HAL open science

A statistical approach to detect disparity prone features in a group fairness setting

Guilherme Dean Pelegrina, Miguel Couceiro, Leonardo Tomazeli Duarte

To cite this version:

Guilherme Dean Pelegrina, Miguel Couceiro, Leonardo Tomazeli Duarte. A statistical approach to detect disparity prone features in a group fairness setting. AI and Ethics, 2023, 10.48550/arXiv.2305.06994 . hal-04096649

HAL Id: hal-04096649

https://inria.hal.science/hal-04096649v1

Submitted on 13 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

arXiv:2305.06994v1 [cs.LG] 11 May 2023

A STATISTICAL APPROACH TO DETECT SENSITIVE FEATURES IN A GROUP FAIRNESS SETTING

A PREPRINT

Guilherme Dean Pelegrina
School of Applied Sciences (FCA)
University of Campinas (UNICAMP)
Limeira, Brazil
guidean@unicamp.br

Miguel Couceiro
Université de Lorraine, CNRS, LORIA
F-54000 Nancy, France
miguel.couceiro@loria.fr

Leonardo Tomazeli Duarte
School of Applied Sciences (FCA)
University of Campinas (UNICAMP)
Limeira, Brazil
leonardo.duarte@fca.unicamp.br

May 12, 2023

ABSTRACT

The use of machine learning models in decision support systems with high societal impact raised concerns about unfair (disparate) results for different groups of people. When evaluating such unfair decisions, one generally relies on predefined groups that are determined by a set of features that are considered sensitive. However, such an approach is subjective and does not guarantee that these features are the only ones to be considered as sensitive nor that they entail unfair (disparate) outcomes.

In this paper, we propose a preprocessing step to address the task of automatically recognizing sensitive features that does not require a trained model to verify unfair results. Our proposal is based on the Hilbert–Schmidt independence criterion, which measures the statistical dependence of variable distributions. We hypothesize that if the dependence between the label vector and a candidate is high for a sensitive feature, then the information provided by this feature will entail disparate performance measures between groups. Our empirical results attest our hypothesis and show that several features considered as sensitive in the literature do not necessarily entail disparate (unfair) results.

1 Introduction

The use of Machine Learning (ML) methods to deal with real-world problems has grown exponentially in the last decades [7, 34]. In particular, ML is currently integrated in every decision support (DS) system with critical societal impacts. This fact has raised increasing concerns about the (un)fairness of such 'artificial deciders' in several contexts [27], such as in recidivism risk prediction [3], facial recognition systems [32], platforms for job applications [18], hate speech and abusive language detection [9]. This motivated several efforts towards both the detection and mitigation of unfairness in ML and DSs. The former are mostly centered in the detection of undesirable bias, while the later seek to reduce such biases. These methods fall into four main categories, namely, pre-processing [8, 33, 30, 29], in-processing [45, 1, 20], post-processing [19, 11, 5], and hybrid-processing[1] [44, 21, 2].

[1]These methods combine different fairness interventions such as in-processing and post-processing.

Fairness of decision systems is commonly assessed via metrics that rely on the outcomes[2] of decision models, and inspired by anti-discrimination laws *under which decision policies or practices can be declared as discriminatory based on their effects on people belonging to certain sensitive demographic groups* [16], *e.g.*, gender, race, age and sexual orientation. Still, fairness is a social construct [22] and an ethical concept [38], and its definition remains prone to subjectivity.

To overcome subjectivity, several fairness notions have been recently proposed to objectively measure and to capture different aspects of fairness. These include group based notions (*e.g.*, [10, 19]), individual-based notions (*e.g.*, [10, 6]), and causal and counterfactual based notions (*e.g.*, [28, 24]). Even though these metrics would replace human subjectivity by objective metrics, they are still dependent on subjective decisions.

Indeed, most of the above metrics rely on the choice of particular features (variables) with respect to which fairness will be measured. For instance, in the case of group based metrics, fairness is mostly measured with respect to disparate performance and decision results among different subpopulations. For example, when predicting recidivism risk [3], disparate results among the white and black subpopulations, or among men and women. In most of the fairness analysis in ML and DS problems, it is assumed that the population is divided into privileged and unprivileged groups on which disparate results are expected. In other words, the *sensitive features* that discriminate these groups are known *a priori*. Examples of commonly considered sensitive features include race, gender, age, sexual orientation, etc. See, *e.g.*, [25] for a set of ML problems and the sensitive features frequently considered in the literature.

However, the choice of sensitive features also remains subjective and, as they are chosen *a priori*, it is dataset agnostic and does not necessarily entail outcome disparities in a given use case scenario. Indeed, consider an ML problem in which both gender and race are among the set of features. Depending on problem addressed, one may have uneven results with respect to race but similar results regardless the gender. In this situation, one may only be aware of race as a sensitive feature when training the ML model. As further illustration, consider one of the above mentioned platforms for job applications. Ethically, physical disabilities of candidates should not constitute a discriminating (*i.e.*, it should be considered sensitive) of candidates. However, if the job is indeed of physical nature, then some physical abilities may be required and, therefore, this information should not be considered as sensitive.

In this paper, we address this critical issue of identifying sensitive feature or subfeatures (*i.e.*, corresponding protected groups or subgroups), and we propose to defining them as *features that entail disparate (unfair) outcomes*. However, in order to verify disparities among different (sub)groups, one usually needs to train the ML model and calculate the pertaining fairness metrics. This step may be computationally costly depending on the ML model is adopted. Thus this raises the question: how to detect possible disparities before the training step?

To tackle this issue, we propose a statistical approach based on the Hilbert–Schmidt independence criterion (HSIC) [14], which is commonly used in data science [41] to measure the dependence between two matrices (or vectors). The HSIC has been used as a feature selection method [36, 37], for instance, in classification or regression tasks. The HSIC is usually used under the assumption that "good features" are those that maximize the dependence degree between these features and the vector of labels or values, and one selects the subset of features with the highest dependence degrees, as they are more relevant to the output. In [4], the authors used the empirical estimate of HSIC in order to derive a supervised version of Principal Component Analysis (PCA). The idea was to perform dimensionality reduction while maximizing the dependence between the projected data and the vector of outcomes. HSIC has also been used when defining the loss-function for regression of classification task [40, 13]. In this context, a noteworthy application addresses fairness in automatic decisions by learning models that reduce disparities between sensitive groups [31, 26].

Our hypothesis in this work is that, for a given feature, if

1. this feature brings information that splits people into different groups, and
2. the HSIC[3] between this feature and vector of labels is high,

then *there will be disparate outcomes among the (sub)groups discriminated by this feature*. The intuition is that the features with high HSIC are important to the classification task as they carry key information to discriminate (sub)classes. Hence, it is likely to observe disparate outcomes among the discriminated (sub)groups.

Note that not all features that split the population into two groups can be considered as sensitive. For example, the charge degree in recidivism risk prediction [3] may split the population between those who committed a felony or a misdemeanor, but it is not assumed as a sensitive one. Moreover, sensitive features may split the population into more

[2]Some works also focus on procedural fairness, *i.e.*, of the decision making processes (means) that lead to the outcomes [15]. However, this concern will remain outside the scope of this paper.

[3]In fact, as will be further discussed in this paper, we adopted a normalized version of HSIC.

than two groups. For instance, as race, one may have whites, blacks and Asians. For each subfeature (*e.g.*, race_whites, race_blacks and race_Asians), a high HSIC will indicate (i) if the feature is sensitive and (ii) which group division may lead to unfair (disparate) results.

As we will see, HSIC indicates both the sensitive features and the associate groups of individuals that may lead to unfair outcomes. More precisely, high HSIC values between features (or subfeatures, in the case of categorical data) and the label vector may entail disparate results. We attest our hypothesis with empirical evaluations on several datasets frequently used in the literature, and that relate high HSIC values to high group fairness measures. This HSIC based preprocessing approach thus constitutes a noteworthy tool researchers and practitioners to automatically detect those features that should be considered sensitive in an ML or DSs tasks.

It is important to stress that our approach deviates from previous attempts such as [16, 15] in two major aspects, namely, by seeking an automatic approach to choosing which features should be considered sensitive, and by claiming that the choice of sensitive features should take into account both the task and use case scenario. In particular, we show that certain features often considered as sensitive from ethical and social standpoints do not result in outcome disparaties, and thus their use in ML and DSs should not be prohibited. In fact, our approach also enables proxy detection as we will discuss in Section 5.

The rest of this paper is organized as follows. In Section 2, we outline the notations used in this paper. Section 3 discusses the Hilbert–Schmidt independence criterion and its normalized version. The proposed approach to automatically detect sensitive features and sensitive groups is presented in Section 4. In Section 5, we conduct the numerical experiments and discuss the obtained results. Finally, in Section 6, we present our conclusions and future perspectives.

Main contributions. Here, we simply highlight the main contributions of the paper.

- We address the problem of detecting sensitive features in group fairness settings, and propose a statistical approach based on the HSIC that does not require a trained model to verify disparate outcomes.
- We present preliminary empirical results on four well known datasets that support our hypothesis. The higher is the HSIC-based dependence measure between a feature and the labels, the higher are the disparate results entailed by using the information provided by such a feature.
- The analysis of our results also shows that several features taken as sensitive in the literature do not necessarily entail disparate (unfair) results.
- We also raise some perspectives for future work. Our proposed approach can be easily extended to decide whether a combination features (*e.g.*, the use of features describing gender and race) can lead to disparate results. Moreover, we will investigate how to adapt our framework to multiclass classification problems and to individual fairness settings.

2 Notation

Throughout this paper, we assume an underlying binary classification problem such that each m-dimensional sample is assigned to a class $y \in \{-1, 1\}$. The set of n samples and the associated vector of classes are represented by $\mathbf{X}$ and $\mathbf{y}$, respectively. Each column of $\mathbf{X}$, defined by $\mathbf{X}^{(j)}$, describes a feature G_j, $j = 1, \ldots, m$.

Very often some features in ML problems are categorical. However, as most of the learning algorithms require numbers as input features, we consider in this paper the one-hot encoding method to transform the categorical features into binary ones. Mathematically, a vector $\mathbf{X}^{(j)}$ describing a categorical feature G_j with feature space $\{G_{j,1}, G_{j,2}, \ldots, G_{j,q}\}$ can be replaced by[4] q binary vectors $\mathbf{X}^{(j,k)}$, $k = 1, \ldots, q$, that indicate which category each sample belongs to. We illustrate this scenario with the 3-dimensional dataset

$$\mathbf{X} = \begin{bmatrix} male & 2 & \text{30-60} \\ female & 0 & < 30 \\ female & 1 & > 60 \end{bmatrix}, \tag{1}$$

whose features are

G_1: $gender$ with possible values $male$ or $female$,

G_2: number $ncrimes$ of committed crimes in the last 5 years, and

[4]One generally replaces by $q - 1$ binary features in order to avoid a redundant column. However, in our analysis, the use of either q or $q - 1$ binary features lead to the same results.

G_3: age that can be less than 30, between 30 and 60 or greater than 60 years old.

After encoding the categorical features, one achieves the extended 5-dimensional dataset

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

where $G_{1,1} = gender_{male}$, $G_{1,2} = gender_{female}$, $G_{3,1} = age_{\leq 30}$, $G_{3,2} = age_{30-60}$ and $G_{3,3} = age_{\geq 60}$ are the novel subfeatures. Note that, after conducting one-hot encoding for all categorical features, one increases the dataset dimension from m to $\tilde{m}$.

3 Hilbert–Schmidt independence criterion

There are several measures adopted in feature selection tasks to evaluate the relation between inputs and outputs [23]. As examples, one may cite the correlation [17] and mutual information [39] between variables. Methods based on correlation are generally straightforward to be deployed as this measure is easy to be calculated. However, correlation only provides second-order information between variables and, therefore, does not necessarily imply dependence. On the other hand, mutual information brings the knowledge about the dependence degree between variables. The price to be paid here is that, in order to calculate the mutual information, one needs to estimate the probability density function of such variables. This task is very impractical in most cases.

Another measure that has been used in recent works in the literature is the Hilbert–Schmidt independence criterion [41]. The HSIC measures the dependence degree between finite number of observations[5] (x_i, y_i), $i = 1, \ldots, n$. Assume $\mathbf{K_x}$ and $\mathbf{K_y}$ as the kernel matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively. In our analysis, we consider the radial basis function (RBF) kernel, defined by

$$K^{RBF}(z_i, z_{i'}) = e^{-\frac{1}{n}(z_i - z_{i'})^2},$$

and the linear kernel, defined by

$$K^{linear}(z_i, z_{i'}) = z_i z_{i'}.$$

However, other kernels can also be considered (see [35] for details). An empirical calculation of HSIC is given by

$$HSIC(\mathbf{x}, \mathbf{y}) = \frac{\text{tr}\left(\mathbf{K_x H K_y H}\right)}{(n-1)^2},$$

where $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{e}\mathbf{e}^T$ is the centering matrix and $\mathbf{e}$ is a n-dimensional column vector of ones. Note that, even if the dataset $\mathbf{x}$ is centered, this is not necessarily true for the kernel $\mathbf{K_x}$. In order to centralize this kernel, one applies the centering matrix $\mathbf{H}$ on both sides of $\mathbf{K_x}$. Recall that

$$\text{tr}\left(\mathbf{K_x H K_y H}\right) = \text{tr}\left(\mathbf{H K_x H K_y}\right) = \text{tr}\left(\tilde{\mathbf{K}}_\mathbf{x} \mathbf{K_y}\right),$$

where $\tilde{\mathbf{K}}_\mathbf{x}$ is the centered kernel. Moreover, when we multiply $\mathbf{K_x}$ on the left (resp. right) by $\mathbf{H}$, one removes its columns (resp. rows) mean.

As highlighted in [26], the HSIC calculation is sensitive to the scale of the observations and, therefore, an appropriate normalization should be conducted in order to compare relative dependence degrees. We thus consider a normalized version of HSIC called NOCCO (NOrmalized Cross-Covariance Operator) [12]. It is defined as follows:

$$NOCCO(\mathbf{x}, \mathbf{y}) = \text{tr}\left(\mathbf{R_x R_y}\right),$$

where

- $\mathbf{R_x} = \mathbf{H K_x H}\left(\mathbf{H K_x H} + n\epsilon \mathbf{I}_n\right)^{-1}$,
- $\mathbf{R_y} = \mathbf{H K_y H}\left(\mathbf{H K_y H} + n\epsilon \mathbf{I}_n\right)^{-1}$,
- ϵ is a regularization parameter (*e.g.*, 10^{-6}), and
- $\mathbf{I}_n$ is a $n \times n$ identity matrix.

It is important to highlight that the use of NOCCO has two main advantages. Firstly, as the calculation is based on the trace of matrix products, it is easily calculated (we do not need to estimate, for instance, probability density functions). Secondly, the use of kernels brings more information than second-order moment to measure the relation between variables. Therefore, it is useful to estimate the dependence degree between them.

[5] We here consider $\mathbf{x} = \{x_i\}_{i=1}^n$ and $\mathbf{y} = \{y_i\}_{i=1}^n$ as vectors. However, HSIC can be also used to calculate the dependence degree between matrices.

4 Proposed approach

The purpose of this work is to automatically detect sensitive features and the associated impacted groups. Our hypothesis is that if a feature that provides information that splits groups of people (*e.g.*, if the individual is white or black) is important to increase the performance of a machine learning model (measure by means of NOCCO), then this feature may create disparities between such groups in terms of performance measures.

Consider the extended dataset $\tilde{\mathbf{X}}$, with columns $\tilde{\mathbf{X}}^{(1)}, \ldots, \tilde{\mathbf{X}}^{(m)}$, obtained after encoding categorical features. As mentioned in Section 3, HSIC can be used to measure the dependence between two vectors. For each numerical feature and encoded features (*i.e.*, for all columns of $\tilde{\mathbf{X}}$), we calculate NOCCO with respect to the vector of outcomes $\mathbf{y}$. For numerical features, the dependence degree with respect to the vector of labels is the NOCCO itself. In the case of categorical features, if this feature is a binary one, then the NOCCO for both groups will be the same and, therefore, we only need to calculate itonce. Otherwise, if the categorical features have three or more categories, we assume as the dependence degree the maximum NOCCO among the associated encoded features. This procedure leads to m measures of dependence

$$d_j = \text{tr}\left(\mathbf{R}_{G_j}\mathbf{R}_{\mathbf{y}}\right), \quad j = 1, \ldots, m, \tag{3}$$

where $\mathbf{R}_{G_j} = \mathbf{H}\mathbf{K}_{G_j}\mathbf{H}\left(\mathbf{H}\mathbf{K}_{G_j}\mathbf{H} + n\epsilon\mathbf{I}_n\right)^{-1}$ and $\mathbf{K}_{G_j}$ is the kernel matrix associated with feature G_j. In the example of Section 2, if NOCCO for age_¡30 is greater than the NOCCO for both age_30-60 and age_¿60, we define d_3 as the dependence measure between $\tilde{\mathbf{X}}^{(4)}$ and $\mathbf{y}$.

Based on all d_j, $j = 1, \ldots, m$, we may evaluate whether a feature should be considered as a sensitive one. In this paper, we assume an automatic procedure that takes into account a predefined threshold t. Algorithm 1 presents a pseudo-code of our proposal. For each feature G_j, if $d_j \geq t$, G_j is considered sensitive and the $G_{j,k'}$ that leads to the higher NOCCO is the harmed group. Otherwise, we do not consider G_j as sensitive. In our empirical setting, we take t to be the median value among all $d_1, \ldots, d_m$. Note that this can be easily adapted to integrate other values of t.

Algorithm 1 (*Automatic approach for sensitive features detection*)

Input: $\mathbf{X}$ and $\mathbf{y}$.
Output: Set of sensitive features $\mathcal{S} = \{G_{j'}, G_{j''}, \ldots\}$ and sensitive groups $\mathcal{S}^g = \{G_{j',k'}, G_{j'',k''}, \ldots\}$.
1: **Encoding categorical features:** Encode the categorical features and obtain the extended dataset $\tilde{\mathbf{X}} \leftarrow encode\left(\mathbf{X}\right)$.
2: **Calculate the kernel matrix of** $\mathbf{y}$: $\mathbf{K}_{\mathbf{y}} = kernel\left(\mathbf{y}\right)$
3: **Calculate the NOCCO for all features and subfeatures**:
for $j \in \{1, \ldots, m\}$ **do**
  if G_j is either a numerical or binary feature **then**
    Calculate the kernel matrix associated with G_j: $\mathbf{K}_{G_j} = kernel\left(G_j\right)$
    Calculate the dependence measure: $d_j = \text{tr}\left(\mathbf{R}_{G_j}\mathbf{R}_{\mathbf{y}}\right)$
  else
    Calculate the maximum NOCCO for the categorical variable:
    for $k \in \{1, \ldots, q\}$ **do**
      Calculate the kernel matrix associated with $G_{j,k}$: $\mathbf{K}_{G_{j,k}} = kernel\left(G_{j,k}\right)$
      Calculate the dependence measure: $d_j^k = \text{tr}\left(\mathbf{R}_{G_{j,k}}\mathbf{R}_{\mathbf{y}}\right)$
    end for
    Define the dependence measure for feature G_j: $d_j \leftarrow \max\left(d_j^1, \ldots, d_j^q\right)$
  end if
end for
4: **Create the set of sensitive features and sensitive groups**: $\mathcal{S} = \emptyset$ and $\mathcal{S}^g = \emptyset$
5: **Calculate the median value among all** d_j**,** $j = 1, \ldots, m$: $t = median\left(d_1, \ldots, d_m\right)$
for $j \in \{1, \ldots, m\}$ **do**
  if G_j is candidate for sensitive feature and $d_j \geq t$: **then**
    Update the set of sensitive features: $\mathcal{S} \leftarrow \{\mathcal{S}, G_j\}$
    Update the set of sensitive groups: $\mathcal{S}_g \leftarrow \{\mathcal{S}_g, G_{j,k^*}\}$ such that $d_j^{k^*} = \max\left(d_j^1, \ldots, d_j^q\right)$
  end if
end for

5 Empirical Evaluation

In this section, we present the experimental setup to evaluate our proposal on several real datasets. We identify possible sensitive features as well as the associated sensitive groups. Moreover, we provide a comparison between the obtained NOCCO values and the fairness measures with respect to those features with high NOCCO values.

5.1 Datasets

In order to assess our proposal, we consider four datasets frequently used in the literature: Adult income[6], COMPAS recidivism risk [3], Law School Admission Council (LSAC) [42] and Taiwanese credit default [43]. A brief description and how we preprocessed some features are provided in the sequel. See [25] for further details as well as the list of features frequently considered as sensitive ones in the literature.

- **Adult income dataset**: In this dataset, the aim is to predict whether a person makes over 50K a year based on the following features: age, workclass, educational-num (educational degree), marital-status, occupation, relationship (husband, not in family, other relative, own child, unmarried or wife), race (Indian-Eskimo, Asian-Pacific Islander, black, white or other), gender (male or female), capital-gain, capital-loss, hours-per-week and native-country. After removing samples with missing values, we achieved 45222 samples. We also rearranged some categorical features: age = $\{< 25$, 25-60, $> 60\}$, workclass = {private, non-private}, marital-status = {married, never-married, other}, occupation = {office, heavy-work, service, other} and native-country = {US, non-US}. One typically assumes age, race and gender as sensitive features.
- **COMPAS recidivism risk dataset**: This dataset has the goal of classifying individuals as a potential criminal recidivist. There are 6167 samples and 8 features, namely sex (male or female), age_cat (age category - less than 25, between 25 and 45 or greater than 45), race, juv_fel_count (number of juvenile felony), juv_misd_count (number of juvenile misdemeanor), juv_other_count (number of others infractions), priors_count (number of priors) and c_charge_degree (charge degree - felony or misdemeanor). We here rearranged race {Caucasian, African-American, Other}. Race and sex are assumed as sensitive features.
- **Law School Admission Council (LSAC) dataset**: In this dataset, there are 23726 candidates described by 11 features: decile1b (decile given the grades in the 1st year), decile3 (decile given the grades in the 3rd year), lsat (score), ugpa (undergraduate GPA), zfygpa (1st year law school GPA), zgpa (cumulative law school GPA), fulltime (full-time or part-time work), fam_inc (family income), male (whether the student is male or female), race and tier (tier of the law school). The goal is to predict whether a candidate would pass the bar exam. One assumes gender and race as sensitive features.
- **Taiwanese credit default dataset**: In this dataset, the aim is to predict customers' default payments in a Taiwanese institution. The 23 features are the following: limit_bal (amount of given credit), sex (male or female), education (graduate school, university, high school or others), marriage (married, single or others), age (less than 35 or at least 35), repayment status (six features, from April to September), amount of bill statement (six features, from April to September) and amount paid (six features, from April to September). We assumed sex, education and marriage as sensitive features.

5.2 Fairness metrics

In order to attest our hypothesis that the NOCCO value can be associated with fairness measures (and, hence, used as a preprocessing approach to detect sensitive features), we compare the obtained values of fairness measures with respect to features with high NOCCO values. These measures are based on the following performances extracted from the trained ML model:

- True positive (TP): # of instances correctly classified as class 1.
- True negative (TN): # of instances correctly classified as class -1.
- False positive (FP): # of instances wrongly classified as class 1.
- False positive (FN): # of instances wrongly classified as class -1.

Aiming at evaluating some fairness measures when splitting the dataset into different groups of people, we restrict the aforementioned performance measures to such groups. Suppose, for example, that feature G_j contains information that splits people into groups $\{G_{j,1}, \ldots, G_{j,q}\}$. As the NOCCO for group $G_{j,k}$ will indicate the dependence degree

[6] https://archive.ics.uci.edu/ml/datasets/adult

between such a group and the vector of labels, when evaluating fairness, we consider the disparities when people is divided between those belonging to group $G_{j,k}$ and those that does not. Therefore, we restrict the performance measures such that $FP_{G_{j,k}}$ means the number of instances belonging to group $G_{j,k}$ wrongly classified as class 1 and $TN_{-G_{j,k}}$ is the number of instances not belonging to group $G_{j,k}$ correctly classified as class -1 (we referred to $-G_{j,k}$ as the instances not belonging to group $G_{j,k}$).

Given the performance measures conditioned on groups, we may define the fairness measures considered in our experiments:

- Predictive equality (PE): A classifier satisfies predictive equality if both groups have equal false positive rate. In other words, we should have low vales of

$$f_{PE} = \left| \frac{FP_{G_{j,k}}}{FP_{G_{j,k}} + TN_{G_{j,k}}} - \frac{FP_{-G_{j,k}}}{FP_{-G_{j,k}} + TN_{-G_{j,k}}} \right|.$$

- Equal opportunity (EP): A classifier satisfies equal opportunity if both groups have equal true positive rate. In other words, we should have low vales of

$$f_{EP} = \left| \frac{TP_{G_{j,k}}}{TP_{G_{j,k}} + FN_{G_{j,k}}} - \frac{TP_{-G_{j,k}}}{TP_{-G_{j,k}} + FN_{-G_{j,k}}} \right|.$$

- Equalized odds (EO): A classifier satisfies equalized odds if both groups have equal true positive and false positive rates. In other words, we should have low vales of

$$f_{EO} = \left| \frac{TP_{G_{j,k}}}{TP_{G_{j,k}} + FN_{G_{j,k}}} - \frac{TP_{-G_{j,k}}}{TP_{-G_{j,k}} + FN_{-G_{j,k}}} \right| + \left| \frac{FP_{G_{j,k}}}{FP_{G_{j,k}} + TN_{G_{j,k}}} - \frac{FP_{-G_{j,k}}}{FP_{-G_{j,k}} + TN_{-G_{j,k}}} \right|.$$

- Overall accuracy equality (OAE): A classifier satisfies overall accuracy equality if both groups have equal correct classification for both classes -1 and 1. In other words, we should have low vales of

$$f_{OAE} = \left| \frac{TP_{G_{j,k}} + TN_{G_{j,k}}}{n_{G_{j,k}}} - \frac{TP_{-G_{j,k}} + TN_{-G_{j,k}}}{n_{-G_{j,k}}} \right|,$$

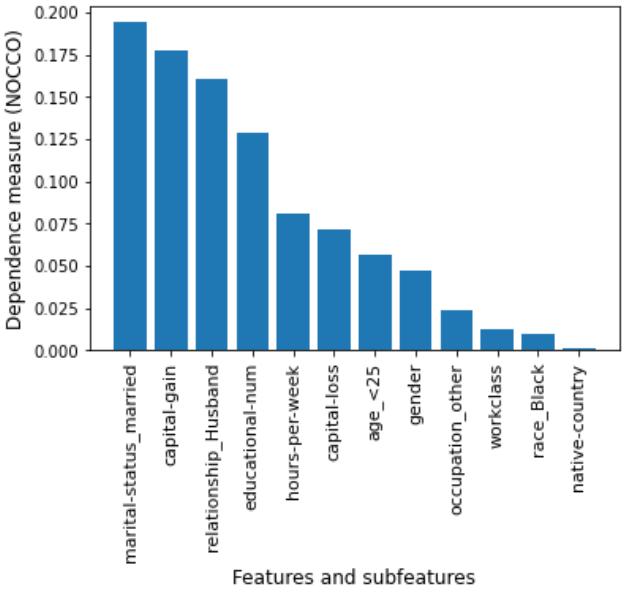
  where $n_{G_{j,k}}$ and $n_{-G_{j,k}}$ are the number of instances that belong and do not belong to group $G_{j,k}$.

5.3 Result analysis

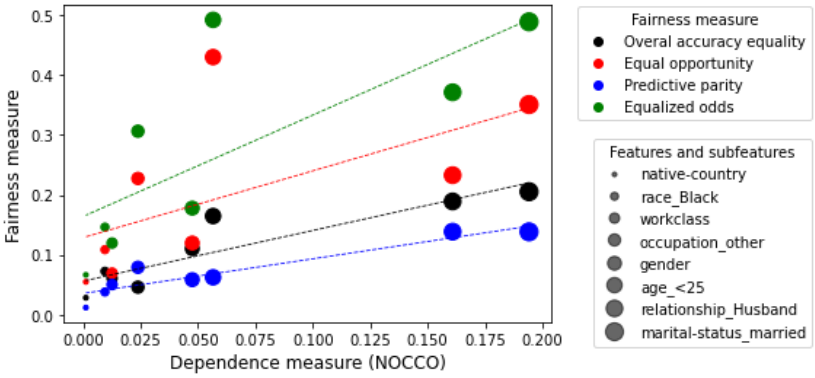
In our experiments, we first calculated the NOCCO for each feature and encoded feature. Thereafter, we calculate the median NOCCO and define the features that could be considered as sensitive ones. Aiming at attesting the usefulness of NOCCO as a measure to detect sensitive features, we comparing the NOCCO values with a set of fairness measures. For this purpose, we trained a Random Forest classifier using k-fold cross validation (with 10 folds). For each fold and encoded features, we calculated performances and fairness measures (equal opportunity, predictive parity, equalized odds and overall accuracy equality, by taking the absolute difference). For categorical features with more than two groups, we calculated these measures by assuming a one-vs-all strategy.

We present the obtained results for the Adult income, COMPAS, LSAC and Taiwanese credit default datasets in Figures 1, 2, 3 and 4, respectively. In NOCCO calculation, we assumed the RBF kernel. Note that, for all datasets and all fairness measures (see Figures 1b, 2b, 3b and 4b), the higher is the NOCCO, the higher is the fairness measures (*i.e.*, the higher are the unfair results).

Results on the Adult income dataset In the results with the Adult income dataset, we obtained a median NOCCO equals to 0.0642. Features whose NOCCO is greater than this threshold are marital-status, capital-gain, relationship, educational-num, hours-per-week and capital-loss (see Figure 1a). We consider that splitting people based on either education-num or hours-per-week do not create unfair concerns. However, the other two features can be seen as sensitive ones. Although age and gender are frequently pointed as sensitive features in this dataset, marital-status and relationship appeared as higher candidates. The higher NOCCO was obtained when splitting people between married and non-married. Moreover, NOCCO is high when we divide the instances in those who are a husband in a relation and those who are not. As in this dataset, husband in a relation is a proxy to know that the person is a man, this information is directly associated with gender. Therefore, not only gender should be consider as a sensitive feature when looking at men/women, but we should also be aware of the information provided by relationship. Another interesting aspect in this dataset is that race led to a very small NOCCO. This suggested that, although frequently considered as a sensitive feature, in this problem, it will not lead to disparate results. We attested this conclusion in the fairness measures presented in Figure 1b. The fairness measures are very small when splitting people by race.



(a) Dependence measure (NOCCO).



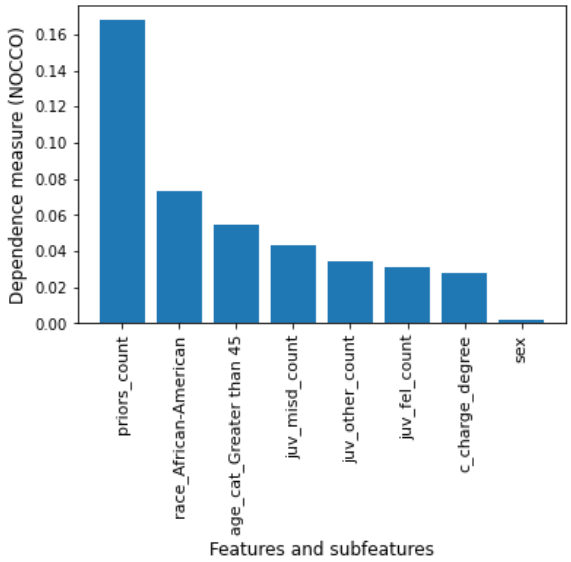
(b) Relation between NOCCO and fairness measures.

Figure 1: Results for the Adult income dataset (with RBF kernel).

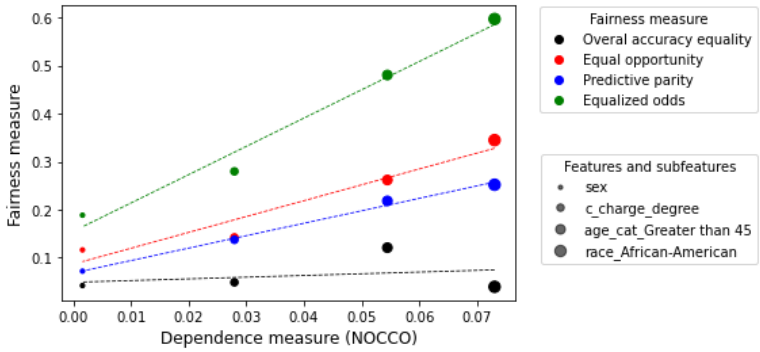
Results on the COMPAS dataset In COMPAS dataset, the literature assumes gender and race as sensitive features. In the NOCCO values presented in Figure 2a and with $t = 0.0390$, only race (when dividing people between African-American and other ones) and age category (greater than 45) could be considered as sensitive features. Although sex is frequently considered as sensitive, the obtained NOCCO value indicates that, in this dataset, it does not entail disparities between male and female. See in Figure 2b that fairness measures when splitting people by sex are much lower in comparison with race or age category. Another interesting remark presented in Figure 2b is that the the relation between NOCCO and most of the fairness measures are nearly linear.

Results on the LSAC dataset The NOCCO values for the LSAC dataset are presented in Figure 3a. Only race and sex is considered as candidates for sensitive features in this dataset. With a median value of 0.0635, race should be considered as a sensitive feature and male should not. Indeed, as can be attested in Figure 3b, while splitting people by race creates unfair results (we obtained high values of fairness measures), a division by gender does not (the fairness measures are very small).

Results the Taiwanese credit default dataset Another interesting result was obtained with the Taiwanese credit default dataset. One may see in Figure 4a that features generally considered as sensitive ones (sex, education and marriage) have very small NOCCO values. Features associated with payment status have a much higher dependence with class labels than the considered sensitive features. Therefore, with a median NOCCO value of 0.2789, no feature



(a) Dependence measure (NOCCO).



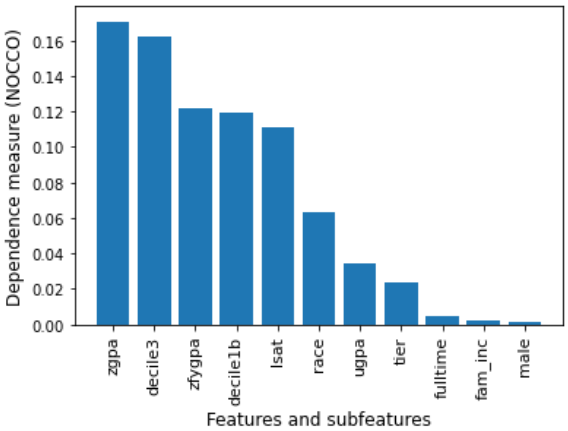
(b) Relation between NOCCO and fairness measures.

Figure 2: Results for the COMPAS dataset (with RBF kernel).

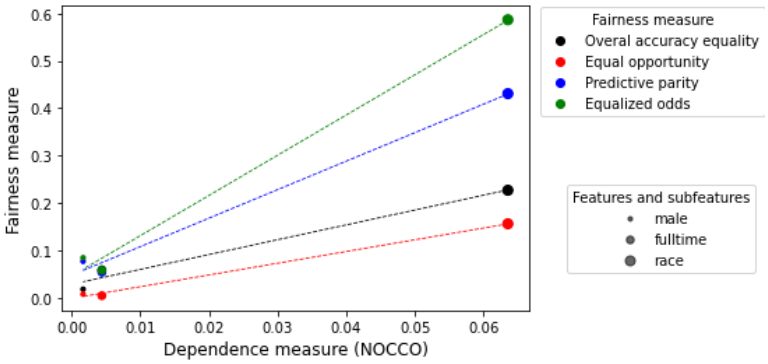
could be defined as sensitive. This result is attested by the small values (with most of them lower than 0.10) of the fairness metrics presented in Figure 4. Indeed, if one compares these fairness metrics with the ones achieved with the previous datasets, one may assume that they are small.

5.4 Consistency with linear kernel

In order to verify the consistency of the previous results by assuming another kernel, we evaluate the relation between linear kernal-based NOCCO and the fairness measures. The results for the Adult income, COMPAS, LSAC and Taiwanese credit default datasets are presented in Figures 5a, 5b, 5c and 5d, respectively. Clearly, all the results with the linear kernel are consistent with the ones obtained by using the RBF kernel. The only difference can be seen in Adult income dataset with respect to the detected sensitive group for race. Instead of race_White (as detected by using the RBF kernel), we here assigned race_Black as the sensitive group. The reason for this disparity lies in the number of samples for each race category. As the number of samples whose categories are different from white or black is less than 5%, when using one-vs-all strategy, the results for race_White and race_Black are practically the same. Therefore, we can safely consider that there was not a disparity between the obtained results.



(a) Dependence measure (NOCCO).



(b) Relation between NOCCO and fairness measures.

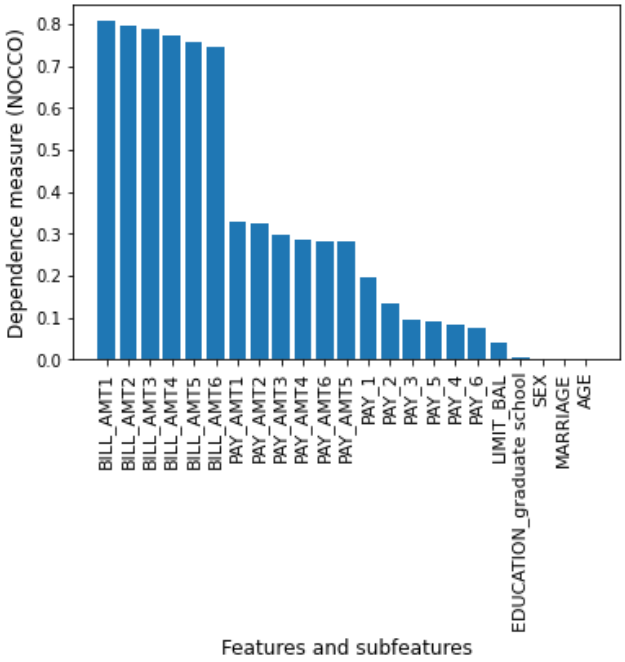
Figure 3: Results for the LSAC dataset (with RBF kernel).

6 Conclusions and perspectives

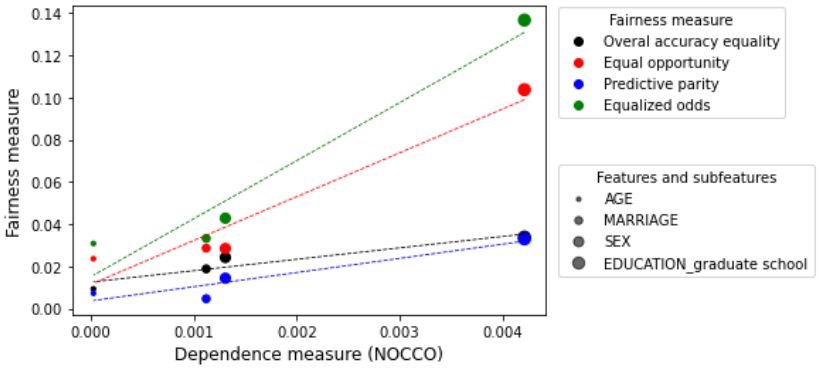
In this paper, we proposed an automatic approach to detect sensitive features based on a normalized version of the Hilbert–Schmidt independence criterion. As our method only requires the dataset and labels, we are able to detect sensitive features without the need of training a machine learning model and evaluating the performance measures. The numerical experiments on several datasets and different kernels attested that the higher is the dependence between a sensitive feature and the vector of labels, the higher are the outcome disparities with respect to groups discriminated by such a feature. Therefore, the statistical dependence calculated from the dataset and labels proved to be a measure that can guide researchers and practitioners in detecting sensitive features, especially in situations and case scenarios where the pertaining fairness notions seem hard to identify.

We highlight that an automatic procedure to detect sensitive features is of importance in practical problems. Generally, one subjectively defines which features are sensitive. However, we showed that features frequently considered as sensitive in the literature may not entail disparate results. This finding can save effort when developing a machine learning model that takes into account information from predetermined sensitive features in order to mitigate disparate results provided by them.

This work opens interesting perspectives for future works. Although we used the NOCCO to calculate the dependence degree between labels and vectors of features that describe groups in the population (*e.g.*, if the person is black or not), this measure can also be applied to coalitions of features (*e.g.*, subgroups). In this case, we aim at verifying the dependence degree between tow or more sensitive information (*e.g.*, if the person is a black woman with less than 30 years old) and the labels. This may show that combining information provided by two or more features may increase the chance of achieving disparate results. As another perspective, we intend to extend the proposed approach to deal with multiclass classification problems. In such scenarios, for each available class, we may detect which features



(a) Dependence measure (NOCCO).



(b) Relation between NOCCO and fairness measures.

Figure 4: Results for the Taiwanese credit default dataset (with RBF kernel).

should be considered as sensitive.

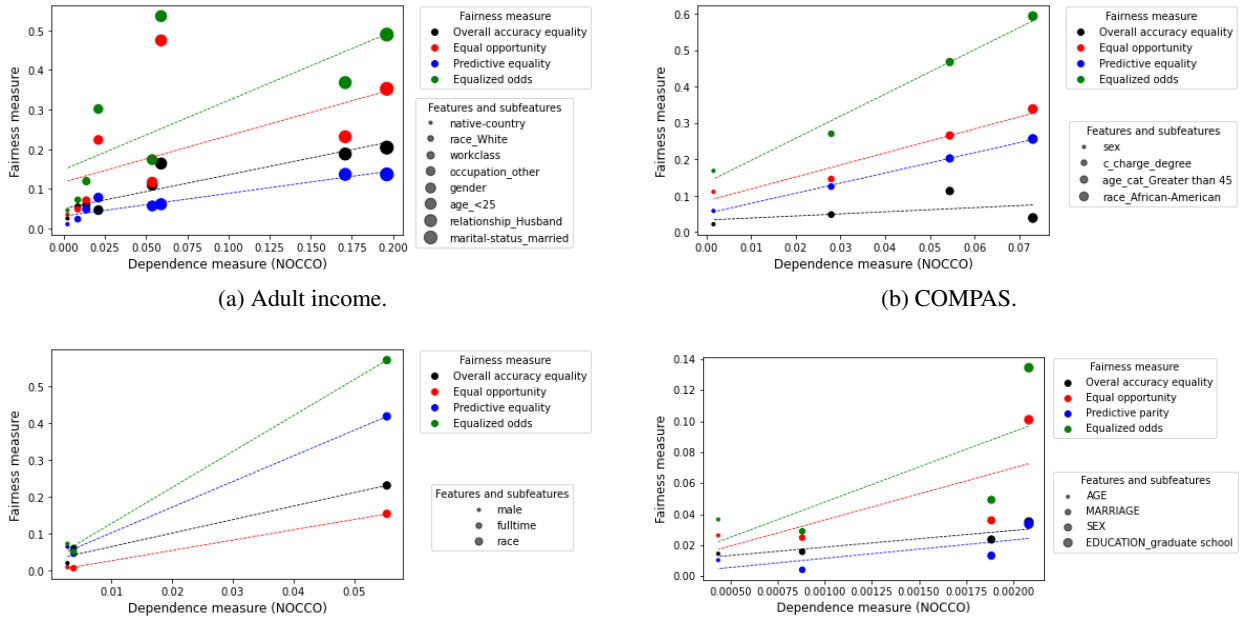
Acknowledgements

Work supported by São Paulo Research Foundation (FAPESP) under the grants #2020/09838-0 (BI0S - Brazilian Institute of Data Science), #2020/10572-5 and #2021/11086-0. The research of the second named author was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215, and the Inria Project Lab "Hybrid Approaches for Interpretable AI" (HyAIAI).

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] Guilherme Alves, Maxime Amblard, Fabien Bernier, Miguel Couceiro, and Amedeo Napoli. Reducing unintended bias of ML models on tabular and textual data. In *8th IEEE International Conference on Data Science and Advanced*



(a) Adult income.

(b) COMPAS.

(c) LSAC.

(d) Taiwanese credit default.

Figure 5: Relation between NOCCO (with linear kernel) and fairness measures for all datasets. We compare 5a (Adult income), 5b (COMPAS), 5c (LSAC), 5d (Taiwanese credit default) with Figures 1b, 2b, 3b and 4b, respectively.

Analytics, DSAA 2021, pages 1–10. IEEE, 2021.

[3] Julia. Angwin, Jeff. Larson, Surya. Mattu, and Lauren. Kirchner. Machine Bias - ProPublica, 2016.

[4] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44:1357–1371, 2011.

[5] Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. Limeout: An ensemble approach to improve process fairness. In *ECML PKDD 2020 Workshops - Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): XKDD 2020 Proceedings*, volume 1323 of *Communications in Computer and Information Science*, pages 475–491. Springer, 2020.

[6] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, pages 405–414. ACM, 2018.

[7] Henrik Brink, Joseph Richards, and Mark Fetherolf. *Real-world machine learning*. Simon and Schuster, 2016.

[8] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[9] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, 2019.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[11] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 144–152. SIAM, 2016.

[12] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20 (NIPS)*, 2007.

[13] Daniel Greenfeld and Uri Shalit. Robust learning with the hilbert-schmidt independence criterion. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3759–3768. PMLR, 2020.

[14] Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory, 16th International Conference, ALT 2005*, volume 3734 of *Lecture Notes in Computer Science*, pages 63–77. Springer, 2005.

[15] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 51–60. AAAI Press, 2018.

[16] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. 2016.

[17] Mark A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[18] Dominik Hangartner, Daniel Kopp, and Michael Siegenthaler. Monitoring hiring discrimination through online recruitment platforms. *Nature*, 589(7843):572–576, 2021.

[19] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[20] Vasileios Iosifidis and Eirini Ntoutsi. Adafair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 781–790, 2019.

[21] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1375–1380. IEEE, 2019.

[22] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.

[23] Sotiris B. Kotsiantis. Feature selection for machine learning classification problems: A recent overview. *Artificial Intelligence Review*, 42(1):157–176, 2011.

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[25] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):1–59, 2022.

[26] Zhu Li, Adrián Pérez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *Pattern Recognition*, 132:108922, 2022.

[27] Ninareh. Mehrabi, Fred. Morstatter, Nripsuta. Saxena, Kristina. Lerman, and Aram. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[28] Judea Pearl. *Causality*. Cambridge university press, 2009.

[29] Guilherme Dean Pelegrina and Leonardo Tomazeli Duarte. A novel approach for Fair Principal Component Analysis based on eigendecomposition. *ArXiv ID: 2208.11362*, 2022.

[30] Guilherme Dean Pelegrina, Renan Del Buono Brotto, Leonardo Tomazeli Duarte, Romis Attux, and João Marcos Travassos Romano. Analysis of trade-offs in fair principal component analysis based on multi-objective optimization. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Padua, Italy, 2022. IEEE.

[31] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017. Lecture Notes in Computer Science*, volume 10534, pages 339–355. Springer, Cham, 2017.

[32] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

[33] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *9th International Conference on Learning Representations, ICLR 2021*.

[34] Iqbal H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):160, 2021.

[35] Bernhard Schölkopf, Alexander J. Smola, and Francis Bach. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[36] Le Song, Alex Smola, Arthur Gretton, Karsten M. Borgwardt, and Justin Bedo. Supervised feature felection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830, 2007.

[37] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13:1393–1434, 2012.

[38] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, 37(1):215–230, 2022.

[39] Jorge R. Vergara and Pablo A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24:175–186, 2014.

[40] Hao Wang, Yijie Ding, Jijun Tang, and Fei Guo. Identification of membrane protein types via multivariate information fusion with hilbert–schmidt independence criterion. *Neurocomputing*, 383:257–269, 2020.

[41] Tinghua Wang, Xiaolu Dai, and Yuze Liu. Learning with hilbert–schmidt independence criterion: A review and new perspectives. *Knowledge-Based Systems*, 234:107567, 2021.

[42] Linda F. Wightman. LSAC national longitudinal bar passage study. Technical report, 1998.

[43] I-Cheng Yeh and Che hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36:2473–2480, 2009.

[44] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[45] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.