# Fairness In Classification / Regression Models

LARA MARZINZIK and ANILCAN POLAT

The field of Machine Learning (ML) and Artificial Intelligence (AI) is a ubiquitous topic in modern society. Today, many decision processes are aided by AI. Classification and regression models are utilized to make crucial decisions like deciding on reoffending rates of criminals and therefore determining individual's just chances for revision [1].

It is crucial for such models to not only make accurate decision, but also to make fair predictions. However, sensitive attributes like race, gender, sexuality and social background can influence such decisions in an unfair way Yet more and more incidents of unfairly made decisions by ML models occur. Decision aid models often learn from data sets containing past decisions. Such data includes biases that humans reflected through their decisions throughout history. Those biases are adapted and can lead to unfair predictions. Those biases need to be found and erased so that a classifier's decisions do not discriminate against minority groups. This report tackles the issue of fairness in ML models by giving an overview of different approaches to measuring fairness in a mathematical context. We then focus on existing work trying to improve fairness by reducing statistical disparity, a commonly used notion of fairness. In doing so, it becomes apparent that this issue is a complicated one to solve. Especially regarding a trade-off between accuracy and fairness, that each model needs to face.

**ACM Reference Format:**
Lara Marzinzik and Anilcan Polat. 2024. Fairness In Classification / Regression Models. 1, 1 (March 2024), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

In the realm of Artificial Intelligence and Machine Learning, the quest for fairness is entangled with challenges stemming from biases in data collection, algorithm design, and the interaction between users and these systems. Such biases significantly influence the outcomes of predictive models, impacting decisions in sectors like criminal justice and hiring. An illustrative example is observed in recidivism prediction tools, where biases in historical arrest data as proxies for future criminal behavior disproportionately target minority groups, leading to unfairly biased algorithmic predictions against these communities. [1]

Biases manifest in various forms, affecting model outcomes in distinct ways. Issues such as how data is measured, collected, and represented can lead to skewed decisions by algorithms. A notable instance of algorithmic bias was seen in a hiring algorithm by a major corporation, which inadvertently favored resumes with phrasing more typical of one gender over another, reflecting historical hiring biases. This case underscores the nuanced nature of algorithmic bias, where attempts to create neutral models can unintentionally perpetuate societal biases. [4]

Fairness within machine learning models is assessed through several metrics, including demographic parity, equality of opportunity, and equalized odds. However, the application of these metrics underscores the complexity of achieving fairness, as they can sometimes be in conflict with each other, indicating that no single measure fully addresses the nuances of fairness in every scenario.

Authors' address: Lara Marzinzik, lara.marzinzik@tu-dortmund.de; Anilcan Polat, anilcan.polat@tu-dortmund.de.


XXXX-XXXX/2024/3-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

To tackle these challenges, a nuanced understanding of both the technical and ethical dimensions of machine learning is required. It involves continuous efforts to identify and mitigate biases at every stage of model development, from data collection to algorithm design. By acknowledging and addressing these issues, the goal is to advance towards more equitable and fair machine learning applications.

## 2 THE SOURCE OF UNFAIRNESS

Before diving deeper into the issue of measuring and improving fairness in classification and regression models, we specify the root of unfairness and motivate the issue by looking at the diversity of fairness itself.

### 2.1 Fairness perceived by society

Fairness itself is a social construct and therefore complicated to define. As a social construct, fairness is a subjective matter, and what some people perceive as fair may be deemed unfair by others. Fairness is entangled with justice and influenced by moral beliefs, culture and religion. As such a diverse field, it is the focus of many philosophical works and proves to be difficult to define [16].

In some cases, however, it is important to find definitions that can be applied to evaluate the fairness of a decision or a setting. This is for example the case in sports, especially in competitions, with Fair Play being an important aspect of a good game. A competition's rules are designed to protect fairness and thereby establish an environment of equal opportunities [9].

### 2.2 Unfairness in Machine Learning

The cases given in the Introduction show that unfair algorithms are a common occurrence in ML models designed to make crucial decisions. The first step toward decimating this unfairness is to locate the root of unfairness in ML models.

It is believed that unfair predictions are caused by human biases that influenced decisions and laws throughout history. Such biases are adapted by ML models in their learning phase and cause them to disadvantage certain groups of people [11]. Such disadvantaged groups are commonly referred to as *minority groups*. The people in those groups all share at least one attribute that is being treated unfairly. Some examples for such attributes are race, gender, sexuality, social background. In classification models, those attributes are a part of the model's feature set and are referred to as the *sensitive attributes*, commonly displayed as the set $A$.

Another occurrence caused by biases in the training data is the so-called *feedback-loop phenomenon*. Feedback loops are a feature to AI models and applications. They can occur int the training phase or when used repeatedly. When a classification model is trained on a data set containing unfair biases, the feedback loop amplifies those biases and leads to unfair predictions. Even models that are deemed fair at some point can pick up on biases and grow unfair through the feedback loop [5, 11].

In their work, Kamishima et al. [8] describe the source of unfairness in classification and regression models in more detail. They identify three causes of unfairness being 1) *prejudice*, 2) *underestimation*, and 3) *negative legacy*. These causes can occur alone but also together. Prejudice itself can again be divided into three subcategories: 1.1) *direct prejudice*, 1.2) *indirect prejudice*, and 1.3) *latent prejudice*, all of which are different forms of statistical dependence. The direct type 1.1) describes a statistical dependence between a sensitive attribute and the generated output. The result is direct discrimination in the ML model's output data set. Indirect prejudice 1.2) describes a dependence between the sensitive attribute and the variable that shall be decided by the model. This dependence indirectly influences the model's output. Latent prejudice 1.3) describes a dependence between a sensitive attribute and another non-sensitive attribute. This does not directly result in

an unfair prediction but might violate regulations, which can also be seen as a form of unfairness. Underestimation entails a too small set of training data. As a model can only be trained on finite sets, the learning progress is also limited. This can lead to a model making more unfair decisions than included in the training set. The last source of unfair predictions, negative legacy, describes unfair sampling or labeling. Historical biases, that were described above, fall under the category of such unfair labeling. Unfair sampling, on the other hand, occurs if minority groups are underrepresented in the training data set.

## 3 APPROACHES TO MEASURING FAIRNESS IN MACHINE LEARNING

As fairness is not only a technical but also a sociological issue, there are multiple approaches to measuring and improving fairness. Some of those approaches are introduced in this section.

In the context of Machine Learning, most measurements take a statistical approach. Those can mainly be described using three main criteria: *independence* (1), *separation* (2), and *sufficiency* (3). Introduced by Barocas et al. [2] and modified by Steinberg et al. [15] as follows:

$$S \perp A \Leftrightarrow P(S, A) = P(S) \cdot P(A) \tag{1}$$

$$S \perp A \mid Y \Leftrightarrow P(S, A \mid Y) = P(S \mid Y) \cdot P(A \mid Y) \tag{2}$$

$$Y \perp A \mid S \Leftrightarrow P(Y, A \mid S) = P(Y \mid S) \cdot P(A \mid S) \tag{3}$$

In this statistical description, $S = f(X)$ is a prediction covered by the prediction function $f : \mathcal{X} \rightarrow \mathcal{S}$. $\mathcal{X}$ is a feature set and $\mathcal{A} \subseteq \mathcal{X}$ is a set containing sensitive attributes. $\mathcal{Y}$ is the target set for the predictions, and $\mathcal{S} = \mathcal{Y}$. $S$, $A$ and $X$ are drawn from the distribution over $\mathcal{Y} \times \mathcal{X}$. $S$, $A$ and $Y$ are simplified as random variables.

Independence (1) promises that the resulting prediction is not influenced by the actual value of the sensitive attribute $A$. Separation (2) and sufficiency (3) are both described as conditional independence. In the scenario for sufficiency, the target variable $Y$ divides the probability distribution into subgroups. In the context of an application process, one such group could be a group of all applicants who are qualified for the position. Separation aims to equalize the acceptance and rejection rate within each of those groups. Following the example, the rejection rate within the group of qualified applicants should be equal for all values of the sensitive attribute $A$ [13]. Sufficiency on the other hand aims to equalize rejection and acceptance rates for groups divided by the prediction $S$. In our example, that entails that the outcome should only be influenced by the qualification of each applicant. In other words, if two applicants with a different value for $A$ and the same qualification should lead to the same prediction $S$. This measure is often already fulfilled before taking any action to ensure it, as an accurately trained model always aims to fulfill this measure [2, 13].

The most commonly used fairness measure is statistical disparity ($SD$), or statistical parity contrariwise. Statistical disparity is easily computed by subtracting the possibility of a prediction $p_a$ from the possibility of a prediction $p_b$, so that

$$SD = p_a - p_b \tag{4}$$

In the case of our application example, $p_b$ could be the possibility of a female applicant being accepted and $p_a$ being the possibility of a male applicant being accepted. In this case, the sensitive attribute $A$ is gender. A large statistical disparity implies an unfair bias against female applicants. Contrariwise, a large negative disparity implies an unfair bias against male applicants. A perfectly fair model would have to achieve a statistical disparity score of 0 [3]. Some attempts to fulfilling this condition are discussed in Section 4.

Recently, new approaches to measuring fairness have been made that question the general definition of fairness, which in the context of ML prediction tasks is often defined as equality. It is

proposed that equity, which would entail equalizing out the unfair biases of the past, is a better measure of fairness altogether [11]. This thesis and how to measure equity is further explained in Section 5.

## 4 IMPROVING FAIRNESS USING STATISTICAL DISPARITY

As mentioned in Section 3, statistical disparity is one of the most commonly used fairness measures for ML models. Therefore, it is only to be expected that many approaches to maximize fairness using this measure exist. Some of those approaches are presented in this section.

### 4.1 Logistic Regression

Logistic regression, a staple for binary classification tasks, encounters significant fairness challenges when sensitive attributes are involved. Zafar et al.'s seminal work introduces fairness constraints into logistic regression models to mitigate bias and ensure equitable outcomes across groups defined by sensitive attributes [18].

Central to their approach is the concept of decision boundary fairness, quantified by the covariance between sensitive attributes and the signed distance from feature vectors to the decision boundary. The fairness metric is mathematically defined as:

$$\mathrm{Cov}(z, d_\theta(x)) = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i) \tag{5}$$

where $z_i$ are the sensitive attributes, $d_\theta(x_i)$ the signed distance to the decision boundary, $\bar{z}$ the average of the sensitive attributes, and $N$ the number of instances. This formulation aims to ensure the decision boundary does not unfairly favor or disadvantage any group based on sensitive attributes.

The authors propose two optimization strategies: maximizing accuracy under fairness constraints and maximizing fairness while maintaining acceptable accuracy. The fairness constraint in the optimization problem ensures that the decision boundary's absolute covariance does not exceed a predefined threshold, striking a balance between fairness and model performance.

For maximizing fairness under accuracy constraints, the goal is to minimize the absolute decision boundary covariance, subject to constraints on the classifier's loss function:

$$\text{minimize} \left|\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)\right| \text{ subject to } L(\theta) \leq (1+\delta)L(\theta^*) \tag{6}$$

where $L(\theta^*)$ is the optimal loss for the training set by the unconstrained classifier, and $\delta$ is the allowable increase in loss for the sake of fairness.

### 4.2 Fair Logistic Regression

Fair Logistic Regression (Fair LR) is a specialization of logistic regression. The approach proposed by Kamishima et al. [8] focuses on reducing indirect prejudice. As described in section 2.2, indirect prejudice is the result of a sensitive attribute influencing the output variable. The general idea of this approach is to couple any logistic regression model with a *prejudice remover*, which is implemented as a regularizer. This prejudice remover targets data points where sensitive attribute have a large influence on a certain output variable. The influence of the attributes is then lessened by the regularizer.

### 4.3 Ensemble Learning

Another approach suggests using ensemble models, also refereed to as committee machines, to reduce statistical disparity. Such models consist of multiple singular models of the same type, in our case multiple classification or regression models. The most common methods for creating such a model are *boosting* and *bagging* [14]. Freund and Schapire's AdaBoost Framework [6] is the most frequently used implementation of a boosting algorithm. Originally those models were developed to improve accuracy, however Grgić-Hlača et al. [7] hinted that such models can also be used to increase fairness, implying that bagging multiple unfair models may lead to a fair model if the unfair biases are cancelled out.

*4.3.1 AdaBoost And Bagging.* Bagging and boosting are generally developed and proven to increase ML model's accuracy. Many different methods of boosting exist, the first boosting algorithm is based on boosting by filtering. This approach relies on big data sets for training, as some are bound to be filtered out.

AdaBoost is a follow up approach that requires fewer data to produce satisfactory results and therefore can be applied in more scenarios. AdaBoost combines different models that in itself might not be very accurate. The resulting ensemble model weights every model's mapping function, that maps the input to the output. This is mathematically described as:

$$f(x) = \sum_{t=1}^{m} \alpha_t f_t(x) \tag{7}$$

with $x$ being the model's input, $f(x)$ the ensemble model's mapping function and $f_t(x)$ each model's mapping function. By applying a lower weight to functions producing inaccurate predictions and higher weights to more accurate functions, the model balances out the inaccurate predictions for each input $x$ [3, 14].

*4.3.2 k-NN Situation Testing.* The underlying method relies on *situation testing*. Situation testing looks at a pair of people that share similar characteristics except for a certain sensitive attribute. This pair is put in the same situation and evaluated separately. For example, two people with the same qualifications apply for the same job. The only thing differentiating those two, is one attribute, like their gender or skin color. Then the outcome of the situation is inspected. The comparison of the outcomes of the situation for each person can indicate discrimination against the sensitive attribute. This technique can be modeled by a *k nearest neighbor* (K-NN) classifier. This classifier views tuples of instances and their labels and compares them to their k nearest neighbors. In Luong et al's adaption, an instance is classified as discriminated against, and therefore unfairly labeled, if three conditions hold: (1) the neighborhood shows large statistical disparity, (2) the instance contains a sensitive attribute, and (3) the label is interpreted as a disadvantage [3, 10].

*4.3.3 Specialized Ensemble Strategy To Improve Fairness.* Based on Grgić-Hlača et al's work, Bhaskaruni et al. [3] propose a new ensemble strategy that adapts the AdaBoost Framework. While the AdaBoost Framework was originally created to improve accuracy by eliminating false predictions, this adaption aims to use AdaBoost's algorithm to eliminate unfair predictions. The unfair labels are identified using a simplification of Luong et al's k-NN based testing technique [10].

To provide a better understanding of Bhaskaruni et al's adaptation of AdaBoost and the k-NN based situation testing, this paragraph describes the used terminology: Each instance is represented by a triple $(x, s, y)$ where $x$ is a non-sensitive attribute, $s$ is a sensitive attribute and $y$ is the instance's label. A training set $L$ consists of $n$ triples, such that $L = \{(x_i, s_i, y_i)\}_{i=1,\dots,n}$. $x$ is the model's input and the label $y$ is its output. The output for each model of the ensemble is generated by mapping

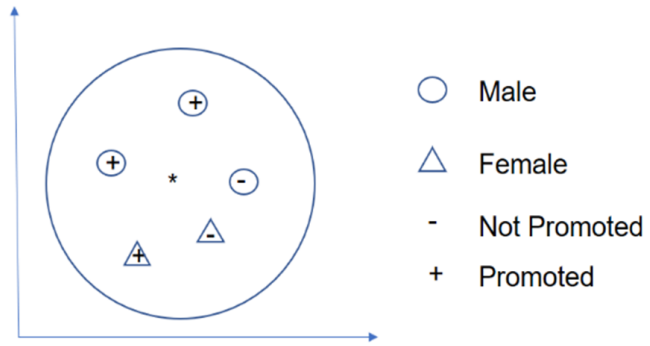


Fig. 1. Example for a k-NN fairness testing scenario. Here the Statistical Dependency SD results to 0.17. Depending on the threshold $r$, the instance is deemed fair or unfair. If $r$ is set to 0.1, the instances prediction is labeled unfair. Bhaskaruni et al. [3]

the input $(x, s)$ to the output $y$ using the corresponding mapping function $f_t$. In an ensemble with $m$ single models $t = 1, ..., m$. Those models are all variants of logistic regression models, with the mapping functions explained below.

To simplify the processes, in the mathematical description, the input is reduced to $x$, as the case with $(x, s)$ can be interpreted as just a specialization without loss of generality.

Bhaskaruni et al. only consider the first condition in their adaptation of the method to detect unfair labels. Another difference is that instead of basing the decision on true labels, the detection algorithm operates on the predicted labels of each model in the ensemble. If the statistical disparity in the neighborhood of an instance $*$ is greater than a fixed threshold $r$, the instance's label $y_i$ is considered unfair. An example is given in figure 1.

The adaption of the k-NN method relies on a definition for statistical disparity. As mentioned in section 3, statistical disparity is measured by subtracting the possibility of an event for the inspected minority group from the possibility of the same event for someone who does not belong to this minority group. Here, $SD$ needs to be defined in a set of the k nearest neighbors $N_{i,k}$. For the context of promotions with the focus on the sensitive attribute gender, $SD$ is defined as follows:

$$SD(f_t; N_{i,k}) = Pr\{x \text{ is promoted } | x \text{ is male }, x \in N_{i,k}\} - Pr\{x \text{ is promoted } | x \text{ is female }, x \in N_{i,k}\} \tag{8}$$

As mentioned earlier, a high statistical disparity reflects a higher level of unfairness, therefore the prediction for $x_i$ is deemed unfairly predicted if $SD$ is bigger than a threshold $r$.

Adapting AdaBoost's algorithm, the mapping function for the ensemble model is formed by building a weighted sum of each model's mapping function $f_t(x)$ for each input $x$ as described in (7). The model weight $\alpha_t$ from (7) is defined as

$$\alpha_t = ln\{\frac{1 - \epsilon_t}{\epsilon_t}\} \tag{9}$$

with

$$\epsilon_t = \frac{\sum_{i=1}^{n} w_i^{(t)} \delta_i^{(t)}}{\sum_{i=1}^{n} w_i^{(t)}} \tag{10}$$

with $w_i^{(t)}$ being the *instance weight* and $\delta_i^{(t)}$ being the indicator to an unfair label, as described below. In this approach, the base models are trained subsequently. Each mapping function is described as a minimized weighted loss:

$$f_t = \mathbf{arg}\min_h \sum_{i=1}^{n} w_i^{(t)} \cdot \text{loss}(h(x_i), yi) \tag{11}$$

with loss(·) being a loss function and $w_i^{(t)}$, as mentioned, being the assigned weight to the input $x_i$., called the *instance weight*. The instance weights for the first model $w_i^{(1)}$ are all set to one. The function $\delta_i^{(t)}$ is used to indicate unfairly predicted labels for each input, as follows:

$$\delta_i^{(t)} = \begin{cases} 1, & x_i \text{ unfairly predicted by } f_t \Leftrightarrow SD(f_t; N_{i,k}) > r \\ 0, & \text{otherwise} \Leftrightarrow SD(f_t; N_{i,k}) \leq r \end{cases} \tag{12}$$

In other words, if the k-NN method described above determines a mapping to be unfair, $\delta_i^{(t)}$ is assigned the value one. As $\delta_i^{(t)} = 0$ means the prediction is deemed fair, a fair mapping also entails a smaller $\epsilon_t$ and therefore a larger model weight $\alpha_t$ (9), (10). Subsequently, a mapping $f_t(x)$ receives a larger weight if it is a fairer prediction, resulting in the fair predictions weighting more in the ensemble model. Ultimately, this leads to a single model's fair prediction to influence the ensemble's prediction more, leading it to an overall fairer prediction. The instance weight plays its part in ensuring a fair prediction for each possible input $x_i$ by assigning higher weights to prior unfairly predicted labels $y_i$. Therefore, if $f_t$ leads to an unfair prediction for $x_i$, the next model's mapping function $f_{t+1}$ is assigned a higher instance weight $w_i^{(t+1)}$:

$$w_i^{(t+1)} = w_i^{(t)} \cdot \exp\{\alpha_t \cdot \delta_i^{(t)}\} \tag{13}$$

Altogether, the instance weights $w_i^{(t)}$ ensure that prior unfairly detected instances are focused on to even out the biases they suffered. Then, after each single model is trained, the model weights $\alpha_t$ determine the influence each single model has on the predictions of the ensemble model. Meanwhile, the threshold $r$ plays a key role in determining which predictions contain unfair biases and need to be evened out.

### 4.4 Comparison Of Different Methods

The priory described method claims to achieve better results than singular models like logistic regression but also standard ensemble models like AdaBoost and Bagging alone. Bhaskaruni et al. [3] conducted an experiment comparing standard Logistic Regression (LG), Fair Logistic regression (FairLG) as proposed by Kamishima et al. [8], Bagging, AdaBoost, and their proposed method. All models were measured using three metrics: 1) statistical disparity, 2) equalized odds, which is a variant of statistical disparity, and 3) classifier error, which entails accuracy. All models were applied to two existing data sets which are known to enable unfair model predictions [12, 17]. Table 1 and Table 2 show the results of the experiment, all data is taken from Bhaskaruni et al. [3].

The results are similar on both data sets. The new proposed method shows improved statistical disparity and equalized odds. The error rate on the other is second highest in the crime data set (Table 1) and highest in the credit card data set (Table 2). This hints to the trade-off between accuracy and fairness, which will be later discussed in section 6.

## 5 IMPROVING FAIRNESS BY EQUITY

The so far viewed and discussed methods all try to minimize statistical disparity or a variation of sorts. The definition of statistical disparity used in those methods relies on the concept of *equality*.

| Method | SD | EO | Error |
|---|---|---|---|
| LR | .1000 ± .0000 | .4695 ± .0000 | .1883 ± .0000 |
| FairLR | .0898 ± .0971 | .0620 ± .0882 | .1166 ± .0189 |
| Bagging | .2267 ± .2025 | .2855 ± .1867 | .1187 ± .0987 |
| AdaBoost | .0746 ± .0124 | .3712 ± .0889 | .1013 ± .0117 |
| Proposed Method | .0293 ± .0247 | .0593 ± .0367 | .1604 ± .0445 |

Table 1. Results using the Community Crime data set

| Method | SD | EO | Error |
|---|---|---|---|
| LR | .1531 ± .0000 | .1161 ± .0000 | .3438 ± .0000 |
| FairLR | .0779 ± .0571 | .1256 ± .0112 | .2412 ± .0469 |
| Bagging | .1915 ± .1766 | .1267 ± .1024 | .3066 ± .0277 |
| AdaBoost | .1697 ± .0335 | .1104 ± .0625 | .2877 ± .0238 |
| Proposed Method | .0213 ± .0171 | .0019 ± .0000 | .4486 ± .0589 |

Table 2. Results using the Credit Data Default set

The overall goal is to establish an environment of equal chances and possibilities. Still, equality is not believed to be the best way to measure fairness by all. Mehrabi et al. [11] claim that the concept of *equity* is a better notion of fairness and suggest going beyond equality to design new algorithms and unfairness detection tools. The proposed strategy also considers the feedback loop, which is described in section 2.2.

To utilize the definition of equity, it first needs to be translated to a mathematical model. Mehrabi et al. [11] propose the following definition: The set $Y$ contains the possible outcomes of an ML model's prediction, and the set $A$ contains the values of a sensitive attribute. A classifier is trained with the set $D$ of all decisions made in the past. Those decisions are statistically represented through the joint distribution $p_D(Y, A)$. The classifier that shall be trained to produce predictions fulfilling the equity conditions operates on a set of instances $M$. The joint distribution $p_M(Y, A)$ entails the statistical characteristics of the classifier model's predictions.

The fairness of the classifiers decisions is measured in groups for each sensitive attribute $A$, if each decision outcome $y$ for a group $A$ fulfills the following criterion:

$$\begin{gathered} p_D(Y = y|A = a)p_D(A = a) + p_M(Y = y|A = a)p_M(A = a) \\ = \\ p_D(Y = y|A = b)p_D(A = b) + p_M(Y = y|A = b)p_M(A = b) \end{gathered} \tag{14}$$

This equation shows the definition of equity: in the case that a value of a sensitive attribute suffered discrimination in the past, it is accounted for in the model's decision-making process. In other words, if a certain value for a sensitive attribute was disadvantaged in the past, it now is an assigned a higher possibility in the model's joint distribution.

To ensure that such disadvantages are correctly detected, it is important to train the classifier with the same amount of data for each vale of the group $A$, such that

$$p_D(Y = y|A = a) + p_M(Y = y|A = a) = p_D(Y = y|A = b) + p_M(Y = y|A = b) \tag{15}$$

Otherwise, the equity definition could detect an underrepresented value of group $A$ to be discriminated against, even if it is not, as the level of representation influences the joint possibility.

Besides the fairness objection, Mehrabi et al. [11] also follow an accuracy objective. Both objectives are represented through a loss term, and the combination of both is regulated by a term $\beta$. The fairness objective is covered by $F_{equity}$, with

$$\begin{gathered} F_{equity}(\theta) = \\ \sum_y ([\ \tfrac{1}{n}\sum_{i=1}^{n} p_M(Y_i = y|A = a) + p_D(Y = y|A = a)\ ] - \\ [\ \tfrac{1}{n}\sum_{i=1}^{n} p_M(Y_i = y|A = b) + p_D(Y = y|A = b)\ ])^2 \end{gathered} \tag{16}$$

The accuracy objective is covered by a loss term $L(\theta)$, which can be any desired loss function. In the following experiment results, $L(\theta)$ is defined as a cross-entropy loss, which is also known as log-loss.

The two terms are combined by the regularization term $\beta$ as follows:

$$\min_{\theta} \beta F_{equity}(\theta) + (1 - \beta)L(\theta) \tag{17}$$

By choosing a larger $\beta$, the fairness constraint grows stronger and is enforced more than the accuracy objective, and vice versa.

For performance testing purposes, another classifier using statistical disparity as a notion of fairness is designed. Instead of using equity loss, it uses statistical parity loss. The loss function is similar to the definition of statistical disparity from section 3 and section 4.3. In the context of the proposed method, the statistical parity loss is described as :

$$F_{parity}(\theta) = \sum_y (\frac{1}{n}\sum_{i=1}^{n} p_M(Y_i = y|A = a) - \frac{1}{n}\sum_{i=1}^{n} p_M(Y_i = y|A = b)) \tag{18}$$

The equity based classifier, parity based classifier and a third classifier ignoring the fairness objective, are compared regarding fairness and accuracy. For that purpose a Fairness Gain is defined as:

$$FG = [|p(Y = y|A = a) - p(Y = y|A = b)|]_{classifier} - [|p(Y = y|A = a) - p(Y = y|A = b)|]_l \tag{19}$$

with a loss function $l \in$ *Equity, Parity, Classifier*, and the classifier being a simple classifier with no fairness constraints. The three classifiers are tested on two different data sets, COMPAS and Adult. Both data sets can be found on GitHub. [1] The results are displayed in figure 2. It is clear to see, that the equity notion archives higher fairness gain but also suffers the most accuracy wise.

Another goal of the equity based classifier is to target the feedback loop and minimize its negative effect biases. The feedback loop is simulated by iterating over a set of training data from the same data sets as before. The experiment consists of 10 iterations performed on 10 random subsets of the overall data set. The results are visualized in figure 3. Again, the equity based classifier archives the best score regarding bias reduction. The results from COMPAS data set also shows, that a classifier without any fairness constraints is prone to biases growing through the feedback loop.

## 6 DISCUSSION

Many approaches to measuring and improving fairness in ML models exist. While some argue about the best way to maximize fairness using a certain measurement, like statistical disparity, some question which approach to measuring is the best. It is easy to see that the topic of fairness is not an easy one. The definition alone leaves potential for discussion. This comes t no surprise as it is a topic rooted in the sociological field. Before improving fairness, it is crucial to find a fitting definition, but what one individual deems fair can be seen as unfair by another. While many agree that *equality*, and therefore equal chances for all, is a good notion of fairness, some go further and

[1] https://ashryaagr.github.io/Fairness.jl/dev/datasets/

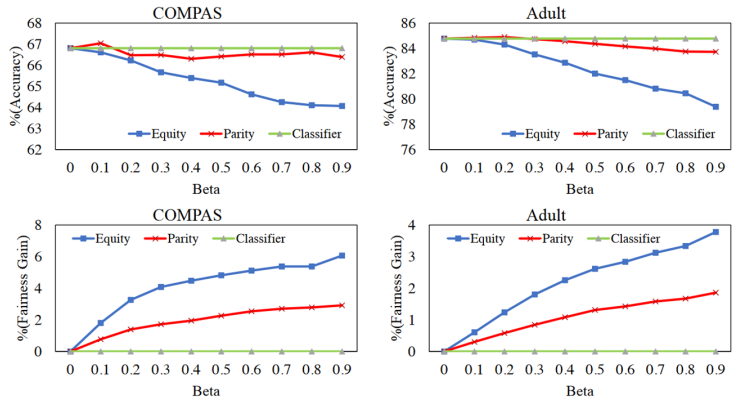


Fig. 2. Results of the experiment. On the left side is the COMPAS data set and on the right the Adult data set. The results are displayed for different values of $\beta$. Mehrabi et al. [11].

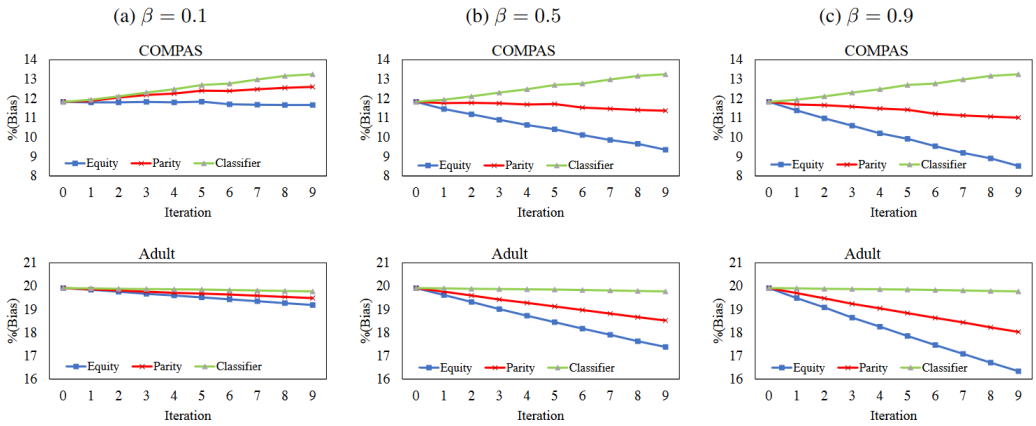


Fig. 3. The results of the Feedback Loop experiment are displayed for 3 different values for $\beta$. The y-axis show the level of biases while the x-axis displays the iterations. Mehrabi et al. [11].

claim that *equity* is a more fitting description of fairness. But not everyone agrees with that notion. In fact, the opinion of the broad majority seems to shift in different scenarios, as Mehrabi et al. [11] show with a questionnaire. The results are shown in figure 4.

Furthermore not many papers regarding equity as a fairness notion exist, which shows the preference of the equality term. However, equity based models are a promising approach as they succeed in diminishing the feedback loop's effect on biases more than other approaches.

One of the most commonly used measurements is statistical disparity. It is a rather simple approach, still it is heavily discussed. Logistic Regression is easily utilized to improve fairness but compared to other methods it shows significantly worse results [3, 18]. Other approaches suggest using ensemble models with a fitting ensemble strategy and achieve better results. However, finding the right ensemble strategy proves to be more difficult. While Bhaskaruni et al's approach [3] from section 4.3 looks promising, it also leads to more issues that need to be solved and parameters that need to be optimized. The neighborhood size k influences the statistical disparity. Too small

| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 |
|---|---|---|---|---|
| Equity | **134** | **115** | 59 | 44 |
| Parity | 16 | 35 | **91** | **106** |

Fig. 4. Results of Mehrabi et al's [11] questionnaire regarding the preference of Equity or Equality as a fair base for decisions. The results show that depending on the scenario, the opinions shift toward one or the other.

k and too big k both can lead to unfairness not being detected, as too broad neighborhoods can include data that is not representative and therefore weaken the effect of the sensitive attribute that could be measured in a smaller neighborhood. Too small k on the other hand can lead to under-representation of important decisions and hide the effect of the sensitive attribute [3].

Another issue appears more commonly. Whenever a model is trained to fulfill fairness constraints or unlearn biases, a trade-off with the model's accuracy is ubiquitous in all models. For models using an unfairness detection tool, the threshold $r$ determines the ratio between fairness and accuracy. Higher thresholds lead to some unfair biases not being detected, while too small thresholds cause accuracy to diminish. Using loss functions, it is easy to adjust the level of fairness and accuracy. But it is not easy to determine the perfect ratio, if it even exists. Accuracy and fairness are both crucial concepts to trustworthy machine interactions, but to this day cannot be fulfilled at the same time.

## 7 CONCLUSIONS

Having viewed the different approaches to measuring fairness and improving the resulting notions of fairness, it becomes apparent that the topic of fairness in classification and regression models is a complex issue. There exists no perfect notion of fairness and no perfect approach to maximizing it. It is important to keep the trade-off between accuracy and fairness, which is bound to occur, in mind.

While most work acknowledges the existence of this trade off, it is rarely the focus of the work. Still, it is a crucial factor in ensuring reliable and satisfactory results. Future work could focus on maximizing both fairness and accuracy. Furthermore it is notable that, while more approaches to explaining the source of unfairness exist, most work focuses on historical biases only. In order to diminish unfairness it might be of more importance to focus on the sources of the issue and try to regulate those.

## REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[3] Dheeraj Bhaskaruni, Hui Hu, and Chao Lan. 2019. Improving prediction fairness via model ensemble. In *2019 IEEE 31st International conference on tools with artificial intelligence (ICTAI)*. IEEE, 1810–1814.

[4] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters, October* 10 (2018).

[5] Laurence Devillers, Françoise Fogelman-Soulié, and Ricardo Baeza-Yates. 2021. AI & Human Values: Inequalities, Biases, Fairness, Nudge, and Feedback Loops. *Reflections on Artificial Intelligence for Humanity* (2021), 76–89.
[6] Yoav Freund and Robert E. Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*, Paul Vitányi (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 23–37.
[7] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2017. On fairness, diversity and randomness in algorithmic decision making. *arXiv preprint arXiv:1706.10208* (2017).
[8] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.
[9] Sigmund Loland. 2013. *Fair play in sport: A moral norm system*. Routledge.
[10] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 502–510.
[11] Ninareh Mehrabi, Yuzhong Huang, and Fred Morstatter. 2020. Statistical equity: A fairness classification objective. *arXiv preprint arXiv:2005.07293* (2020).
[12] Michael Redmond and Alok Baveja. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141, 3 (2002), 660–678.
[13] Rahul Shekhar. 2020. An Introduction to Fairness in Machine Learning. *Medium* (2020).
[14] Dimitri P Solomatine and Durga L Shrestha. 2004. AdaBoost. RT: a boosting algorithm for regression problems. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, Vol. 2. IEEE, 1163–1168.
[15] Daniel Steinberg, Alistair Reid, and Simon O'Callaghan. 2020. Fairness measures for regression via probabilistic classification. *arXiv preprint arXiv:2001.06089* (2020).
[16] Tom Tyler, Robert J Boeckmann, Heather J Smith, and Yuen J Huo. 2019. *Social justice in a diverse society*. Routledge.
[17] I-Cheng Yeh and Che-hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications* 36, 2 (2009), 2473–2480.
[18] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 962–970. https://proceedings.mlr.press/v54/zafar17a.html

Annotation:
Parts produced by Anilcan : 1 Introduction, 4.1 Logistic Regression;
Parts produced by Lara: rest