# A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set

P. Besse[1]  &  E. del Barrio[2]  &  P. Gordaliza[3,2]  &  J-M. Loubes[1]  &  L. Risser[1]

1*: Institut de Mathématiques de Toulouse
INSA, Université Toulouse 3, CNRS
Toulouse, France

2: Instituto de Matemáticas de la Universidad de Valladolid
Dpto. de Estadistica e Investigacion Operativa
Universidad de Valladolid
Valladolid, Spain

3: Basque Center for Applied Mathematics
Bilbao, Spain

November 18, 2020

### Abstract

Applications based on Machine Learning models have now become an indispensable part of the everyday life and the professional world. A critical question then recently arises among the population: Do algorithmic decisions convey any type of discrimination against specific groups of population or minorities? In this paper, we show the importance of understanding how a bias can be introduced into automatic decisions. We first present a mathematical framework for the fair learning problem, specifically in the binary classification setting. We then propose to quantify the presence of bias by using the standard *Disparate Impact* index on the real and well-known *Adult income* data set. Finally, we check the performance of different approaches aiming to reduce the bias in binary classification outcomes. Importantly, we show that some intuitive methods are ineffective with respect to the Statistical Parity criterion. This sheds light on the fact that trying to make fair machine learning models may be a particularly challenging task, in particular when the training observations contain some bias.

*Keywords:* Fairness, Disparate Impact, Machine Learning, Tutorial.

---

# 1    Introduction

Fairness has become one of the most popular topics in machine learning over the last years and the research community is investing a large amount of effort in this area. The main motivation is the increasing impact that the lives of Human beings are experiencing due to the generalization of machine learning systems in a wide variety of fields. From the first artificial neural network in the late 50's, machine learning algorithms are now becoming an inseparable part of our daily lives since more and more companies are integrating Artificial Intelligence (AI) into their existing practice or products. While some of these uses may involve leisure, with trivial consequences (Amazon or Netflix use recommender systems to present a customized page that offers their products according to the order of preference of each user), other ones entail particularly sensitive decisions such as in Medicine, where patient suitability for treatment is considered; in Human Resources, where candidates are sorted out on an algorithmic decision basis; in the Automotive industry, with the release of self-driving cars; in the Banking and Insurance industry, which characterize customers according to a risk index; in Criminal justice, where the COMPAS algorithm is used in the United States for recidivism prediction... For a more detailed background on these facts see for instance Romei and Ruggieri (2014), Berk et al. (2018) Pedreschi et al. (2012) or Friedler et al. (2018), and references therein.

The technologies that AI offers certainly make life easier. It is however a common misconception that they are absolutely objective. In particular, machine learning algorithms which are meant to automatically take accurate and efficient decisions that mimic and even sometimes outmatch human expertise, rely heavily on potentially biased data. It is interesting to remark that this bias is often due to an inherent social bias existing in the population that is used to generate the training dataset of the machine learning models. A list of potential causes for the discriminatory behaviours that machine learning algorithms may exhibit, in the sense that groups of population are treated differently, is given in Barocas and Selbst (2016). Various real and striking cases that can be found in the literature are the following. In Angwin et al. (2016), it was found that the algorithm COMPAS used for recidivism prediction produces much higher rate of false positive predictions for black people than for white people. Later in Lahoti et al. (2019), a job platform similar to Linkedin called XING was found to predict less highly ranked qualified male candidates than female candidates. Publicly available commercial face recognition online services provided by Mi-

crosoft, Face++, and IBM respectively were also recently found to suffer from achieving much lower accuracy on females with darker skin color in Buolamwini and Gebru (2018). Although a discrimination may appear naturally and could be thought as acceptable, as in Kamiran et al. (2010) for instance, quantifying the effect of a machine learning predictor with respect to a given situation is of high importance. Therefore, the notion of fairness in machine learning algorithms has received a growing interest over the last years. We believe this is crucial in order to guarantee a fair treatment for every subgroup of population, which will contribute to reduce the growing distrust of machine learning systems in the society.

Yet providing a definition of fairness or equity in machine learning is a complicated task and several propositions have been formulated. First described in terms of law (Winrow and Schieber, 2009), fairness is now quantified in order to detect biased decisions from automatic algorithms. We will focus on the issue of biased training data, which is one of the several possible causes of such discriminatory outcomes in machine learning mentioned above. In the fair learning literature, fairness is often defined with respect to selected variables, which are commonly denoted *protected* or *sensitive attributes*. We note that throughout the paper we will use both terms indistinctly. This variables encode a potential risk of discriminatory information in the population that should not be used by the algorithm. In this framework, two main streams of understanding fairness in machine learning have been considered. The probabilistic notion underlying this division is the independence between distributions. The first one gives rise to the concept of Statistical Parity, which means the independence between the protected attribute and the outcome of the decision rule. This concept is quantified using the Disparate Impact index, which is described for instance in Feldman et al. (2015). This notion was firstly considered as a tool for quantifying discrimination as the so-called $4/5^{th}$-rule by the State of California Fair Employment Practice Commission (FEPC) in 1971. For more details on the origin and first applications of this index we refer to Biddle (2006). The second one proposes the Equality of Odds, which considers the independence between the protected attribute and the output prediction, conditionally to the true output value. In other words, it quantifies the independence between the error of the algorithm and the protected variable. Hence, in practice, it compares the error rates of the algorithmic decisions between the different groups of the population. This second point of view has been originally proposed for recidivism of defendants in Flores et al. (2016). Many other criteria (see for instance in Berk et al. (2018) for a review) have been proposed leading sometimes to incompatible formulations as stated in Chouldechova (2017). Note

finally that the notion of fairness is closely related to the notion of privacy as pointed out in Dwork et al. (2012).

In this paper, our goal is to present some comprehensive statistical results on fairness in machine learning studying the statistical parity criterion through the analysis of the example given in the *Adult Income* dataset. This public dataset is available on the UCI Machine Learning Repository[1]. Here, the goal consists in forecasting a binary variable (low or high income) which corresponds to an income lower or higher than 50k$ a year. This decision could be potentially used to evaluate the credit risk of loan applicants, making this dataset particularly popular in the machine learning community. It is considered here as potentially sensitive to a discrimination with respect to the *Gender* and *Ethnic origin* variables. The co-variables used in the prediction as well as the true outcome are available in the dataset, hence supervised machine learning algorithms will be used.

Section 2 describes this dataset. It specifically highlights the existing unbalance between the income prediction and the *Gender* and *Ethnic origin* sensitive variables. We note that a preprocessing step is needed in order to prepare the data for further analyses and the performed modifications are detailed in the Appendix A.1.1. In Section 3, we then explain the statistical framework for the fairness problem, by particularly focusing on the binary classification setting. Among the wide array of definitions for fairness existing in the literature, we follow the approach of the *Statistical Parity* to quantify the fairness and we thus present the Disparate Impact as our preferred index for measuring the bias. Note that the bias is present in this dataset, so the machine learning decision rules learned in this paper will be trained by using a biased dataset. Although, many criteria have been described in the fair learning literature, they are often used as a score without statistical control. In the cases where test procedures or confidence bounds are provided, they are obtained using a resampling scheme to get standardized Gaussian confidence intervals under a Gaussian assumption which does not correspond to the distribution of the observations. In this work, we promote the use of confidence intervals to control the risk of false discriminatory assessment. We then show in the Appendix A.2 the exact asymptotic distribution of the estimates of different fairness criteria obtained through the classical approach of the Delta method described in Van der Vaart (1998). Then, Section 4 is devoted to present some naive approaches that try to correct the discriminatory behaviour of machine learning algorithms or to test possible

---

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

discriminations. Finally, Section 5 is devoted to studying the efficiency of two easy ways to incorporate fairness in machine learning algorithms: building a different algorithm for each class of the population or adapting the decision of a single algorithm in a different way for each subpopulation. We then in Section 6 present some conclusions for this work and thus provide a concrete pedagogical example for a better understanding of bias issues and fairness treatment in machine learning. It is convenient to note at this point the frequent use throughout the paper of term "bias" referring to "social bias", in stead of "statistical bias", and the term "discrimination" referring to any harmful or illegal practice, in stead of the usual sense in classification. In both cases, we ask the reader to infer the intended use from context. Proofs and more technical details are included in the Appendix. Relevant code in Python to preprocess the Adult Income dataset and reproduce all the analysis and figures presented in this paper are available at the link `https://github.com/XAI-ANITI/StoryOfBias/blob/master/StoryOfBias.ipynb`. We also provide the French version of this Python notebook at `https://github.com/wikistat/Fair-ML-4-Ethical-AI/blob/master/AdultCensus/AdultCensus-R-biasDetection.ipynb`.

# 2    Machine learning algorithms for the attribution of bank loans

One of the applications for which machine learning algorithms have already become firmly established is credit scoring. In order to minimize its risks, the banking industry uses machine learning models to detect the clients who are likely to deal with a credit loan. The FICO score in the US or the SCHUFA score in Germany are examples of these algorithmically determined credit rating scores, as well as those used by a number of Fintech startups, who are also basing their loan decisions entirely on algorithmic models (Hurley and Adebayo, 2016)[2]. Yet, credit rating systems have been criticized as opaque and biased in Pasquale (2015), Rothmann et al. (2014) or Hurley and Adebayo (2016).

In this paper, we use the Adult Income dataset as a realistic material to reproduce this kind of analyses for credit risk assessment. This dataset was built by using a database containing the results of a census made in the United States in 1994. It has been largely used among the fair learning community as a suitable benchmark to compare the performance

---

[2]See, e.g., `https://www.kreditech.com/`.

of different machine learning methods. It contains information from about 48 thousands of individuals, each of them being described by 14 variables as detailed in Table 2. This dataset is often used to predict the binary variable *Anual Income higher or not than* $50k\$$. Such a forecast does not convey any discrimination itself, but it illustrates what can be done in the banking or insurance industry since the machine learning procedures are similar to those made by banks to evaluate the credit risk of their clients. The fact that the true value of the target variable is known, in contrast to the majority of the datasets available in the literature (e.g. the German Credit Data), as well as the value of potential protected attributes such as the ethnic origin or the gender, makes this dataset one of the most widely used to compare the properties of the fair learning algorithms. In this paper, we will then compare supervised machine learning methods on this dataset. A graphic representation of the distribution of each feature can be found in `https://www.valentinmihov.com/2015/04/17/adult-income-data-set/`. This representation gives a good overview of what this dataset contains. It also makes clear that it has to be pre-processed before its analysis using black-box machine learning algorithms. In this work, we have deleted missing data, errors or inconsistencies. In some cases, this type of pre-processing practices could have fairness implications. This is a very interesting issue, yet beyond the scope of our study. We also have merged highly dispersed categories and eliminated strong redundancies between certain variables (see details in Supplementary material A.1.1). In Figure 1, we represent the data set after our pre-treatments, and show the number of occurrences for each categorical variable as well as the histograms for each continuous variable.

## 2.1 Unbalanced Learning Sample

After pre-processing the dataset, standard preliminary exploratory analyses first show that the dataset obviously suffers from an unbalanced repartition of low and high incomes with respect to two variables: *Gender* (male or female) and *Ethnic origin* (caucasian or non-caucasian). These variables therefore seem to be potentially sensitive variables in our data. Figure 2 shows this unbalanced repartition of incomes with respect to these variables. It is of high importance to be aware of such unbalanced repartitions in reference datasets since a bank willing to use an automatic algorithm to predict which clients should have successful loan applications could be tempted to train the decision rules on such unbalanced data.
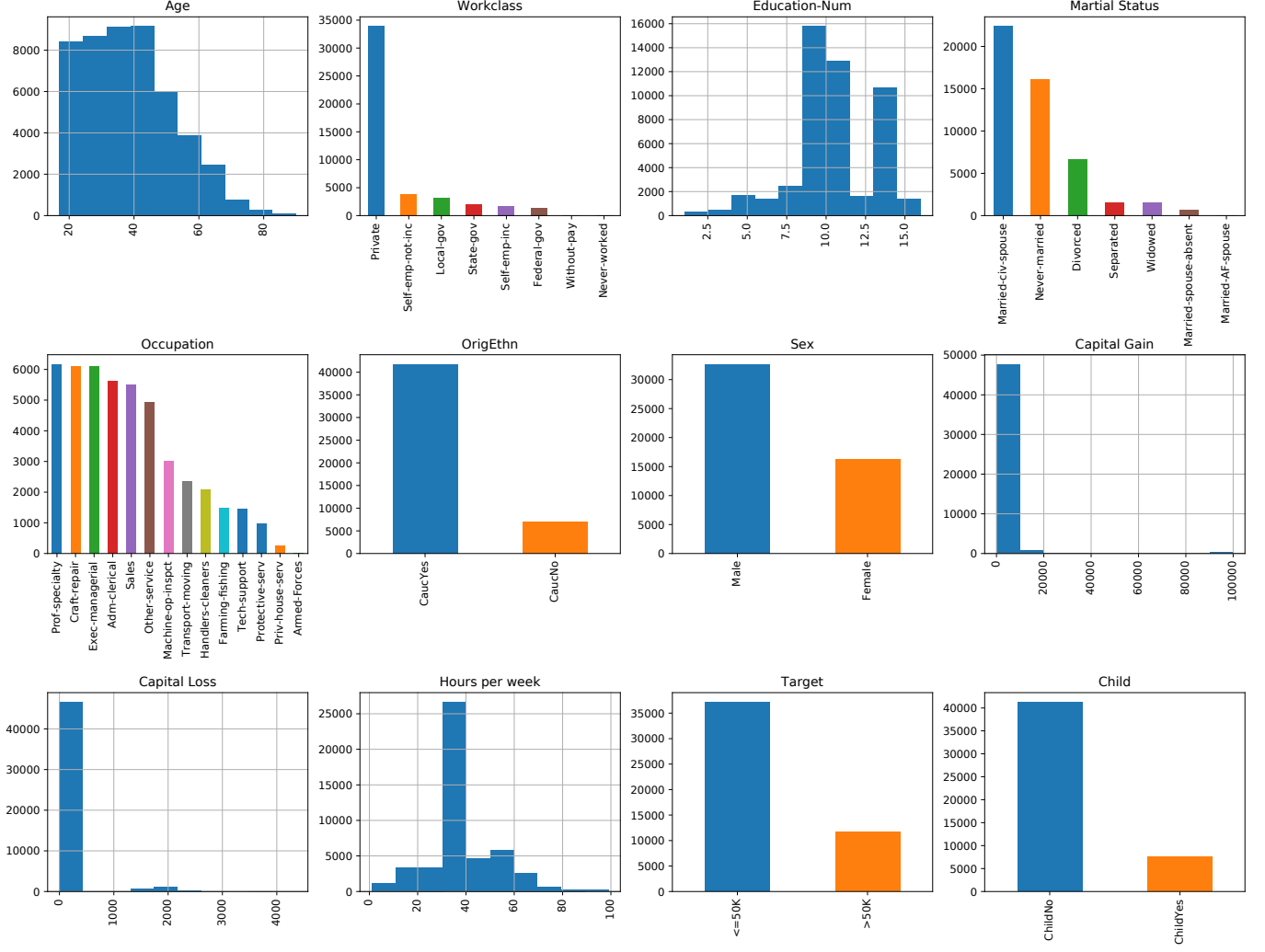
6

Figure 1: Adult Income dataset after pre-processing phase

This fact is at the heart of our work and we question its effect on further predictions on other data. What information will be learnt from such unbalanced data: a fair relationship between the variables and the true income that will enable socially reasonable forecasts; or biased relations in the repartition of the income with respect to the sensitive variables? We explore this question in the following section.

## 2.2 Machine Learning Algorithms to forecast income

We study now the performance of four categories of supervised learning models: Logistic Regression (LR) (Cramer, 2002), Decision Trees (DT) (Mitchell et al., 1997), Gradient Boosting (GB) (Sutton, 2005), and Neural Network (NN). We used the *Scikit-learn* implementations of the LR and DT, and the *lightGBM* implementation of the GB) algorithm.
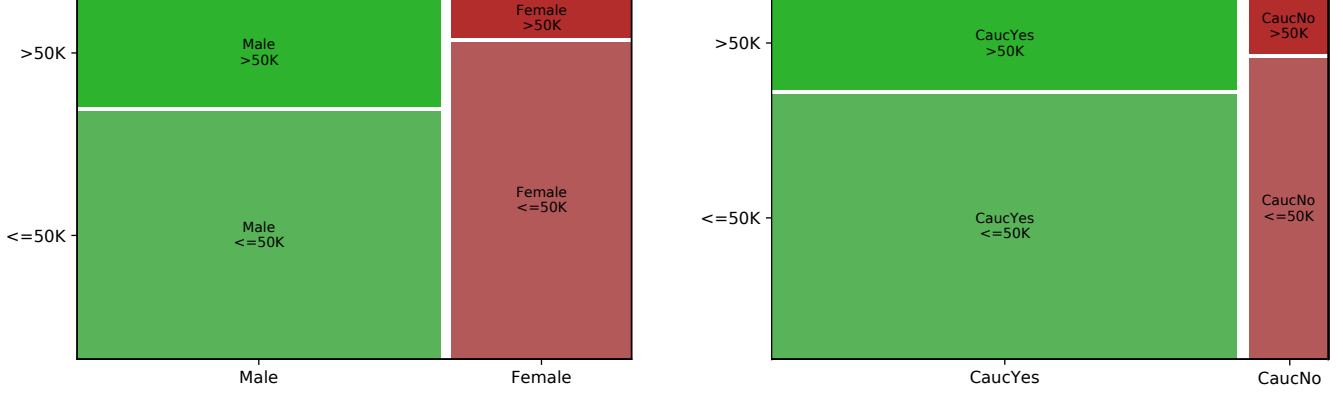
Figure 2: Enbalancement of the reference decisions in the *Adult Income* dataset with respect to the *Gender* and *Ethnic origin* variables.

The NN was finally coded using *PyTorch* and contains four fully connected layers with Rectified Linear Units (ReLU) activation functions.

In order to analyze categorical features using these models, the binary categorical variables were encoded using zeros and ones. The categorical variables with more than two classes were also transformed into one-hot vectors, *i.e.* into vectors where only one element is non-zero (or hot). We specifically encoded the target variable by the values $Y = 0$ for an income below $50K\$$, and $Y = 1$ for an income above $50K\$$. We used a 10-fold cross-validation approach in order to assess the robustness of our results. The average accuracy as well as its true positive (TP) and true negative (TN) rates were finally measured for each trained model. Figure 3 summarizes these results.

We can observe in Fig. 3 that the best average results are obtained by using Gradient Boosting. More interestingly, we can also remark that the prediction obtained using all models for $Y = 0$ (represented by the true negative rates) are clearly more accurate than those obtained for $Y = 1$ (represented by the true positive rates), which contains about 24% of the observations. All tested models then make more mistakes in average for the observations which should have a successful prediction than a negative one. Note that the tested neural network is outperformed by other methods in these tests in term of prediction accuracy. Although we used default parametrizations for the Logistic Regression model as well as the Gradient Boosting model, and we simply tuned the decision tree to have a maximum depth of 5 nodes, we tested different parametrizations of the Neural Network model (number of epochs, mini-batch sizes, optimization strategies) and kept the best performing one. It therefore appears that the neural network model we tested was clearly not adapted
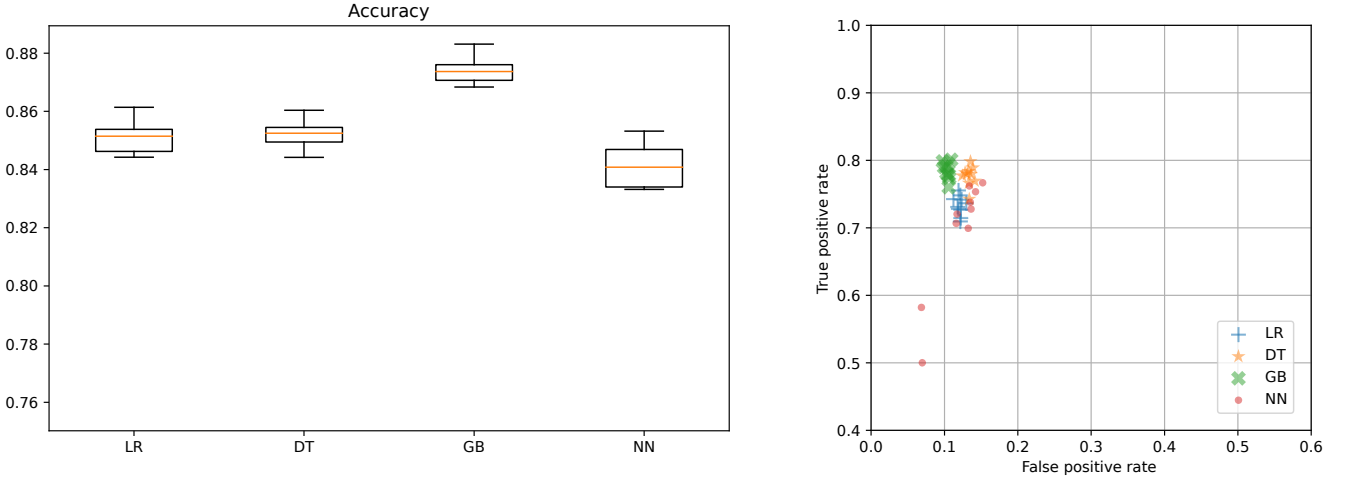
Figure 3: Prediction accuracies, true positive rates and true negative rates obtained by using no specific treatment. Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB) and Neural Network (NN) models were tested with 10-folds cross validation on the *Adult Income* dataset.

to the *Adult Income* dataset.

Hence we have built and compared several algorithms ranging from completely interpretable models to black box models involving optimization of several parameters. Note that we could have used the popular Random Forest algorithm that could lead to equivalent but we privileged boosting models whose implementation is easier using Python.

# 3 Measuring the Bias with Disparate Impact

## 3.1 Notations

Among the criteria proposed in the literature to reveal the presence of a bias in a dataset or in automatic decisions (see *e.g.* Hardt et al. (2016) for a recent review), we focus in this paper on the so-called Statistical Parity. This criterion deals with the differences in reference decisions or the outcome of decision rules with respect to a sensitive attribute. Note that we only consider the binary classification problem with a single sensitive attribute for the sake of simplicity, although we could consider other tasks (*e.g.* regression) or multiple sensitive attributes (see Hébert-Johnson et al. (2018) or Kearns et al. (2018)). Here is a summary of the notations we use:

- $Y$ is the variable to be predicted. We consider here binary variables where $Y = 1$ is

a positive decision (here a high income) while $Y = 0$ is a negative decision (here a low income);

- $g(X) = \hat{Y}$ is the prediction given by the algorithm. As for $Y$, this is a binary variable interpreted such that $\hat{Y} = 0$ or $\hat{Y} = 1$ means a negative or a positive decision, respectively. Note that most machine learning algorithms output continuous scores or probabilities. We consider in this case that this output is already thresholded.

- $S$ is the variable which splits the observations into groups for which the decision rules may lead to discriminative outputs. From a legal or a moral point of view, $S$ is a sensitive variable that should not influence the decisions, but could lead to discriminative decisions. We consider hereafter that $S = 0$ represents the minority that could be discriminated, while $S = 1$ represents the majority. We specifically focus here on estimating the disproportionate effect with respect to two sensitive variables: gender (male vs. female) and ethnic origin (caucasian vs. non-caucasian).

Statistical Parity is often quantified in the fair learning literature using the so-called Disparate Impact (DI). The notion of DI has been introduced in the us legislation in 1971[3]. It measures the existing bias in a dataset as

$$DI(Y, S) = \frac{\mathbb{P}(Y = 1 | S = 0)}{\mathbb{P}(Y = 1 | S = 1)} \, , \tag{3.1}$$

and can be empirically estimated as

$$\frac{n_{10}}{(n_{00} + n_{10})} \Big/ \frac{n_{11}}{(n_{01} + n_{11})} \, , \tag{3.2}$$

where $n_{ij}$ is number of observations such that $Y = i$ and $S = j$. The smaller this index, the stronger the discrimination over the minority group. Note first that this index supposes that $\mathbb{P}(Y = 1 | S = 0) < \mathbb{P}(Y = 1 | S = 1)$ since $S$ is defined as the group which can be discriminated with respect to the output $Y$. It is also important to remark that this estimation may be unstable due to the unbalanced amount of observations in the groups $S = 0$ and $S = 1$ and the inherent noise existing in all data. We then propose to estimate a confidence interval around the Disparate Impact in order to provide statistical guarantees of this score, as detailed in the Supplementary material A.2. These confidence intervals will

---

[3]https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

be used later in this section to quantify how reliable are two disparate impacts computed on our dataset. This fairness criterion can be extended to the outcome of an algorithm by replacing in Eq. (3.1) the true variable $Y$ by $g(X) = \hat{Y}$, that is

$$DI(g, X, S) = \frac{\mathbb{P}(g(X) = 1 | S = 0)}{\mathbb{P}(g(X) = 1 | S = 1)}. \tag{3.3}$$

This measures the risk of discrimination when using the decision rules encoded in $g$ on data following the same distribution as in the test set. Hence, in Gordaliza et al. (2019) a classifier $g$ is said not to have a Disparate Impact at level $\tau \in (0, 1]$ when $DI(g, X, S) > \tau$. Note that the notion of DI defined Eq. (3.1) was first introduced as the $4/5^{th}$-rule by the State of California Fair Employment Practice Commission (FEPC) in 1971. Since then, the threshold $\tau_0 = 0.8$ was chosen in different trials as a legal score to judge whether the discriminations committed by an algorithm are acceptable or not (see e.g. Feldman et al. (2015) Zafar et al. (2017), or Mercat-Bruns (2016)).

## 3.2 Measures of disparate impacts

The disparate impact $DI(g, X, S)$ should be close to 1 to claim that $g$ makes fair decisions. A more subtle, though critical, remark is that it should at least not be smaller than the general disparate impact $DI(Y, S)$. This would indeed mean that the decision rules $g$ reinforce the discriminations compared with the reference data on which it was trained. We will then measure hereafter the disparate impacts $DI(Y, S)$ and $DI(g, X, S)$ obtained on our dataset.

In Table 1, we have quantified confidence intervals for the bias already present in the original dataset using Eq. (3.1) with the sensitive attributes *Gender* and *Ethnic origin*. They were computed using the method of Appendix A.2 and represent the range of values the computed disparate impacts can have with a 95% confidence (subject to standard and reasonable hypotheses on the data). Here the DI computed on the *Gender* variable then appears as very robust and the one computed on the *Ethnic origin* variable is relatively robust. It is clear from this table that both considered sensitive attributes generate discriminations. These discriminations are also more severe for the *Gender* variable than for the *Ethnic origin* variable.

We have then measured the disparate impacts Eq. (3.3) obtained using the predictions made by the four models in the 10-folds cross-validation of Section 2.2. These disparate

Table 1: Bias measured in the original dataset

| Protected attribute | DI | CI |
|---|---|---|
| Gender | 0.3597 | $[0.3428, 0.3765]$ |
| Ethnic origin | 0.6006 | $[0.5662, 0.6350]$ |

impacts are presented in Fig. 4. We can see that, except for the decision tree with the *Ethnic origin* variable, the algorithms have smaller disparate impact than for the true variable. The impact is additionally clearly worsened with the *Gender* variable using all trained predictors. These predictors therefore reinforced the discriminations in all cases by enhancing the bias present in the training sample. Observing the true positive and true negative rates of Fig. 4, which distinguish the groups $S = 0$ and $S = 1$ is particularly interesting here to understand this effect more deeply. As already mentioned Section 2.2, the true negative (TN) rates are generally higher than the true positive (TP) rates. It can be seen Fig. 4 that this phenomenon is clearly stronger in the subplot representing the TP and TN for $S = 0$ than the one representing them for $S = 1$, so false predictions are more favorable to the group $S = 1$ than the group $S = 0$. This explains why the disparate impacts of the predictions are higher than those of the original data (boxplots *Ref* in Fig. 4). Note that these measures are directly related to the notions of equality of odds and opportunity as discussed in Hardt et al. (2016). The machine learning models we used in our experiments were then shown as unfair on this dataset, in the sense that discrimination is reinforced.

As pointed out in Friedler et al. (2018), there may be a strong variability when computing the disparate impact of different subsamples of the data. Hence, we additionally propose in this paper an exact Central Limit Theorem to overcome this effect. The confidence intervals we obtain prove their stability when confronted to bootstrap replications and for this reason, therefore, we cross-validated our results using 10 replications of different learning and test samples on the three algorithms. The construction of these confidence intervals are postponed to Section A.2 while comparison with bootstrap procedures are detailed in Section A.3 of the Appendix. In order to conveniently compare the bias in the predictions with the one in the original data, we show on the left the bias measured in the data. We can see that these boxplots are coherent with the results of Table 1 and Figure 4, and again show that the discrimination was reinforced by the machine learning models in
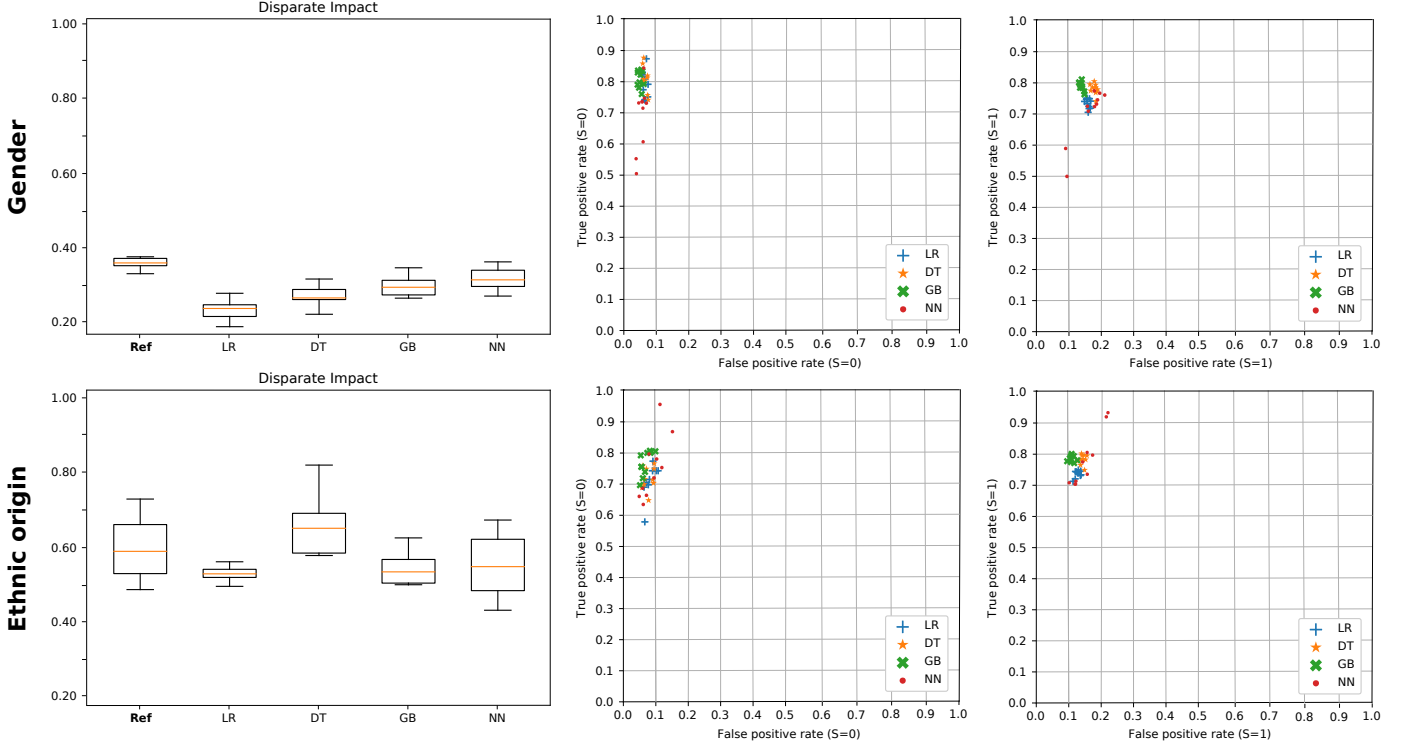
Figure 4: Bias measured in the outputs of the tested machine learning models (LR, DT, GB, NN) using the 10-folds cross validation. The disparate impacts of the reference decisions are represented by the boxplot *Ref* to make clear that the unfairness is almost always re-inforced in our tests by automatic decisions. **(Top)** *Gender* is the sensitive variable. **(Bottom)** *Ethnic origin* is the sensitive variable.

this test.

In all generality, we conclude here that one has to be careful when training decision rules. They can indeed worsen existing discriminations in the original database. We also remark that the majority of works using the Disparate Impact as a measure of fairness rely only on this score as a numerical value with no estimation of how reliable it is. This motivated the definition of our confidence intervals strategy in Appendix A.2, which was shown to be realistic in our experiments when comparing the *Ref* boxplots of Figure 4 with the confidence intervales of Tables 1. Note that we will only focus in the rest of the paper on the protected variable *Gender* since it was shown in Section 3 to be clearly the variable leading to discrimination for all tested machine learning models. We will also only test the Logistic Regression (LR) and Decision Tree (DT) as they are highly interpretable, plus the Gradient Boosting (GB) model which was shown to be the best performing one on the

*Adult Census* dataset.

# 4   A quantitative evaluation of GDPR recommendations against algorithm discrimination

Once the presence of bias is detected, the goal of machine learning becomes to reduce its impact without hampering the efficiency of the algorithm. Actually, the predictions made by the algorithm should remain sufficiently accurate to make the machine learning model relevant in Artificial Intelligence applications. For instance, the decisions $\hat{Y}$ made by a well balanced coin when playing *head or tail* are absolutely fair, as they are independent of any possible sensitive variable $S$. However, they also do not take into account any other input information $X$, making them pointless in practice. Reducing the bias of a machine learning model $g$ therefore ideally consists in getting rid of the influence of $S$ in all input data $(X, S)$ while preserving the relevant information to predict the true outputs $Y$. We will see below that this is not that obvious, even in our simple example.

It is first interesting to remark that the problem cannot be solved by simply having a balanced amount of observations with $S = 0$ and $S = 1$. We indeed reproduced the experimental protocol of Section 3.2 with 16,192 randomly chosen observations representing males (instead of 32,650), so that the decision rules were trained in average with as many males as females. As shown in Fig. 5, the trends of the results turned out to be very similar to those obtained in Fig. 4-*(Gender)*.
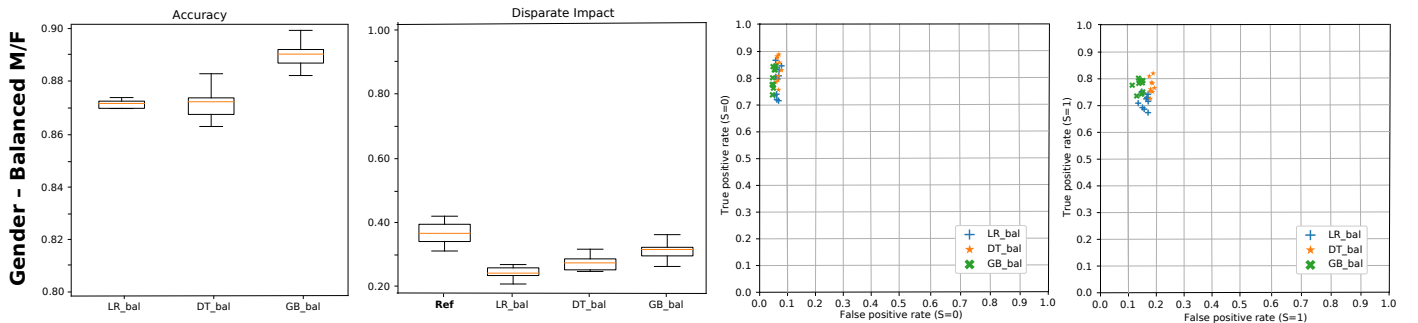


Figure 5: Bias measured in the outputs of the LR, DT and GB machine learning models using the same experimental protocol as in Section 3.2 (see specifically Fig. 4-*(Gender)*), except that we used the same amount of males ($S = 1$) and females ($S = 0$) in the dataset.

We specifically study in section the effect of complying to the European regulations.

From a legal point of view, the GDPR's recommendation indeed consists in not using the sensitive variable in machine learning algorithms. Hence, we simply remove here $S$ from the database in subsection 4.1, and we consider in subsection 4.2 one of the most common legal proof for discrimination called the *legal testing method* (in the following *testing method*, or simply *testing*). It consists in considering the response for the same individual but with a different sensitive variable. We will study whether this procedure enables to detect the group discrimination coming from the decisions of an algorithm.

## 4.1 What if the sensitive variable is removed?

The most obvious idea to remove the influence of a sensitive variable $S$ is to remove it from the data, so we cannot use it when training the decision rules and then obviously when making new decisions. Note that this solution is recommended by GPDR regulations. To test the pertinence of this solution, we considered the algorithms analyzed in Sections 2 and 3 and then used them without using the *Gender* variable. As in Section 3, a 10-fold cross-validation approach was used to assess the robustness of our results.

As shown Figure 6-(top), the disparate impacts as well as the model accuracies remained almost unchanged when removing the *Gender* variable from the input data. Anonymizing database by removing a variable therefore had very little effect on the discrimination that is induced by the use of an automated decision algorithm. This is very likely to be explained by the fact that a machine learning algorithm uses all possible information conveyed by the variables. In particular, if the sensitive variable (here the *Gender* variable) is strongly correlated to other variables, then the algorithm learns and reconstruct automatically the sensitive variable from the other variables. Hence we can deduce that social determinism is stronger than the presence of the sensitive variable here, so the classification algorithms were not impacted by the removal of this variable.

Obtaining fairness is a far more complicated task than this simple trick. It is at the heart of modern research on fair learning. More complex fairness mathematical methods to reduce disparate treatment are discussed for instance in Kleinberg et al. (2016) or in Gordaliza et al. (2019).
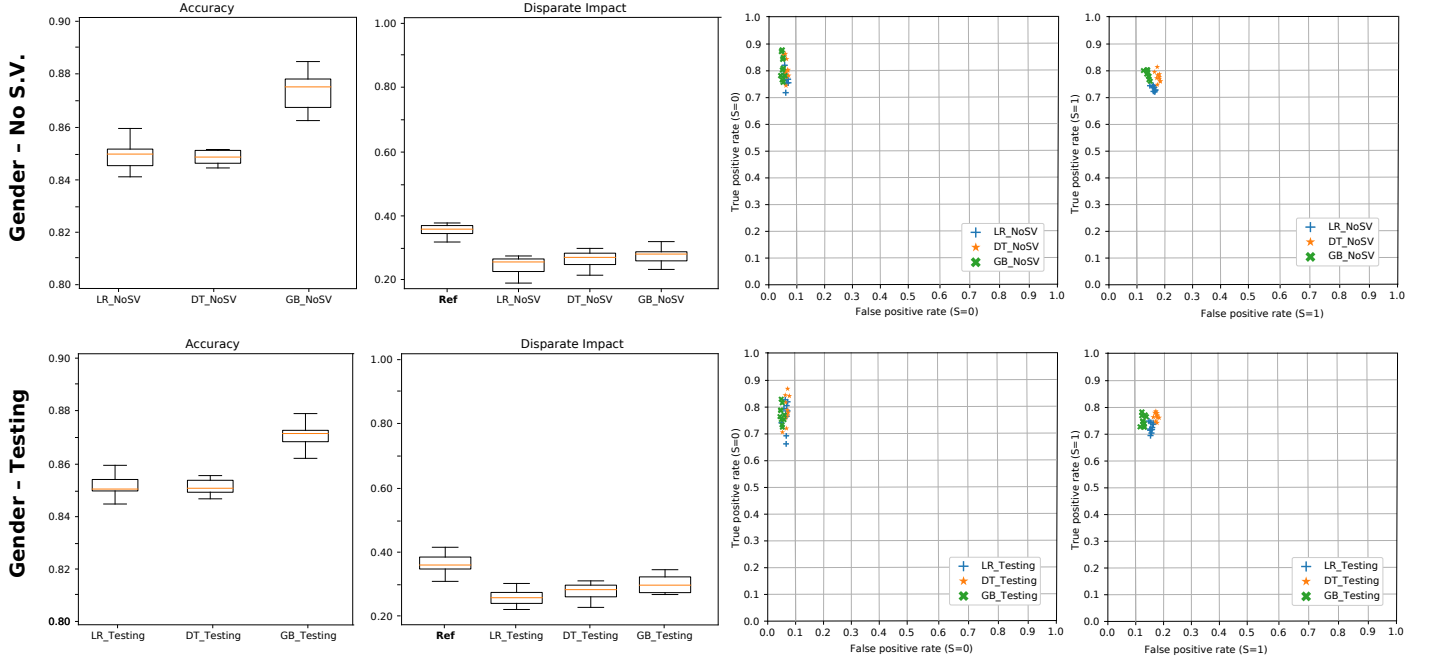
Figure 6: Performance of the machine learning models LR, DT and GB when **(top)** removing the *Gender* variable, and **(bottom)** when using a testing procedure.

## 4.2 From Testing for bias detection to unfair prediction

Testing procedures are often used as a legal proof for discrimination. For a chosen individual $(x, s)$ that suspects a disparate treatment and discrimination, such procedures consist in first creating an artificial individual $(x, s')$ which shares the same characteristics but has a different protected variable. Then it amounts to testing whether this artificial individual has the same prediction $\hat{y}' = f(x, s')$ as the original one $\hat{y} = f(x, s)$. If the predictions differ, that is if $\hat{y}' \neq \hat{y}$, then this conclusion can serve as a legal proof for discrimination.

These methods have existed for a long time (first introduced in 1939[4]) and the French justice has taken them as a proof of biased treatment since 2006, although the testing process itself has been qualified as unfair[5]. Furthermore, this technique has been generalized by sociologists and economists (see Riach and Rich (2002) for a description of such method) to statistically measure group discrimination in housing and labour market by conducting carefully controlled field experiments.

This testing procedure considered as a discrimination test is nowadays a commonly used method in France to assess fairness for sociological studies of *Observatoire des dis-*

---

[4] https://fr.wikipedia.org/wiki/Test_de_discrimination

[5] https://www.juritravail.com/discrimination-physique/embauche/ph-alternative-A-1.html

*criminations*[6] and *laboratoire TEPP* as pointed out in L'Horty, Bunel, Mbaye, Petit, and du Parquet (L'Horty et al.), or governemental studies DARES[7] of French Ministry of Work ISM Corum [8]. Some industries are labeled using such test. An audit quality of recruiting methods is proposed while *Novethic*[9] proposes ethic formations.

Testing is efficient to detect human discrimination specially in labour market but hiring tech is producing more and more softwares or web platforms performing predictive recruitment as in Raghavan et al. (2020). Does testing remains valid in front of machine learning algorithms? This last strategy is evaluated using the same experimental protocol as in the previous sections. The results of these experiments are shown in Figure 6-(bottom). Testing does not detect any discrimination when the sensitive variable is captured by the other variables.

An algorithmic solution to bypass this testing procedure is given by the following trick. Train a classifier as usual using all available information $X, S$ and then build a *testing compliant* version of it as follows : for an individual, the predicted outcome is assigned as the best decision obtained on the actual individual $f(x, s)$ and a virtual individual with exactly the same characteristics as the original one, except for the protected variable $s$ which has the opposite label $s'$ (*e.g.* the *Gender* variable is *Male* instead of *Female*), namely $f(x, s')$. Note that in case of multi-class labels, the outcome should be the most favourable decision for all possible labels. This classifier is fair by design in the sense that no matter their gender, the *testing procedure* can not detect a change in the individual prediction.

Nevertheless, this trick against testing cannot cheat usual evaluation of discrimination by using a disparate impact measure which is usual in the USA by measuring the impact on real and not fictitious recruitment. This is the reason why hiring tech companies add some facilities (Raghavan et al. (2020)) to mitigate ethnic bias of algorithmic hiring for avoiding an enterprise juridical complications. The evaluation of this strategy is evaluated using the same experimental protocol as in the previous sections and these are shown in Figure 6-(bottom).

As expected for previous results, this method has little impact on the classification

---

[6]https://www.observatoiredesdiscriminations.fr/testing

[7]https://dares.travail-emploi.gouv.fr/dares-etudes-et-statistiques/
etudes-et-syntheses/dares-analyses-dares-indicateurs-dares-resultats/testing

[8]http://www.ismcorum.org/

[9]https://www.novethic.fr/lexique/detail/testing.html

errors and the disparate impacts. This emphasises the conclusion of Section 4.1 claiming that the *Gender* variable is captured by other variables. Removing the effect of a sensitive variable can therefore require more advanced treatments than those described above.

# 5 Differential treatment for fair decision rules

## 5.1 Strategies

As we have seen previously, bias may induce discrimination of an automatic decision rule. Although many complex methods have been developed to tackle this problem, we investigate in this section the effects of two easy and maybe naive modifications of machine learning algorithms. We present in this section the effect of two alternative strategies to build fair classifiers. They have in common the idea of considering different treatments according to each group $S = \{0, 1\}$. These strategies are the following :

1. **Building a different classifier for each class of the sensitive variable:** This strategy consists in training the same prediction model with different parameters for each class of the sensitive variable. We denote *separate treatment* this strategy.

2. **Using a specific threshold for each class of the sensitive variable:** Here, a single classifier is trained for all data to produce a score. The binary prediction is however get using a specific threshold for each sub-group $S = 0$ or $S = 1$. Note that when the score is obtained by estimating the conditional distribution $\eta(x) = P(Y = 1|X = x)$ then the threshold used is often 0.5. Here this threshold is made $S$-dependent and is adapted to avoid any possible discrimination. In practice, we keep a threshold of 0.5 for the observations in the group $S = 1$ but we adapt the corresponding threshold for the observations in the group $S = 0$. In our tests, we automatically set this threshold on the training set so that the disparate impact is close to 0.8 in the cases where it was originally lower to this socially accepted threshold. The classifier and the potentially adapted threshold are then used for further predictions. This corresponds in a certain way to favor the minority class by changing equality to equity. We denote this strategy as *positive discrimination* since this procedure corresponds to this purpose.

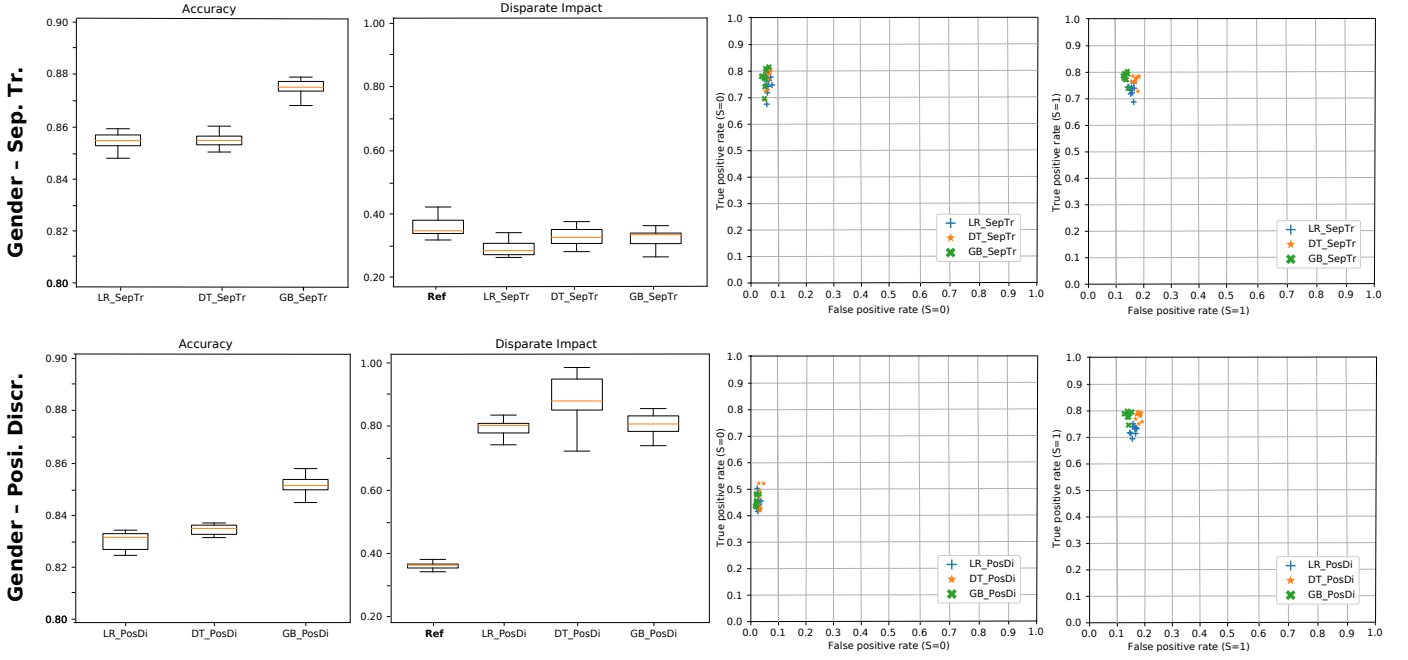## 5.2 Results obtained using the *Separate Treatment* strategy



Figure 7: Performance of the machine learning models LR, DT and GB when **(top)** using a *Separate Treatment* for the groups $S = 0$ and $S = 1$, and **(bottom)** when using a *Positive Discrimination* strategy for the groups $S = 0$.

Splitting the model parameters into parameters adapted to each group reduces the bias of the predictions when compared to the initial model, but it does not remove it. As we can see in Figure 7-*(top)*, where the notations are analogous to those in the above figures, it improved the disparate impact in all cases for relatively stable prediction accuracies. Note that the improvements are more spectacular for the basic Logistic Regression and Decision Tree models than for the Gradient Boosting model. This last model is indeed particularly efficient to capture fine high order relations between the variables, which gives less influence to the strong non-linearity generated when splitting the machine learning model into two class-specific models. Hence building different models reduces but does not solve the problem, the level of discrimination in the decisions being only slightly closer to the level of bias in the initial dataset.

## 5.3   Results obtained using the *Positive Discrimination* strategy

Results obtained using the *positive discrimination* strategy are shown in Figure 7-*(bottom)*. They clearly emphasize the spectacular effect of this strategy on the disparate impacts, which can be controlled by the data scientist. By adjusting the threshold, it is possible to adjust the levels of discriminations in the dataset, as in this example where the socially acceptable level of 0.8 can be reached. In this case we see a decrease in the performance of the classifier, but yet being reasonable.

These results should however be tempered for a main reason. Although the average error receives little changes, the number of false positive cases of women is clearly increased when introducing positive discrimination. In our tests more than half of the predictions that should have been false in the group $S = 0$ are even true. These false positive decisions have a limited impact on the average prediction accuracy as they where obtained in the group $S = 0$ which has less observations than $S = 1$ and that there are clearly less true predictions with $Y = 1$ than $Y = 0$. Yet, many definitions of fairness such as *equality of odds* or *equality of opportunity* (see e.g. Hardt et al. (2016)) are based on false and true positive errors, and thus this increase may be very harmful for a decision maker considering such a criteria. Moreover, from a legal point of view, the fact that this strategy favors the minority class by looking for equity rather than equality may be judged as unfair or rises political issues that are far beyond the scope of this paper.

## 6   Conclusions

In this paper, we provided a case-study of the use of machine learning technics for the prediction of the well-known *Adult Income* dataset. We focused on a specific fairness criterion, the statistical parity, which is measured through the Disparate Impact. This metric quantifies the difference of the behaviour of a classification rule applied for two subgroups of the population, the minority and the majority. Fairness is achieved when the algorithm behaves in the same way for both groups, hence when the sensitive variable does not play a significant role in the prediction. Main results are summarized in Figure 8.

In particular, we convey the following take-home messages: **(1)** Bias in the training data may lead to machine learning algorithms taking unfair decisions, but not always. While, in this widely studied data set, there was a clear increase of bias using the tested

machine learning algorithms with respect to the *Gender* variable, the *Ethnic Origin* did not lead in this case to a severe bias. For this reason, as mentioned above in section 3.2, we have focused in the rest of the paper on the protected variable *Gender*." **(2)** As always in Statistics, computing a mere measure is not enough but confidence intervals are needed to determine the variability of such indexes. Hence, we proposed an ad-hoc construction of confidence intervals for the Disparate Impact. **(3)** Standard regulations that promote either the removal of the sensitive variable or the use of testing techniques appeared as irrelevant when dealing with fairness as Statistical Parity of machine learning algorithms.

Note also that different notions of fairness (local and global) are at stake here. We first point out that testing methods focus on individual fairness while statistical methods such as the Disparate Impact Analysis tackle the issue of group fairness. These two notions if related to the similar notion of discrimination with respect to an algorithmic decision are yet different. In this work, we showed that an algorithm can be designed to be individually fair while still presenting a strong discrimination with respect to the minority group. This is mainly due to the fact that testing methods are unable to detect the discrimination hidden in the algorithmic decisions that are due to the training on an unbalanced sample. Testing methods detect discrimination if individuals with the same characteristics but different sensitive variables are treated in a different way. This corresponds to trying to find counterfactual explanation to an individual with a different sensitive variable. This notion of counterfactual explanations to detect unfairness has been developed in Kusner et al. (2017). Yet the testing method fails in finding a counterfactual individual since it is not enough to change only the sensitive variable but a good candidate should be the closest individual with a different sensitive variable but with the variables that evolve depending on $S$. For this, following some recent work on fairness with optimal transport theory as in Gordaliza et al. (2019) developing an idea from Feldman et al. (2015), some authors propose a new way of testing discrimination by computing such new counterfactual models in Black et al. (2020). Finally, we tested two a priori naive solutions consisting either in building different models for each group or in choosing different rules for each group. Only the latter that can be considered as *positive discrimination* proves helpful in obtaining a fair classification. Note that if some errors are increased (false positive rate), this method has a good generalization error. Yet in other cases, the loss of efficiency could be greater and this method may lead to unfair treatment.

This data set has been extensively studied in the literature on fairness in machine

learning and we are well aware of the numerous solutions that have been proposed to solve this issue. Even with standard methods, it is possible for a data scientist, when confronted to fairness in machine learning, to design algorithms that have very different behaviors and yet achieving a good classification error rate. Some algorithms hamper discrimination in the society while others just maintain its level, and some others correct this discrimination and provide gender equity. It is worth noting that the most explainable algorithms, such as the logistic regression, do not protect from discrimination. On the contrary, the capture of gender bias is inmediate due to its simplicity, while more complex algorithms might be more protected from this spurious correlation or, since the variable is discrete, better said spurious dependency.

The choice of a model should not be driven only by its performance with respect to a generalization error but should also be explainable in terms of bias propagation. For this, measures of fairness should be included in the evaluation of the model. In this work, we only considered statistical parity type fairness but many other definitions are available, without any consensus on the better choice for such a definition neither from a mathematical or a legal point of view. A strong research effort in data science is hence the key for a better use of Artificial Intelligence type algorithms. This will allow data scientists to describe precisely the algorithmic designing process, as well as their behaviour, in terms of precision and propagation of bias.

In closing, note that helpfulness of complex algorithms is due to their ability to find unintended bias and hidden correlations in large data sets. Hence algorithmic bias removal should be handled carefully because there is a part of such information crucial for the performance of the machine learning model, while the other part is potentially dangerous and may lead to prejudicial outcomes. Therefore, explainability should not be understood in terms of explainability of the whole algorithm, but maybe one line of future research in machine learning should focus on explainability of the inner bias of an algorithm, or its explainability with respect to some legal regulations.

# References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Barocas, S. and A. D. Selbst (2016). Big data's disparate impact. *Calif. L. Rev. 104*, 671.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0049124118782533.

Biddle, D. (2006). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing.* Gower Publishing, Ltd.

Black, E., S. Yeom, and M. Fredrikson (2020). Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, New York, NY, USA, pp. 111–121. Association for Computing Machinery.

Buolamwini, J. and T. Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*(2), 153–163.

Cramer, J. S. (2002). The origins of logistic regression.

Del Barrio, E., P. Gordaliza, and J.-M. Loubes (2019). A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA 8*(4), 817–849.

Dieterich, W., C. Mendoza, and T. Brennan (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap.* CRC press.

Feldman, M.and Friedler, S. A., J. Moeller, C. Scheidegger, and S. Venkatasubramanian (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM*

*SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM.

Flores, A. W., K. Bechtel, and C. Lowenkamp (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation 80*, 38.

Friedler, S. A., C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth (2018, February). A comparative study of fairness-enhancing interventions in machine learning. *ArXiv e-prints*.

Gordaliza, P., E. Del Barrio, F. Gamboa, and J.-M. Loubes (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pp. 2357–2365.

Hardt, M., E. Price, and N. Srebro (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323.

Hébert-Johnson, U., M. P. Kim, O. Reingold, and G. N. Rothblum (2018). Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948.

Hurley, M. and J. Adebayo (2016). Credit scoring in the era of big data. *Yale JL & Tech. 18*, 148.

Kamiran, F., T. Calders, and M. Pechenizkiy (2010, Dec). Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pp. 869–874.

Kearns, M., S. Neel, A. Roth, and Z. S. Wu (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076.

Lahoti, P., K. P. Gummadi, and G. Weikum (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1334–1345. IEEE.

L'Horty, Y., M. Bunel, S. Mbaye, P. Petit, and L. du Parquet. Discriminations dans l'accès à la banque et à l'assurance.

Mercat-Bruns, M. (2016). *Discrimination at Work*. University of California Press.

Mitchell, T. M. et al. (1997). Machine learning.

Morris, S. B. and R. E. Lobsenz (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology 53*(1), 89–111.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Pedreschi, D., S. Ruggieri, and F. Turini (2012). A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 126–131. ACM.

Raghavan, M., S. Barocas, J. Kleinberg, and K. Levy (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, New York, NY, USA, pp. 469–481. Association for Computing Machinery.

Riach, P. A. and J. Rich (2002). Field experiments of discrimination in the market place. *The economic journal 112*(483), F480–F518.

Romei, A. and S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review 29*(5), 582?638.

Rothmann, R., J. Krieger-Lamina, and W. Peissl (2014, 07). Credit scoring in Österreich.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics 24*, 303–329.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3. Cambridge university press.

Winrow, B. and C. Schieber (2009). The disparity between disparate treatment and disparate impact: An analysis of the ricci case. *Academy of Legal, Ethical and Regulatory Issues*, 27.

Zafar, M. B., I. Valera, M. Gomez Rodriguez, and K. P. Gummadi (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee.

# A   Appendix

## A.1   The Adult Income dataset

### A.1.1   Data preparation

As discussed in the introduction of Section 2, the study has started with a detailed pre-processing of the raw data to give a more clear interpretation to further analyses. First, we noticed that the variable `fnlwgt` (Final sampling weight) has not a very clear meaning so it has been removed. For a complete description of such variable access the link `http://web.cs.wpi.edu/~cs4341/C00/Projects/fnlwgt`. We have also performed a basic and multidimensional exploration (MFCA) in order to represent the possible sources of bias in the data in `https://github.com/wikistat/Fair-ML-4-Ethical-AI/blob/master/AdultCensus/AdultCensus-R-biasDetection.ipynb`.

This exploration led to a deep cleaning of the data set and highlighted difficulties present on certain variables, raising the need to transform some of them before fitting any statistical model. In particular, we have deleted missing data, errors or inconsistencies; grouped together certain highly dispersed categories and eliminated strong redundancies between certain variables. This phase is notoriously different from the strategy followed by Friedler et al. (2018) who analyze raw data directly. Some of these main changes are listed below:

- Variable 3 `fnlwgt` is removed since it has little significance for this analysis.

- The binary variable `child` is created to indicate the presence or absence of children.

- Variable 8 `relationship` is removed since it is redundant with `gender` and `mariStat`.

Table 2: The Adult Income dataset

| N° | Label | Possible values |
|---|---|---|
| 1 | Age | Real |
| 2 | workClass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Withoutpay, Never-worked |
| 3 | fnlwgt | Real |
| 4 | education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| 5 | educNum | integer |
| 6 | mariStat | Married-civ-spouse, Divorced, Nevermarried, Separated, Widowed, Marriedspouse- absent, Married-AF-spouse |
| 7 | occup | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transportmoving, Priv-house-serv, Protective-serv, Armed-Forces |
| 8 | relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| 9 | origEthn | White, Asian-Pac-Islander, Amer-Indian- Eskimo, Other, Black |
| 10 | gender | Female, Male |
| 11 | capitalGain | Real |
| 12 | capitalLoss | Real |
| 13 | hoursWeek | Real |
| 14 | nativCountry | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying- US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad and Tobago, Peru, Hong, Holand- Netherlands |
| 15 | income | $> 50k$, $\leq 50k$ |

- Variable 14 `nativCountry` is removed since it is redundant with variable `origEthn`.

- Variable 9 `origEthn` is transformed into a binary variable: CaucYes vs. CaucNo.

- Varible 4 `education` is removed as redundant with variable `educNum`.

- Additionally clean-up the $< 50K$, $\leq 50K$, $> 50K$ and $\geq 50K$ in variable "Target"

## A.2 Testing lack of fairness and confidence intervals

Let $\left(X_i, S_i, \hat{Y}_i = g(X_i)\right), i = 1, \ldots, n$, be a random sample of independent and equally distributed variables. The previous criterion can be consistently estimated by its empirical version. Yet the value of the criterion may depend on the data sample. Due to the importance of obtaining an accurate proof of unfairness in a decision rule it is important to obtain confidence intervals in order to control the error of detecting unfairness. In the literature it is often achieved by computing the mean over several sampling of the data. We provide in the following the exact asymptotic behaviors of the estimates in order to build confidence intervals.

**Theorem A.1 (Asymptotic behavior of the Disparate Impact estimator)** *Set the empirical estimator of DI(g) as*

$$T_n := \frac{\sum_{i=1}^{n} \mathbb{1}_{g(X_i)=1} \mathbb{1}_{S_i=0} \sum_{i=1}^{n} \mathbb{1}_{S_i=1}}{\sum_{i=1}^{n} \mathbb{1}_{g(X_i)=1} \mathbb{1}_{S_i=1} \sum_{i=1}^{n} \mathbb{1}_{S_i=0}}.$$

*Then the asymptotic distribution of this quantity is given by*

$$\frac{\sqrt{n}}{\sigma}\left(T_n - DI(g, X, S)\right) \xrightarrow{d} N(0, 1), \ as \ n \to \infty, \tag{A.1}$$

*where $\sigma = \sqrt{\nabla \varphi^T \left(\mathbb{E} Z_1\right) \Sigma_4 \nabla \varphi \left(\mathbb{E} Z_1\right)}$ and*

$$\nabla \varphi^T \left(\mathbb{E} Z_1\right) = \left(\frac{\pi_1}{p_1 \pi_0}, -\frac{p_0 \pi_1}{p_1^2 \pi_0}, -\frac{p_0 \pi_1}{p_1 \pi_0^2}, \frac{p_0}{p_1 \pi_0}\right)$$

$$\Sigma_4 = \begin{pmatrix} p_0(1-p_0) & & & \\ -p_0 p_1 & p_1(1-p_1) & & \\ \pi_1 p_0 & -\pi_0 p_1 & \pi_0 \pi_1 & \\ -\pi_1 p_0 & \pi_0 p_1 & -\pi_0 \pi_1 & \pi_0 \pi_1 \end{pmatrix},$$

*where we have denoted $\pi_s = \mathbb{P}(S_1 = s)$ and $p_s = \mathbb{P}(g(X_1) = 1, S_i = s)$, $s = 0, 1,$ .*

**Proof:**

Consider for $i = 1, \ldots, n$, the random vectors

$$
Z_i = \begin{pmatrix} \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0} \\ \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1} \\ \mathbb{1}_{S_i=0} \\ \mathbb{1}_{S_i=1} \end{pmatrix},
$$

where $\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=s} \sim B(\mathbb{P}(g(X_i)=1, S_i=s))$ and $\mathbb{1}_{S_i=s} \sim B(\mathbb{P}(S_i=s))$, $s = 0, 1,$. Thus, $Z_i$ has expectation

$$
\mathbb{E}Z_i = \begin{pmatrix} \mathbb{P}(g(X_i)=1, S_i=0) \\ \mathbb{P}(g(X_i)=1, S_i=1) \\ \mathbb{P}(S_i=0) \\ \mathbb{P}(S_i=1) \end{pmatrix}.
$$

The elements of the covariance matrix $\Sigma_4$ of $Z_i$ are computed as follows:

$$
Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}, \mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}\right) = \mathbb{E}\left(\mathbb{1}^2_{g(X_i)=1}\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(g(X_i)=1, S_i=1)
$$

$$
\begin{aligned}
Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}, \mathbb{1}_{S_i=0}\right) &= \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}^2_{S_i=0}\right) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(S_i=0) \\
&= \mathbb{P}(g(X_i)=1)\mathbb{P}(S_i=0) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(S_i=0) \\
&= [1 - \mathbb{P}(S_i=0)]\,\mathbb{P}(g(X_i)=1, S_i=0)
\end{aligned}
$$

$$
Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}, \mathbb{1}_{S_i=1}\right) = \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(g(X_i)=1, S_i=0)\mathbb{P}(S_i=1)
$$

$$
Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}, \mathbb{1}_{S_i=0}\right) = \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(g(X_i)=1, S_i=1)\mathbb{P}(S_i=0)
$$

$$
\begin{aligned}
Cov\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}_{S_i=1}, \mathbb{1}_{S_i=1}\right) &= \mathbb{E}\left(\mathbb{1}_{g(X_i)=1}\mathbb{1}^2_{S_i=1}\right) - \mathbb{P}(S_i=1)\mathbb{P}(g(X_i)=1, S_i=1) \\
&= \mathbb{P}(g(X_i)=1, S_i=1) - \mathbb{P}(S_i=1)\mathbb{P}(g(X_i)=1, S_i=1) \\
&= \mathbb{P}(g(X_i)=1, S_i=1)\,[1 - \mathbb{P}(S_i=1)] \\
&= \mathbb{P}(g(X_i)=1, S_i=1)\mathbb{P}(S_i=0)
\end{aligned}
$$

and finally,

$$
Cov(\mathbb{1}_{S_i=0}, \mathbb{1}_{S_i=1}) = \mathbb{E}\left(\mathbb{1}_{S_i=0}\mathbb{1}_{S_i=1}\right) - \mathbb{P}(S_i=0)\mathbb{P}(S_i=1) = -\mathbb{P}(S_i=0)\mathbb{P}(S_i=1).
$$

From the Central Limit Theorem in dimension 4, we have that

$$\sqrt{n}\left(\bar{Z}_n - \mathbb{E}Z_1\right) \xrightarrow{d} N_4\left(\mathbf{0}, \Sigma_4\right), \ as \ n \to \infty.$$

Now consider the function

$$\varphi: \qquad \mathbb{R}^4 \qquad \longrightarrow \qquad \mathbb{R}$$
$$(x_1, x_2, x_3, x_4) \quad \longmapsto \quad \frac{x_1 x_4}{x_2 x_3}$$

Applying the Delta-Method (see in Van der Vaart (1998)) for the function $\varphi$, we conclude that

$$\sqrt{n}\left(\varphi(\bar{Z}_n) - \varphi(\mathbb{E}Z_1)\right) \xrightarrow{d} \nabla\varphi^T\left(\mathbb{E}Z_1\right) N_4\left(\mathbf{0}, \Sigma_4\right), \ as \ n \to \infty,$$

where $\varphi(\bar{Z}_n) = T_n$, $\varphi(\mathbb{E}Z_1) = DI(g, X, S)$.  $\square$

Hence, we can provide a confidence interval when estimating the disparate impact over a data set. Actually $\left(T_n \pm \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}\right)$ is a confidence interval for the parameter $DI(g, X, S)$ asymptotically of level $1 - \alpha$.

Previous theorem can be used to test the presence of disparate impact at a given level.

$$H_{0,\beta}: \ DI(g, X, S) \leqslant \beta \quad vs. \quad H_{1,\beta}: \ DI(g, X, S) > \beta \tag{A.2}$$

aims at checking if $g$ has Disparate Impact at level $\beta$. We want to check whether $DI(g, X, S) \leq \beta$. Under $H_0$, the inequality $T_n - \beta \leqslant T_n - DI(g, X, S)$ holds, and so

$$\frac{\sqrt{n}}{\sigma}(T_n - \beta) \leqslant \frac{\sqrt{n}}{\sigma}(T_n - DI(g, X, S)).$$

Finally, from the inequality above and Eq. (A.1), we have that

$$\mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}(T_n - \beta) < Z_{1-\alpha}\right) \geqslant \mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}(T_n - DI(g, X, S)) < Z_{1-\alpha}\right) \longrightarrow 1 - \alpha,$$

as $n \to \infty$ and, equivalently,

$$\mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}(T_n - \beta) \geqslant Z_{1-\alpha}\right) \leqslant \mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}(T_n - DI(g, X, S)) \geqslant Z_{1-\alpha}\right) \longrightarrow \alpha,$$

as $n \to \infty$, where $Z_{1-\alpha}$ is the $(1-\alpha)$-quantile of $N(0,1)$. In conclusion, the test rejects $H_0$ at level $\alpha$ when

$$\mathbb{P}_{H_0}\left(\frac{\sqrt{n}}{\sigma}(T_n - \beta) \geqslant Z_{1-\alpha}\right) \geqslant \alpha.$$

When dealing with Equality of Odds, we want to study the asymptotic behavior of the estimators of the True Positive and True Negative rates across both groups. The reasoning is similar for the two rates, so we will only show the convergence of the True Positive rate estimator, denoted in the following by $TP(g)$.

**Theorem A.2** *Set the following estimate of the True Positive rate of a classifier g:*

$$R_n := \frac{\sum_{i=1}^n \mathbb{1}_{g(X_i)=1} \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=0} \sum_{i=1}^n \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=1}}{\sum_{i=1}^n \mathbb{1}_{g(X_i)=1} \mathbb{1}_{g(X_i)=1} \mathbb{1}_{S_i=1} \sum_{i=1}^n \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=0}}.$$

*Then, the asymptotic distribution of this quantity is given by*

$$\frac{\sqrt{n}}{\sigma} \left(R_n - TP(g)\right) \xrightarrow{d} N(0,1), \ \text{as } n \to \infty, \tag{A.3}$$

*where* $\sigma = \sqrt{\nabla \varphi^T \left(\mathbb{E}Z_1\right) \Sigma_4 \nabla \varphi \left(\mathbb{E}Z_1\right)}$ *and*

$$\nabla \varphi^T \left(\mathbb{E}Z_1\right) = \left(\frac{r_1}{p_1 r_0}, -\frac{p_0 r_1}{p_1^2 r_0}, -\frac{p_0 r_1}{p_1 r_0^2}, \frac{p_0}{p_1 r_0}\right)$$

$$\Sigma_4 = \begin{pmatrix} p_0(1-p_0) & & & \\ -p_0 r_1 & p_1(1-p_1) & & \\ p_0(1-r_0) & -p_1 r_0 & r_0(1-r_0) & \\ p_0 r_1 & p_1(1-r_1) & -r_0 r_1 & r_1(1-r_1) \end{pmatrix},$$

*where we have denoted* $p_s = \mathbb{P}(g(X_1) = 1, Y_1 = 1, S_1 = s)$, *and* $r_s = \mathbb{P}(Y_1 = 1, S_1 = s)$, *for* $s = 0, 1$.

Proof of Theorem A.2 The proof follows the same guidelines of previous proof. We set here

$$Z_i = \begin{pmatrix} \mathbb{1}_{g(X_i)=1} \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=0} \\ \mathbb{1}_{g(X_i)=1} \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=1} \\ \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=0} \\ \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=1} \end{pmatrix},$$

where $\mathbb{1}_{g(X_i)=1} \mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=s} \sim B(\mathbb{P}(g(X_i) = 1, Y_i = 1, S_i = s))$ and $\mathbb{1}_{Y_i=1} \mathbb{1}_{S_i=s} \sim B(\mathbb{P}(Y_i = 1, S_i = s))$, $s = 0, 1,$. From the Central Limit Theorem, we have that

$$\sqrt{n} \left(\bar{Z}_n - \mathbb{E}Z_1\right) \xrightarrow{d} N_4 \left(\mathbf{0}, \Sigma_4\right), \ \text{as } n \to \infty.$$

with

$$\Sigma_4 = \begin{pmatrix} p_0(1-p_0) & & & \\ -p_0 r_1 & p_1(1-p_1) & & \\ p_0(1-r_0) & -p_1 r_0 & r_0(1-r_0) & \\ p_0 r_1 & p_1(1-r_1) & -r_0 r_1 & r_1(1-r_1) \end{pmatrix}. \tag{A.4}$$

Now consider the function

$$\varphi : \qquad \mathbb{R}^4 \qquad \longrightarrow \quad \mathbb{R}$$
$$(x_1, x_2, x_3, x_4) \quad \longmapsto \quad \frac{x_1 x_4}{x_2 x_3}$$

Applying the Delta-Method for the function $\varphi$, we conclude that

$$\sqrt{n} \left( \varphi(\bar{Z}_n) - \varphi(\mathbb{E}Z_1) \right) \xrightarrow{d} \nabla \varphi^T (\mathbb{E}Z_1) \, N_4 \left( \mathbf{0}, \Sigma_4 \right), \ as \ n \to \infty,$$

where $\varphi(\bar{Z}_n) = R_n$, and $\varphi(\mathbb{E}Z_1) = TP(g)$. $\qquad \square$

## A.3 Bootstraping vs. Direct Calculation of IC interval

The estimation of the Disparate Impact is unstable. In this paper we promote the use of the theoretical confidence interval based on the well known Delta method to control its variability. Contrary to Morris and Lobsenz (2000), it does not rely on Gaussian approximation. We compare the stability of this confidence interval to bootstrap simulations, see for instance in Efron and Tibshirani (1994) for more details on bootstrap methods.

For this we build 1000 bootstrap replicates and estimate the disparate impact. Figure 9 presents the simulations. We can see that the bootstrap simulations remain in the confidence interval. Moreover, if we build a confidence interval for the bootstrap estimator, the confidence intervals are the same. We obtain by the theoretical confidence interval $[0.349, 0.384]$ while the bootstrap's confidence interval is $[0.349, 0.385]$.

Hence the theoretical confidence is a reliable measure of fairness for the data set and should be preferred due to its small computation time compared to the 1000 bootstrap replication.

Note that in this paper, for sake of clarity, we have chosen to focus only on the disparate impact criterion. Yet all other fairness criteria should be given with the calculation of a confidence interval. For instance in Del Barrio et al. (2019) we propose confidence intervals for Wasserstein distance which is used in many methods in fair learning.

## A.4 Application to other real datasets

To illustrate these tests we have also considered another two well-known and real data sets.

1. **German Credit data.** This data set is often claimed to exhibit some origin discrimination in the success of being given a credit by the German bank. Hence we

compute the disparate impact w.r.t Origin. We obtain

$$DI = 0.77 \in [0.68, 0.87].$$

Hence here confidence intervals play an important role. Actually the disparate impact is not statistically significantly lower than 0.8, which entails that the discrimination of the decision rule of the German bank can not be shown, which promotes the use of a proper confidence interval.

2. **COMPAS Recidivism data**. A third data set is composed by the data of the controversial COMPAS score detailed in Dieterich et al. (2016). The data is composed of 7214 offenders with personal variables observed over two years. A score predicts their level of dangerosity which determines whether they can be released while a variable points out if there has been recidivism. Hence Recidivism of offenders is predicted using a score and confronted to possible racial discrimination which corresponds to the protected attribute. The protected variable separates the population into caucasian and non caucasian. To evaluate the level of discrimination we first compute the disparate impact with respect to the true variable and the COMPAS score seen as a predictor.

$$DI = 0.76 \in [.72, .81]; \quad DI(\text{COMPAS}) = 0.71 \in [0.68; 0.74].$$

In both cases, the data are biased but the level of discrimination is low. Yet as mentioned in al the studies on this data set, the level of errors of prediction is significantly different according to the ethnic origin of the defender. Actually the conditional accuracy scores and their corresponding confidence intervals show clearly the unbalance treatment received by both populations.

$$TPR = 0.6 \in [0.54, 0.65]$$

$$TNR = 3.38 \in [2.46, 4.3]$$

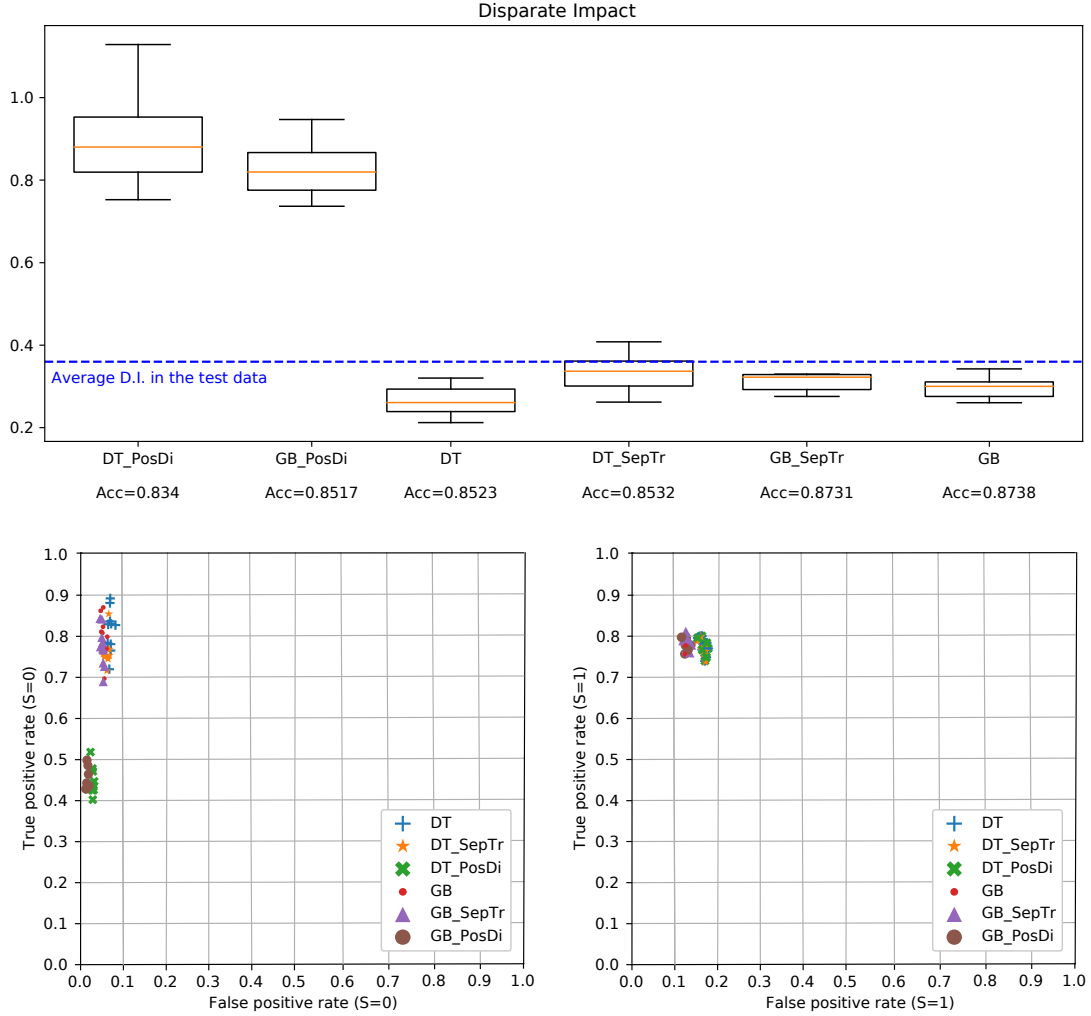This unbalanced treatment is clearly assessed with the confidence interval.

Figure 8: Summary of the main results: The best performing algorithms of Sections 3 and 5 are compared here. (**top**) Boxplots of the disparate impacts from the least accurate method on the left, to the most accurate method on the right, and (**bottom**) corresponding true positive and true negative rates in the groups $S = 0$ and $S = 1$. For $S = 0$, there is a better balance between the true and the false positive decisions when the results are closer to the dashed blue line.
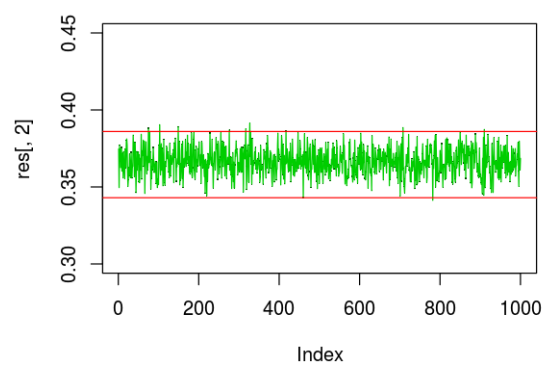
Figure 9: Comparison with bootstrap computations