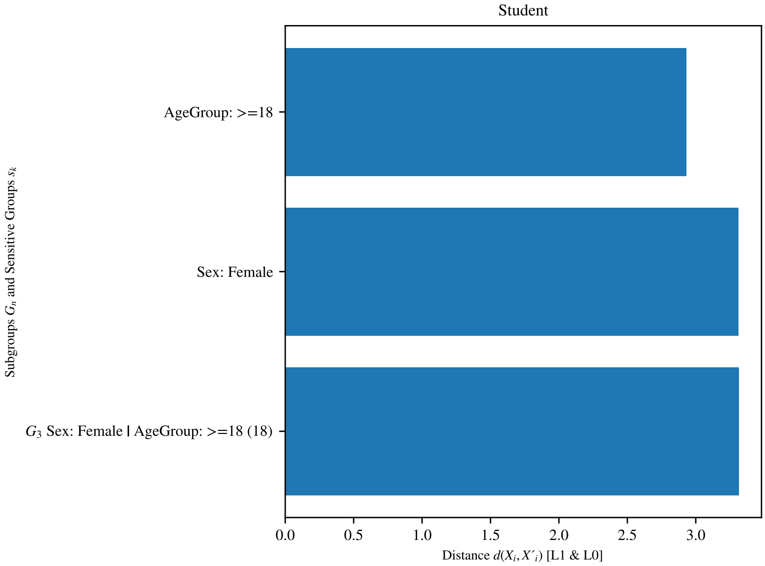
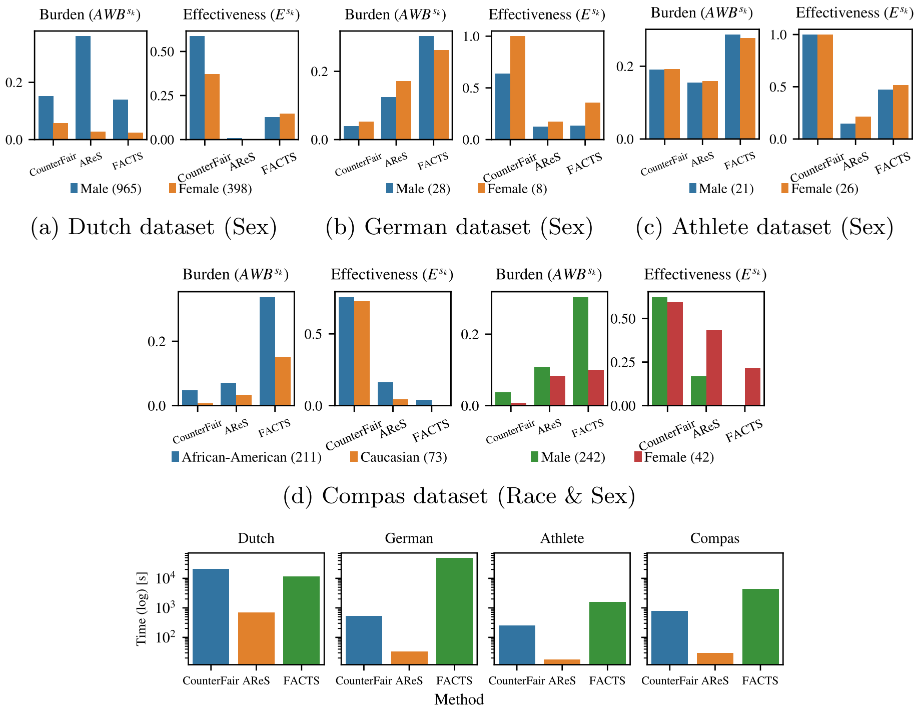


**Fig. 31** Bar plot showing the $L1$ and $L0$-norm for the subgroup $G_3$ of the Student dataset with respect to the forming sensitive feature groups



(a) Dutch dataset (Sex) (b) German dataset (Sex) (c) Athlete dataset (Sex)

(d) Compas dataset (Race & Sex)

(e) Run times for CounterFair, AReS and FACTS

**Fig. 32** CounterFair, AReS and FACTS performance. Lower AWB, higher effectiveness and lower times are better. CounterFair is run with $Z_1$ and $alpha = 1.0$ for AWB, and with $Z_3$ for effectiveness

can be explained by the lack of feasibility constraints on the CFs, leading sometimes to infeasible CFs. Timewise, AReS is the fastest, and CounterFair is at least 10 times faster than FACTS except in the Dutch dataset (ran with 10%). We excluded Adult and Student since the recommended threshold of 1% overshot the run time beyond the week, or it had to be raised beyond 10%, hindering the performance.

## 6 Conclusions and future work

We propose CounterFair, an MP-based, model-agnostic CF generation algorithm that can detect biases, mitigate them, and identify relevant subgroups in the data, all via group CF generation. The generation of group CFs requires only the input of the feature properties of mutability, directionality and possible values. CounterFair is, as demonstrated, adaptable to generate CFs based on different cost functions thanks to its flexibility in cost and constraints definitions. An example is analyzed with group effectiveness, and it is the only group CF generation method, to the best of our knowledge, that is also able to reduce the burden biases among sensitive groups by selecting CFs that decrease the difference in aggregated burden among them. From a holistic perspective, having a tool that is not only able to detect biases, but also extract fair recommendations based on the trained ML models is useful for scientists and developers but also useful for users who are looking to find ways to improve their condition without them turning to be unfair with their peers. We have additionally extended the discussion by analyzing the subgroups identified on the datasets that presented more than 2 sensitive features, i.e., Adult, Compas and Student datasets, with the goal of detecting whether the biases, based on the measure of distance to the counterfactuals (the main component of the AWB measure) was different across the different subgroups, or between the different subgroups and their corresponding sensitive feature groups. In this regard, the study opened the discussion regarding the intersection of sensitive feature groups, leading to the notion of subgroup counterfactual fairness. As part of future work, other cost functions could be formulated based on the literature on CF explanations quality measures, such as likelihood or sparsity, as well as the usage of other commonly used fairness measures. Additionally, the introduction of intersectional fairness: the study of fairness across the specific found subgroups of interest, is a natural step forward. Moreover, the inclusion of the classifiers as nonlinear constraints in the mathematical programming formulations could be researched. Furthermore, the consideration of non-binary datasets, which should be easy to tackle using, for example, a one-versus-the-rest approach, is also a logical progression, while the scalability and complexity is a good topic to focus on. Finally, the balance between the objectives of identifying biases and subgroup identification could be explored, i.e., either performing an ablation study on the weights given to each term in the cost function depicted in Eq. (9). It could also be interesting to find a way to reformulate the bias mitigation so that the formulation of the problem can remain an integer program, in order to integrate the term $Z_2$ found in Eq. (17) into (9). In this way, the objectives of identification, mitigation of bias and subgroup search could all be balanced and weighted in a further ablative study.

**Acknowledgements** This work has been supported by project MIS 5154714, and under framework of the H.F.R.I call “Basic Research Financing” (H.F.R.I. Project Number: 016636), both under the National Recovery and Resilience Plan “Greece 2.0” funded by the European Union - NextGenerationEU.

**Author Contributions** Alejandro developed the main manuscript text. Panayiotis helped with the definition of subgroups. Zed helped with the presentation and the cover letter. All authors reviewed the manuscript.

**Funding** Open access funding provided by Stockholm University.

**Data Availability** https://archive.ics.uci.edu/.

## Declarations

**Competing interests** The authors declare no competing interests.

## References

1. Guidotti R (2022) Counterfactual explanations and how to find them: literature review and benchmarking. Data Min Knowl Discov:1–55

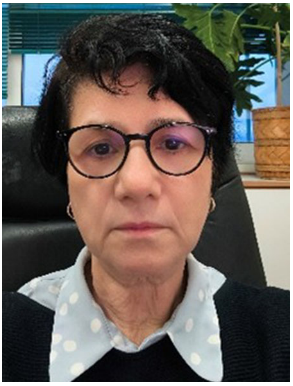
2. Karimi A-H, Barthe G, Balle B, Valera I (2020) Model-agnostic counterfactual explanations for consequential decisions. In: International conference on artificial intelligence and statistics. PMLR, pp 895–905
3. Molnar C (2021) Interpretable machine learning: a guide for making black-box models explainable. https://christophm.github.io/interpretable-ml-book/limo.html
4. Karimi A-H, Barthe G, Schölkopf B, Valera I (2022) A survey of algorithmic recourse: contrastive explanations and consequential recommendations. ACM Comput Surv 55(5):1–29
5. Sharma S, Henderson J, Ghosh J (2020) CERTIFAI: counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society, pp 166–172. https://doi.org/10.1145/3375627.3375812 . arXiv:1905.07857. Accessed 05 March 2022
6. Kuratomi A, Pitoura E, Papapetrou P, Lindgren T, Tsaparas P (2022) Measuring the burden of (un) fairness using counterfactuals. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 402–417
7. Kavouras L, Tsopelas K, Giannopoulos G, Sacharidis D, Psaroudaki E, Theologitis N, Rontogiannis D, Fotakis D, Emiris I (2024) Fairness aware counterfactuals for subgroups. Adv Neural Inf Process Syst 36:58246
8. Carrizosa E, Ramírez-Ayerbe J, Morales DR (2024) Mathematical optimization modelling for group counterfactual explanations. Eur J Oper Res 319(2):399–412
9. Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. Adv Neural Inf Process Syst 30:1–11
10. Rawal K, Lakkaraju H (2020) Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. Adv Neural Inf Process Syst 33:12187–12198
11. Kuratomi A, Lee Z, Chaliane Junior GD, Lindgren T, Papapetrou P, Pitoura E, Tsaparas P (2024) CounterFair: Group counterfactuals for bias detection, mitigation and subgroup identification. In: IEEE international conference on data mining (ICDM)
12. Spangher A, Ustun B, Liu Y (2018) Actionable recourse in linear classification. In: Proceedings of the 5th workshop on fairness, accountability and transparency in machine learning
13. Wang X, Li Q, Yu D, Li Q, Xu G (2024) Counterfactual explanation for fairness in recommendation. ACM Trans Inf Syst 42(4):1–30
14. Pawelczyk M, Broelemann K, Kasneci G (2020) Learning model-agnostic counterfactual explanations for tabular data. In: Proceedings of the web conference 2020, pp 3126–3132
15. Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 607–617
16. ReVelle CS, Eiselt HA (2005) Location analysis: a synthesis and survey. Eur J Oper Res 165(1):1–19
17. Avella P, Sassano A, Vasil'ev I (2007) Computational study of large-scale *p*-median problems. Math Program 109:89–114
18. Verma S, Dickerson J, Hines K (2020) Counterfactual explanations for machine learning: a review. arXiv:2010.10596 [cs, stat]. arXiv: 2010.10596. Accessed 05 March 2022
19. Coston A, Mishler A, Kennedy EH, Chouldechova A (2020) Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. ACM, Barcelona, Spain, pp 582–593. https://doi.org/10.1145/3351095.3372851. Accessed 05 March 2022
20. Lenstra HW Jr (1983) Integer programming with a fixed number of variables. Math Oper Res 8(4):538–548
21. Kannan R, Monma CL (1978) On the computational complexity of integer programming problems. In: Optimization and operations research, pp 161–172. Chap. 17
22. Papadimitriou CH (1981) On the complexity of integer programming. J ACM (JACM) 28(4):765–768
23. Cohen MB, Lee YT, Song Z (2021) Solving linear programs in the current matrix multiplication time. J ACM (JACM) 68(1):1–39
24. Basu A, Conforti M, Di Summa M, Jiang H (2022) Complexity of branch-and-bound and cutting planes in mixed-integer optimization. Math Program 198:1–24
25. Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E (2022) A survey on datasets for fairness-aware machine learning. Wiley Interdiscip Rev Data Min Knowl Discov 12(3):1452

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Alejandro Kuratomi** is an Assistant Professor at the Department of Computer and Systems Sciences of Stockholm University. He holds a Ph.D. from the same department, a M.Sc. in Mechatronics from KTH Royal Institute of Technology, Sweden, a B.Sc. in Mechanical Engineering, and a B.Sc. in Industrial Engineering from Los Andes University, Colombia. His research focuses on interpretable machine learning, algorithmic fairness, optimization, and multivariate time-series classification.

**Zed Lee** received his doctoral degree from Stockholm University in Computer and Systems Sciences. He has also a Masters in Computer Science from KTH Royal Institute of Technology. He is currently a senior manager at Hyundai Motor Company. His research interests include interpretable time-series classification and anomaly detection.

**Panayiotis Tsaparas** is an Associate Professor in the Department of Computer Science and Engineering at the University of Ioannina, Greece, and a Senior Researcher at the Archimedes Research Unit, Athena Research Center. He received his Ph.D. in Computer Science from the University of Toronto and has held research positions at Sapienza University of Rome, the University of Helsinki, and Microsoft Research. His research interests include data mining, machine learning, social network analysis, and algorithmic fairness. He is a Senior Member of the ACM and has served as PC and Senior PC member for leading data mining and database conferences, as well as PC co-chair for WSDM 2023. He has published over 75 peer-reviewed papers and holds 12 patents, eight of which have been awarded.



**Evaggelia Pitoura** is a Professor at the University of Ioannina and Lead Researcher at the Archimedes Research Unit, ATHENA RC, Greece. She holds a BSc degree from the University of Patras, Greece, and MSc and PhD degrees from Purdue University, USA. She has held visiting positions at the University of Pittsburgh, the University of Cyprus, and Georgia Tech. Her current research focuses on social networks and on responsible data management.

**Tony Lindgren** is an associate professor at the Department of Computer and Systems Sciences (DSV), Stockholm University. In 2006, he received his Ph.D. degree in computer and systems sciences. He has worked both in academia and industry, and he is the inventor of numerous patents and the author of numerous scientific articles. His research is aimed at exploring machine learning in general, with a special interest in using data-driven methods for predictive maintenance problems. Since 2019, Tony Lindgren has been the head of the Systems Analysis and Security Unit at DSV.

**Guilherme Dinis Junior** is undertaking an industrial PhD at Stockholm University, specializing in bridging theoretical knowledge with practical applications of sequential decision-making through reinforcement learning. Their research investigates challenges such as learning with scarce signals and developing interpretable solutions, with a focus on designing and implementing novel algorithms for recommender systems in an industrial setting.

**Panagiotis Papapetrou** is a professor at Stockholm University's Department of Computer and Systems Sciences and adjunct professor at Aalto University, Finland. His research focuses on explainable machine learning for temporal data, such as time series and event sequences. He earned his PhD from Boston University and held positions at Aalto and Birkbeck University. Panagiotis has led and contributed to numerous national and international research projects and serves as action editor for DMKD and MACH, and a board member of the Swedish AI Society.