Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4677
This work is licensed under a Creative Commons Attribution 4.0 International License.

Influence of data stratification criteria on fairer classifications

Diego Minatel [**University of São Paulo** | ***dminatel@usp.br***]
Nícolas Roque dos Santos [**University of São Paulo** | ***nrsantos@usp.br***]
Angelo Cesar Mendes da Silva [**University of São Paulo** | ***angelo.mendes@usp.br***]
Mariana Cúri [**University of São Paulo** | ***mcuri@icmc.usp.br***]
Ricardo Marcondes Marcacini [**University of São Paulo** | ***ricardo.marcacini@usp.br***]
Alneu de Andrade Lopes [**University of São Paulo** | ***alneu@icmc.usp.br***]

Institute of Mathematics and Computer Sciences, University of São Paulo, Av. Trabalhador São Carlense, 400, Centro, São Carlos, São Paulo, Brazil, 13566-590.

Received: 6 July 2024 • **Published:** 20 June 2025

Abstract Data stratification by class is a prominent strategy to enhance the accuracy of model evaluation in unbalanced scenarios. This type of strategy, added to other stratification criteria, can also be effective in a significant issue with machine learning systems, which is their potential to propagate discriminatory effects, harming specific people groups. Therefore, it is crucial to assess whether these systems' decision-making processes are fair across the diversity present in society. This assessment requires stratifying the test set not only by class but also by sociodemographic groups. Furthermore, applying stratification by class and group during the validation step can contribute to developing fairer models. Despite its importance, there is a lack of studies analyzing the influence of data stratification on fairness in machine learning. We address this gap and propose an experimental setup to analyze how different data stratification criteria influence the development of impartial classifiers. Our results suggest that stratifying data by class and group aids develop fairer classifiers, thereby minimizing the spread of discriminatory effects in decision-making processes.

Keywords: Analysis, Binary Classification, Data Bias, Discriminatory Effects, Fairness, Machine Learning, Supervised Learning

1 Introduction

In many circumstances, Machine Learning (ML) model decision-making can benefit or harm a specific type of individual. For example, several reports exist regarding cases where the models propagated discriminatory bias with relevant societal impact [Alikhademi *et al*., 2022; Minatel *et al*., 2023b]. The best-known example of this situation is the COMPAS [Angwin *et al*., 2016], used in the American court to support their decisions regarding parole, which yields almost twice as many false positives in the classification of criminal recidivism for black people compared to false positives for white people. Furthermore, according to Buolamwini and Gebru [2018], the likelihood of a black woman being accused of a crime she did not commit is higher if the police use the main commercial facial recognition tools (*e.g.*, IBM Watson Visual Recognition) to solve crimes, as these tools have lower accuracy in recognizing black women. Moreover, web search engines are known to perpetuate social stereotypes and prejudices with their ML models [Howard and Borenstein, 2018].

Considering these situations, a dataset can be interpreted as a social mirror and reflect the prejudices, stereotypes, social inequalities, injustices, and other types of discrimination integrated into society [Barocas and Selbst, 2016; Mehrabi *et al*., 2021; Pessach and Shmueli, 2022]. Therefore, developing fairer models is a relevant challenge in the ML area because their applications are data-driven and can reproduce these social biases [Goodman and Flaxman, 2017; Le Quy *et al*., 2022]. The research topic of Fairness in Machine Learning tackles this issue by integrating fairness notions in the learning process to develop non-discriminatory ML decision-making while preserving the models' performance as much as possible [Barocas *et al*., 2023].

One factor that hinders the induction of fairer models is unbalanced data. In addition to the already known class imbalance, social data has an aggravating factor in this regard, as certain sociodemographic groups (*e.g.*, white men) may be overrepresented while others are underrepresented. This imbalance arises from population bias, where historical and social conditions limit the representation of certain groups in specific contexts, or from biased data collection practices that fail to capture the diversity of the population [Barocas and Selbst, 2016; Mehrabi *et al*., 2021].

In population bias cases, data augmentation techniques have been shown to mitigate imbalance effects [Yucer *et al*., 2020; Xu *et al*., 2020; Pastaltzidis *et al*., 2022]. Alternatively, training the model to capture the inherent imbalance present in the problem representation and subsequently incorporating notions of fairness into the learning process is another effective strategy [Kamiran and Calders, 2012; Zhang *et al*., 2018; Celis *et al*., 2019; Minatel *et al*., 2023d,e]. In this situation, whether or not to use data stratification in the model validation stage can be a crucial aspect in selecting the hyperparameter values that best fit a fairer model, especially in small datasets.

Although stratification methods have been studied for decades in the machine learning field, the recent emphasis

on fairness has stimulated new research into evaluating and adapting these methods to mitigate underrepresentation, not only in data labels but also among instances associated with marginalized groups. This paper handles this gap and proposes an experimental setup to evaluate and analyze the impact of the different data stratification criteria on selecting the fairest model. By bridging the gap between stratification and addressing fairness in machine learning, our paper provides empirical evidence and practical recommendations for effectively applying stratification methods to handle imbalanced datasets to build more equitable models. Therefore, our study comprises the following contributions:

- It proposes a novel and robust experimental setup capable of evaluating data stratification criteria;
- It analyzes the influence of data stratification in inducing fairer classifiers;
- It analyzes the influence of data stratification by different classification algorithms;
- It emphasizes that when group information is available in classification tasks involving people, it is essential to use the class and group criteria to stratify the test set.

This study is an extended version based on previous work [Minatel *et al*., 2023a]. In this extension, we expand the experimental setup by adding more classification algorithms and settings. We have added more details of our experimental protocol and expanded the background and related work section. Furthermore, we carried out a more in-depth analysis of the results and also modified the way we evaluated the stratification criteria to identify which one contributed the most to the selection of more impartial models.

Our experimental evaluation indicates that stratifying the data by matching the original distribution of groups and classes in cross-validation selects models that minimize discriminatory effects in binary classification tasks. Therefore, we recommend using stratification based on group and class criteria, as integrating these straightforward details into the validation process can improve the development of fairer models.

The remaining of this paper contains four other sections. Section 2 summarizes the main topics related to this work and introduces fundamental concepts for understanding our proposal. Section 3 describes our proposal and experimental settings. Section 4 presents and discusses the results of the experiments. Finally, Section 5 presents our concluding remarks and future work.

2 Background and Related Work

This section describes the terminology and fundamental concepts required to understand our proposal in Section 3.

2.1 Basic concepts

Protected attributes are features that include sensitive data such as gender, nationality, race, religion, and sexual orientation. From protected attributes derive groups, which, regardless of value, require equal treatment [Mehrabi *et al*., 2021]. Thus, in a dataset with the protected attributes gender and nationality (where the domain only considers: Argentina and Brazil), we have the following groups: `Argentine man`, `Argentine woman`, `Brazilian man`, and `Brazilian woman`. *Privileged group* is a group or set of groups that historically obtained advantageous treatment than other groups, called *unprivileged groups*.

Adverse treatment occurs when the protected attributes support a decision in part or in full. In many countries, adverse treatment is forbidden by law, as is the case in Brazil, which in Item IV of the third article of its Constitution says: "to promote the well-being of all, without prejudice as to the origin, race, sex, color, age and any other forms of discrimination." [BRASIL, 1988]. *Adverse impact* occurs when there are disproportionate outcomes that harm or benefit a particular group [Barocas and Selbst, 2016]. In the ML domain, adverse treatment occurs when the protected attributes are used to train a model. As well as we verify the adverse impact when there are disparities in results between groups, such as accuracy.

2.2 Data stratification

Data stratification is a sampling defined by one or more criteria that creates a subset of an original dataset. In the context of ML, this process splits a dataset into subsets used as input in a model's training, validation, and testing steps. These subsets retain the original sample proportions based on predefined criteria, such as preserving the distribution of target classes across all new subsets [Valentim *et al*., 2019]. Furthermore, when the goal is to minimize discriminatory effects, it is important to apply specific additional criteria during the dataset-splitting process to achieve a balanced and fair distribution of samples, such as stratification by group or group and target class simultaneously [Hanna *et al*., 2020; Gerdon *et al*., 2022].

Figure 1 shows an example of data stratification, where the criterion used is to maintain the color proportion. Therefore, the ratio of orange and blue colors is maintained in the generated subset (bottom of the image).

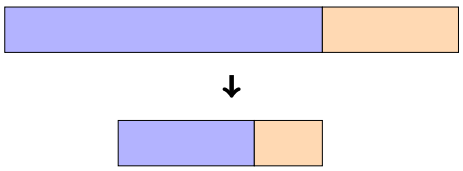


Figure 1. Illustration of data stratification using the color criterion so that the new subset maintains the proportion of blue and orange colors.

2.3 Group fairness analysis

Group fairness analysis focuses on verifying disproportionate results between groups, that is, identifying adverse impact. Some of the main group fairness notions applied in binary classification tasks are presented as follows:

- **Demographic parity:** each group has an equal likelihood of being classified with a positive label (selection rate) [Dwork *et al*., 2012].

- **Equal opportunity:** all groups have equal true positive rates, i.e., each group has the same recall score [Hardt *et al*., 2016].
- **Equalized odds:** all groups have the same true positive rate and false positive rate [Hardt *et al*., 2016].

To achieve these fairness notions, the results between all groups must be equal. However, according to Chouldechova [2017], achieving parity across all fairness metrics is impossible. Thus, a more accessible way to evaluate a classifier by these different definitions is to transform them into measures of group fairness. This typically involves calculating the ratio of scores between privileged and unprivileged groups, allowing us to quantify how much the result deviates from the ideal.

Note that we can transform any performance measure into a group fairness measure. For example, it is expected to do this with the performance measure chosen to evaluate the classifier, such as Macro F1-Score. Thus, it is possible to measure the asymmetry of Macro F1-Score scores between the analyzed groups.

To facilitate the interpretation of these metrics' results, we used the higher of the two scores in the denominator so that the result is between 0 and 1, with the ideal value being equal to 1. Note that this formulation only evaluates disproportionality between the results, not indicating which group is being harmed. Table 1 shows the acronym and description of each group fairness measure used in this study.

Table 1. Group fairness measures: the ratio of the calculated scores is between the privileged and unprivileged groups.

Acronym	Description
RDP	ratio of scores associated with demographic parity
REO	ratio of scores associated with equal opportunity
RDO	ratio of scores associated with equalized odds
RMF1	ratio of Macro F1-Score

2.4 Related work

Different works have explored the insertion of fairness notions in ML-based systems. In [Valentim *et al*., 2019], the authors investigated how data preparation influences the effectiveness of the learned model and its fairness. It was identified that the transformations applied to a dataset affect the analyzed fairness notions. Additionally, removing sensitive attributes helps achieve fairer models, but it is not enough to remove the unfairness from the predictions made by a model. Karimi *et al*. [2022] proposed a framework for fairer predictions using a causal analysis method and devised a new measure to quantify how fair a model is to different individuals.

Generally, these algorithms differ by stage in the learning process, whether preprocessing, in-processing, or post-processing [Barocas *et al*., 2023]. Preprocessing algorithms, such as data augmentation and sample reweighting techniques, act directly to reduce the discriminatory bias of the dataset [Calmon *et al*., 2017; Pastaltzidis *et al*., 2022; Minatel *et al*., 2023c,d]. In-processing methods modify the classification algorithms by including fairness constraints for model induction [Narasimhan, 2018; Zhang *et al*., 2018; Celis *et al*., 2019]. Finally, post-processing algorithms transform the classifier's responses to make them more impartial [Hardt *et al*., 2016; Pleiss *et al*., 2017].

Our work is similar to theirs in the sense that we also investigate strategies to improve the fairness of ML-based decisions to different groups. However, our work is focused on identifying which data stratification criteria yield fairer models in this context. Additionally, we propose a novel and robust experimental setup capable of evaluating data stratification criteria in Fairness in Machine Learning.

3 Proposal

This section presents our proposal to analyze the influence of different data stratification criteria on the propagation of discriminatory effects in binary classification. We designed an experimental setup to perform this analysis. Figure 2 shows the overview of this experimental setup.

Firstly, we selected and preprocessed a collection of binary classification datasets (Section 3.1). After, we applied the holdout sampling on the dataset to split it into the train (70%) and test (30%) subsets. The original ratio of the data — the focus of our analysis — is maintained. In other words, we stratified the data by group and class to maintain each group's positive and negative class ratio (privileged and unprivileged). The main idea is to test the models with the original ratio of the data in order to simulate the distribution found in a real decision-making situation. We randomly split the datasets using three different seeds: 31415, 42, and 4321. Thus, for each dataset, the steps to follow are executed three times, considering different train and test subsets.

In sequence, we used nine distinct classification algorithms in the validation stage with sixteen settings each (Section 3.2). We applied a five-fold cross-validation sampling process on the training set in these different settings using the following data stratification criteria: `(none)`, `(class)`, `(group)`, and `(group, class)`. We adopted five-fold cross-validation because some of the selected datasets have few instances. Below, we describe the data stratification criteria used:

`(none)` — does not use data stratification.
`(class)` — data stratification by target class (positive and negative).
`(group)` — data stratification by group, which are privileged and unprivileged in this experiment.
`(group, class)` — data stratification by group and target class.

Section 3.1 provides more detailed information about the classes and composition of each dataset's privileged and unprivileged groups.

At the end of the validation stage, we applied a multicriteria measure (more details in Section 3.3) to jointly evaluate different group fairness measures and select the best hyperparameter values for each classification algorithm. Next, we retrained the classifiers with the best hyperparameters selected for each data stratification criteria with the entire training set.

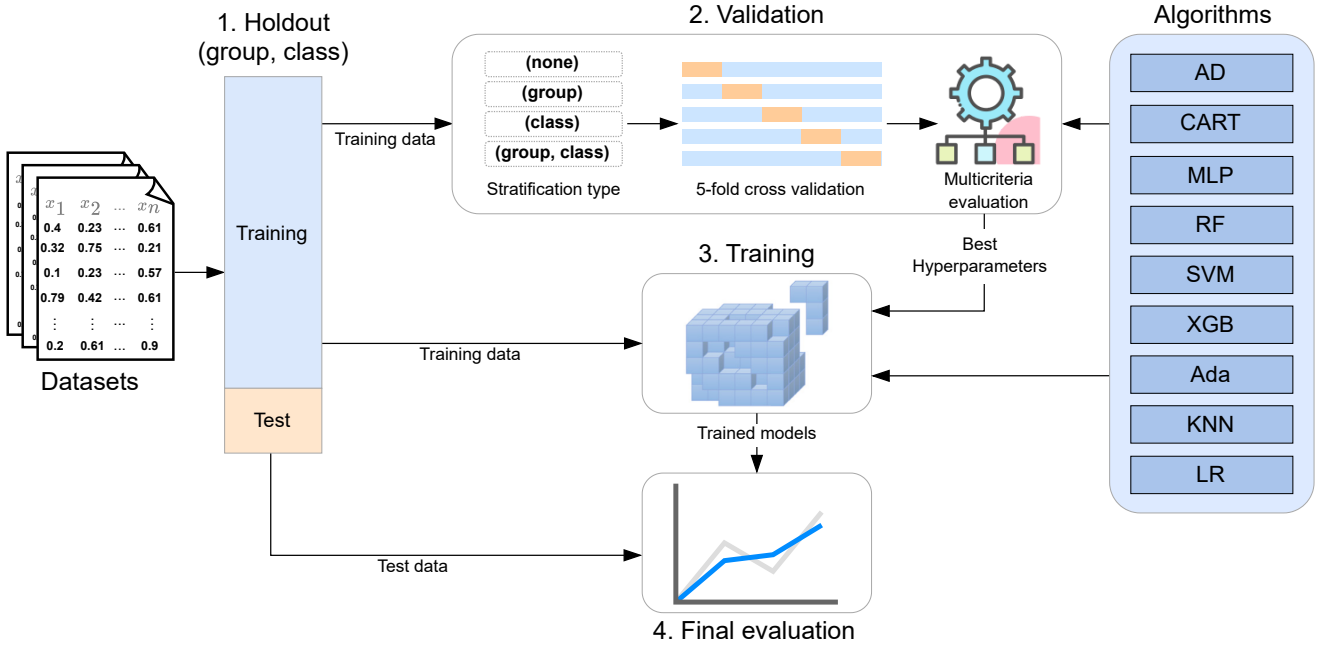


Figure 2. Overview of the performed analysis. Initially, we split the dataset into training and testing sets. In sequence, we conducted a multicriteria evaluation using nine classification algorithms and five-fold cross-validation stratified sampling on the training data to identify the best hyperparameters for each classifier. Thus, we then trained each classifier using the obtained hyperparameters. Finally, we assess the influence of data stratification criteria on fairer classifications in the testing set.

In the final evaluation, we assessed the selected models in the training set using the multicriteria measure (fairness analysis) and one performance measure, detailed in Section 3.3. For each tuple *{dataset, classification algorithm}*, we calculated the average of the results on three different test sets (three different seeds applied).

To confront the results statistically, we used Friedman's non-parametric hypothesis test for paired data and multiple comparisons at a significance level of 5% (p-value < 0.05), followed by the Nemenyi *post-hoc* test [Demšar, 2006]. Therefore, with the results of this experiment, we can discuss new approaches and help in elaborating fairer models.

In the remainder of this section, we detail all the datasets, classification algorithms, and the evaluation process used in the experiment proposed.

3.1 Datasets

We selected the relevant binary classification benchmark datasets used in the Fairness in Machine Learning research community for this work, which are described as follows:

Arrhythmia: consists of clinical records of patients with the purpose of predicting the absence or presence of one of sixteen groups of cardiac arrhythmia. The cardiac arrhythmia groups were consolidated into a single category to make it compatible with binary classification. Therefore, the classes in this dataset for this study are defined as 'absence' or 'presence' of cardiac arrhythmia [Dua and Graff, 2017].

Bank Marketing: contains information about marketing campaigns related to term deposits from a banking institution. Therefore, the objective is to predict whether or not the customer will sign a term deposit [Dua and Graff, 2017].

Census Income: comprises information from the US census carried out in 1994. The goal is to predict whether a person earns less or more than fifty thousand dollars annually [Dua and Graff, 2017].

Contraceptive: contains information from the National Indonesia Contraceptive Prevalence Survey that was carried out in 1987. Examples refer to married women who were not pregnant during the interview. The goal is to predict whether or not a woman uses contraceptive methods [Dua and Graff, 2017].

Drug: includes responses from a survey on drug use, allowing us to predict whether a person has used or never used a variety of 18 types of drugs, covering both legal and illicit drugs. For this work, the data was used to classify the use or non-use of the following drugs: Alcohol, LSD, and Nicotine [Dua and Graff, 2017].

German Credit: includes personal data and credit history, aiming to classify whether a person presents a good or bad credit risk [Dua and Graff, 2017].

Heart: contains patient information related to heart disease. Thus, the goal is to predict whether or not a patient has heart disease [Dua and Graff, 2017].

Recidivism: contains data on criminal history, prison time, demographic information, and COMPAS risk scores. The goal is to predict whether an individual will commit criminal recidivism two years after the first arrest. This data set was divided into Recidivism Female (female examples) and Recidivism Male (male examples) [Larson *et al.*, 2016].

Titanic: contains information about the passengers who boarded the Titanic. The classification aims to predict whether a given person survived its sinking [Van-

schoren *et al.*, 2013].

Table 2 summarizes the datasets, showing their amount of instances (#I), number of attributes (#A), which protected attributes (#PA) are analyzed, the privileged group (#PG) of each task related to the dataset, and reference. It is important to note that, as discussed in Section 2, the unprivileged group comprises all groups that are not contained in the privileged group. For example, in the Recidivism Female, where the protected attribute analyzed is race, and the privileged group is white people, the unprivileged group corresponds to non-white individuals. Furthermore, Figure 3 shows the ratio of each subset `(group, class)` in the datasets.

Table 2. Dataset information

ID	Dataset	#I	#A	#PA	#PG
1	Alcohol	1,885	11	ethnicity	caucasian
2	Arrhythmia	452	278	sex	male
3	Bank	45,211	42	age	over 25 years
4	Census Income	48,842	76	race and gender	white-male
5	Contraceptive	1,473	10	religion	non-islam
6	German Credit	1,000	36	gender	male
7	Heart	303	13	age	middle-aged
8	LSD	1,885	11	ethnicity	caucasian
9	Nicotine	1,885	11	ethnicity	caucasian
10	Recid. Female	1,395	176	race	white
11	Recid. Male	5,819	375	race	white
12	Titanic	1309	6	gender	female

For Arrhythmia and Heart, the protected attributes 'gender' and 'age' were used in training, respectively, as they play essential roles in disease prediction. For Titanic, the protected attribute 'gender' was used due to known selection bias: women and children had priority in rescue.

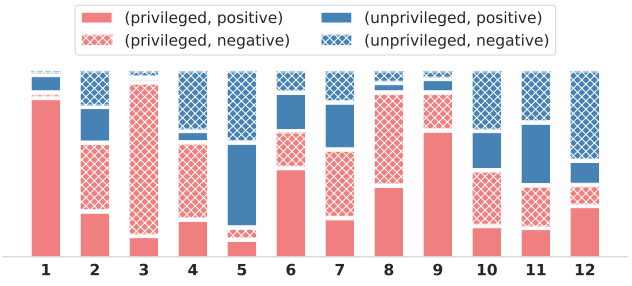


Figure 3. Ratio of each subset `(group, class)` in the datasets.

3.2 Classification algorithms

We used the following classification algorithms for the experiment: AdaBoost (ADA), Classification Trees (CART), K-Nearest Neighbors (KNN), Logistic Regression (LR), Multilayer Perceptron (MLP), Random Forest (RF), Support Vector Machines (SVM), and eXtreme Gradient Boosting (XGB). We also used the well-known classification algorithm to minimize discriminatory effects called Adversarial Debiasing (AD) [Zhang *et al.*, 2018]. Table 3 shows each classification algorithm and the numerical variation range for its hyperparameters used in this experiment[1]. We tested sixteen parametrization settings per classification algorithm.

[1]For hyperparameters not mentioned, we used the default values of the classification algorithms from the following Python language libraries: scikit-learn (ADA, CART, KNN, MLP, RF, and SVM), aif360 (AD), and xgboost (XGB).

Table 3. Algorithms and variation of values for their hyperparameters. Numerical variations in the format $(i : f : p)$ indicate that the initial and final values are i and f, respectively, and p indicates the increment used.

Algorithm	Parameter	Fixed Value	Value Variation
AD	Number of epochs for which to train	—	`(50 : 530 : 30)`
ADA	The number of trees	—	`(120 : 440 : 20)`
CART	The function to measure the quality of a split	`gini`	—
	The minimum of samples required to be at a leaf node	—	`(2 : 18 : 2)`
	The minimum of samples required to split an internal node	—	`(4 : 5 : 1)`
KNN	Number of neighbors	—	`(1 : 17 : 2)`
	Power parameter for the Minkowski metric	—	`(1 : 2 : 1)`
LR	Regularization	—	0.8 : 1.2 : 0.025
MLP	The number of neurons in the hidden layer	—	`(5 : 21 : 1)`
RF	The number of trees	—	`(120 : 440 : 20)`
SVM	Kernel	`rbf`	—
	Regularization	1	—
	Gamma	—	`(0.001 : 1.2 : 0.075)`
XGB	The number of trees	—	`(120 : 440 : 200`

3.3 Evaluation measures

In this work, we prioritize the group fairness analysis between the privileged and unprivileged groups (described for each dataset in Table 2), and we also evaluated the performance of the classifiers using the Macro F1-Score; we selected this measure due to the imbalance of classes present in the datasets (see Figure 3).

We use the Multi-Criteria Performance Measure (MCPM) proposed in [Parmezan *et al.*, 2017] for group fairness analysis. Figure 5 shows an example of the multicriteria measure, where three measures (to facilitate visualization of the calculation) were selected in the evaluation. Thus, each algorithm's area of each irregular triangle formed by the meeting of the edges with vertices that represent each pair of measurements is calculated. Therefore, the value of MCPM is given by the sum of these areas.

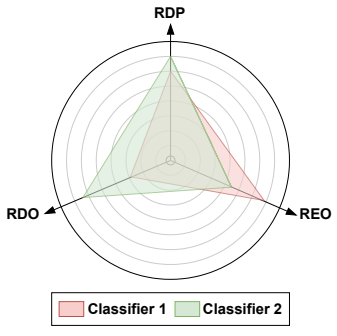


Figure 4. Example of how we calculate the MCPM score: the value of the multicriteria measure is given by the sum of the area of each irregular triangle formed by the meeting of the edges with vertices representing each pair of measurements.

To perform group fairness analysis, we apply MCPM, combining four fairness measures. Herein, we set the three main group fairness measures, RDP, REO, and RDO, and also the fairness measure associated with the classifier performance: RMF1. As described in Section 2.3, all these metrics range from 0 to 1 and have an ideal value of 1. Consequently,

the most impartial classifier has the highest MCPM score.

4 Results and Discussion

This section presents the results obtained in the evaluation process for the approach proposed in Section 3.

Table 4 shows the average values of MCPM on the test set for each data stratification criterion. The average for each dataset is made up of nine results, which refer to each classification algorithm. The highest value, highlighted in green, indicates the best average value per dataset, while the worst average value is highlighted in red. Values in parentheses indicate the standard deviation.

Table 4. Average MCPM values on the test set reflect the combination of the four following group fairness metrics: (i) RDP, (ii) REO, (iii) RDO, and (iv) RMF1. The best result for each dataset is highlighted in green, while the worst is highlighted in red.

Dataset	`(none)`	`(class)`	`(group)`	`(group, class)`
Alcohol	2.53 (0.05)	2.54 (0.02)	2.55 (0.01)	2.55 (0.01)
Arrhythmia	1.45 (0.39)	1.41 (0.40)	1.38 (0.36)	1.39 (0.38)
Bank	1.23 (0.23)	1.23 (0.24)	1.22 (0.23)	1.25 (0.24)
Census Income	1.13 (0.11)	1.14 (0.12)	1.14 (0.12)	1.14 (0.12)
Contraceptive	2.06 (0.19)	2.03 (0.16)	2.09 (0.20)	2.08 (0.20)
German	2.19 (0.22)	2.19 (0.21)	2.16 (0.22)	2.21 (0.24)
Heart	1.34 (0.29)	1.31 (0.24)	1.33 (0.28)	1.38 (0.27)
LSD	1.77 (0.19)	1.73 (0.25)	1.75 (0.24)	1.83 (0.25)
Nicotine	2.29 (0.17)	2.31 (0.16)	2.32 (0.14)	2.30 (0.19)
Recid. Female	1.82 (0.29)	1.89 (0.30)	1.88 (0.30)	1.89 (0.27)
Recid. Male	1.55 (0.17)	1.54 (0.18)	1.56 (0.18)	1.57 (0.18)
Titanic	0.55 (0.27)	0.55 (0.26)	0.53 (0.24)	0.54 (0.25)
Average	1.66 (0.21)	1.66 (0.21)	1.66 (0.21)	1.68 (0.22)

As shown in Table 4, using only the group or class criterion in data stratification resulted in the worst average MCPM value in 8 of the 12 datasets, and they also obtained the worst overall average score. In contrast, when the criterion passes the combination of group and class, the best result is obtained in 8 of the 12 datasets, in addition to the best overall average MCPM. Moreover, it is important to highlight that the criterion `(group, class)` did not have the worst average in any dataset.

Through the low average MCPM values, it is possible to identify some data sets in which it is more challenging to achieve the desired fairness concepts in the classifiers. Among them, the Titanic dataset stands out, which has the worst MCPM average. In a hypothetical situation of use in the real world, it would probably propagate discriminatory effects, regardless of the classification algorithm and configuration applied.

To complement the MCPM analysis, we applied Friedman's non-parametric statistical test for paired data, followed by Nemenyi's *post-hoc* test, to check whether there is a statistical difference in the results in the test sets. Each tuple *{dataset, classification algorithm}* was considered in the test, resulting in the analysis of 108 classifiers (12 datasets $\times$ 9 classification algorithms $\times$ 1 best configuration). Figure 5 shows the CD diagram representing MCPM on test sets. At the top of the diagram, we can observe the Critical Difference (CD), and the horizontal axis represents the average ranks of the model selection strategies, with the best-ranked data stratification criterion on the left. A black line connects criteria when no significant difference is detected between them.

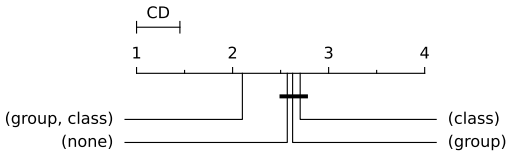


Figure 5. Nemenyi *post-hoc* test applied to the MCPM results on the test sets.

The statistical results suggest, with a statistically significant difference, that among the analyzed data stratification criteria in the validation set, the criterion `(group, class)` is the best option for selecting more impartial classifiers. The average rank for criterion `(group, class)` was 2.10, while criteria `(none)`, `(class)`, and `(group)` had similar average rank, 2.57, 2.70, and 2.63, respectively.

As the MCPM measure encapsulates four different group fairness metrics, it is also interesting to analyze the individual results of these measures. For this purpose, we individually applied Nemenyi's *post-hoc* test to the measurements: RDP, REO, RDO, and RM1. Figure 6 shows the CD diagrams for each of them.

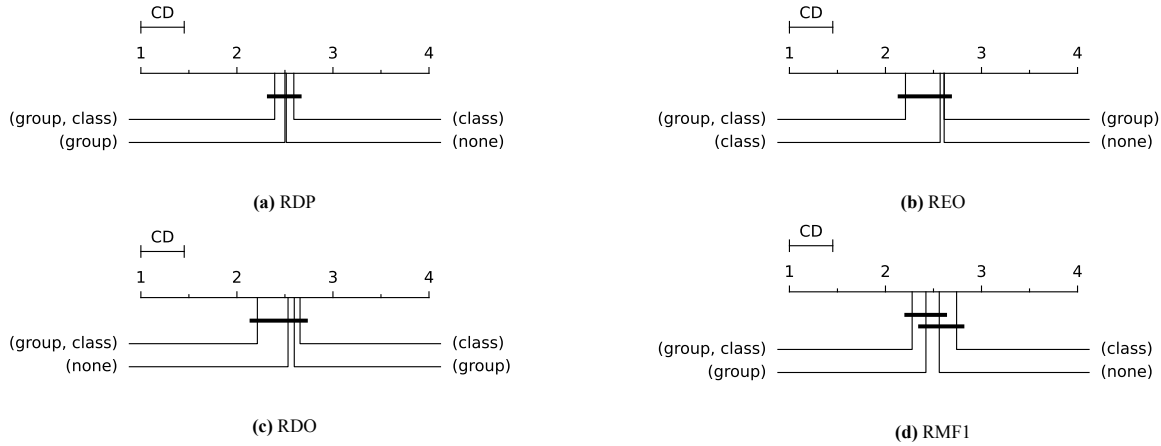
As seen in Figures 6a, 6b, 6c, and 6d, the criterion `(group, class)` was ranked first in all metrics analyzed, emphasizing the RMF1 measure, which had a statistically significant difference in the results about using the stratification criterion by class. These results show that the criterion `(group, class)` facilitates the selection of classifiers that reduce the disparity in performance between the privileged and unprivileged groups.

There is no point in having better results in the group fairness metrics if there is a considerable loss in Macro F1-Score. Therefore, it is essential to perform a performance analysis of the analyzed classifiers. Table 5 shows the average percentages of the Macro F1-Score on the test set for each data stratification criterion. The average for each dataset comprises nine results, which refer to each classification algorithm. The highest value, highlighted in green, indicates the best average value per dataset, while the worst average value is in red. Values in parentheses indicate the standard deviation.

Table 5. Average Macro F1-Score (%) on the test set. The best result for each dataset is highlighted in green, while the worst is highlighted in red.

Dataset	`(none)`	`(class)`	`(group)`	`(group, class)`
Alcohol	49.55 (0.45)	49.49 (0.18)	49.52 (0.06)	49.53 (0.08)
Arrhythmia	69.47 (5.65)	69.66 (5.56)	69.29 (6.25)	69.95 (5.49)
Bank	70.61 (4.75)	70.72 (4.69)	70.65 (4.76)	70.68 (4.81)
Census Income	78.00 (2.53)	77.97 (2.49)	78.04 (2.52)	78.00 (2.53)
Contraceptive	64.50 (5.96)	65.48 (5.42)	64.44 (5.77)	64.63 (6.05)
German	60.92 (7.79)	61.35 (7.96)	61.10 (8.02)	60.66 (7.85)
Heart	81.08 (4.16)	80.34 (4.54)	81.37 (3.78)	81.23 (4.16)
LSD	72.32 (2.26)	71.25 (3.57)	72.15 (3.44)	71.98 (3.21)
Nicotine	53.40 (5.23)	52.92 (5.26)	53.10 (5.40)	52.82 (5.25)
Recid. Female	59.34 (4.02)	59.83 (4.26)	59.73 (3.89)	59.99 (4.12)
Recid. Male	64.16 (2.50)	64.09 (2.49)	64.13 (2.51)	64.12 (2.50)
Titanic	77.80 (6.18)	78.20 (5.13)	77.39 (6.83)	77.73 (6.19)
Average	66.76 (4.29)	66.77 (4.30)	66.74 (4.44)	66.78 (4.35)

Sometimes, enhancing various fairness concepts leads to a loss of prediction performance. Fortunately, this did not happen in the case of the criterion `(group, class)`, as it had the best overall average Macro F1-Score. However, the average Macro F1-Score results were very close across all tested criteria. Also worth noting is the data stratification



(a) RDP **(b)** REO **(c)** RDO **(d)** RMF1

Figure 6. Nemenyi *post-hoc* test applied on the test sets individually to each group fairness measure that makes up the MCPM multi-criteria measure.

by class criterion, which had the best results on 4 out of 12 datasets while also having 5 of the worst results.

In some datasets, such as Alcohol and Nicotine, the classifiers had Macro F1-Score results that were far below expectations. One reason for this is the large class and group imbalance in both datasets, as seen in Figure 3. In cases like these, it is also essential to test data augmentation techniques focusing on minimizing the effects of the data unbalance. However, this type of analysis is outside the scope of this work.

We also apply Nemenyi's *post-hoc* test to the Macro F1-Score results. Figure 7 shows the CD diagram representing Macro F1-Score on test sets. The ranking obtained was as follows: first `(group)`, followed in ranking order by `(group, class)`, `(class)`, and `(none)`. There was no statistically significant difference between the results of the analyzed criteria.

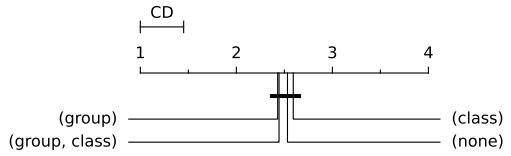


Figure 7. Nemenyi *post-hoc* test applied to the Macro F1-Score results on the test sets.

Table 5 and Figure 7 provide a clear path to answering the crucial question we posed earlier about the significance of improving outcomes on fairness metrics while also maintaining good results in the selected predictive performance measures. These results confirm no loss of predictive power with criterion `(group, class)`; its performance in Macro F1-Score is similar to the other criteria analyzed and even better in some situations. This confirmation further highlights this criterion's results in improving notions of group fairness.

Our last analysis of the experimental results concerns the influence of the types of classifiers on the average results highlighted so far. From now on, for each result related to a classification algorithm, we calculated the average score of the 12 analyzed datasets. Figures 8 and 9 present the MCPM and Macro F1-Score results per the classification algorithm. The red dashed line is a reference and indicates the highest average achieved for the measurement in question. The colors of the bars indicate the data stratification criterion.

Figure 8 shows that the algorithms based on artificial neural networks (AD and MLP) and KNN had the best average performance in MCPM. It is worth highlighting that AD was created to mitigate bias. In these three algorithms, the criterion `(group, class)` was among the two best performers. In contrast, ADA obtained the worst average MCPM results for all criteria, which is not the ideal algorithm for generating impartial classifiers in the collection of datasets analyzed.

With Figure 9, it is possible to contrast the results of Figure 8, as the three algorithms with the best average in Macro F1-Score (ADA, RF, and XGB) did not have good results in the fairness metrics. However, the algorithms that improved performance in MCPM, especially in AD and MLP, also had good results in Macro F1-Score and can be considered the best options for balancing accuracy and fairness in the experimental configuration adopted.

The criterion `(group, class)` had the best MCPM average across five algorithms tested (ADA, CART, MLP, SVM, and XGB). Regarding the Macro F1-Score results, the criteria results were very similar between the algorithms, except for CART and KNN. Finally, we did not notice any criteria plus algorithm results that influenced the average results, statistical evidence, and insights presented previously.

5 Conclusion and Future Work

This paper introduced a robust experimental setup with the ability to evaluate the impact of different data stratification criteria on model selection. The main objective of this study was to associate which data stratification criteria help in a fairer selection of models. According to the experimental results, stratifying the data by class and group of people (in the case of this paper, they are privileged and unprivileged groups) selects more impartial classifiers, contributing to fairer classifications and minimizing the spread of discriminatory effects. Furthermore, we reinforce that in classification tasks that involve people, when group information is available, it is essential to use the class and group criteria to stratify the test set, as this way, there is a more accurate evaluation of the group fairness metrics. In conclusion, the findings of this study highlight that a simple yet effective stratification method can serve as a straightforward pathway

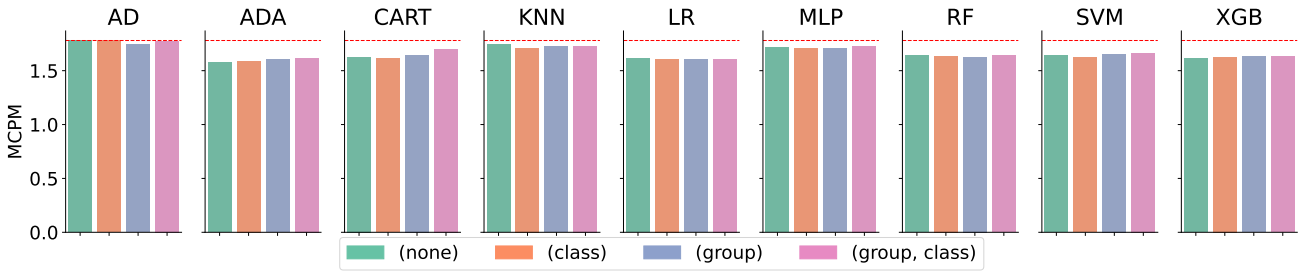


Figure 8. Average MCPM per classification algorithm on the test set. The red dashed line serves as a reference, indicating the highest average score.

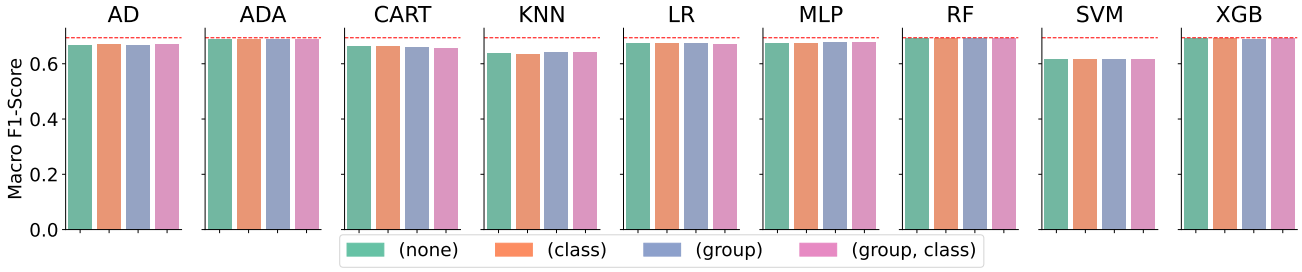


Figure 9. Average Macro F1-Score per classification algorithm on the test set. The red dashed line serves as a reference, indicating the highest average score.

to incorporate fairness into machine learning models.

In future work, we intend to expand the experimental setup to include multiclass classification and add datasets with unstructured data, such as text and images, to evaluate the generalization power of data stratification by group and class. We also intend to evaluate data stratification criteria combined with different data preparation and pre-processing techniques present in the Fairness in Machine Learning literature to identify best practices for inducing fairer classifiers.

Funding

This study was supported in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil* (CAPES) - Finance Code 001. Scholarships from CAPES support authors Angelo Cesar Mendes da Silva, Diego Minatel, and Nícolas Roque dos Santos. Author Alneu de Andrade Lopes is supported by the Brazilian National Council for Scientific and Technological Development (CNPq) grant #303588/2022-5. This work is supported by the National Institute of Artificial Intelligence (IAIA) of CNPq with grant number 406417/2022-9. The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation.

Competing interests

The authors declare that they do not have competing interests.

Availability of data and materials

The raw and preprocessed datasets analyzed, codes, and complete results are available in: `https://github.com/diegominatel/JIDM-Fairness-Analysis-In-Data-Stratification-Criteria`

References

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., and Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, pages 1–17.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: Risk assessments in criminal sentencing.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

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

BRASIL (1988). *Constituição da República Federativa do Brasil*. Brasília, DF: Centro Gráfico.

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

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, 30:3992–4001.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328.

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

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of*

the 3rd innovations in theoretical computer science conference, pages 214–226.

Gerdon, F., Bach, R. L., Kern, C., and Kreuter, F. (2022). Social impacts of algorithmic decision-making: A research agenda for the social sciences. *Big Data & Society*, 9(1):20539517221089305.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Howard, A. and Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33.

Karimi, H., Akbar Khan, M. F., Liu, H., Derr, T., and Liu, H. (2022). Enhancing individual fairness through propensity score matching. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. DOI: 10.1109/DSAA54385.2022.10032333.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Minatel, D., da Silva, A. C. M., dos Santos, N. R., Curi, M., Marcacini, R. M., and de Andrade Lopes, A. (2023a). Data stratification analysis on the propagation of discriminatory effects in binary classification. In *Anais do XI Symposium on Knowledge Discovery, Mining and Learning*, pages 73–80. SBC.

Minatel, D., dos Santos, N. R., da Silva, A. C. M., Cúri, M., Marcacini, R. M., and Lopes, A. d. A. (2023b). Unfairness in machine learning for web systems applications. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, pages 144–153.

Minatel, D., dos Santos, N. R., da Silva, V. F., Cúri, M., and de Andrade Lopes, A. (2023c). Item response theory in sample reweighting to build fairer classifiers. In *Annual International Conference on Information Management and Big Data*, pages 184–198. Springer.

Minatel, D., Parmezan, A. R., Cúri, M., and de A. Lopes, A. (2023d). Dif-sr: A differential item functioning-based sample reweighting method. In *Iberoamerican Congress on Pattern Recognition*, pages 630–645. Springer.

Minatel, D., Parmezan, A. R., Cúri, M., and Lopes, A. D. A. (2023e). Fairness-aware model selection using differential item functioning. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1971–1978. IEEE.

Narasimhan, H. (2018). Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654.

Parmezan, A. R. S., Lee, H. D., and Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Systems with Applications*, 75:1–24.

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., and Tzovaras, D. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2302–2314, New York, NY, USA. Association for Computing Machinery.

Pessach, D. and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, 30:5680–5689.

Valentim, I., Lourenço, N., and Antunes, N. (2019). The impact of data preparation on the fairness of software systems. In *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, pages 391–401. IEEE.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). Openml: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60. DOI: 10.1145/2641190.2641198.

Xu, T., White, J., Kalkan, S., and Gunes, H. (2020). Investigating bias and fairness in facial expression recognition. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 506–523, Berlin, Heidelberg. Springer-Verlag.

Yucer, S., Akcay, S., Al-Moubayed, N., and Breckon, T. P. (2020). Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.