Item Response Theory In Sample Reweighting To Build Fairer Classifiers

Diego Minatel, Nícolas Roque dos Santos, Vinícius Ferreira da Silva, Mariana Cúri, and Alneu de Andrade Lopes

Institute of Mathematics and Computer Science
University of São Paulo, São Carlos, Brazil
{dminatel, nrsantos, vfsilva}@usp.br, {mcuri, alneu}@icmc.usp.br

Abstract. Currently, one of the biggest challenges of Machine Learning (ML) is to develop fairer models that do not propagate prejudices, stereotypes, social inequalities, and other types of discrimination in their decisions. Before ML faced the problem of unfair decision-making, the field of educational testing developed several mathematical tools to decrease bias in selections made by tests. Thus, the Item Response Theory is one of these main tools, and its great power of evaluation helps make fairer selections. Therefore, in this paper, we use the concepts of Item Response Theory to propose a novel sample reweighting method named IRT-SR. The IRT-SR method aims to assign weights to the most important instances to minimize discriminatory effects in binary classification tasks. According to our results, IRT-SR guides classification algorithms to fit fairer models, improving the main group fairness notions such as demographic parity, equal opportunity, and equalized odds without significant performance loss.

Keywords: Data Bias · Fairness · IRT · Machine Learning · Preprocessing Algorithm.

1 Introduction

Machine Learning (ML) algorithms significantly influence consequential decisions in various domains, including credit transactions, advertising targeting, credit assessment, translation, and content recommendation [26]. As these algorithms possess the power to shape people's lives, it becomes imperative to acknowledge the accompanying responsibilities. It is crucial to exercise caution and ensure that these models do not perpetuate societal biases and discrimination that already exist within our society [23, 24].

One notable example of discrimination arising from learning models is exemplified by the utilization of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system to predict recidivism risk and aid judges in determining sentences within the criminal justice system of the United States. This case gained attention due to the identification of potential biases and adverse impacts on certain racial and socioeconomic groups, particularly

against black defendants [2]. These biases resulted in a troubling outcome where low-risk black defendants were twice as likely to be misclassified as high-risk compared to their white counterparts, consequently depriving them of parole rights.

In this context, one of the biggest challenges in the field of ML is to develop fairer decision-making models, especially when these decisions involve people's futures [22]. An inherent challenge in this pursuit arises from the nature of data itself, as they often serve as a reflection of societal realities [3]. Thus, prejudices, stereotypes, and inequalities present in society are often contained in the data. As models are data-driven, training them on biased datasets unintentionally reproduces undesirable behaviors.

One way to minimize the biases contained in the data is through the preprocessing algorithms, which, when developed for this purpose, aim to transform the data set used in training, thereby incorporating some fairness notions or eliminating explicit discriminatory biases [25]. One of these preprocessing strategies is the sample reweighting method, the focus of this paper, which aims to assign weights to the instances used in model training. This helps to determine which instances are more important to be classified correctly to minimize the discriminatory effects of these models. Consequently, these weights support the classification algorithms' search for fairer solutions.

Before ML faced the problem of unfair decision-making, the educational testing field developed solutions to decrease bias in the applicant selection process. One of these solutions is the Item Response Theory (IRT), which has a great power of evaluation and is a fairer tool for evaluating tests than the Classical Test Theory [19]. IRT is used as an assessment model in some of the world's leading exams, such as the SAT[1], TOEFL[2], and ENEM[3]. Therefore, given its success in educational testing, IRT can be a promising path to developing fairer models [16].

In this scenario, we propose a novel sample reweighting method named IRT-SR based on Item Response Theory concepts to be applied in binary classification tasks. Our experimental results show that IRT-SR can improve key group fairness metrics, making classifiers fairer without significant performance loss. Complementarily, we highlight two main contributions of this paper. Firstly, it introduces Item Response Theory concepts into solutions to minimize discriminatory effects in machine learning models. Secondly, it introduces our IRT-SR sample reweighting method to guide classification algorithms to fit fairer models.

The remaining of this paper is divide as follows: Section 2 provides an overview of the background and related work about the topic. Section 3 presents our proposed methods and approaches. In Section 4, we detail the experimental settings and methodology employed to evaluate our proposed methods. Section

[1] The SAT is an educational exam given to high school students in the United States, which serves as a criterion for admission to American universities.

[2] TOEFL is the acronym for Test of English as a Foreign Language.

[3] ENEM is the exam that evaluates high school students in Brazil. The students use their ENEM scores to try to enter public and private universities in the country.

5 outlines the results obtained from our experiments and analyzes their implications. Finally, in Section 6, we conclude the paper by summarizing our findings and discussing their significance.

2 Background

This section presents the key terms and fundamental concepts of group fairness analysis and item response theory necessary to understand our proposal.

2.1 Group fairness analysis

Protected attributes are characteristics that hold sensitive information, like gender, race, nationality, religion, and sexual orientation. These attributes should receive equal treatment, regardless of their value. A *group* is a collection of individuals who share the same protected attributes, such as males and females in the case of gender. Moreover, a *privileged group* refers to a group or set of groups historically receiving better treatment than *unprivileged groups*.

One form of discrimination is to use protected attributes in decision-making. This practice is called *adverse treatment* and is typically forbidden by law in democratic countries. However, *adverse impact* occurs when certain groups are either advantaged or disadvantaged by outcomes, irrespective of whether adverse treatment is present or not [3]. In machine learning, adverse treatment arises when protected attributes are incorporated into model training, while adverse impact pertains to uneven results (*e.g.*, F1-score) across various groups.

Group fairness analysis aims to identify any potential unfair outcomes between different groups, with a particular focus on identifying adverse impacts. Three group fairness notions are discussed when we want to ensure that the adverse impact does not occur: demographic parity, equal opportunity, and equalized odds. *Demographic parity* means that every group has an equal chance of receiving a positive label [13]. *Equal opportunity* ensures that each group has an equal true positive rate [15]. Finally, *equalized odds* ensure all groups share the same true and false positive rates [15].

Therefore, group fairness analysis is determined by comparing group outcomes. This implies that any performance metric can be analyzed as group fairness. However, achieving equal rates is sometimes infeasible. Thus, we typically calculate the score ratio between privileged and unprivileged groups to determine any disparities in results based on these group fairness notions.

2.2 Item response theory

Item Response Theory (IRT) is a collection of mathematical models that are utilized in test evaluation, primarily in educational and psychometric applications. These models depict the relationship between the responses to test items and the abilities of the examinees, which enhances the assessment's effectiveness [10]. The IRT models stand out for their evaluation power and can be considered a

fairer form of evaluation since they are able to detect unwanted behaviors of an examinee, such as correct answers by guessing.

Dichotomous item response models are characterized by evaluating tests in which the correctness of the test questions is in the right and wrong format, regardless of the number of answer options. This form of evaluation resembles the evaluations of binary classifiers, where to calculate the metrics derived from the confusion matrix, the correct and incorrect classifications of the classifier are used. For this reason, we use dichotomous item response models in this work, specifically the two-parameter logistic model. Table 1 shows an example of the modeling of dichotomous items, where the data structure U represents a test with k items (columns) and n examinees (rows), where $U_{ij} = 1$ indicates that examinee i correctly answered question j, and $U_{ij} = 0$ indicates an incorrect answer.

Table 1: Data structure for modeling dichotomous items.

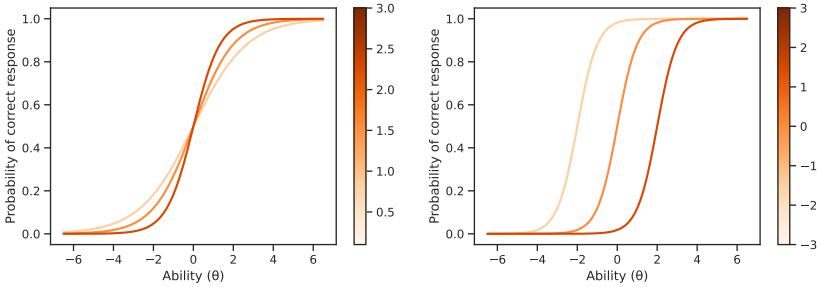
Individual	Item 1	Item 2	Item 3	$\cdots$	Item k
Examinee 1	1	1	0	$\cdots$	1
Examinee 2	0	0	0	$\cdots$	1
Examinee 3	0	1	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Examinee n	1	1	0	$\cdots$	1

The two-parameter logistic model (2PL) is formulated by Equation 1. If we have a data structure U that contains k items and n examinees, then $P(U_{ij} = 1 \mid \theta_i)$ represents the probability of examinee i answering item j correctly, which depends on their ability θ_i. The parameters a_j and b_j define the logistic curve associated with item j, known as the Item Characteristic Curve (ICC).

$$P(U_{ij} = 1 \mid \theta_i) = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}} \tag{1}$$

Fig. 1 shows examples of ICCs, where the y-axis represents the probability of correctly answering the item and the x-axis represents the ability θ. Fig. 1a shows the influence of parameter a, and it can be seen that parameter a acts directly on the ICC slope. Therefore, the parameter a is proportional to the derivative of the logistic curve at its inflection point. In contrast, Fig. 1b shows an example with three items, each with a different b-value. The b-value indicates the location of the ICCs on the ability scale, at which the probability of providing a correct answer is 50%. As the value of b increases, θ's ability to answer the item correctly increases. In general, ability values are commonly assumed to follow a normal distribution with an average of 0 and a standard deviation of 1. Therefore, the θ values typically fall within the range of -4 to $+4$, and b values fall between -2 and $+2$ [14].

In the context of ML, IRT has already been applied in classification tasks. In [20, 21], the authors find a strong correlation between the abilities (θ value)



(a) parameter a values (with $b = 0$) (b) parameter b values (with $a = 2$)

Fig. 1: Example of Item Characteristic Curves

of the classifiers and their accuracy; with this result, they suggest some applications for IRT in ML, such as model selection and classifier evaluation. Another application is [9], in which the authors proposed a new weighted voting in an ensemble of classifiers. Thus, the voting weight of a classifier is given by its θ values. These works use the same item modeling, with classifiers modeled as examinees and instances as items. However, this work inverts this modeling, modeling classifiers as items and instances as examinees, to make it possible to assign one θ value to each instance and thus formulate our sample reweighting method.

3 Proposal

This section presents our sample reweighting method, IRT Sample Reweighting (IRT-SR), based on Item Response Theory concepts. In order to transpose them to the domain of ML, we model a set of base classifiers as items and the sample as examinees. Then, we can define the weight of each instance by calibrating IRT parameters. The underlying idea of this method is that as the IRT is a fairer evaluation method, the weights defined using this mathematical model can contribute to developing fairer classifiers. Our proposed method comprises four stages, which we discuss in detail in the subsequent sections.

3.1 Stage 1: Base classifiers predictions

In the first stage of our method, we train a set of k base classifiers using the training sample. Next, we perform predictions on that same sample with the same k classifiers. This step is necessary as it allows us to model this set of predictions as a dichotomous test problem for using Item Response Theory, as described in Section 2.2.

As we model the set of classifiers as items in the next stage, the number k of base classifiers must satisfy the following condition $k > 2$ [10]. Thus, we need more than two items to estimate the parameters a and b of the ICCs. We selected

the following four classifiers from the k-Nearest Neighbors (kNN) classification algorithm: 1NN, 3NN, 5NN, and 7NN, which use 1, 3, 5, and 7 as the value of the hyperparameter of nearest neighbors, respectively. We opted for kNN in this stage due to its simplicity, performance, and ease of interpreting its decisions. Upon the completion of this stage, given a sample D_m with m instances, we have the following set of predictions $[\hat{Y}_{1NN}, \hat{Y}_{3NN}, \hat{Y}_{5NN}, \hat{Y}_{7NN}]$.

3.2 Stage 2: Item modeling

Since we want to include a weight for each instance of the training sample, we model in this stage the training sample as examinees and the set of base classifiers as items. Thus, we are able to associate one θ value for each instance, which can be used to determine the weight of the instance. Therefore, to transform the classifier predictions $[\hat{Y}_{1NN}, \hat{Y}_{3NN}, \hat{Y}_{5NN}, \hat{Y}_{7NN}]$ into items $[I_{1NN}, I_{3NN}, I_{5NN}, I_{7NN}]$, we have the modeling of items U, where $U_{ij} = 1$ indicates a correct prediction of classifier i in instance j and 0 otherwise.

Table 2 illustrates the functioning of the item modeling matrix U. Each entry U_{ij} indicates if a specific instance i is correctly classified by the classifier j. A value of 1 in the table indicates a correct classification, while a value of 0 indicates the opposite. For example, the item I_{7NN} shows that the 7NN correctly classified instances 1, 2, and 5 and misclassified instances 3 and 4.

Table 2: The classifier's predictions are modeled as a right or wrong test. The value of cell ij indicates correct (equal to 1) or incorrect (equal to 0) prediction of instance i by the classifier in column j.

Instances	I_{1NN}	I_{3NN}	I_{5NN}	I_{7NN}
Instance 1	1	1	0	1
Instance 2	0	0	0	1
Instance 3	0	1	1	0
Instance 4	0	1	0	0
Instance 5	1	1	0	1

3.3 Stage 3: IRT parameters calibration

In the third stage, we employ the 2PL model in the item modeling matrix U to estimate the θ values of the instances and the ICCs associated with the trained base classifiers. To calibrate these parameters, we utilize the expectation-maximization (EM) algorithm [5].

To better understand the behavior of an ICC, Fig. 2 presents an illustrative example with parameters $a = 2$ and $b = 0$ within the classification context. Notably, the classifier has a high probability of misclassifying instances with θ values lower than the value of its parameter b, as exemplified by the red vertical line. Conversely, instances with θ values greater than parameter b (indicated by the green vertical line) are more likely to be correctly classified by the classifier.

Moreover, instances with θ values equal to parameter b are characterized by the vertical gray line, signifying a 50% probability of being classified correctly.

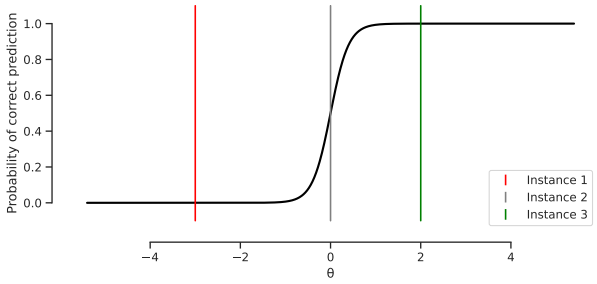


Fig. 2: Example of the Item Characteristic Curve, with $a = 2$ and $b = 0$, associated with a classifier. Note that instance 3 (green line) has a high probability of being correctly predicted by the classifier, while instance 1 (red line) has a high probability of being incorrectly predicted by the classifier. Finally, instance 2 (gray line) has a 50% probability of being predicted correctly.

3.4 Stage 4: Sample reweighting

Lastly, we can assign the sample weight with the estimated θ values for each instance. However, as seen in Fig. 2, the smaller the θ value of an instance, the more difficult it is to predict it correctly. Thus, these instances should have greater weight in the sample reweighting. Therefore, we rescaled the θ values to a range between 1 and 5. We use this new scale to maintain the tradeoff between improvement in group fairness measures without significant performance losses. Finally, the sample weight of IRT-SR is given by Equation 2, where $\theta_{rescaled}$ is the θ value translated to the new value scale.

$$Sample\,Weight = \frac{1}{\theta_{rescaled}} \tag{2}$$

4 Experimental settings

This experiment evaluates IRT-SR's capacity to aid the selected classification algorithms to fit fairer classifiers. The first step of the experiment is to separate the dataset into a training set (80%) and a test set (20%). We employ 5-fold cross-validation on the training set. This configuration was chosen because some selected datasets have few instances, as shown in Table 3. Moreover, we apply the selected sample reweighting methods for each training fold and use the sample weight generated in the selected classification algorithms. We also apply the

classification algorithms without using sample reweighting methods to have a benchmark for comparison.

At the end of the validation step, based on the demographic parity, equal opportunity, and equalized odds fairness measures, we select the best configuration for each type of classification algorithm of each sample reweighting method tested. Then we reapply the sample reweighting methods and retrain the classifiers with the entire training set and its best settings. Thus, in the end, we can compare the sample reweighting methods on the test set, verifying which method best guides the classification algorithms to develop fairer classifiers. An overview of the experiment we performed can be seen in Fig. 3.

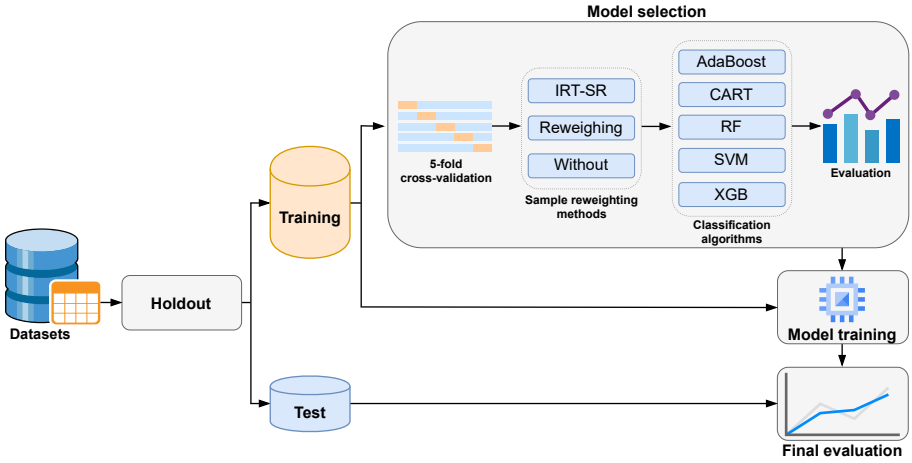


Fig. 3: Overview of the performed analysis. Initially, we split the dataset into training and testing sets. In sequence, we use 5-fold cross-validation and apply sample reweighting methods to each training fold to assess fairness metrics using five classification algorithms to identify the best hyperparameters for each classifier. Thus, we then trained the classifiers with the entire training set using the obtained hyperparameters. Finally, we assess the discriminatory effects of each classifier on the test set.

The source code and benchmark datasets utilized in the evaluation process are available in a public code repository[4]. The experiments were conducted using Python and R programming languages, with the help of the following libraries: scikit-learn (classification algorithms) [6], xgboost [8], aif360 (fairness metrics and Reweighing method) [4], and mirt (IRT calibration) [7]. In the remainder of this section, we detail the datasets, algorithms, and the evaluation approach used in the experiments.

[4] https://github.com/diegominatel/irt-sample-reweighting-method

4.1 Datasets

For this work, we selected the relevant binary classification benchmark datasets used in the Fairness in Machine Learning research community. Table 3 summarizes the datasets, showing their amount of instances (#I), number m of attributes (#A), which protected attributes are analyzed, the privileged group of each task related to the dataset, and reference.

Table 3: Dataset information. Here #I denotes amount of instances and #A represents the number of attributes

Dataset	#I	#A	Protected Attributes	Privileged Group	Ref.
Arrhythmia[6][7]	452	278	sex	male	[12]
Bank Marketing	45,211	42	age	over 25 years old	[12]
Census Income	48,842	76	race and sex	white-male	[12]
Contraceptive[6][8]	1,473	10	religion	non-islam	[12]
Crack[9]	1,885	11	race	non-white	[12]
German Credit	1,000	36	sex	male	[12]
Heart[6]	303	13	age	middle-aged	[12]
Heroin[9]	1,885	11	race	non-white	[12]
Recidivism[10] Female	1,395	176	race	white	[18]
Recidivism Male[10]	5,819	375	race	white	[18]
Student	480	46	sex	male	[1]
Titanic	1309	6	sex	female[11]	[27]

4.2 Algorithms

We used the following classification algorithms that allow the application of sample reweighting for the experiment: AdaBoost (ADA), Classification Trees (CART), Random Forest (RF), Support Vector Machines (SVM), and XGBoost (XGB). We tested fifteen parameterization settings for each of them. Table 4 shows each classification algorithm and the numerical variation range for their hyperparameters used in this experiment.

In addition to comparing the results without using sample reweighting, we used the well-known Reweighing [17] method in our experiment. This method

[6] Age and gender are protected attributes that can play a crucial role in predicting health datasets, which is why they are included in class prediction. Nevertheless, does not preclude the analysis of adverse impact.

[7] We binarize the output between the absence and presence of cardiac arrhythmia, ignoring the different arrhythmia groups.

[8] We binarize the output to predict whether or not a woman uses contraception.

[9] It is the Drug Consumption dataset just changing the target class.

[10] We split this dataset into two: Recidivism Female (female examples) and Recidivism Male (male examples).

[11] There was selection bias in the rescue operation during the Titanic disaster, as women and children were given priority. As a result, the protected attribute is utilized in making predictions using the Titanic dataset.

Table 4: Algorithms and ranges of numeric variation defined for their hyperparameters.

Algorithm	Hyperparameter	Variation Range (initial : final : step)
ADA	Number of trees	100 : 500 : 25
CART	Minimum number of samples to be a leaf node	2 : 30 : 2
RF	Number of trees	100 : 500 : 25
SVM	Gamma	0.0025 : 1.075 : 0.075
XGB	Number of trees	100 : 500 : 25

aims to enhance fairness in classification tasks by assigning appropriate sample weights W. Specifically, Reweighing computes:

$$W(A = i, Y = j) = \frac{P(A = i)P(Y = j)}{P(A = i, Y = j)}, \tag{3}$$

where $P(A = i)$ is the probability of occurrence of group i and $P(Y = j)$ is the probability of occurrence of class j. Additionally, $P(A = i, Y = j)$ is the probability of occurrence of group i with class j in dataset D. The underlying idea of this method is to assign greater weight to instances with less frequent `(group, class)` pairs.

4.3 Evaluation

As previously mentioned, we use the fairness measures of demographic parity, equal opportunity, and equalized odds to select the best hyperparameters of each classification algorithm. Also, we use these measures in the final evaluation to compare the effectiveness of the reweighting methods. Furthermore, due to the class imbalance of some datasets, we use the macro F1-score to analyze the performance of the selected classifiers.

To simplify the categorization process of fairness metrics, we use the highest score as the denominator for calculating the ratio between privileged and unprivileged groups of a specific fairness metric, as described in Section 2.1. As a result, the ratio of group fairness metrics, such as the demographic parity ratio, will always be in the interval $[0, 1]$, with the ideal outcome being a score of 1.

5 Results

In this section, we present the results obtained in our experiments, as explained in Section 4. We provide a summary of fairness metrics results on the test set for selected models with or without applying reweighting method. Additionally, we discuss the macro F1-score results.

Table 5 shows the average result of the demographic parity ratio on the test set applying demographic parity as a criterion for model selection. The "Without" column indicates the results without using the sample reweighting method.

Bold values indicate the best score by datasets, and the value in parentheses indicates the standard deviation. IRT-SR performed best in 7 of the 12 datasets and also had the best average demographic parity ratio, being more than 5% ahead of the second-best average. Reweighing performed better in 4 of the 12 datasets, while not using sample reweighting had better results in only one dataset.

Table 5: Average demographic parity ratio results on the test set.

Dataset	Without	IRT-SR	Reweighing
Arrhythmia	70.75% (19.80%)	**82.31% (12.91%)**	75.97% (15.09%)
Bank Marketing	50.57% (8.89%)	56.31% (16.51%)	**61.81% (5.73%)**
Census Income	31.83% (2.02%)	36.48% (4.01%)	**36.53% (3.45%)**
Contraceptive	94.73% (4.61%)	**94.83% (3.94%)**	92.01% (4.35%)
Crack	52.49% (39.93%)	**59.21% (34.13%)**	23.40% (21.47%)
German Credit	**83.61% (9.46%)**	83.41% (10.33%)	82.78% (10.78%)
Heart	53.40% (13.82%)	52.21% (5.28%)	**63.15% (16.83%)**
Heroin	56.55% (19.08%)	**77.04% (15.03%)**	46.23% (14.66%)
Recidivism Female	82.44% (22.03%)	**85.64% (5.31%)**	80.71% (17.37%)
Recidivism Male	72.36% (5.89%)	**83.59% (11.03%)**	76.39% (6.43%)
Student	86.82% (6.46%)	**88.01% (4.88%)**	86.82% (6.46%)
Titanic	19.59% (4.55%)	28.58% (12.15%)	**36.92% (19.69%)**
Average	62.93% (26.73%)	**68.97% (24.44%)**	63.56% (25.04%)

Table 6 shows the average equal opportunity ratio on the test set applying equal opportunity as a criterion for model selection. Reweighing performed best on 7 out of 12 datasets in this assessment criteria. However, IRT-SR had a better average performance and the best performance on 4 datasets. It is important to note that in the Crack dataset, not using sample reweighting methods obtained a result almost three times better than the selected methods.

Table 6: Average equal opportunity ratio results on the test set.

Dataset	Without	IRT-SR	Reweighing
Arrhythmia	81.17% (10.29%)	85.61% (6.42%)	**87.77% (3.48%)**
Bank Marketing	88.34% (11.40%)	**93.92% (6.42%)**	82.36% (9.08%)
Census Income	86.19% (1.89%)	90.25% (4.34%)	**91.08% (2.97)**
Contraceptive	93.35% (5.99%)	91.45% (5.39%)	**95.69% (5.02%)**
Crack	**59.80% (42.80%)**	21.08% (29.54%)	16.72% (37.40%)
German Credit	90.33% (5.83%)	90.86% (6.10%)	**91.91% (6.94%)**
Heart	78.72% (9.33%)	77.59% (3.12%)	**85.16% (5.76%)**
Heroin	9.80% (15.10%)	**50.65% (33.72%)**	26.14% (21.02%)
Recidivism Female	83.69% (22.09%)	**96.15% (3.19%)**	82.08% (23.17%)
Recidivism Male	78.62% (3.77%)	**88.56% (8.20%)**	80.89% (3.58%)
Student	98.32% (1.62%)	98.52% (1.72%)	**98.76% (0.94%)**
Titanic	51.52% (3.60%)	57.31% (19.88%)	**71.02% (21.85%)**
Average	74.99% (27.57%)	**78.50% (26.35%)**	75.80% (29.41%)

The last group fairness metric evaluated is equalized odds, shown in Table 7. Once again, IRT-SR had the best average performance, in addition to having the best performance in 7 of the 12 datasets. Reweighing obtained the best

performance of the five datasets. In contrast, not using reweighting method did not perform better in any of the datasets. We highlight that the IRT-SR obtained better average results for equalized odds in all datasets when compared to not using the reweighting method.

Table 7: Average equalized odds ratio results on the test set.

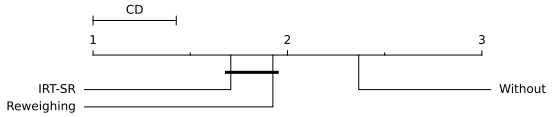
Dataset	Without	IRT-SR	Reweighing
Arrhythmia	67.67% (16.05%)	**78.35% (9.29%)**	74.87% (16.33%)
Bank Marketing	65.53% (4.13%)	72.68% (6.09%)	**79.11% (10.77%)**
Census Income	54.02% (2.78%)	60.41% (4.34%)	**61.10% (4.64%)**
Contraceptive	88.09% (0.35%)	**91.37% (13.89%)**	88.61% (3.63%)
Crack	38.85% (40.02%)	**44.18% (29.54%)**	23.68% (23.82%)
German Credit	76.64% (14.87%)	80.29% (11.99%)	**81.89% (10.65%)**
Heart	53.00% (5.83%)	**56.67% (6.10%)**	55.31% (11.67%)
Heroin	51.19% (26.37%)	**60.12% (29.06%)**	42.72% (17.57%)
Recidivism Female	81.71% (16.18%)	**81.86% (9.74%)**	78.39% (20.61%)
Recidivism Male	77.66% (6.84%)	**85.90% (7.92%)**	82.67% (7.04%)
Student	91.38% (10.64%)	92.93% (6.52%)	**93.13% (8.43%)**
Titanic	33.38% (2.92%)	43.90% (13.53%)	**58.14% (25.75%)**
Average	64.93% (23.58%)	**70.72% (21.31%)**	68.30% (24.10%)

We apply a Nemenyi posthoc test [11] to verify if there is a statistically significant difference in the results of demographic parity, equal opportunity, and equalized odds. For the Nemenyi posthoc test, we consider all classifiers selected with better hyperparameters per classification algorithm. Fig. 4 shows the results of the Nemenyi posthoc test. The top of the diagram indicates the critical difference (CD), and the horizontal axes indicate the average ranks of the group fairness metric, with the best-ranked algorithms to the left. A black line connects the algorithms when it is not detected a significant difference between them. For this experiment, with a significance level of 5% (p-value < 0.05), the critical difference is 0.4278.

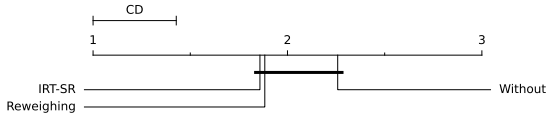
The IRT-SR method ranked first in the three evaluated group fairness metrics, as illustrated in Fig. 4. Reweighing was ranked second in all group fairness metrics. Figs. 4a and 4c show that IRT-SR and Reweighing with statistically significant differences compared to using no sample reweighting method on demographic parity and equalized odds metrics. Finally, the non-use of the sample reweighting method was ranked last in all fairness metrics tested.

Table 8 shows the macro F1-score averages for each group fairness metrics used in model selection. As expected, not using the sample reweighting method has the best macro F1-score averages. In contrast, the models developed with IRT-SR had the worst performance in the macro F1-score. IRT-SR had only a maximum mean difference of 1.50% for the best means. However, this small performance loss is compensated by improving fairness metrics.

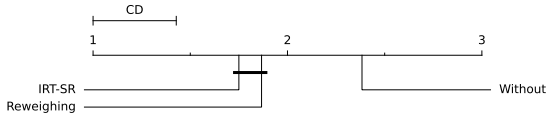
The results demonstrate that using the IRT-SR can be a great option to improve demographic parity, equal opportunity, and equalized odds. This is without a significant performance loss, as shown in Table 8. We note that both for average results (Tables 5, 6, and 7) and for methods ranking (Fig. 4) of the three



(a) Demographic Parity



(b) Equal Opportunity



(c) Equalized Odds

Fig. 4: Nemenyi posthoc test applied to the results of demographic parity, equal opportunity, and equalized odds.

Table 8: Macro F1-score averages on the test set for each criterion used in model selection.

Group fairness metric	Without	IRT-SR	Reweighing
Demographic parity	**69.53% (11.98%)**	68.09% (11.64%)	68.42% (11.74%)
Equal opportunity	**69.36% (11.85%)**	68.23% (11.68%)	68.69% (12.09%)
Equalized odds	**69.83% (11.76%)**	68.23% (11.26%)	68.86% (11.92%)

group fairness metrics, IRT-SR stood out as the best option among the options tested in this experiment. Finally, experimental results indicate that our IRT-SR sample reweighting method can guide classification algorithms to fit fairer models.

6 Conclusion

This paper introduced a novel sample reweighting method named IRT-SR that uses concepts from the Item Response Theory. We aimed to model the sample reweighting problem as a test to benefit from the IRT's evaluative power and use it to improve the group fairness notions through sample reweighting. The experimental results indicate that our method is more effective in maximizing demographic parity, equal opportunity, and equalized odds metrics than not using sample reweighting and the Reweighing method. In conclusion, the findings of this study highlight that IRT-SR effectively guides the classification algorithms to fit fairer classifiers.

In future work, we intend to optimize the hyperparameters of the base classifiers set and also test other classification algorithms in this set, which enables

the merging of classifiers from different paradigms. With this, we aim to improve further the group fairness notions presented in this work.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001; the São Paulo Research Foundation [grants #20/09835-1 and #22/09091-8]; and the Brazilian National Council for Scientific and Technological Development [grant #303588/2022-5].

References

1. Amrieh, E.A., Hamtini, T., Aljarah, I.: Preprocessing and analyzing educational data set using x-api for improving student's performance. In: 2015 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT). pp. 1–5. IEEE (2015)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: Risk assessments in criminal sentencing (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
4. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias (Oct 2018)
5. Bock, R.D., Aitkin, M.: Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. Psychometrika **46**(4), 443–459 (1981)
6. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)
7. Chalmers, R.P.: mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software **48**(6), 1–29 (2012). https://doi.org/10.18637/jss.v048.i06
8. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785, http://doi.acm.org/10.1145/2939672.2939785
9. Chen, Z., Ahn, H.: Item response theory based ensemble in machine learning. International Journal of Automation and Computing **17**(5), 621–636 (2020)
10. De Ayala, R.J.: The theory and practice of item response theory. Guilford Publications, New York City (2013)

11. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (Dec 2006), http://dl.acm.org/citation.cfm?id=1248547.1248548
12. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
14. Hambleton, R.K., Swaminathan, H., Rogers, H.J.: Fundamentals of item response theory, vol. 2. SAGE Publications, Thousand Oaks, CA, US (1991)
15. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29**, 3315–3323 (2016)
16. Hutchinson, B., Mitchell, M.: 50 years of test (un) fairness: Lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 49–58 (2019)
17. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (Oct 2012). https://doi.org/10.1007/s10115-011-0463-8, https://doi.org/10.1007/s10115-011-0463-8
18. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the compas recidivism algorithm (2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
19. van der Linden, W.J., Hambleton, R.K.: Handbook of modern item response theory. Springer Science & Business Media (2013)
20. Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Making sense of item response theory in machine learning. In: Proceedings of the Twenty-second European Conference on Artificial Intelligence. pp. 1140–1148 (2016)
21. Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Item response theory in ai: Analysing machine learning classifiers at the instance level. Artificial Intelligence **271**, 18–42 (2019)
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
23. Minatel, D., dos Santos, N.R., da Silva, A.C.M., Cúri, M., Marcacini, R.M., Lopes, A.d.A.: Unfairness in machine learning for web systems applications. In: Proceedings of the 29th Brazilian Symposium on Multimedia and the Web. pp. 144–153 (2023)
24. Minatel, D., da Silva, A.C.M., dos Santos, N.R., Curi, M., Marcacini, R.M., de Andrade Lopes, A.: Data stratification analysis on the propagation of discriminatory effects in binary classification. In: XI Symposium on Knowledge Discovery, Mining and Learning. pp. 73–80. SBC (2023)
25. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Computing Surveys (CSUR) **55**(3), 1–44 (2022)
26. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN computer science **2**(3), 160 (2021)
27. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: networked science in machine learning. SIGKDD Explorations **15**(2), 49–60 (2013). https://doi.org/10.1145/2641190.2641198, http://doi.acm.org/10.1145/2641190.264119