A Methodology based on Rebalancing Techniques to measure and improve Fairness in Artificial Intelligence algorithms

Ana Lavalle
Lucentia Research (DLSI)
University of Alicante (Spain)
alavalle@dlsi.ua.es

Alejandro Maté
Lucentia Research (DLSI)
University of Alicante (Spain)
amate@dlsi.ua.es

Juan Trujillo
Lucentia Research (DLSI)
University of Alicante (Spain)
jtrujillo@dlsi.ua.es

Jorge García
Lucentia Research (DLSI)
University of Alicante (Spain)
jorge.g@ua.es

ABSTRACT

Artificial Intelligence (AI) has become one of the key drivers for the next decade. As important decisions are increasingly supported or directly made by AI systems, concerns regarding the rationale and fairness in their outputs are becoming more and more prominent nowadays. Following the recent interest in fairer predictions, several metrics for measuring fairness have been proposed, leading to different objectives which may need to be addressed in different fashion. In this paper, we propose (i) a methodology for analyzing and improving fairness in AI predictions by selecting sensitive attributes that should be protected; (ii) We analyze how the most common rebalance approaches affect the fairness of AI predictions and how they compare to the alternatives of removing or creating separate classifiers for each group within a protected attribute. Finally, (iii) our methodology generates a set of tables that can be easily computed for choosing the best alternative in each particular case. The main advantage of our methodology is that it allows AI practitioners to measure and improve fairness in AI algorithms in a systematic way. In order to check our proposal, we have properly applied it to the COMPAS dataset, which has been widely demonstrated to be biased by several previous studies.

1 INTRODUCTION

The use of Artificial Intelligence (AI) systems is rapidly spreading across many different sectors and organizations. More and more important decisions are being made supported by AI algorithms. Therefore, it is essential to ensure that these decisions do not reflect discriminatory behavior towards certain groups. However, given the lack of an adequate methodology, creating fair AI systems has proven to be a complex and challenging task [16].

As it is becoming more and more used, big companies and governments are delegating responsibilities to AI systems which have not been thoroughly evaluated. In turn, some taken decisions have often been biased and unfair (e.g. the AI system from Amazon to qualify job applicants [22] or the granting of credit for the Apple credit card [20]). One of the most notorious cases where AI tools have acted in a biased and unfair way is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). This software has been used by judges in order to decide whether to grant parole to criminals or keep them in prison. The output is provided by an algorithm that evaluates the probability that a criminal defendant becoming a recidivist. Unfortunately, several studies have shown that the recidivism prediction scores are biased [1, 2]. This algorithm showed discriminatory behavior towards African-American inmates, which were almost three times more likely to be classified as high risk inmates than Caucasian inmates [1].

As a result of this trend, AI research communities have recently increased their attention towards the issue of AI algorithm's fairness. The IEEE Standards Association pays attention to the meaning and impact of algorithmic transparency [18]. Moreover, these issues are also aligned to the ethical guidelines for a trustworthy AI presented by the European Commission [8].Therefore, it is essential to ensure that the decisions made by AI solutions do not reflect discriminatory behavior.

Nevertheless, to the best of our knowledge, most of the approaches are mainly focused on improving the accuracy of algorithms in the prediction, while the fairness of the output is relegated to a second-class metric [5, 11, 14]. Thus, there has not been any proposal or methodology that guides the AI practitioners in choosing the best features to avoid unfair and discriminatory outputs from AI algorithms.

In this paper, we propose a methodology that considers fairness as a first-class citizen. Our methodology measures and evaluates the impact of the dataset rebalancing techniques on AI fairness. The novelty of our methodology is that it introduces new steps with respect to the traditional process of AI development such as: (i) a bias analysis, (ii) fairness definition and (iii) fairness evaluation. Moreover, another novelty of our methodology is that it helps to improve fairness by applying rebalancing approaches considering not only the target variable/s, but also sensitive attributes in the dataset that should be protected from discrimination. In order to both exemplify our approach and test the impact of each rebalancing alternative, we implement a classifier over the COMPAS dataset, calculating the degree of fairness obtained according to three different fairness definitions.

2 RELATED WORK

Bias can appear in many forms. [16] groups and lists different types of biases that can affect AI solutions according to where they appear: from **Data to Algorithm**, when AI algorithms are trained with biased data, the output of these algorithms might be also biased. From **Algorithm to User**, when bias arises as a result of an algorithm output it affects users' behavior. Or from **User to Data**, when data sources used for training AI algorithms are generated by users, historical socio-cultural issues can be introduced into the data even when perfect sampling and feature selection are carried out.

To tackle these situations, researchers have proposed different techniques that can be grouped into the next perspectives. **Data Perspective** when class distribution is artificially rebalanced by sampling the data. This rebalancing can be done by: Oversampling [14], creating more data in the minority classes. Undersampling [11], eliminating data from the majority classes or other like SMOTE [5], where minority classes are oversampled by interpolating between neighboring data points. However, these techniques must be used with tremendous care as they can lead to the loss of certain characteristics of the data. An alternative perspective is the **Algorithmic Perspective**, these solutions adjust the hyperparameters of the learning algorithms. Or, the **Ensemble Approach** that mixes aspects from both the data and algorithmic perspectives.

Most of these approaches mainly focus on improving the accuracy of algorithms in the prediction, while the fairness of the output is relegated to a second-class metric. As [21] states, accuracy is no longer the only concern when developing models. Fairness must be taken into account as well in order to avoid more cases as those presented in the introduction.

Moreover, as [9] argues, modifying data sources or restricting models in order to improve the fairness can harm the predictive accuracy. The fairness of predictions should be evaluated in the context of data. Unfairness induced by inadequate samples sizes or unmeasured predictive variables should be addressed through data collection, rather than by constraining the model [6].

Thus, differently from the above-presented proposals, we propose a novel methodology that considers fairness as a first-class citizen from the very beginning of the AI process. We drive the whole process considering protected attributes during the rebalance step and leading the AI practitioner to a conscious decision on the trade-off (if necessary) between accuracy and fairness.

3 IMPROVING FAIRNESS IN ARTIFICIAL INTELLIGENCE

Tackling AI challenges requires awareness of the context where algorithms will be not only trained, but also, where they generate outputs. Biases and errors that go unnoticed lead into wrong or unfair decisions. Moreover, since training AI algorithms is a time-consuming task (several days or weeks), developing them without a clear direction may result in considerable waste of resources.

For this reason, we propose the methodology shown in Fig. 1. By following this methodology, AI practitioners will be able to analyze and improve fairness in AI predictions.

The first step in our methodology (Fig. 1) starts with the definition of the **Target Variable** by AI practitioners. Then, during the **Bias Analysis** step, the algorithm proposed in [15] is executed in order to detect existing biases in the dataset. This algorithm output will provide an overview of how biased the attributes of the dataset are. Moreover, this information will help practitioners to select the **Protected Attribute/s** such as race, gender, or any other that requires special attention to ensure fair treatment. Whether protected attributes have been detected in the dataset, a **Definition of Fairness** will be launched in order to allow practitioners to measure whether the AI system is really being fair. Then, a **Data Rebalancing** (whether necesary) will be accomplished and AI practitioners will proceed to the **Algorithm Training**. Finally, we propose a set of tables and visualizations in order to interpret the **Algorithm Results**.

In the following, we will further describe all the steps of our methodology by applying it in a real case study.

3.1 Dataset Description

The dataset chosen in order to apply our methodology in a real case study has been the ProPublica COMPAS dataset available in [17]. This dataset includes information about criminal defendants who were evaluated with COMPAS scores in the Broward County Sheriff's Office in Florida, during 2013 and 2014.

For each accused (case), this dataset contain information related to their demographic information (race, gender, etc), criminal history and administrative information. Finally, the dataset also contains information about whether the accused was really a recidivists or not in the next 2 years. This dataset is highly imbalanced, the representation of the different races is heavily skewed. Then, we will apply our methodology step by step.

3.2 Target Variable Definition

The first step of our proposed methodology is to define the target variable. In this case, the target variable is "v_score_text" which uses 3 attributes (Low, Middle, High) to classify the risk of recidivism. For the sake of simplicity, we will binarize the target variable by mapping the Low class to Non-Recividist, and the Middle and High classes to the Recividist class thereby facilitating following the analysis presented. Therefore the **Target variable** is defined as Risk of recidivism (0 Non-recividist, 1 Recividist).

3.3 Bias Analysis

The second step is to perform a Bias Analysis. As previously-summarized in Section 2, and according to [16], the different data bias that can be used in our case study context are (i) Data to Algorithm, (ii) Algorithm to Use and (iii) User to Data. As in our particular case, we are analyzing how biased data sets affect AI algorithms, we will apply the **Data to Algorithm** bias.

In order to analyze how data bias affect the behavior of AI algorithms, firstly we apply our previously published algorithm [15] that automatically detects and visualizes bias in data analytics.

This algorithm examines the dataset returning us as output a number between 0 and 10 that establishes the bias ratio of the attributes (being 0 equally distributed and 10 very biased). This number is visually represented in order to present an overview of the data bias for a better understanding and exploration.

In this case of study, the bias ratios were (Race: 9.95, Sex: 7.60 and Age category: 6.28), the most biased attribute was race and it was selected as a protected attribute. The main reason is that the race of the accused should never be a characteristic that influences the classification of risk of recidivism (the target variable). Therefore the **Protected attribute** is defined as Race.

Furthermore, a visualization (Fig. 2) that groups the predicted target variable (risk of recidivism) by the attributes selected as protected (race) is created. As cleary observed, there is a high risk in accused of African-American race than in the rest of races.

Once the dataset has been analyzed and the bias has been located, AI practitioners will have more detailed knowledge in order to detect the types of bias that might arise. Among the types of bias which can appear, those relevant for our methodology are categorized in Data to Algorithm bias as described by [16]:

- **Measurement Bias:** Arises when we choose and measure features of interest. If a group is monitored more frequently, more errors will be observed in that group.
- **Omitted Variable Bias:** When important variables are left out of the model.
- **Representation Bias:** Arises in the data collection process when data does not represent the real population.

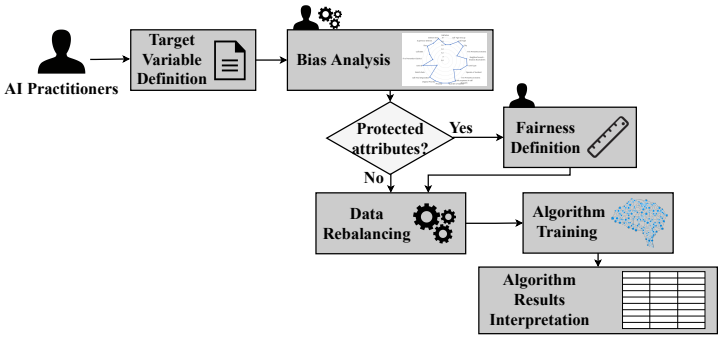

Figure 1: Methodology to mitigate bias in AI algorithms

- **Aggregation Bias:** When false conclusions are drawn about individuals from observing the entire population. Data from several groups (i.e. cities, races, age groups, etc.) can be correlated differently across classes. However, if an aggregation is performed, the general correlation of the aggregated data could be completely different from the earlier correlations.
- **Sampling Bias:** Trends estimated for one population may not generalize to data collected from a new population.
- **Longitudinal Data Fallacy:** When temporal data is modeled using a cross-sectional analysis, which combines multiple groups at a single point in time.
- **Linking Bias:** When network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users.

In this specific case study, the bias analysis leads to consider as potential biases both **Measurement Bias**, some individuals tend to live in zones with high criminal activity, hence a higher level of surveillance by the police is needed and it could derive into a feedback loop. And **Representation Bias** since data presents a significantly different distribution compared to the demographic distribution of Florida state [3] (where the data was collected).

3.4 Fairness Definition

The presence of bias can eventually derive into unfair results, especially when the bias is present in protected attributes. Thus, analyzing which biases might be present in the current problem is essential to determine which fairness metrics are more important.

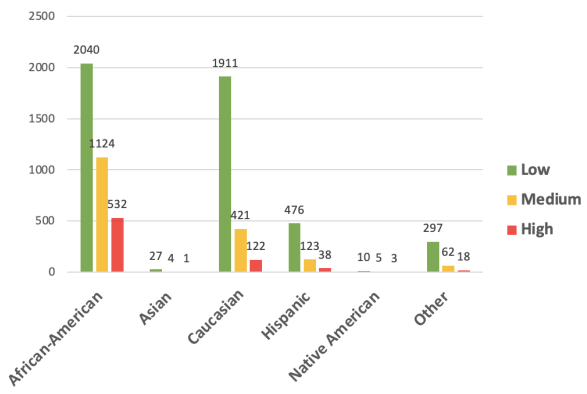

Figure 2: Risk of recidivism score by race

In our case study, the race attribute was considerably biased. As this is a protected attribute, it is important to define one or more metrics that quantify the fairness of the results.

As [19] argues, in general terms, fairness can be defined as the absence of any prejudice or favoritism towards an individual or a group. However, although fairness is a quality highly desired by society, it can be surprisingly difficult to achieve in practice [16].

Therefore, with the aim of defining, limiting and being able to measure whether fairness is being achieved, our proposed method makes AI practitioners reflect on the type of justice that they want to achieve. Among the types of justice we can find:

- **Individual Fairness:** Give similar predictions to similar individuals [10], i.e. points that are closer to each other in the feature space should have similar predictions.
- **Group Fairness:** Treat different groups equally [10].
- **Subgroup Fairness:** Try to obtain the best properties of the group and individual notions of fairness. It picks a statistical fairness constraint (like equalizing false positive rates across protected groups) and asks whether this constraint holds over a large collection of subgroups [13].

In this case of study, "Group Fairness" has been selected, since the race attribute has been selected as protected and fairness is sought between the different groups of races. Specifically, the following definitions of "Group Fairness" have been followed:

- **Equalized Odds:** Groups within protected attributes must have the same ratio of true and false positives [12]. As equality of odds can be really difficult to achieve, it can be decomposed into two more relaxed versions:
- **Equal Opportunity:** Groups within protected attributes must have equal true positive rates [12].
- **Predictive Equality:** False positive rates must be equal across all groups of the protected attribute [4].

Depending on the problem, one definition could be more important than the other. For example, in building a model to predict if a subject is eligible for a grant, it is relevant for the rate of true positives of both sexes to be equal, i.e. equal opportunity should be achieved. On the other side, a risk assessing model should focus on having the same false positive rates across protected groups, as missclassifying an individual as high risk can be really harmful, hence the importance to prioritize predictive equality.

3.5 Data Rebalancing

By choosing and studying which fairness metrics are more suitable for the current problem, AI practitioners are now able to focus on applying several techniques and evaluate its impact based on these fairness definitions.

In this case, different data rebalancing techniques will be used to modify the dataset distribution in terms of race and recidivism rate, in order to assess its impact in terms of fairness.

Usually, data rebalancing techniques are used in problems of imbalanced classification, where the target variable to be predicted has a majority and a minority class.

In this case study, the dataset could be rebalanced to be composed of 50% non-recidivists and 50% recidivists, which is the target variable. However, this approach does not take into account the different groups where fairness has to be assessed and preserved. Therefore, as an alternative view on the problem, we propose to treat the bias and unfairness in the protected attributes as a rebalancing problem. In this sense, we extend the rebalancing methods to consider the protected attribute in addition to their associated target variable, thus allowing to control the proportion of each group in the sample.

In other words, by extending the rebalancing techniques, the dataset of this case study can be modified as follows: 25% African-American non-recidivists, 25% African-American recidivists, 25% Caucasian non-recidivists and 25% Caucasian recidivists.

As there are several techniques for rebalancing, in this case study we will focus on three different data rebalancing techniques: **Undersampling** [11], **Oversampling** [14], and **SMOTE** [5].

3.6 Algorithm Training

In this case, the XGBoost [7] classifier has been used with the default hyperparameters. In order to complement the experiments related to rebalancing using the previous techniques, three extra experiments have been carried out to provide further insights:

- **Baseline:** It is important to evaluate the model obtained without applying any rebalancing so that it acts as a baseline model in order to compare the results.
- **Split by race:** Two separate classifiers will be trained, one for each of the race studied.
- **Remove race attribute:** Same experiment as baseline, but omitting the race attribute.

Regarding the accomplished experiments, the whole training process can be described as follows: (1) Split the dataset into a training and test sets, (2) Rebalance the dataset by using the forementioned techniques (depending on the experiment, either the training set or both sets are rebalanced), (3) Train the classifier to predict the risk of recividism given variables such as sex, age, race and prior criminal history of the subject, and (4) Once the classifier is trained, it is evaluated on the test set by computing the metrics above-mentioned.

In total, nine experiments will be performed: the baseline, training one separate model for each race, completely omitting the race variable, and six related to rebalancing either the training set or both the training and test set, with each of the rebalancing techniques presented: undersampling, oversampling and SMOTE. The code of the experiments is publicly available in https://gitlab.com/lucentia/DOLAP2022.

3.7 Algorithm Results Interpretation

Finally, in order to compare the output of the XGBoost classifier algorithm and to be able to measure if it has been fair, we have created Table 1 and Table 2. It should be noted that this tables can be easily replicated in any Artificial Intelligence challenge.

First, Table 1 represents the True Positive Rates (TPR) and False Positive Rates (FPR) for Caucasian and African-American groups. In this specific case, False Positive Rates (FPR) were the most sensitive classification, since classifying non-recidivists as a high risk of recidivism can bring them negative consequences.

As we can see in Table 1, the techniques that achieve the best FPR for the Caucasian race are *Original Train - Original Test* and *Remove race attribute* with a 0.172 rate. Meanwhile, *Remove race attribute* obtains the best FPR for African-American race with a 0.347 rate. It is remarkable how the Caucasian race obtains the best results when the data is original, while the African-American race obtains the best results when the race attribute is removed.

However, even though using these techniques we get better False Positive rates, the difference between getting a 17,2% of Caucasian defendants wrongly accused as a recidivists and that the 34,7% of African-American defendants were wrongly accused as a recidivists would still be considered highly unfair.

Additionally, our methodology generates Table 2 that calculates and compares the fairness definitions chosen in Section 3.4. Using Table 2 is possible to know, depending on the type of fairness pursued, which technique will bring better results. We have marked the best (green) and worst (red) techniques for each definition of fairness and for the overall accuracy. We should clarify that a lower fairness number represents that there is less difference between the protected groups, i.e. it is more fair. However, the accuracy is better when its value is higher, since it means that there have been fewer errors in the classification.

As we can observe, the technique that gets the best score in terms of Equal Opportunity, Equalized Odds and Accuracy is to remove the protected attribute, in this case the attribute race. However, other highly used techniques as SMOTE gets the worst results in terms of Equal Opportunity and Equalized Odds.

4 CONCLUSIONS AND FUTURE WORK

The use of Artificial Intelligence (AI) systems is rapidly spreading across different sectors and organizations. More and more important decisions are being made supported by AI systems which have not been thoroughly evaluated. It is essential to ensure that these decisions do not reflect discriminatory behavior towards certain groups. Nevertheless, most of the approaches mainly focus on improving the prediction accuracy of algorithms without considering fairness in their development.

Thus, in this paper we have presented a methodology that allows AI practitioners to measure and improve fairness in AI algorithms in a systematic way. Our novel methodology considers fairness as a first-class citizen and introduces new steps with respect to the traditional process of AI development such as: (i) a bias analysis, (ii) fairness definition and (iii) fairness evaluation. We have also analyzed how the most common data rebalancing approaches affect the fairness of AI predictions taking into account both (i) the target variable and (ii) the protected attributes. Furthermore, our methodology generates a set of tables for choosing the best rebalancing alternative for each particular definition of fairness. Both our methodology as well as the interpretation of the algorithms results (tables and visualizations) can be easily replicated in any AI algorithm.

In order to both exemplify our approach and test the impact of each rebalancing alternative, we have applied it in a real case of study. We have implemented a classifier over the COMPAS dataset, calculating the degree of fairness obtained according to three different fairness definitions.

Given the obtained results, we consider that by following our proposed methodology we can avoid falling into the usual pitfalls that lead to controversial outputs when the input datasets include

Table 1: True Positive and False Positive Rates for African-American and Caucasian races

	TPR Cauc.	TPR Afr.	FPR Cauc.	FPR Afr.
Original Train - Original Test	0.356	0.714	**0.172**	0.381
SMOTE Train - Original Test	0.340	0.716	0.178	0.384
SMOTE Train - SMOTE Test	0.294	0.716	0.188	0.391
Over Train - Original Test	0.397	0.701	0.215	0.370
Over Train - Over Test	0.372	0.701	0.206	0.380
Under Train - Original Test	0.371	0.721	0.188	0.418
Under Train - Under Test	0.371	0.722	0.227	0.454
Split by race	0.371	0.703	0.198	0.372
Remove race attribute	0.407	0.674	**0.172**	**0.347**

Table 2: Fairness rates comparison

	Eq. Oportunity	Pred. Equality	Eq. Odds	Accuracy
Original Train - Original Test	0.358	0.209	0.567	0.659
SMOTE Train - Original Test	0.376	0.206	0.582	0.654
SMOTE Train - SMOTE Test	**0.422**	0.203	**0.625**	0.608
Over Train - Original Test	0.304	**0.155**	0.459	0.654
Over Train - Over Test	0.328	0.174	0.503	0.622
Under Train - Original Test	0.350	**0.230**	0.580	0.649
Under Train - Under Test	0.351	0.227	0.577	**0.603**
Split by race	0.332	0.174	0.506	0.654
Remove race attribute	**0.267**	0.175	**0.442**	**0.664**

biased protected attributes. In addition, it allows us to discover which is the most appropriate data rebalancing techniques to try to maximize different definitions of fairness.

Regarding the limitations of our proposal, we should take into account that our proposal has achieved successful results when protected attributes are individual and binary. However, when as the number of protected attributes increases, rebalancing becomes more difficult. Future work is needed in order to study the best approach to carry out rebalancing techniques in the cases where there are several protected attributes defined and the classes contain a large number of different attribute groups.

ACKNOWLEDGMENTS

This work has been co-funded by the AETHER-UA project (PID 2020-112540RB-C43), funded by Spanish Ministry of Science and Innovation and the BALLADEER (PROMETEO/2021/088) projects, funded by the Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital (Generalitat Valenciana).

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Matias Barenstein. 2019. ProPublica's COMPAS Data Revisited. *CoRR* abs/1906.04711 (2019). arXiv:1906.04711

[3] The U.S. Census Bureau. 2010. Population percent change. https://www.census.gov/quickfacts/FL.

[4] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, and Daniele Regoli. 2021. The zoo of Fairness metrics in Machine Learning. *CoRR* abs/2106.00467 (2021).

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory?. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc.

[7] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug 2016).

[8] European Commission. 2021. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. (2017), 797–806.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. Association for Computing Machinery, 214–226.

[11] Salvador García and Francisco Herrera. 2009. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary computation* 17, 3 (2009), 275–306.

[12] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[13] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 2564–2572.

[14] György Kovács. 2019. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing* 83 (2019), 105662.

[15] Ana Lavalle, Alejandro Maté, and Juan Trujillo. 2020. An Approach to Automatically Detect and Visualize Bias in Data Analytics. In *Proceedings of the 22nd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, DOLAP@EDBT/ICDT 2020*, Vol. 2572. CEUR-WS.org, 84–88. http://ceur-ws.org/Vol-2572/short11.pdf

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[17] Broward County Clerk's Office, Broward County Sherrif's Office, Florida Department of Corrections, and ProPublica. 2021. COMPAS Recidivism Risk Score Data and Analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

[18] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

[19] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[20] Neil Vigdor. 2019. Apple Card Investigated After Gender Discrimination Complaints. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

[21] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *CoRR* abs/1807.00199 (2018). arXiv:1807.00199

[22] Jordan Weissmann. 2018. Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women. https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html.