Identifying risks in datasets for automated decision–making

Mariachiara Mecati$^{1[0000-0002-0041-1809]}$, Flavio Emanuele Cannavò^{1}, Antonio Vetrò$^{1[0000-0003-2027-3308]}$, and Marco Torchiano$^{1[0000-0001-5328-368X]}$

Politecnico di Torino, Turin, Italy

Abstract. Our daily life is profoundly affected by the adoption of automated decision making (ADM) systems due to the ongoing tendency of humans to delegate machines to take decisions. The unleashed usage of ADM systems was facilitated by the availability of large-scale data, alongside with the deployment of devices and equipment. This trend resulted in an increasing influence of ADM systems' output over several aspects of our life, with possible discriminatory consequences towards certain individuals or groups. In this context, we focus on input data by investigating measurable characteristics which can lead to discriminating automated decisions. In particular, we identified two indexes of heterogeneity and diversity, and tested them on two datasets. A limitation we found is the index sensitivity to a large number of categories, but on the whole results show that the indexes reflect well imbalances in the input data. Future work is required to further assess the reliability of these indexes as indicators of discrimination risks in the context of ADM, in order to foster a more conscious and responsible use of ADM systems through an immediate investigation on input data.

Keywords: Bias, data quality, data ethics, imbalance measures, algorithm fairness

1 Introduction: background and motivations

Our daily life is profoundly affected by the development and adoption of automated decision making (ADM) systems [15]. This is due to the ongoing tendency of humans to make decisions based on software-elaborated recommendations or even to entirely delegate decision-making to machines. The adopted technical approaches range from sophisticate neural networks to simpler software systems that calculate and sort data according to predefined sets of rules.

A crucial enabling factor for these systems is the wide availability of data: ADM systems are widely used to predict behaviours and classify individuals depending on patterns extracted from the data collected about them or other persons. The growing employment of these systems gives rise to both opportunities and risks at the same time. Opportunities usually concern improved efficiency of the automated decision processes; on the other hand, one of the main risks is represented by data and algorithm bias, which usually induces systemic discrimination. Generally speaking, discrimination can be defined as an "unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category" [10]. Since biased software is software that exposes a group (e.g. an ethnic minority, gender, or type of worker) to an unfair treatment [21], an algorithm - often in order to achieve its optimization purposes - might discriminate and filter between people under consideration, with the result of a disparate impact on different population groups.

A large amount of evidence of discrimination by ADM systems has been recently collected in both scientific literature and journalistic investigations. Herein we rely on such body of evidence and we select only a few cases of discrimination caused by ADM systems; we are not proving here a complete review of the literature on automated discrimination, simply, we wish to highlight their impact on citizens' life. In a recent analysis of **risk assessment for juvenile justice** in Catalonia [22], the machine learning model marked male defendants and people of some specific national group as recidivist more frequently than others. Based on this discrimination problem, researchers suggested a method to assess predictive performance and unfairness in Machine Learning algorithms employed in the prediction of juvenile recidivism; then, the obtained results have been compared to Structured Assessment of Violence Risk in Youth (SAVRY), a widespread risk assessment tool employed to assess the risk of violence in juvenile justice. Researchers proposed two metrics in order to evaluate fairness: demographic parity and error rate balance. They discovered that machine learning algorithms become discriminatory when adopting SAVRY demographic features: male defendants were more likely to be classified as recidivists and foreigners were more likely to be labelled as high risk although they were non-recidivists.

Another example is represented by the "**Black box Schufa**" [18]. Schufa, which is the most well-known credit agency in Germany, asserts to have information on more than 67 million consumers and to output a score for each of them. Telecom providers, retailers, and even banks rely on these scores to support their business, ranging from determining which customers might get a loan, to which user get to see a certain ad. Thanks to a crowdsourcing project which involved $2,800$ volunteers who asked Schufa for their free personal credit report, researchers reverse-engineered how Schufa works and found that younger people are often evaluated worse than older people. The same happens to males, worse ranked than females. The problem derives from the fact that the General Equal Treatment Act, whose purpose is to protect consumers from discrimination based on gender and age, is ineffective with regard to credit bureaus, so age and gender are legally but unfairly included in the score.

A third representative case is in the field of image classification, and specifically **facial recognition systems**, which has collected a lot of critics not only for the problem of discrimination, but also for the technology per se. The case concerns commercial gender classification: in [8] the authors revealed how automated facial image analysis is affected by performance disparity in gender and race: in particular, the gender classification on female faces works significantly worse than classification on male faces, and performance are better on lighter skin tones than darker ones.

Lastly, it is worthy to mention the "Report of the Special rapporteur on extreme poverty and human rights" [1], released by the United Nations clearly disapprove the way governments are actually automating welfare management, because the collected evidence showed that these systems systematically discriminate the weakest segments of society and exacerbate existing inequalities.

To conclude, the diffusion of ADM systems in a wide range of application domains has raised serious concerns about discriminatory impact towards certain individuals or social groups. Despite existing anti-discrimination laws in several countries forbid – for certain business and government services – unfair treatment of people based on the so-called sensitive attributes (i.e. specific traits of a person such as gender and race), this is not enough to mitigate the problem: fairness and bias in ADM systems remain an open and relevant issue [17]. These problems emerge in large part due to imbalanced datasets [6], mainly because machine learning systems search for certain patterns in the input data and apply them to new data.

In order to investigate this crucial aspect, in this paper we lay the foundations of a risk assessment approach based on quantitative measures to evaluate imbalance in the input datasets of ADM systems. Specifically, we use two indexes of heterogeneity and diversity to identify imbalance in two datasets that lead to disparate impact when used to feed an ADM. Our preliminary observations confirm that the indexes can be used as indicators of risks in datasets for ADM. The paper is organized as follows. In section 2 we discuss the idea of imbalance metrics as risk indicators, motivated by the conceptual framework provided by the series of ISO/IEC standards on Software Quality Requirements and Evaluation (SQuaRE). In Section 3 we describe our approach, the selected metrics and the datasets we employed to conduct our exploratory study. Then, in Section 4 we report the results and discuss then. We take into account the limitations of this approach in Section 6 and, eventually, we highlight conclusions and potential future work in Section 7.

2 Imbalance in datasets as risk indicator

Discrimination carried out by ADM systems is often due to imbalance in the frequencies of certain sensitive attributes in the datasets used as input (e.g., for training a machine learning algorithm) [6]. This chain of effects can be interpreted in light of the conceptual model of the series of standards ISO/IEC 25000, also known as SQuaRE [12], where both the internal software quality and the data quality have an impact on the external software quality, which in turn impact the users -active or passive- in the contexts in which the software is employed (quality in use). Figure 1 represents the chain of effects formalized in SQuaRE. A simplification of this concept is the well-known GIGO principle (i.e. "garbage in, garbage out"), which states that flawed input data produces garbage as output.

We use this concept and we propose that bias in the input datasets should be measured because it has the same propagation effects that data quality is-

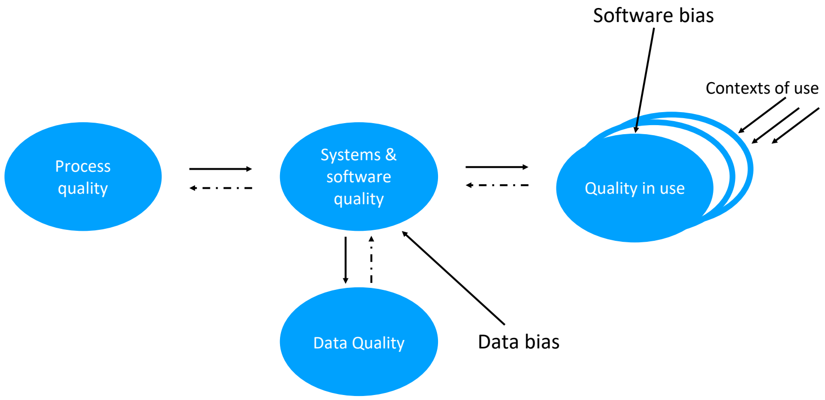

Fig. 1: Data/software bias in the context of quality effects conceptualized in SQuaRE

sues have, and can cause biased outputs, as most of today software-automated decisions are based on the analysis of historical data.

In practice, imbalanced datasets may lead to imbalanced results, generating problems of representativity when the data are sampled - therefore leading to an underestimation or an overestimation of the population groups - and of imbalance when the dataset used has not been generated using classical sampling methods.

Taking as reference this simple conceptual framework, we propose a metric-based approach to evaluate imbalance in a given dataset as a proxy of risk of biased output from ADM systems.

3 Exploratory study

We conducted a study aimed at answering the following research question:

> *Are imbalance measures on a dataset able to reveal a discrimination risk when an ADM is trained with such data?*

3.1 Metrics

In this study we focus on *categorical* and we propose two indicators of imbalance in data. The first one is the Gini index [9], a measure of *heterogeneity* used in many disciplines and often discussed with different designations: examples are political polarization, market competition, ecological diversity. Heterogeneity reflects how many different types (such as protected groups) are represented. In statistics, the heterogeneity of a discrete random variable R, which assumes m categories c_j with frequency f_j (with $j = 1, ..., m$), varies between a degenerate (=minimum value of heterogeneity) and an equiprobable case (=maximum value of heterogeneity, since categories are all equally represented). This means that for a given m, the heterogeneity increases if probabilities become as equal as possible, i.e. the different protected groups have similar representations. The Gini index is computed as follows:

$$I = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} f_i^2\right) \tag{1}$$

where we multiply by $\frac{m}{m-1}$ in order to normalize the index. According to the formula, the closer the index to 1 and the higher the heterogeneity is (i.e. categories have similar frequencies), and viceversa index closer to 0 means more concentration of frequencies in few categories, thus lower heterogeneity.

The second aspect used to measure imbalance is the *diversity*. Diversity indexes provide information about community composition taking the relative amounts of different species (classes) into account. We use the Shannon index, which is a measure of species diversity in a community and is calculated in this way: the proportion of species j relative to the total number of species (p_j) is computed, and then multiplied by the natural logarithm of this proportion ($\ln p_j$). The resulting product is summed across species and multiplied by -1. Then, we divide by $\ln m$ in order to normalize the index as follows:

$$H = -\left(\frac{1}{\ln m}\right) \sum_{j=1}^{m} p_j \ln p_j \tag{2}$$

Values of normalized Shannon index close to 1 indicate higher diversity (classes have similar frequencies) while values closer to 0 indicate less diversity (because frequencies are concentrated in fewer classes). To summarize:

> the closer the Gini/Shannon index to 0, the more data is affected by imbalance and the higher is the risk that such imbalance would cause effects in the output of an ADM system.

3.2 Data

We tested two datasets, each referring to a different application domain.

- **Credit card default dataset**. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005 [14]. The dataset is composed by 25 variables of which four have demographic character and can be considered as protected attributes (i.e.: sex, age, education and marital status).
- **COMPAS Recidivism racial bias dataset**. Data contains variables used by the COMPAS algorithm in scoring criminal defendants in Broward County (Florida), along with their outcomes within two years of the decision. The original dataset contains 28 variables; in particular, eight of such variables

are considered as protected attributes: first name, last name, middle name, sex, race, date of birth, spoken language, marital status. [2].

We chose the COMPAS dataset because it is well-known it the scientific communities that study measures of algorithmic bias and related mitigation strategies. It was provided by the U.S. non-profit organization ProPublica that showed that the COMPAS algorithm is distorted in favor of white individuals, thus exposing black people to a risk of distorted recidivism. ProbPublica investigation [1] showed that one of the motivations for discriminations was that input data is highly imbalanced (e.g., black defendants are many more than white defendants).

The credit card default dataset was chosen because of the high impact of using ADM software in this domain (see motivations in the Introduction), and that particular dataset because of popularity: at the time of the research, it was ranked as the second most voted dataset on credit cards on Kaggle[2] and it fits better our study than the one ranked first (Credit Card Fraud Detection), which is based on transactions, while we are interested on datasets that collect data on persons. Differently than COMPAS, the credit card default dataset does not contain a pre-computed classification, so we trained a classifier with a portion of the data and ran it on the remainder, observing also in this case a problem of discrimination (although less evident than in COMPAS): details are reported in the next section and for reproducibility purposes we share data, code and environment in a permanent location [16].

4 Results and Discussion

We report results of applying the indexes on selected columns of the datasets in Table 1: the first two columns indicate the input dataset and sensitive attributes, while Gini index and Shannon index are reported in the third and fourth columns. Note that we normalize these two indexes between 0 and 1 in order to ensure their comparability, and we exclude missing values (NA) from our analysis. We report also the distributions of each class, as basis for interpretation and discussion in figures from 2 to 5.

For the credit card dataset, we report data for the following variables: sex, education, marital status and age, expressed as a percentage for each of their categories. For the COMPAS dataset we report data for the attributes ethnicity, sex and age category (which includes three age classes, i.e. "Less than 25", "25-45", "Greater than 45").

The rightmost two columns of Table 1 report the fairness test based on the separation criterion (equivalence of true positive and false positive for each level of the protected attributed under analysis, as formalized in [4]) and computed

[1] `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm` , last visited on May 29, 2020

[2] `https://www.kaggle.com/datasets?search=credit+card&sort=votes` , last visited on May 29, 2020

with respected to the protected attributes which lowest indexes (hence, higher risk of discrimination if bias is propagated): “marital status” for the credit card default dataset and “ethnicity” for COMPAS, which has a tie with “sex” but it is preferred because of the findings of the Pro Publica study. We found that the separation criterion was only partially met in the case of “marital status” -no difference between False Positive rates, but 8% difference for true positive- and not met at all in the case of “ethnicity” in COMPAS: computation is reproducible at [16].

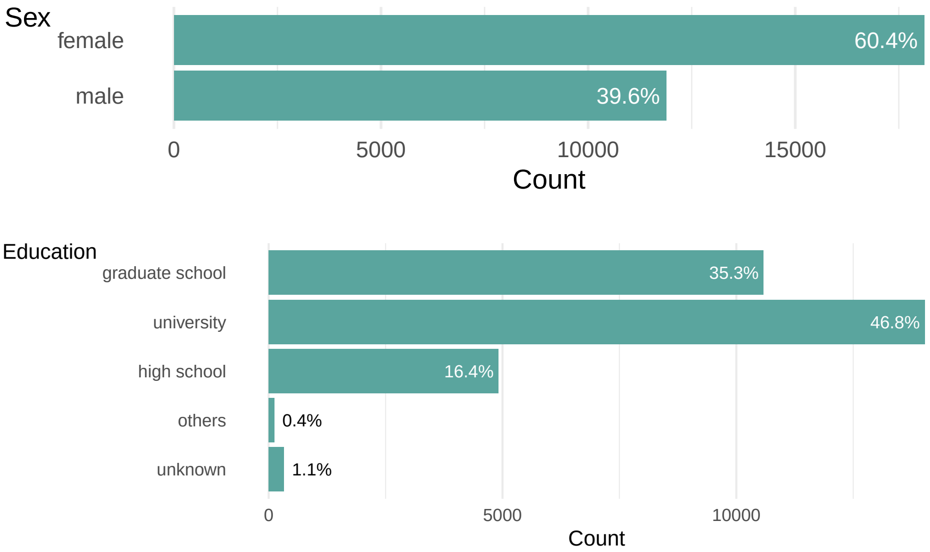

Fig. 2: Frequency histograms for the classes of the protected attributes SEX and EDUCATION in the Credit card default dataset.

Table 1: Gini index and Shannon index for each protected attribute in the Default of credit card clients dataset and in the COMPAS dataset.

Dataset	**Protected attribute**	**Gini index**	**Shannon index**	**Fairness (separation)**	
				difference TP rates	**difference FP rates**
Default of credit card clients	*sex*	0.96	0.97		
	education	0.79	0.68		
	marital status	0.76	0.68	0.08	0
	age	0.98	0.88		
COMPAS	***ethnicity***	0.73	0.62	0.23	0.18
	sex	0.62	0.70		
	age category	0.87	0.89		

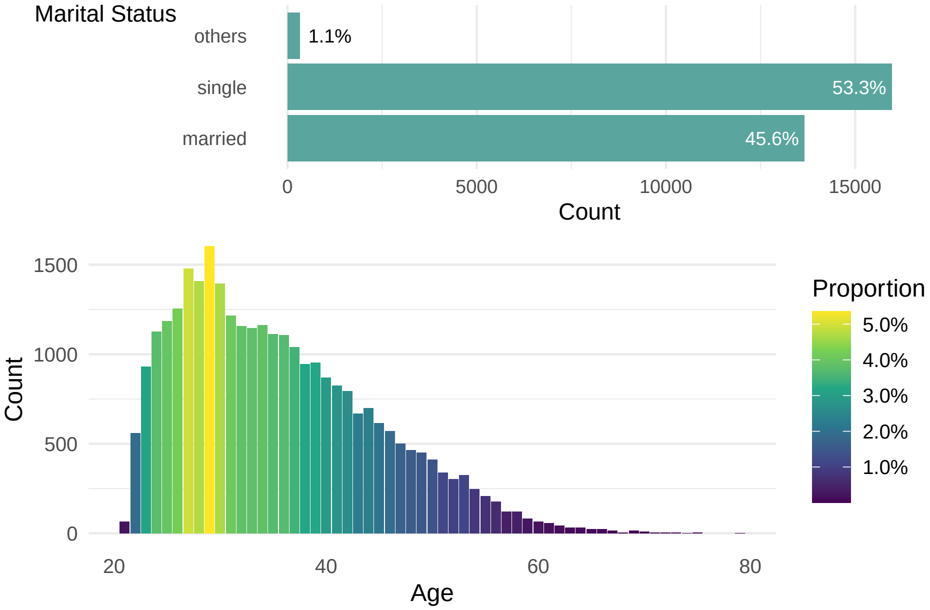

Fig. 3: Frequency histograms for the classes of the protected attributes MARITAL STATUS and AGE in the Credit card default dataset.

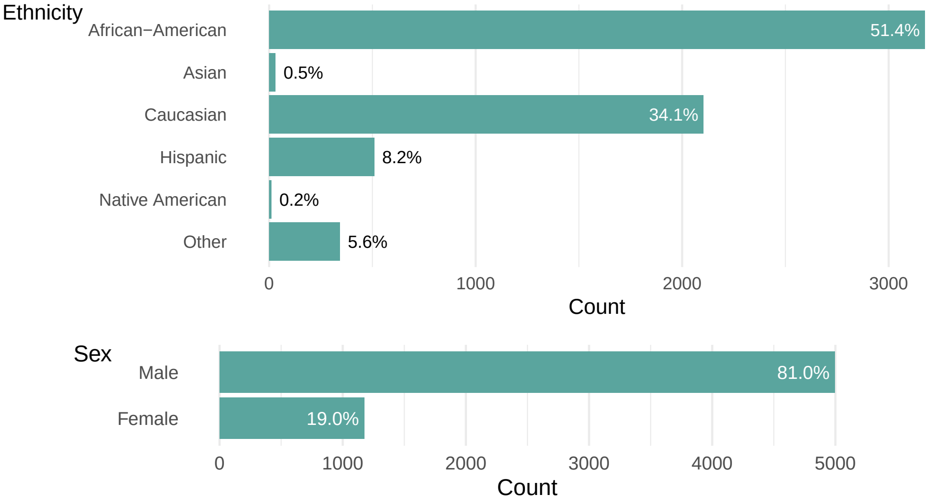

Fig. 4: Frequency histograms for the classes of the protected attributes ETHNICITY and SEX in the COMPAS dataset.

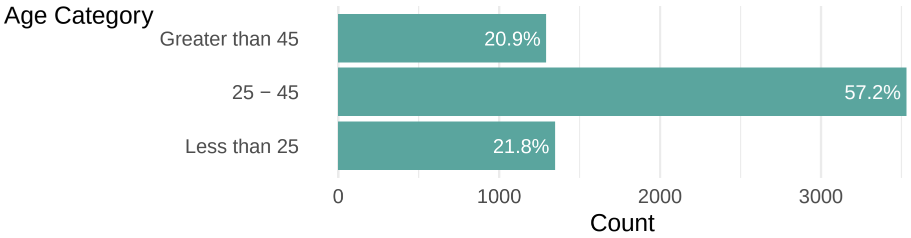

Fig. 5: Frequency histogram for the classes of the protected attribute AGE CATEGORY in the COMPAS dataset.

Credit card dataset. In the literature, issues related to ethical decisions often appear alongside the field of creditworthiness [19][23]. For this reason, some studies have been conducted and have recently shown that access to credit is indirectly modulated by certain attributes such as race, rather than by information about the payer's status [5][7]. The dataset that we analyzed does not contain the protected attribute race, but contains other personal information, notably sex, level of education, age, marital status. The data show that 60.4% of individuals are women, 46.8% of individuals have attended university, the proportion of single individuals is predominant, and the most represented age is 29 years old. Looking at the single classes, for the protected attribute "**sex**" we found very high and similar indexes, suggesting a certain balance between the two classes (around 0.6 and 0.4 for female and male respectively). Then, we observe that categories "**education**" and "**marital status**" have most of data concentrated in certain categories (three out of five for education, two out of three for marital status): the corresponding two indexes assume values between 0.68 and 0.79, thus reflecting less heterogeneity and diversity than variable "sex". Finally, concerning the protected attribute "**age**", we note the presence of several categories with very different frequencies, which would lead to low values for Gini and Shannon indexes. By contrast, we obtained very high indexes: a feasible explanation is given by the fact that a very large number of categories is taken into account; as a consequence, occurrences are spread on all these categories, so that none of them present a distinct frequency on the whole. Indeed, in our frequency histogram we observe the typical shape of a right-skewed distribution. So, looking at such normal distribution, we can see the histogram as composed by several categories with high but similar frequencies (in the central region between around 20-25 and 40-45 years old), which may explain the high value indexes. Therefore, in presence of a large number of categories the behaviour of the indexes could be misleading and it would be recommended to aggregate - in a meaningful manner for the context at hand - the classes in fewer groups.

COMPAS Recidivism racial bias dataset. Previous research has shown that the data in the COMPAS dataset is imbalanced in favor of white people and our investigation confirms the previous study [13]: the highest levels of reoffending are observed in black individuals. Indeed, analysis show very imbalanced data considering any of the selected protected classes. Concerning "**ethnicity**", about 34% of the dataset's observations refer to white people, while 51.4% refer to black people, indicating that there may be an overestimation of the race attribute - against black people - which would contribute to the estimation of recidivism, as suggested by the two indexes: 0.73 and 0.62 respectively for Gini and Shannon confirmed medium level of heterogeneity and diversity. With respect to the protected attribute "**sex**", we note very different frequencies (81% and 19% for male and female respectively) that result in low value indexes, which reveal low heterogeneity and diversity. Finally, as regards "**age category**" we obtained 0.87 and 0.89 respectively for Gini and Shannon indexes: indeed, we observe that the distribution is principally concentrated on the class "25-45" (57.2%) but is almost equally distributed on the other two categories ($\sim$ 21% each one), so the indexes tend to be higher than expected.

5 Related work

To date, approaches similar to ours are in the direction of labelling datasets for ethical indication purposes. One consists in a collaboration between MIT Media Lab and the Berkman Klein Center at Harvard University, which resulted in "The Dataset Nutrition Label Project" [20]. This initiative aims at avoiding that flawed, incomplete, skewed or problematic data have a negative impact on automated decision systems.

A similar approach is the "Ethically and socially-aware labeling" (EASAL) [6], in which authors propose a conceptual and operational framework to label datasets and identify possible risks of discriminatory output when used in decision making or decision support systems. Thus, it aims to plan and develop datasets metadata to help software engineers to be aware of the risks of discrimination.

Yet another labelling approach is proposed by Gebru et al., "Datasheets for Datasets" [11]: with respect to the previous proposals, this research work consists of more discursive technical sheets for the purpose of encouraging an increasingly clear and comprehensive communication between users of a dataset and its creators.

Finally, although in a different field, it is worth to mention "DataTags - Share Sensitive Data with Confidence" [3], a project conducted by members of the Privacy Tools project in collaboration with the IQSS Dataverse team. The goal of DataTags is to support researchers who are not legal or technical experts in investigating considerations about proper handling of human subjects data, and make informed decisions when collecting, storing, and sharing sensitive data.

6 Limitations

As limitations of our approach we could highlight two main aspects. The first issue is related to the amount of data that has been taken into consideration: it would be recommended to retrieve a wider number of datasets with all the concerning information, with the aim of further understanding and assessing the reliability of this approach. We are confident that investigating on a wider amount of data could help to interpret more profoundly the suitability of the adopted indexes. In the second place, for the purpose of improving our approach based on quantitative measures, it would be advantageous to take into account other kinds of metrics, examining and comparing their performance with the Gini and Shannon indexes. More datasets and more metrics are necessary to generalize the findings of this exploratory work and being able to deeply understand whether indicators of imbalance are reliable to anticipate the risk for potential emergence of discriminatory behavior by ADM software, as it should be accurately tested the chain of effects through which imbalance propagates.

7 Conclusions and future work

We presented a metric-based approach for detecting risks of biased output from automated decision making systems due to imbalance in the input data. The rational of this approach builds upon the conceptual framework of ISO SQuaRE standard series, in which a chain of quality effects is described: internal software quality and data quality have effect on external software quality, which in turn has effect on the quality in use and in the socio-technical context where software is used. Following this line of reasoning, and in line with the well known GIGO principle ("garbage in garbage out"), our hypothesis is that bias on input data will probably cause biased output data: in terms of automated decision systems, this would lead to potential discriminatory outputs.

We identified a heterogeneity index and a diversity index to measure the level of imbalance of the values of sensitive attributes and tested them on two datasets. The results showed that these indexes are able to represent imbalance in datasets that exposed problems of biased outputs. We also observed that the indexes are sensitive to the number of categories in the data: the indexes were not well reflecting the imbalance of attributes with a large number of classes.

In future work we want to test the indexes on a much larger number of datasets. Furthermore, we will enlarge the set of indexes and we aim to test this approach on real cases of automated decision making, with a view to understanding when the underlying hypothesis of the chain of effects holds.

References

1. Alston, P.: Report of the special rapporteur on extreme poverty and human rights (Oct 2019), https://undocs.org/A/74/493
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: COMPAS Recidivism Racial Bias, propublica/compas-analysis (2016), https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv
3. Bar-Sinai, M., Sweeney, L., Crosas, M.: Datatags, data handling policy spaces and the tags language. In: 2016 IEEE Security and Privacy Workshops (SPW). pp. 1–8 (May 2016). https://doi.org/10.1109/SPW.2016.11
4. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019), http://www.fairmlbook.org

5. Bartlett, R., Morse, A., Stanton, R., Wallace, N.: Consumer-lending discrimination in the fintech era. Tech. rep., National Bureau of Economic Research (2019)
6. Beretta, E., Vetrò, A., Lepri, B., De Martin, J.C.: Ethical and Socially-Aware Data Labels. In: Lossio-Ventura, J.A., Muñante, D., Alatrista-Salas, H. (eds.) Information Management and Big Data, vol. 898, pp. 320–327. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-11680-4_30
7. Breiling, L.: Opinion | The Race-Based Mortgage Penalty - The New York Times (2018), https://www.nytimes.com/2018/03/07/opinion/mortage-minority-income.html
8. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91 (2018)
9. Capecchi, S., Iannario, M.: Gini heterogeneity index for detecting uncertainty in ordinal data surveys. Metron **74**(2), 223–232 (2016)
10. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. **14**(3), 330–347 (Jul 1996). https://doi.org/10.1145/230538.230561, http://doi.acm.org/10.1145/230538.230561
11. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumeé III, H., Crawford, K.: Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018)
12. ISO: ISO/IEC 25000:2005 (2005), http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/56/35683.html
13. Julia Angwin, Jeff Larson, S.M., Kirchner, L.: Machine Bias — ProPublica (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
14. Learning, U.M.: Default of Credit Card Clients Dataset. https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset (2016)
15. Matzat, L.: Atlas of Automation. Tech. rep., AlgorithmWatch (2019)
16. Mecati, M., Cannavò, F.E., Vetrò, A., Torchiano, M.: Reproducibility package for identifying risks in datasets for automated decision–making. https://dx.doi.org/10.24433/CO.5067135.v2 (2020)
17. O'Neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books (2016)
18. Pauly, M.: Black box Schufa - Data Journalism Awards (2019), https://datajournalismawards.org/projects/black-box-schufa/
19. Rice, L., Swesnik, D.: Discriminatory effects of credit scoring on communities of color. Suffolk UL Rev. **46**, 935 (2013)
20. Ross, H., Bassoff, N.W.: The "Dataset Nutrition Label Project" Tackles Dataset Health and Standards. https://medium.com/berkman-klein-center/the-dataset-nutrition-label-project-tackles-dataset-health-and-standards-658dc162dfbb (2019)
21. Rovatsos, M., Mittelstadt, B., Koene, A.: Landscape summary: Bias in algorithmic decision-making. Tech. rep., Centre for Data Ethics and Innovation (CDEI) (2019)
22. Tolan, S., Miron, M., Gómez, E., Castillo, C.: Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In: Proc. of the 17th Int. Conf. on Artificial Intelligence and Law. p. 83–92. ICAIL '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3322640.3326705
23. Yang, T.: Choice and fraud in racial identification: The dilemma of policing race in affirmative action, the census, and a color-blind society. Mich. J. Race & L. **11**, 367 (2005)