The Crucial Role of Sensitive Attributes in Fair Classification

1st Maryam Amir Haeri
Algorithm Accountability Lab
University of Kaiserslautern
Kaiserslautern, Germany
haeri@cs.uni-kl.de

2nd Katharina Anna Zweig
Algorithm Accountability Lab
University of Kaiserslautern
Kaiserslautern, Germany
zweig@cs.uni-kl.de

***Abstract*—In many countries, it is illegal to make certain decisions based on sensitive attributes such as gender or race. This is because historically, sensitive attributes of individuals were exploited to abuse the rights of individuals, leading to unfair decisions. This view is extended to algorithmic decision-making systems (ADMs) where similar to humans, ADMs should not use sensitive attributes for input. We reject the extension of law from humans to machines, since contrary to humans, algorithms are explicit in their decisions, and the fairness of their decision can be studied independently of their input. The main purpose of this paper is to study and discuss the importance of using sensitive attributes in fair classification systems. Specifically, we suggest two statistical tests on the training dataset, to evaluate whether using sensitive attributes may have an impact on the quality and fairness of prospective classification algorithms. These statistical tests compare the distribution and data complexity of the training dataset between groups identified by the same value for sensitive attributes (e.g., men vs. women). We evaluated our fairness tests on several datasets. It was shown that, the removal of sensitive attributes may result in the decrease of the fairness of ADMs. The results were confirmed by designing and implementing simple classifiers on each dataset (with and without the sensitive attributes). Therefore, the use of sensitive attributes must be evaluated per dataset and algorithm, and ignoring them blindly may result in unfair ADMs.**

I. Introduction

Recently, machine learning is widely used in different aspects of human life for making decisions more easily. However, using algorithmic decision-making (ADM) systems for deciding about humans needs more attention. Incorrect decisions may be against human rights or may affect their life.

There are some observations regarding unfair automated decisions made by machine learning methods. For example, Angwin et al. [1] reported on the unfairness of recidivism risk assessment against Afro-Americans. In another research, Barocas and Selbst [2] showed that an ADM based hiring system is biased against women. These observations lead to a strong focus on the fairness of machine learning methods. Thus, fair (discrimination-aware) machine learning is an emerging research topic that attracts researchers [3]–[6].

In many countries, there are anti-discrimination laws which generally prohibit unfair treatment with humans based on certain sensitive features such as race, gender, and nationality. From the regulatory point of view, there are two potential approaches to prevent algorithmic discrimination. In the first approach, the algorithm is not allowed to receive sensitive attributes as input. In the second approach, the algorithm is monitored to prevent discrimination from being propagated to the final. However, it seems that the regulatory prefers, or in some cases mandates, the first approach [7].

Thus, usually the law prohibits the use of sensitive attributes such as race, skin color, and sexual orientation in ADM systems [8]. For this reason, the effect of using sensitive attributes in the learning systems is not studied well. Based on these rules, most of the proposed fairness aware classification methods do not use the sensitive attribute as one of the inputs of the models (do not use in an explicit manner). However, many of these methods use sensitive attributes in an implicit manner. For example, several methods use sensitive attributes in computing a fairness-based term of their objective function of the classification algorithm.

Previously Žliobaitė and Custers [7] showed that for the standard regression analysis, sensitive attributes may be required to train a fairer model. Based on their analysis: they stated that from the regulatory point of view there is an important implication that collecting sensitive attributes is a necessity in order to guarantee fairness. However, we still need more research on analyzing using sensitive attributes in machine learning methods such as classification tasks.

It is important to know that machines are different from humans. When humans decide about humans, we prefer that the decision makers do not aware of sensitive attributes to prevent biased decisions. For example, when somebody decides about job hiring, we perceived the system as fair if the admission committee does not aware of gender and race of the applicants. However, machines learn based on the data. It extracts statistical patterns from the data. Thus, there is no guarantee to have a fair system by ignoring sensitive attributes and losing some pieces of information. Neglecting sensitive attributes even may lead to more discrimination and cause inaccurate results especially for the protected (minority) group.

Hence, in this paper, we want to discuss in which situations using sensitive attributes is needed to have fairer classifiers. At first, we bring an example to show that there are cases that using sensitive attributes results in fairer and more accurate so-

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

lutions. Then we suggest two statistical tests over training data to understand the potential role of the sensitive attributes on the fairness of classifiers trained over that data. For example, one of these statistical analyses is to use the Kolmogorov-Smirnov test to check the similarity of the distribution of training data in different groups. If the distribution of data is varied for different sensitive groups, ignoring sensitive attributes may cause classifier bias towards the distribution of the majority group. The other one is based on the comparison of the data complexity of sensitive groups. If the complexity of the training data is different for various sensitive groups, using sensitive attributes may help to find a fairer model.

Moreover, this paper empirically investigates the use of sensitive attributes in classification for three real-world problems. The results of this study show that the use of sensitive attributes can play a significant role in changing the results of the classification and even lead to an increase in the quality and fairness of the classifier. Therefore, it is crucial not to emphasize the omission of sensitive attributes from the input of ADMs. Rather, concentrate on how to design ADMs which are transparent, accountable, fair and privacy preserving.

The rest of the paper is organized as follows: Section II reviews fairness definitions and recent fairness aware classification approaches. Section III tries to illustrate that how using sensitive attributes in an explicit manner helps to have fairer ML models. Section IV introduces statistical analyses to investigate the role of sensitive attributes. Section V is devoted to the experimental results. Section VI concludes the paper.

II. Related Works

This section explains several of the mostly used fairness definitions and describes fairness aware classification methods.

A. Notations and Formalisation

Prior to review previous works let us define the problem formally. Considering a classification problem, let $A_1, \ldots, A_m$ be the sensitive attributes and $X_1, \ldots, X_n$ be the other (insensitive) attributes. Moreover, let Y be the target (desired) output and $\hat{Y}$ be the classifier output. In this paper, in order to simplify the problem (without loss of generality), we assume that m is equal to one; that means there is only one sensitive attribute A in the dataset. Additionally, assume that A has two possible values a_1 and a_2. These notations are used until the end of the paper. We also define all instances with A=a_1 as group G_1, and all instances with $A = a_2$ as group G_2.

B. Fairness Definition

There are numerous fairness definitions in the literature. Four of the most popular fairness criteria are as follows:

- **Independence:** A classifier satisfies the independence fairness criterion, if its output ($\hat{Y}$) is independent of the sensitive attributes A [9]. Considering the binary classification problem, the condition of independence can be represented by the following equality.

$$P\{\hat{Y} = 1|A = a_1\} = P\{\hat{Y} = 1|A = a_2\} \quad (1)$$

If for a classifier the probability of assigning a data instance to the positive class is the same for all sensitive groups then the system is fair. In other words, based on this definition, if the classifier has equal support values ($\frac{TP+FP}{TP+FP+TN+FN}$) for both groups G_1 and G_2, the system is fair. It is possible to consider a discrimination (fairness) measure based on this definition as Equation 2. Where, M is the classification model, and $\mathsf{Support}(M, G_i)$ is the support of model M over the data instances of group G_i. If the value $\mathsf{Discrimination}_{\text{Independence}}$ is near 0, the system is considered fair based on the independence definition.

$$\mathsf{Discrimination}_{\text{Independence}} = |\mathsf{Support}(M, G_1) - \mathsf{Support}(M, G_2)| \quad (2)$$

- **Equalized Odds or Separation:** A classifier satisfies the equalized odds definition of fairness (the other name is separation), if its output ($\hat{Y}$) is conditionally (with respect to the target output Y) independent of the sensitive attribute A [10]. Considering the binary classification problem, the conditions of equalized odds are shown in Equation 3 and Equation 4.

$$P\{\hat{Y} = 1|Y = 1, A = a_1\} = P\{\hat{Y} = 1|Y = 1, A = a_2\} \quad (3)$$

$$P\{\hat{Y} = 1|Y = 0, A = a_1\} = P\{\hat{Y} = 1|Y = 0, A = a_2\} \quad (4)$$

Equation 3 and Equation 4 mean that the classifier should have similar recall and similar FPR (false positive rate) for both groups. Thus, it is possible to consider a discrimination measure based on this definition as follows.

$$\mathsf{discrimination}_{\text{Equalized\_odds}} = |\mathsf{Recall}(M, G_1) - \mathsf{Recall}(M, G_2)| + |\mathsf{FPR}(M, G_1) - \mathsf{FPR}(M, G_2)| \quad (5)$$

- **Overall Accuracy Equality:** According to this definition of fairness, a learning system is fair if its accuracy values for both groups are equal [11]. This definition can be expressed by Equation 6.

$$P\{\hat{Y} = Y|A = a_1\} = P\{\hat{Y} = Y|A = a_2\} \quad (6)$$

The discrimination measure based on this definition can be considered as Equation 7.

$$\mathsf{discrimination}_{\text{Overall\_Accuracy\_Equality}} = |\mathsf{Accuracy}(M, G_1) - \mathsf{Accuracy}(M, G_2)| \quad (7)$$

- **Sufficiency:** A classification system satisfies the sufficiency condition if the target output Y is conditionally

(with respect to $\hat{Y}$ independent of the sensitive attribute A [12]. This criterion is named *calibration* as well.
Considering the binary classification problem, the condition of sufficiency is equivalent with the following equality.

$$P\{Y = 1|\hat{Y} = \hat{y}, A = a_1\} = P\{Y = 1|\hat{Y} = \hat{y}, A = a_2\} \tag{8}$$

If we assume that $\hat{Y}$ can only be zero and one, the Equation 8 means that the classifier should have similar precision and similar FOR (false omission rate) for both groups. Thus, it is possible to consider a discrimination measure based on this definition as follows.

$$\begin{aligned} &\mathsf{discrimination}_{\mathsf{Sufficiency}} = \\ &|\mathsf{Precision}(M, G_1) - \mathsf{Precision}(M, G_2)| \\ &+ |\mathsf{FOR}(M, G_1) - \mathsf{FOR}(M, G_2)| \end{aligned} \tag{9}$$

C. Fairness-aware Classification Approaches

If we take a look at classification methods, they can be divided into three subgroups in terms of using sensitive attributes as well as considering fairness.

The first group includes the common classification systems not considering fairness criteria and using sensitive attributes in the learning model.

The second group is referring to the classification systems which try to provide fair decisions only by ignoring sensitive attributes. This approach is called *fairness through unawareness*.

The third group, which is the largest group of fairness-aware methods, use sensitive attributes implicitly or indirectly (not as the main input) in different phases such as pre-processing, in-processing and post-processing. In pre-processing approaches such as [2], [13], [14], the aim is to remove bias from the data by pre-processing. In these approaches, sensitive attributes are used to verify that the transformed data is independent of sensitive attributes or without any bias. It is also possible, to enhance fairness by making some changes in the learning algorithms; this approach is named in-processing. For example, in most of these approaches, another fairness objective function is added to the main objective function of the learning problem. In this case, the sensitive attributes are used to evaluate the second (fairness-aware) objective function [15]. Many researchers used fairness measures as a regularizer in learning optimization problems to find high quality and fair models such as [16]–[18]. However, in post-processing approaches such as [19], after the output of the model is generated, the output is changed in a way that results in increasing fairness. In this case, sensitive attributes are used to evaluate whether changing the output of the system leads to boost the fairness measures.

So we can express that most of the fairness-aware classification methods do not use sensitive attributes directly in the models (as an input) but they utilize them in an implicit manner. As research result [16]–[18] show that the implicit use of those features can help to increase fairness, the key question here is whether there are situations in which the direct (explicit: i.e. use as the main input of the model) use of sensitive attributes can help to increase fairness in learning systems?

Another important note is that in the above mentioned systems that already implicitly use sensitive features, the impact of using these attributes on the model is not clear to the user. Thus, it is necessary to study the role of sensitive attributes in classification systems and considering using sensitive attributes in transparent and accountable ways.

III. Discussion and Illustration of The Role of Sensitive Attributes

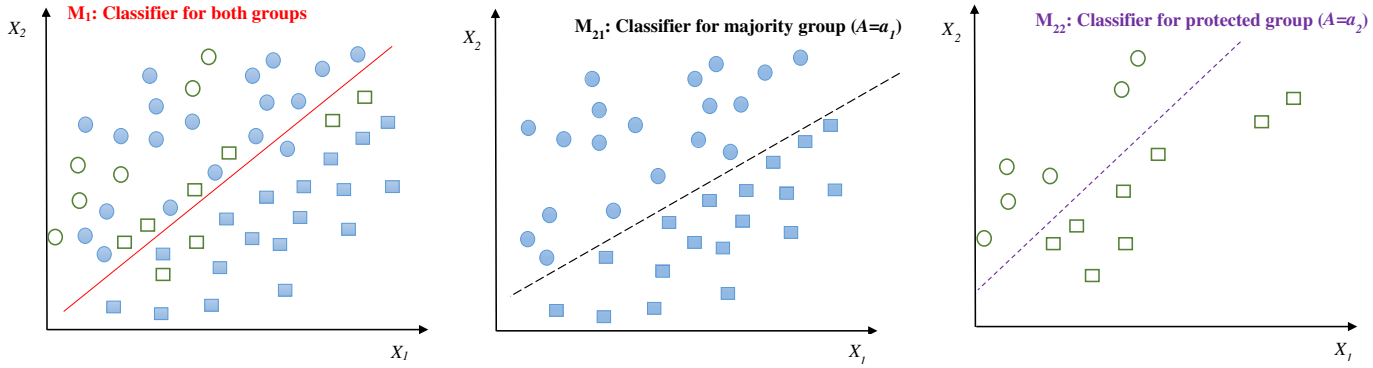
For justification of our perspective, in the following, we bring a simple example showing that using the sensitive attributes in the model may help the fairness. This example is brought in Figure 1. Consider a binary classification problem over a three-dimensional dataset with three attributes X_1, X_2 and A. Let $X = \{X_1, X_2\}$ be the sets of insensitive attributes and A is the sensitive attribute of this problem with two possible values a_1 and a_2. Here, we assume two groups, a majority group with ($A = a_1$) and a minority group with ($A = a_2$). In Figure 1, the red solid line represents an optimal linear classifier which only uses two insensitive features X_1 and X_2. As can be seen, this classifier has an overall good accuracy of 90%, i.e., 90% of all decisions are correct. But this is mainly achieved in the majority group with an accuracy of 97% while the smaller protected group suffers from 29% wrong decisions. This could be deemed unfair. Please note that as mentioned in Section II one of the definitions of the fairness of ADMs is having equal overall accuracy for both groups [20].

However, if we consider a model that involves the sensitive attribute and provide two different models for each group, it can reach higher accuracy and fairness. This method provides two different models, one for each group: if the new data belongs to the majority group ($A = a_1$), it uses model one (M_{21}) and if it belongs to the protected group ($A = a_2$), it uses model two (M_{22}). The accuracy and fairness of both methods are depicted in Figure 1.

This is just a very simple example to show that it is possible to involve the sensitive attributes in the learning models to reach better and fairer classifiers. Thus, it is very important to explore that under which constraints it is better to use sensitive features and how this can be done in an accountable and transparent manner. This paper focuses on understanding the role of the sensitive attributes in a dataset, before the training procedure. These may help to select appropriate approaches to have a fair and accurate classification.

IV. Statistical Test for Understanding the Role of Sensitive Attributes

One of the tools that can help to identify the need to use sensitive attributes for a fair system is statistical analysis. In



Model	Function	Accuracy (over all data instances)	Accuracy for majority group	Accuracy for protected group
M_1	Use M_1	90%	**0.97%**	**0.71%**
M_2	If (A=a_1) use M_{21} else use M_{22}	100%	**100%**	**100 %**

Fig. 1: Difference between two binary classifiers: one using sensitive attributes, and the other not. Here, the problem is a binary classification (rectangle and circle). A is the sensitive attribute, and X_1, X_2 are the insensitive attributes. Filled shapes are from the majority group ($A = a_1$), while empty shapes belong to the minority or protected group ($A = a_2$). Classifier M_1 is a linear classifier which decides based on two attributes X_1 and X_2. However, classifier M_2 first asks for the sensitive attribute and then proceeds with two different models (M_{21} for group a_1 and M_{22} for group a_2). Here, we define fairness as equality of classification accuracy for both groups. Evidently, while the classifier (M_2) uses the sensitive attribute, it is fairer than the classifier M_1 which does not consider the sensitive attribute.

this section, we suggest two statistical tests that will help to understand the behavior and the nature of the data and the underlying problem. Based on these tests, it is possible to have a clue whether using sensitive attributes may help to have fairer systems. Here, for simplicity (without loss of generality), it is assumed that there is only one sensitive attribute at the dataset. The notations used here are the same as mentioned in Section II.

1) Comparing the data distributions of groups: : One of the factors which can indicate the importance of a sensitive attribute in creating a fair model is the difference of data distribution of groups. It means that if the data is grouped on the basis of a sensitive attribute, the data distributions of the sensitive groups (for positive and negative classes) are the same or they are different. Kolmogorov-Smirnov test (two-sample KS- test) can be used to compare the distributions of data in different groups. Kolmogorov–Smirnov test is one of the mostly used non-parametric tests of the equality of one-dimensional probability distributions of two samples. KS-test is sensitive to the differences between the empirical *cumulative distribution functions* (CDF) of the two samples, in terms of both location and shape.

This test is appropriate to test two one-dimensional distributions. However, in most problems, there are more than one insensitive attributes. Thus, in order to compare two multivariate distributions, it is possible to use other versions of KS-test which are appropriate for multivariate distributions such as [21].

For simplifying the test, it is also possible to check the equity of the distributions of each attribute for both groups. In this case, KS-test can be used as follows:

For each attribute X_i

1) Check whether the attribute X_i has the same distribution within positive class for both groups.

$$\text{Use KS-test with } H_0 : \mathbb{P}_{X_i|A=a_1,C=+} = \mathbb{P}_{X_i|A=a_2,C=+}$$
$$\text{and } H_1 : \mathbb{P}_{X_i|A=a_1,C=+} \neq \mathbb{P}_{X_i|A=a_2,C=+}.$$

2) Check whether the attribute X_i has the same distribution within negative class for both groups.

$$\text{Use KS-test with } H_0 : \mathbb{P}_{X_i|A=a_1,C=-} = \mathbb{P}_{X_i|A=a_2,C=-}$$
$$\text{and } H_1 : \mathbb{P}_{X_i|A=a_1,C=-} \neq \mathbb{P}_{X_i|A=a_2,C=-}.$$

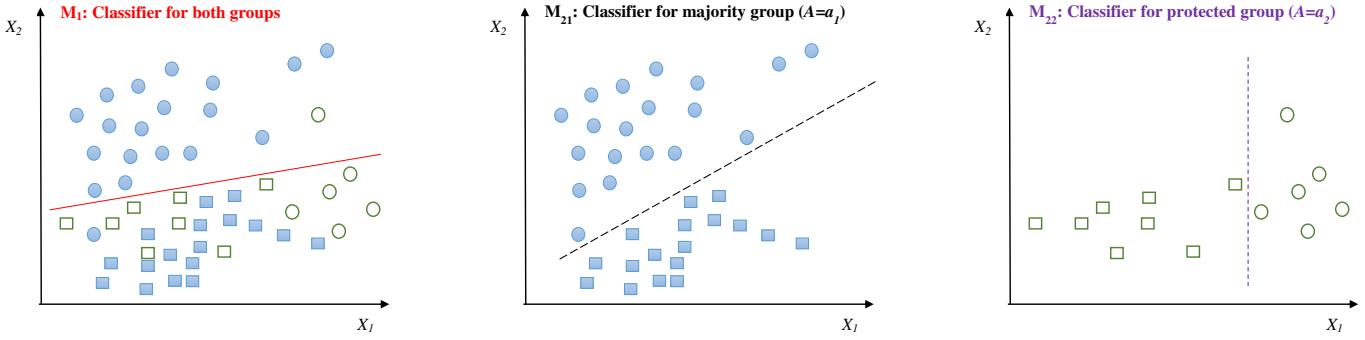
End

** Here, $\mathbb{P}$ denotes the distribution, and C denotes the class label. Thus, $\mathbb{P}_{X_i|A=a_1,C=+}$ means the distribution of X_i in condition to $A = a_1$ and the class label is positive.

Just in case that the distributions of all attributes are the same for both groups, then it is possible to perform the multivariate Kolmogorov-Smirnov test.

Note that for a more complete study it is better to compare the distributions of groups for positive and negative class separately, as it is described in the above pseudo-code.

As an illustrative example, consider a job hiring system. The X_1 attribute is the average university grades and the X_2 is the years of experience, and the sensitive attribute A is the gender. Let the dataset of this problem be as shown in Figure 2. It can be seen that the distribution of the data is different for the groups in both classes. This means that the women who are actually experts (belong to the positive class) are better in the university grades and men are better in years of experience. Thus, if a linear classifier (see M_1 in the Figure 2)



Model	Function	Accuracy (over all data instances)	Accuracy for majority group	Accuracy for protected group
M_1	Use M_1	88%	**0.97%**	**0.64%**
M_2	If (A=a_1) use M_{21} else use M_{22}	100%	**100%**	**100 %**

Fig. 2: Example of the effects of different distributions of sensitive groups. Here, again the problem is a binary classification (rectangle for + and circle for -). A is gender which is a sensitive attribute, and X_1 (average of university grades), X_2 (years of experience) are the insensitive attributes. filled shapes depict data points of men and empty shapes demonstrate group of women. The other settings of this figure are the same as that of figure 1. Women who belong to the positive class are better in the university grades (X_1) and men belonging to positive class are better in years of experience (X_2). As the distributions of the data of the groups are different, the M_2 model which uses sensitive attributes and generates two different classifiers for men and women leads to a fairer and more accurate solution than M_1 which ignores sensitive attributes.

is trained only with these two attributes, as the classifier learns the pattern of the majority group and the number of instances of the men is greater than the women, it cannot find the pattern in the minor group (women). According to this linear classifier M_1, the applicants who has more years of experience (as in the men groups) are accepted for the job. However, for the groups of women who are actually experts in that job the grades are better. In this case, if we use the sensitive attributes, for example using different models for each group, the fairness and performance of the classifier are increased (see model M_2 in in the Figure 2). Thus, it is possible to use KS-test to analyze whether the distributions of sensitive groups are different.

2) Comparing the data complexity of groups: : Another point to note is that there is not necessarily a slight difference in the distribution of the two classes means that using sensitive attributes an inputs, has a big impact on the results of the classifier. Another parameter that can indicate the need to use sensitive attributes is the comparison of data complexity (for classification problem) in different groups.

For illustration, consider the example depicted in Figure 3. The instances belong to the major group can be classified by a linear classifier. However, the minor group cannot be precisely classified by a linear classifier. In other words, the complexity of the minor group data is greater than the complexity of the major group data.

There are several measures for assessing data complexity for classification problems [22], [23]. One of the measures of the geometrical complexity of the classification problem is the *Non-parametric separability of classes* (N2). This measure indicates the ratio of the average distance to the intra-class nearest neighbor and the average distance to the inter-class nearest neighbor. The smaller values of the measurement indicate that the classification problem is easier and the classes can be separated by a smoother discriminant function.

Suppose a classification problem with N samples. For each sample S_i belongs to the training data, it can be found the nearest neighbor $\nu_1^{=(S_i)}$ in the same class called intra-class nearest neighbor, and the nearest neighbor $\nu_1{}^{\neq(S_i)}$ in the opposite class called inter-class nearest neighbor. Then it is possible to compute $N2$ measure by the following formula [23].

$$N2 = \frac{\sum_{i=1}^{N} \delta(\nu_1{}^{=(S_i)}, S_i)}{\sum_{i=1}^{N} \delta(\nu_1{}^{\neq(S_i)}, S_i)} \tag{10}$$

For more complete analysis, it can suggest to test the significant difference between the data complexity of groups by using the following statistical test.

- Select K subsets randomly from each group
- Compute the mean and the standard deviation of data complexity of the subsets for each group.
- Use t-test, to analyze whether there is a significant difference between the mean data complexity of the two groups.

If there is a significant difference between the data complexity of sensitive groups, it means that they need different classifiers (with different complexities) to have accurate solutions.

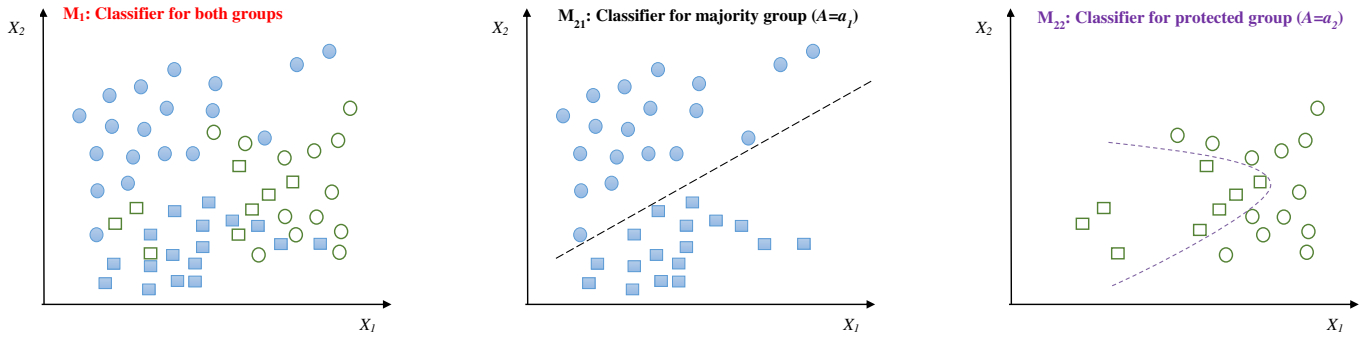


Fig. 3: Different data complexity of sensitive groups. The settings in this figure are similar to that of figures 1 and 2. Here, the complexity of data in groups are different. The complexity of data for one group is less than the other, in a way that one of them can be separated by a linear classifier, and the other cannot.

V. Experimental results

In this section, several experiments are done in order to study the role of sensitive attributes in some real-world problems.

A. Datasets

In these experiments, three popular datasets in the literature (such as [24]–[27]) are utilized. These datasets are as follows:

- **COMPAS Recidivism Dataset:** Recidivism means being charged with a new crime in the future. In this dataset, the goal is to predict which individuals recidivate within two years. This dataset includes records for all offenders in Broward County, Florida in 2013 and 2014 who were assigned a COMPAS score pre-trial [26]. We target variable represents the criminal's likelihood of being involved in the crime (recidivism) and race as a sensitive attribute. We consider two groups based on this attribute (1) Afro-American and (G_1) (2) Non Afro-American (others, G_2).
- **Adult Income Dataset:** In this dataset, the target variable is adult income which indicates whether or not income is greater than 50K dollars [28]. We consider gender as a sensitive attribute, thus we have two groups, men (G_1) and women (G_2).
- **Bank Dataset:** This dataset includes one instance (tuple) for each telephone call in a marketing campaign of a Portuguese banking institution [29]. Each data instance consists of information of a client being contacted by the institution. The target output denotes whether the client subscribes to the bank term deposit (class positive) or not (class negative). The sensitive attribute is the age. Two groups are considered based on this attribute the first group (G_1) includes costumers between 25 and 65 and the second group (G_2) consists of others.

B. Statistical Analyses

At first, the statistical analyses suggested in Section IV are performed over the mentioned datasets, to understand the behavior of each dataset better.

1) Comparison of distributions of other attributes: : In the next step, KS-test has been applied on all datasets, to check whether the distributions of data in two groups are the same. The results of the KS-test over each attribute of all datasets are depicted in Table II. Here for each attribute X_i, the two KS-tests were performed (one for positive class and one for the negative class) the same as that of explained in Section IV.

If the distributions of even one attribute are different it means that the data distributions (joint distribution of all attributes) of sensitive groups are different. Thus, based on the results, the distributions of data for the two groups are different in all datasets. However, analyzing the equality of the distributions of all attributes is very useful to understand that level of difference. the results of these KS-tests reveal that for the two datasets (Adult and COMPAS), the distributions of all or most of the attributes are different for the two groups for both positive and negative classes. If the classification method does not use the sensitive attributes, it should fit a classifier to the mixture of these distributions. Therefore, if these distributions are very different from each other such as the example of Section IV, then it may lead to lower accuracy and lower fairness. The results of such a classifier are mostly against the minority group.

However, if we take a look at the results of the Bank data, for many attributes the p-value is not small, which means that the hypothesis of equal distributions for sensitive groups cannot be rejected for them. So for Bank data, unlike the other two datasets, we do not have different distributions in two groups in many attributes. Thus, we expect that using sensitive attribute as an input for this dataset may not have very important effects on the fairness of the classifier (in comparison with the Adult and COMPAS datasets).

2) Comparison of data complexity: : The next statistical analysis is devoted to investigating the data complexity of the classification problem in each group. Here $N2$ complexity measure is used. For each group of each dataset 30 random sub-samples are selected and $N2$ is calculated for each sub-samples. The average and the standard deviation of the $N2$ complexity measure are reported in Table III. It can be seen that the groups are different, in terms of data complexity, in all

TABLE I: Characteristics of each dataset.

Dataset	# of instances	# of instances of G_1	# of instances of G_2	Base rate of positive class in G_1	Base rate of positive class in G_2
COMPAS	6172	3175	2997	0.532	0.382
Adult	32561	21790	10771	0.305	0.109
Bank	4521	4371	150	0.109	0.293

TABLE II: P-values of KS-tests for the COMPAS dataset. The null hypothesis of KS-test 1 is ($H_0 : \mathbb{P}_{X_i|A=a_1,C=+} = \mathbb{P}_{X_i|A=a_2,C=+}$) and the null hypothesis of KS-test 2 is ($H_0 : \mathbb{P}_{X_i|A=a_1,C=-} = \mathbb{P}_{X_i|A=a_2,C=-}$).

COMPAS dataset	X_1	X_2	X_3	X_4	X_5
KS-test 1	0.00	0.00	0.00	0.00	0.00
KS-test 2	0.00	0.00	0.00	0.00	0.01

Adult dataset.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
KS-test 1	0.00	0.00	0.05	0.31	0.01	0.00	0.00	0.00	0.06	0.07	0.81	0.00	0.98
KS-test 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.88	0.00	0.07

Bank dataset	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
KS-test 2	0.00	1.00	0.01	1.00	0.14	0.01	0.99	0.28	0.30	0.84	0.00	0.19	0.83	1.00	0.88
KS-test 2	0.00	0.01	0.41	1.00	0.02	0.00	0.37	0.15	0.45	0.08	0.69	0.09	0.42	0.51	0.51

TABLE III: Comparison of data Complexity of sensitive groups of each dataset.

Dataset	Group 1	Group 2	P-value of t-test
COMPAS	0.6918 $\pm$0.0873	0.7719 $\pm$0.0818	0.0005
Adult	0.3783 $\pm$0.0485	0.3025 $\pm$0.0739	0.0001
Bank	0.3599 $\pm$0.0711	0.2993 $\pm$0.082	0.0034

datasets. However, COPMAS dataset has the highest difference between the means of data complexity of sensitive groups and the Bank data has the lowest.

C. Comparison of classifiers using and not using sensitive attributes

After the statistical analyses have been done on these datasets, here the effects of sensitive attributes in training classifiers are investigated. Here the main aim is to evaluate the results after using sensitive attributes. In these experiments, we compare three models. One not using sensitive attributes (M_1), one using sensitive attribute as simple input (M_2), and one which uses the sensitive attribute to train separate model for each group (M_3).

We use the random forest method as the main classifier and 70% of each data as the training data and 30% of the data as the test data. The test and training datasets are selected randomly at each run. The results of all three models M_1, M_2 and M_3 are compared based on different quality and fairness measures. All the measures are reported over the test data for 30 independent runs with different random selection of test and train data. We consider all the quality (performance) measures that are involved in computing fairness (discrimination) measures in Section II. These measures include accuracy, precision, recall, FOR, FPR. We also compute the proportion of data assigned to the positive class (here, called support ($\frac{TP+FP}{TP+TN+FP+FN}$). Support is a measure which is considered in the independence definition of fairness. These measures are reported for the two groups. By comparing the results of each method for two groups it is possible to evaluate its fairness.

Figure 5, 6 and 4 depict the performance of each method for both groups. Moreover, Figure 7 illustrate the fairness of each method based on the discrimination measures defined in Section II i.e. $\mathsf{discrimination}_{\text{Independence}}$ (Equation 2), $\mathsf{discrimination}_{\text{Equalized\_Odd}}$ (Equation 5), $\mathsf{discrimination}_{\text{Overall\_Accuracy\_Equality}}$ (Equation 7), $\mathsf{discrimination}_{\text{Sufficiency}}$ (Equation 9).

For COMPAS data, The accuracy and precision of M_2 and M_3 are quite higher than that of M_1, and the FPR of M_2 and M_3 are quite lower than that of M_1. Moreover, the recall for M_2 and M_3 for group 2 is higher than those of M_1 and is very near to the recall of group 1 and the FOR for M_2 and M_3 for group 2 is less than those of M_1 and is very near to the FOR of group 1. Additionally, M_2 and M_3 are fairer (with less discrimination) models according to all fairness definitions except overall accuracy equality.

For Adult data, M_2 and M_3 have better results in terms of accuracy and precision, for both groups. The results (test accuracy and precision) of M_3 are even better than those of M_2. Training two different models for sensitive groups leads to having better quality measures for both groups. This means that by considering the sensitive attributes in training

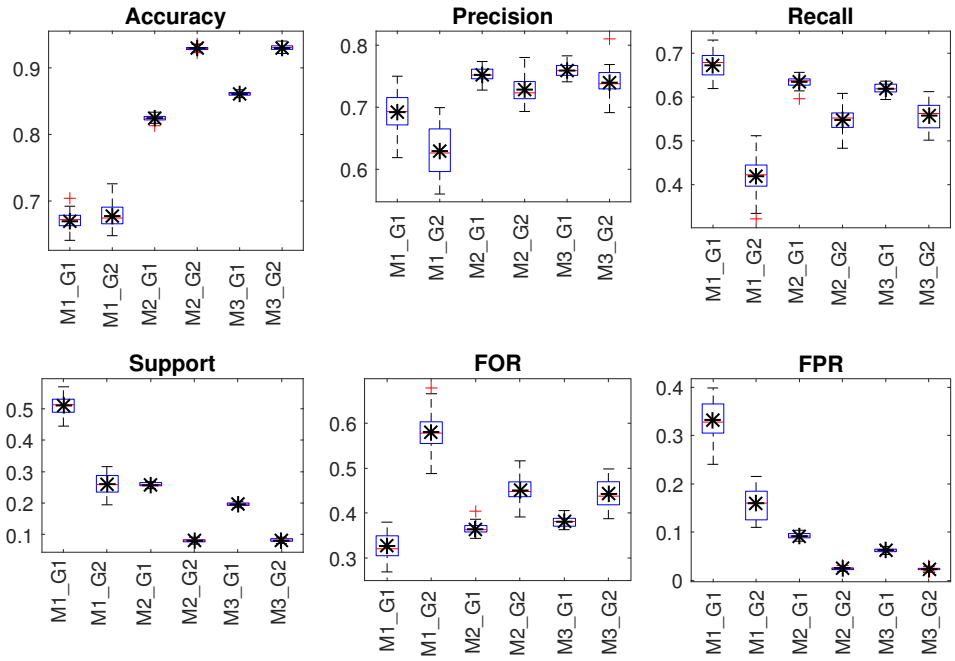


Fig. 4: Comparison of the quality of the methods for COMPAS data.

the classifier in this way, it is possible to have a classifier with higher quality for both groups. Moreover, if we consider the fairness measures, M_3 and M_2 which use the sensitive attribute are fairer than M_1, based on all definitions of fairness.

However, for the Bank data, using sensitive attributes is not helpful to reach classifiers with higher quality and fairness. The quality of M_1, M_2, M_3 are very near to each other. And the fairness of the two models which use sensitive attributes (M_2 and M_3) is a bit worth than that of M_1, based on all fairness definitions. This result is compatible with the results of KS-tests for this dataset. The distributions of some attributes are the same for the two sensitive groups. Moreover, the Bank data had the lowest difference between the means of data complexity of groups. Thus, here using sensitive attributes may not lead to getting better results.

VI. Conclusion

In this paper, we aimed to show that, contrary to common belief, avoiding sensitive attributes in learning classifiers does not necessarily result in fairness. We showed cases where using sensitive attributes is mandatory for achieving fair classifiers. We also suggested two statistical tests to study the role of sensitive attributes in each dataset. These statistical tests can be considered in the data understanding phase that helps to understand the behavior of a dataset. They give us a clue that the sensitive attributes may play an important role in a dataset, and thus it is not possible to remove them from the inputs without caution. If someone wants to use these features in the models, it is very important that their impact in the model is transparent for the end-users. In other words, it is necessary to develop explainable learning methods when incorporating sensitive attributes. Privacy laws often take caution by forbidding the process of sensitive attributes altogether. For instance, European Union's General Data Protection Regulation (GDPR) [30] Article 9 Sentence 1 expresses special categories of personal data (such as ethnicity, race, and religion), whose processing is prohibited. On the other hand, Sentence 2 of the same article lists a number of exceptions to Sentence 2. We believe a new exception is cases where the processing of certain attributes is deemed necessary for providing a fair result, as demonstrated in this paper. The conditions for legally approving an algorithm as such requires close collaboration of legislators and algorithm designers.

Another issue, besides processing sensitive attributes, is their storage. There is always the risk of data being compromised. Fortunately, several data anonymization techniques address this issue. For instance, differential privacy [31] aims at perturbing the data in such a way that any individual has plausible deniability as to whether he/she is included in the dataset, while the result of processing the data remains essentially the same.

In summary, we took a step in justifying the use of sensitive attributes in training fair algorithms. Further discussions are required to measure when using or omitting such attributes results in fairer algorithms, and to legalize the processing of sensitive attributes if they clearly help in fair results.

Acknowledgment

This research has been conducted within the project "Fair and Good ADM" (16ITA203) funded by the Federal Ministry of Education and Research (BMBF) of Germany.

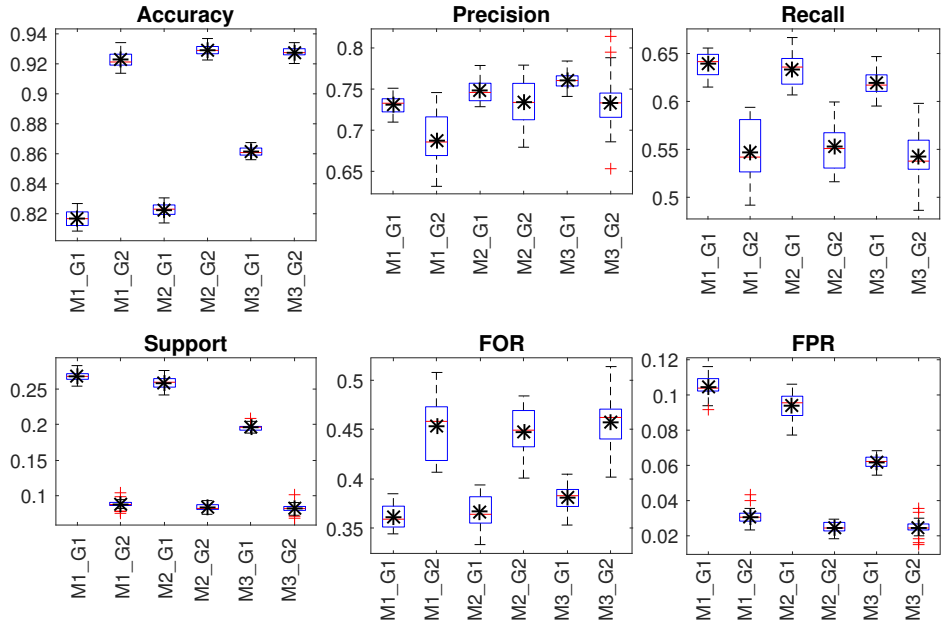


Fig. 5: Comparison of the quality of the methods for Adult data.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

[2] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019.

[3] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017.

[4] Edward Raff, Jared Sylvester, and Steven Mills. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018.

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

[6] 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 *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] Indrė Žliobaitė and Bart Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2):183–201, 2016.

[8] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[9] 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*. ACM, 2012.

[10] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 2016.

[11] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[12] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2), 2017.

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 2017.

[14] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851. IEEE, 2019.

[15] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 2018.

[16] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, 2017.

[17] Mahbod Olfat and Anil Aswani. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[18] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018.

[19] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019.

[20] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 2018.

[21] Ana Justel, Daniel Peña, and Rubén Zamar. A multivariate kolmogorov-smirnov test of goodness of fit. *Statistics & Probability Letters*, 35(3):251–259, 1997.

[22] Ana C Lorena, Luís PF Garcia, Jens Lehmann, Marcilio CP Souto, and Tin Kam Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34, 2019.

[23] José Salvador Sánchez, Ramón Alberto Mollineda, and José Martínez Sotoca. An analysis of how training data complexity affects the nearest neighbor classifiers. *Pattern Analysis and Applications*, 10(3):189–201, 2007.

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

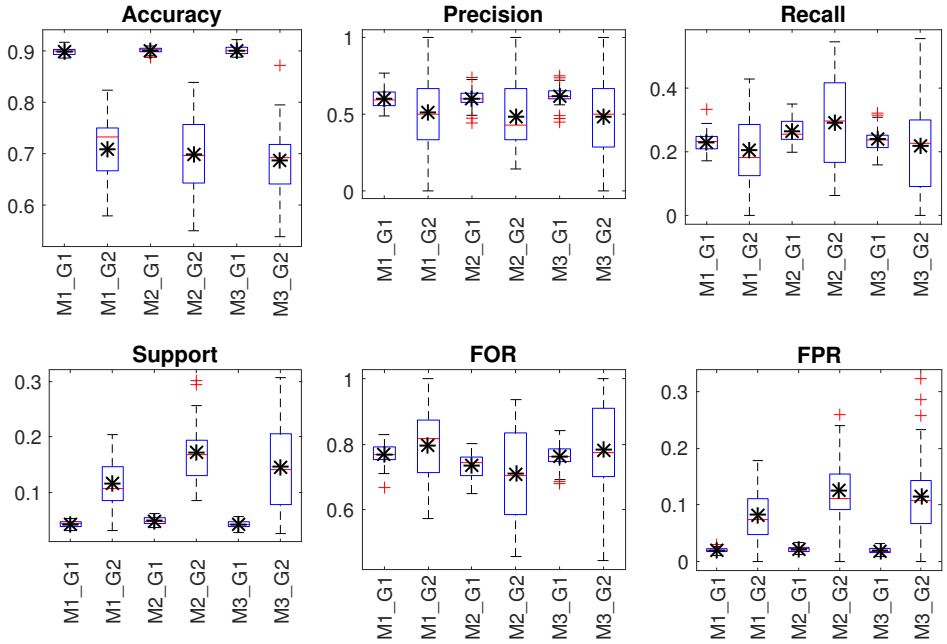
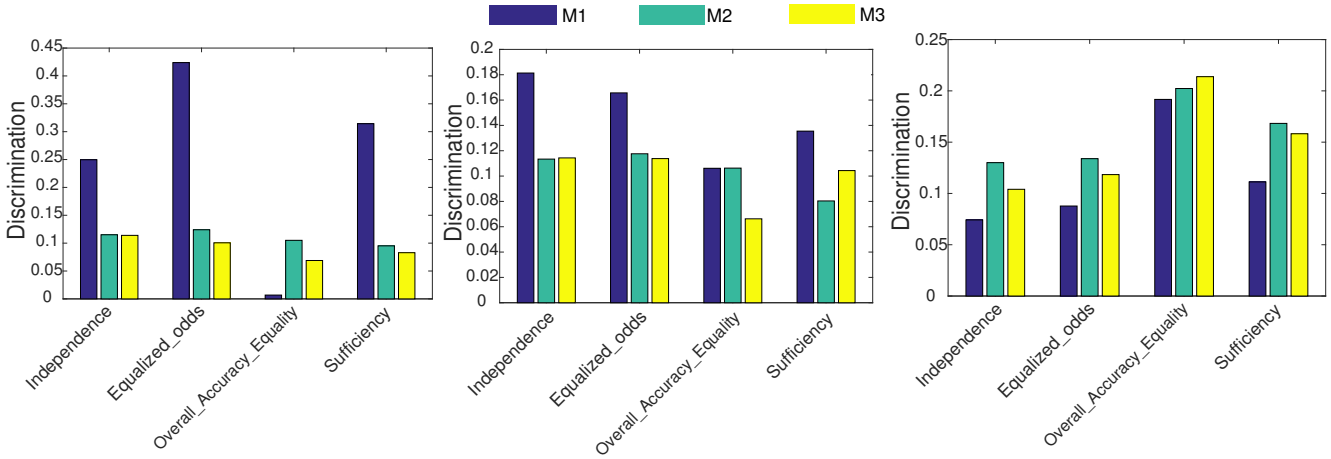


Fig. 6: Comparison of the quality of the methods for Bank data.



(a) Compas data (b) Adult data (c) Bank data

Fig. 7: Comparison of the fairness of the methods based on different definitions of fairness. The methods with less discrimination values are fairer.

[25] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011.

[26] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

[27] Bo Cowgill. The impact of algorithms on judicial discretion: Evidence from regression discontinuities. Technical report, Technical Report. Working paper, 2018.

[28] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, 1996.

[29] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[30] Council of European Union. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR), 2016. https://gdpr-info.eu.

[31] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.