Data stratification analysis on the propagation of discriminatory effects in binary classification

Diego Minatel, Angelo Cesar Mendes da Silva, Nícolas Roque dos Santos, Mariana Curi, Ricardo Marcondes Marcacini, Alneu de Andrade Lopes

Institute of Mathematics and Computer Science, University of Sao Paulo (USP), Brazil
{dminatel, angelo.mendes, nrsantos, ricardo.marcacini}@usp.br
{mcuri, alneu}@icmc.usp.br

Abstract. Unfair decision-making supported by machine learning, which harms or benefits a specific group of people, is frequent. In many cases, the models only reproduce the biases in the data, which does not absolve its responsibility for these decisions. Thus, with the increase in the automation of activities through machine learning models, it is mandatory to prospect solutions that add fairness factors to the models and clarity about the supported decisions. One option to mitigate model discrimination is quantifying the ratio of instances belonging to each target class to build data sets that approximate the actual data distribution. This alternative aims to reduce the responsibility of data on discriminatory effects and direct the function of treating them to the models. In this sense, we propose to analyze different types of data stratification, including stratification by sociodemographic groups that are historically unprivileged, and associate these stratification types to the fairer or unfairer models. According to our results, stratification by class and group of people helps to develop fairer models, reducing the discriminatory effects in binary classification.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: analysis, binary classification, data bias, data stratification, discriminatory effects, fairness, machine learning, unfairness

1. INTRODUCTION

In many circumstances, Machine Learning (ML) model decision-making can benefit or harm a specific type of person. For example, several reports exist regarding cases where the models propagated discriminatory bias with relevant impact on society [Alikhademi et al. 2022]. The best-known example of this situation is the COMPAS [Angwin et al. 2016], used in the American court to support their decisions, 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].

A social 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; Pessach and Shmueli 2022]. Developing non-discriminatory models is a relevant challenge to the ML area because the application is data-driven, reproducing the data biases [Goodman and Flaxman 2017; Le Quy et al. 2022]. In this context, the research topic of Fairness in Machine Learning has emerged. Its primary goal is to aggregate 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. 2017].

Unbalanced datasets in practice can reproduce social injustice and discrimination in model training. For example, an imbalance can underrepresent some social groups through the minority class, perpetuating systemic discrimination. Therefore, the stratification process is a resource for handling unbalanced scenarios in binary classification. This resource ensures that the original class frequencies present in the dataset are preserved in the training and testing sets [Kohavi 1995]. However, in decisions involving people, in addition to maintaining proportion between class frequencies, it is also necessary to preserve the distribution of sociodemographic groups because these groups may be underrepresented in the data. Hence, by incorporating such a crucial factor into a model, we can ensure that it accurately captures the data diversity, which can minimize the discriminatory effects of ML decisions.

Although stratification methods have been explored for decades in the machine learning field, the need to address the recent demand for a rawness in machine learning has stimulated new research in the evaluation and adaptation of stratification methods to mitigate underrepresentation, not only of data labels but also instances associated with socially marginalized minority groups. This paper handles this gap and proposes an experimental setup to evaluate and analyze the impact of the different data stratification types 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.

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. In contrast, not using data stratification penalizes the models by decreasing their performance and increasing the discriminatory effects of their predictions. Therefore, we recommend using stratification by group and class, as by adding these simple detail into validation, we can develop fairer models.

2. BACKGROUND

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

Data stratification is the process that splits a dataset into new subsets used as input in a model's training and testing steps. Thus, a new subset preserves the percentage of samples based on a predefined criterion, such as preserving the proportion of examples with the same target class for each new sampled subset [Valentim et al. 2019]. We commonly note the random process for data stratification in machine learning applications to measure the generalization power of models to handle the downstream tasks [Martin Hirzel and Ram 2021]. However, in the context of smooth discriminatory effect, this process must consider some criteria to balance the class distributions based on settings that aggregate fair sense [Hanna et al. 2020; Gerdon et al. 2022]. Therefore, we aim to explore multiple strategies for data stratification inserting criteria to minimize discrimination scenarios in different tasks.

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 a decision is supported in part or in full by the protected attributes. 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 in the training of a model. As well as we verify the adverse impact when there are disparities in results between groups, such as accuracy.

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].

Therefore, group fairness notions are based on the parity of outcomes between privileged and underprivileged groups. Thus, to check the disproportionality of a specific fairness notion, we typically calculate the ratio of scores between privileged and unprivileged groups to determine whether there is a disproportionateness in the results.

3. PROPOSAL

The proposed work analyzes the influence of different types of data stratification on the propagation of discriminatory effects in binary classification. In this context, we carried out an experiment (see Fig. 1) where in the first step, we applied the holdout sampling on the dataset to split into the train (70%) and test (30%) subsets. The original ratio of the data – the focus of our analysis – is maintained. Then, we stratified the data to maintain the positive and negative classes ratio for each group (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.

In sequence, we used six distinct classification algorithms and applied a five-fold cross-validation sampling process on the training set in these different settings using the following data stratification types: `(none)`, `(class)`, `(group)`, and `(group, class)`. We adopt five-fold cross-validation because some of the selected datasets have a small number of instances. Below we describe the data stratification types used:

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

At the end of the validation step, we apply a multicriteria measure to evaluate jointly different group fairness notions and performance measures to select the best hyperparameter values for each classifier. Finally, we train the classifiers with the best hyperparameters selected for each type of data stratification with the entire training set. Then we evaluate the multicriteria measure of each classifier on the test set to assess the discriminatory effects in their predictions. Therefore, with the results of this experiment, we can discuss new approaches and help in elaborating fairer models.

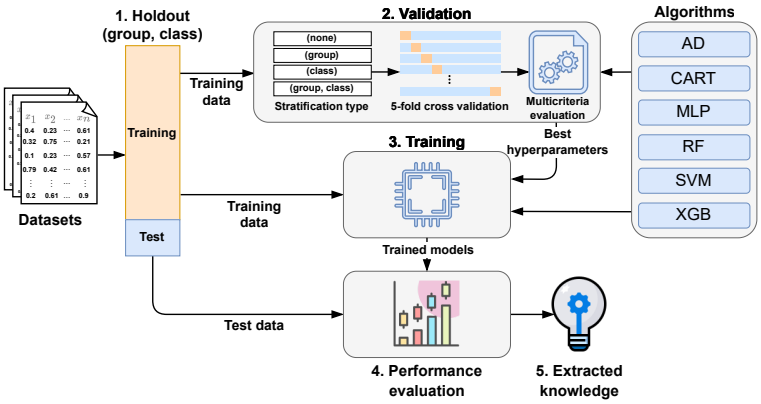

Fig. 1: Overview of the performed analysis. Initially, we split the dataset into training and testing sets. In sequence, we conducted a multicriteria evaluation using six classification algorithms and five-fold cross-validation stratified sampling on the training data to identify the best hyperparameters for each classifier. Thus, we then trained each classifier using the obtained hyperparameters. Finally, we assess the discriminatory effects of each classifier on the testing set.

All developed source code and benchmark datasets used in the evaluation process are available in a public code repository[1]. We also used the following libraries: scikit-learn, which contains classification algorithms and performance measures [Buitinck et al. 2013], and aif360, which contains fairness measures and algorithms to mitigate bias [Bellamy et al. 2018]. In the remainder of this section, we detail all datasets, algorithms, and the evaluation process used in the experiments.

3.1 Datasets

We selected the relevant binary classification benchmark datasets used in the Fairness in Machine Learning research community for this work. Table I summarizes the datasets, showing their amount of instances (#I), number of attributes (#A), which protected attributes are analyzed, the privileged group of each task related to the dataset, and reference. Important to note, as discussed in Section 2, that the unprivileged group is made up of all groups that are not contained in the privileged group. Furthermore, Fig. 2 shows the ratio of each subset `(group, class)` in the datasets.

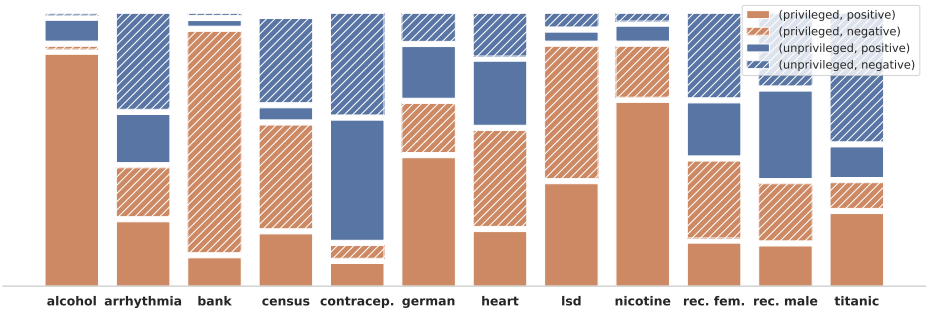

Fig. 2: Ratio of each subset `(group, class)` in the datasets.

[1]https://github.com/diegominatel/fairness-analysis-in-data-stratification

Table I: Dataset information

Dataset	#I	#A	Protected Attributes	Privileged Group	Reference
Arrhythmia[2][3]	452	278	sex	male	[Dua and Graff 2017]
Bank Marketing	45,211	42	age	over 25 years old	[Dua and Graff 2017]
Census Income	48,842	76	race and sex	white-male	[Dua and Graff 2017]
Contraceptive[4]	1,473	10	religion	non-islam	[Dua and Graff 2017]
Drug[5] (Alcohol)	1,885	11	ethnicity	caucasian	[Dua and Graff 2017]
Drug[5] (LSD)	1,885	11	ethnicity	caucasian	[Dua and Graff 2017]
Drug[5] (Nicotine)	1,885	11	ethnicity	caucasian	[Dua and Graff 2017]
German Credit	1,000	36	sex	male	[Dua and Graff 2017]
Heart[2]	303	13	age	middle-aged	[Dua and Graff 2017]
Recidivism Female[6]	1,395	176	race	white	[Larson et al. 2016]
Recidivism Male[6]	5,819	375	race	white	[Larson et al. 2016]
Titanic	1309	6	sex	female[7]	[Vanschoren et al. 2013]

3.2 Classification algorithms and evaluation

We used the following classification algorithms for the experiment: Classification Trees (CART) [Loh 2011], Multilayer Perceptron (MLP) [Hinton 1989], Random Forest (RF) [Breiman 2001], Support Vector Machines (SVM) [Cortes and Vapnik 1995], and eXtreme Gradient Boosting (XGB) [Chen et al. 2015]. We also used the well-known classification algorithm to minimize discriminatory effects called Adversarial Debiasing (AD) [Zhang et al. 2018]. Table II shows each classification algorithm and the numerical variation range for its hyperparameters used in this experiment. We tested fifteen parametrization settings per algorithm classification.

Table II: Algorithms and ranges of numeric variation defined for their parameters.

Algorithm	Parameters	Fixed Value	Variation Range (initial : final : step)
AD	Number of epochs for which to train	-	50 : 500 : 30
CART	The function to measure the quality of a split	gini	-
	The minimum of samples required to split an internal node	5	-
	The minimum of samples required to be at a leaf node	-	2 : 30 : 2
MLP	The number of neurons in the hidden layer	-	5 : 20 : 1
RF	The number of trees	-	100 : 500 : 25
	The minimum of samples required to split an internal node	$\sqrt{\#A}$	-
SVM	Kernel	rbf	-
	Regularization	1	-
	Gamma	-	0.0025 : 1.075 : 0.075
XGB	The number of trees	-	100 : 500 : 25

We use the multicriteria measure proposed in [Parmezan et al. 2017] to analyze the propagation of discriminatory effects in binary classification. Fig. 3 shows an example of the multicriteria measure, where three measures were selected in the evaluation. Thus, for each algorithm, the 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 the multicriteria measure is given by the sum of these areas.

In this work, the analysis of group fairness notions is prioritized, but considering the classifier performance. We want to identify the classifier that maximizes different notions of group fairness

[2]The protected attributes `age` and `gender` can be important in predicting health datasets, so they are used in class prediction. Nevertheless, this does not prevent the analysis of the difference in results between groups.
[3]We binarize the output between the absence and presence of cardiac arrhythmia, ignoring the different arrhythmia groups.
[4]We binarize the output to predict whether or not a woman uses contraception.
[5]In this dataset, it is possible to predict whether a person has never used or has used 18 different legal and illegal drugs.
[6]We split this dataset into two: Recidivism Female (female examples) and Recidivism Male (male examples).
[7]In this case, the selection bias is explicit, as women and children were favored in the rescue. Therefore, in the Titanic dataset, the protected attribute is used in prediction.

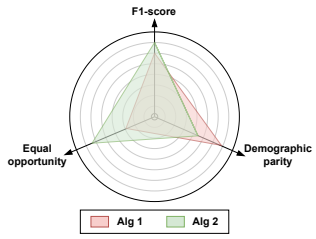

Fig. 3: Example of multicriteria measure: where the value of the multicriteria measure is given by the sum of the area of each irregular triangle formed by the meeting of the edges with vertices representing each pair of measurements.

without reducing performance; in other words, we want to identify the most impartial classifier in decision-making. Thus, the measures selected to compose the multicriteria measure are the F1-score, demographic parity, equal opportunity, and equalized odds. For the group fairness measures demographic equality, equal opportunity, and equalized odds, we calculated the ratio between the privileged and unprivileged groups to measure the disproportionality of the results, as discussed in Section 2. For ease of understanding, we use the highest score value as the denominator to calculate the ratio, so the ratio will always be between 0 and 1, with the optimum result equal to 1. Consequently, the best classifier has the highest value of the sums of the areas of measures produced by multicriteria metrics.

4. RESULTS

This section presents the results obtained in the evaluation process for the approach proposed in Section 3. We list the performance of all classifiers for each data stratification type and discuss the statistical difference aiming at the impact of discriminatory effects in relation to these explored data stratification types.

Table III reports the average result of the multicriteria measure in the test set of the best settings (considering the validation stage) per classifier for each data stratification type. The highlighted value indicates the most satisfactory average multicriteria score per dataset, while the value in parentheses is the standard deviation. For example, the data stratification by `(group, class)` performs best in nine of the twelve evaluated datasets and has the best average result among the tested data stratification types. Stratification by `(class)` obtained the best performance in three of the twelve datasets tested and had the second-best average result. However, not using data stratification did not get the best result in any dataset and had the worst average result. Thus, we noticed that the models were penalized, decreasing their performance and increasing the discriminatory effects of their predictions by not stratifying the data in the validation stage to select the best values of the hyperparameters.

Table III: Average multicriteria score, and standard deviation in parentheses, on the test set for the best hyperparameters of each classifier.

Dataset	Data Stratification Type			
	`(none)`	`(class)`	`(group)`	`(group, class)`
Alcohol	2.5560 (0.0745)	**2.5852 (0.0030)**	2.5825 (0.0067)	**2.5852 (0.0030)**
Arrhythmia	1.3140 (0.2220)	1.3571 (0.2623)	1.2087 (0.2799)	**1.3777 (0.3282)**
Bank	1.0687 (0.1409)	1.0860 (0.1423)	**1.0961 (0.1030)**	1.0584 (0.1587)
Census Income	0.9271 (0.1256)	0.9154 (0.0909)	0.9313 (0.1223)	**0.9333 (0.1290)**
Contraceptive	1.8167 (0.0825)	1.8667 (0.1112)	1.8684 (0.1611)	**1.9313 (0.1309)**
German	2.1140 (0.1297)	2.1361 (0.1121)	2.1424 (0.1429)	**2.1586 (0.1135)**
Heart	1.2557 (0.1327)	1.3549 (0.2625)	1.3542 (0.1684)	**1.4249 (0.1663)**
LSD	1.6796 (0.1527)	**1.7473 (0.1499)**	1.7015 (0.1237)	1.7373 (0.1733)
Nicotine	2.2274 (0.2043)	**2.2423 (0.2088)**	2.2168 (0.1827)	2.2270 (0.2075)
Recidivism Female	1.2161 (0.2004)	1.1800 (0.1833)	1.1891 (0.2162)	**1.2186 (0.2672)**
Recidivism Male	1.2308 (0.1310)	1.2513 (0.1329)	1.2740 (0.1177)	**1.2785 (0.1330)**
Titanic	0.5060 (0.3770)	0.3725 (0.1153)	0.5131 (0.3678)	**0.5165 (0.3671)**
Average	1.4927 (0.5981)	1.5079 (0.6225)	1.5065 (0.6057)	**1.5373 (0.6110)**

We apply a Nemenyi posthoc test [Demšar 2006] to the results to verify if there is a statistically significant difference in the multicriteria measure score among the data stratification types selected for the experiments. 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 stratification strategies, with the best-ranked algorithms to the left. A black line connects the algorithms when it is not detected a significant difference among them. For example, with a significance level of 10% (p-value < 0.10), the critical difference is 0.4930.

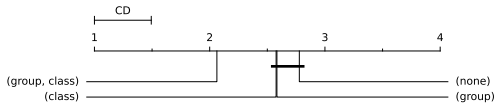

Fig. 4: Nemenyi posthoc test applied to the results of the multicriteria measure.

Fig. 4 shows data stratification by `(group, class)` ranked first with statistically significant differences for the other data stratification types. In contrast, there are no statistically significant differences between stratification by `(none)`, `(class)`, and `(group)`, but the non-use of data stratification was ranked last.

As we can note from the results of Table III and Fig. 4, using stratification by `(group, class)` in cross-validation to select the best hyperparameter values can make the classifier more impartial in its predictions. This experiment shows the importance of maintaining the original ratio, not only for classes but also for privileged and unprivileged groups. Incorporating this simple detail into model validation can significantly lessen discriminatory effects on binary classification tasks.

5. CONCLUSION

This paper introduced an experimental setup to analyze different types of data stratification in cross-validation to analyze the influence of stratification on model selection. Our goal was to associate which data stratification is more related to the choice of fairer or unfairer models. According to the experimental results, stratifying data by class and group of people (in the case of this paper: privileged and unprivileged groups) selects more impartial classifiers, which contributes to minimizing discriminatory effects in binary classification tasks. In conclusion, the findings of this study highlight that a simple yet effective stratification method, which has been investigated for decades, can serve as a straightforward pathway to incorporate fairness into machine learning models.

In future work, we intend to evaluate the generalization power of data stratification by group and class, extending the range of experiments by increasing the number of used classifiers combined with more hyperparameter values variations and dataset categories. Finally, we want explicit fairness notions groups individually and analyze the influence of stratification according to classification algorithm characteristics.

ACKNOWLEDGMENT

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. 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.

REFERENCES

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

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

Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. *Nips tutorial* vol. 1, pp. 2017, 2017.

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

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., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

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

Breiman, L. Random forests. *Machine learning* vol. 45, pp. 5–32, 2001.

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., and 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.

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

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. Xgboost: extreme gradient boosting. *R package version 0.4-2* 1 (4): 1–4, 2015.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning* vol. 20, pp. 273–297, 1995.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* vol. 7, pp. 1–30, Dec., 2006.

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

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

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

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

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

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

Hinton, G. E. Connectionist learning procedures. *Artificial Intelligence* 40 (1): 185–234, 1989.

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

Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*. Morgan Kaufmann Publishers Inc., pp. 1137–1143, 1995.

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

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

Loh, W.-Y. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1 (1): 14–23, 2011.

Martin Hirzel, K. K. and Ram, P. Engineering fair machine learning pipelines. *target* 73 (2.2): 1–028, 2021.

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

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

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

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

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

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