POLITECNICO DI TORINO
Repository ISTITUZIONALE

Detecting Risk of Biased Output with Balance Measures

Original
Detecting Risk of Biased Output with Balance Measures / Mecati, Mariachiara; Vetro', Antonio; Torchiano, Marco. - In: ACM JOURNAL OF DATA AND INFORMATION QUALITY. - ISSN 1936-1955. - 14:4(2022). [10.1145/3530787]

Availability:
This version is available at: 11583/2970220 since: 2023-05-03T12:30:51Z

Publisher:
Association for Computing Machinery

Published
DOI:10.1145/3530787

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
ACM postprint/Author's Accepted Manuscript

© ACM 2022. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in ACM JOURNAL OF DATA AND INFORMATION QUALITY, http://dx.doi.org/10.1145/3530787.

(Article begins on next page)

07 May 2026

Detecting Risk of Biased Output with Balance Measures

MARIACHIARA MECATI, ANTONIO VETRÒ, and MARCO TORCHIANO, Politecnico di Torino, Italy

Data has become a fundamental element of the management and productive infrastructures of our society, fuelling digitization of organizational and decision-making processes at an impressive speed. This transition shows lights and shadows, and the "bias in-bias out" problem is one of the most relevant issues, which encompasses technical, ethical, and social perspectives. We address this field of research by investigating how the balance of protected attributes in training data can be used to assess the risk of algorithmic unfairness. We identify four balance measures and test their ability to identify the risk of discriminatory classification by applying them to the training set. The results of this proof of concept show that the indexes properly detect unfairness of software output. However we found the choice of the balance measure has a relevant impact on the threshold to consider as risky; further work is necessary to deepen knowledge on this aspect.

CCS Concepts: • **General and reference** → **Measurement**; *Experimentation*; • **Information systems** → **Data analytics**; *Decision support systems*; • **Social and professional topics**;

Additional Key Words and Phrases: Data quality, Data bias, Data ethics, Algorithm fairness

ACM Reference Format:
Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. 2022. Detecting Risk of Biased Output with Balance Measures. *ACM J. Data Inform. Quality* 1, 1 (February 2022), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The large availability of data, in conjunction with the widespread use of predictive, classification, and ranking models, has fuelled the ongoing mass digitization of organizational processes in our societies [3]. This is especially true for decision-making processes, which are rapidly turning into automated data-driven decision-making systems in a variety of sectors, both in private and public organizations. Such processes range from predicting debt repayment capability to identifying the best candidates for a job position, from detecting social welfare frauds to suggesting which university to attend; just to mention a few cases [4]. Advantages for using these systems concern scalability of the operations and consequent economic efficiency, as well as the removal of human subjectivity and errors. Though the benefits materialize only if the underlying data is of high quality, otherwise errors could lead to relevant extra costs [18] and also give rise to serious ethical issues: several studies showed that automated data-driven processes replicate or even amplify the same bias of our society, producing systematic discrimination to the weakest people and exacerbating existing inequalities [16]. A recurring cause for unintended but nevertheless dramatic consequence is the use of biased data. From a data engineering perspective, this means imbalanced data, i.e. a condition with an uneven distribution of data between the classes of a given attribute, which causes highly heterogeneous accuracy across the classifications [11]. Imbalance can origin from errors or limitations in the data collection, design, and operations, or simply from the reality that

Authors' address: Mariachiara Mecati, mariachiara.mecati@polito.it; Antonio Vetrò, antonio.vetro@polito.it; Marco Torchiano, marco.torchiano@polito.it, Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, Italy, 10129.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
1936-1955/2022/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the data itself reproduce. When the objects of automated decisions are individuals, such disparate performance of the algorithm represents in practice a systematic discriminatory behavior that causes relevant social, legal and ethical issues [2].

In this paper we investigate whether and to which extent it is possible to assess the risk of unfairness in software output by measuring the imbalance of protected attributes in training data. We describe the design of the proof of concept in Sec. 2 and results in Sec. 3. Then, we position our work in the literature in Sec. 4 and we highlight the limitations of the study in Sec. 5; we conclude with a few final remarks in Sec. 6.

2 PROOF OF CONCEPT

On the basis of the motivations presented above we formulated the following research question:

> **RQ**: is it possible to measure the risk of bias in a classification output by measuring the level of (im)balance in the protected attributes of the training set?

The research question relies on the following definitions:

- we consider software systems as biased when they "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others [by denying] an opportunity for a good or [assigning] an undesirable outcome to an individual or groups of individuals on grounds that are unreasonable or inappropriate" [7];
- we refer to protected attributes as those identified by the characteristics provided in "Article 21 - Non- discrimination" of the EU Charter of Fundamental Rights [6]: *Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.*

With the goal of exploring the research question , we set up a method to deliver a proof of concept:

- we took into account five large *datasets*, available in the literature;
- using a *mutation technique* we generated a number of derived synthetic datasets having different levels of balance;
- we measured the balance of such derived datasets through four different widely used *balance measures*;
- we then trained a new ML model for each dataset, and we applied three distinct *fairness criteria* to the classifications obtained from the model, for a total of five unfairness measures on each output.

To explore our RQ and check whether lower levels of balance – as detected by the selected measures – correspond to higher unfairness levels, we finally assessed the relationship between the unfairness measures and the balance measures.

Datasets. We examined five datasets coming from four distinct sources – summarized in table 1 – belonging to two different application domains. All the datasets include a binomial target variable that we predict with a binary classifier. More specifically we trained a logistic regression model on a training set composed of 70% of the original dataset (randomly selected) and we used the remaining 30% as the test set. We observe that in real datasets we can often find missing values (NA), which we decided to include in the analysis by treating them as a separate category.

Mutation technique. The target for the mutation is the Sex protected attribute, the reasons being that (i) it is present in all five datasets and (ii) it is one of the most common sources of imbalance

and consequent discrimination [16]. In order to generate a variant of an original dataset (mutant) w.r.t Sex attribute, we adopted a widely used re-balancing technique – ROSE [15] – that works specifically with binary attributes. We used the `ROSE-package` in R [1] In particular the `ovun.sample` function that generates samples with different level of balance through a combination of over- and under-sampling of the set of records whose Sex attributes belong to distinct classes. The generated mutated datasets have the same number of rows as the original ones. The mutation is driven by a parameter p that represents the probability of resampling from the rare class. In our mutations we adopted nine different values for such parameter: $p \in \{ 0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5 \}$. Since the Sex attribute has two classes, setting $p = 0.5$ means aiming for the maximum balance while smaller values correspond to less balance.

In order to increase the variability – and reliability – of our method, given the random nature of the resampling, we generated 100 different mutations (for each value of p) using distinct seeds. Overall we applied this technique to the five datasets described above obtaining:
5 datasets × 9 levels of p × 100 seeds = 4500 synthetically mutated datasets.

Balance measures. In this study we limited our attention to categorical attributes and we selected four indexes of data balance retrieved from the literature of different disciplines and reported in table 2. We normalized the measures to satisfy two criteria: i) range in the interval $[0, 1]$; ii) share the same interpretation, that is, the closer the measure to 1 and the higher the balance (i.e., categories have similar frequencies), vice-versa, values closer to 0 indicate an imbalanced distribution (e.g., male 90% - female 10%).

Fairness assessment. We assessed the unfairness of automated classifications relying on three criteria formalized in [1]: we consider a binary sensitive categorical attribute A (corresponding to Sex) that can assume the values a_1 or a_2, a target variable Y and a predicted class R where Y is binary (i.e., $Y = 0$ or $Y = 1$ and thus $R = 0$ or $R = 1$). Hence we checked whether the predictions systematically disadvantaged males or females. The unfairness measures range in the interval $[0, 1]$, where zero is a perfect balance.

- ***Independence***. It requires the acceptance rate to be the same in all groups, i.e.:

$$\mathfrak{U}_I(a_1, a_2) = |P(R = 1 \mid A = a_1) - P(R = 1 \mid A = a_2)|$$

- ***Separation***. It requires the equivalence of True Positive rate and False Positive rate for each level of the protected attribute under analysis, i.e.:

$$\mathfrak{U}_{Sep\_TPR}(a_1, a_2) = |P(R = 1 \mid Y = 1, A = a_1) - P(R = 1 \mid Y = 1, A = a_2)|$$

$$\mathfrak{U}_{Sep\_FPR}(a_1, a_2) = |P(R = 1 \mid Y = 0, A = a_1) - P(R = 1 \mid Y = 0, A = a_2)|$$

[1] https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ROSE-package, last visited on February 18, 2022.

Table 1. Summary of the Datasets and their prominent properties.

Dataset	Size	Domain	Target variable	Source
Default of credit cards clients (Dccc)	30000×29	Financial	default payment next month	https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
Statlog	1000×23	Financial	creditworthiness	https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
Income	32561×16	Welfare	income bracket	https://archive.ics.uci.edu/ml/datasets/adult
Student mathematics **Student portuguese**	Math. 395×37 Port. 649×37	Welfare	final grade (separate for Mathematics and Portuguese)	https://archive.ics.uci.edu/ml/datasets/Student+Performance

Table 2. The *balance measures* with the respective formula, where we consider a discrete random variable with m classes, each with frequency f_i (= proportion of the class i w.r.t. the total) where $i = 1, ..., m$:

Gini	$G = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} f_i^2\right)$	***Simpson***	$D = \frac{1}{m-1} \cdot \left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$
Shannon	$S = -\left(\frac{1}{\ln m}\right) \sum_{i=1}^{m} f_i \ln f_i$	***Imbalance Ratio***	$IR = \frac{\min(\{f_{1..m}\})}{\max(\{f_{1..m}\})}$

- ***Sufficiency***. It implies calibration of the model for the different groups, that is, Parity of Positive/Negative predictive values across all groups:

$$\mathfrak{U}_{Suf\_PP}(a_1, a_2) = |P(Y = 1 \mid R = 1, A = a_1) - P(Y = 1 \mid R = 1, A = a_2)|$$

$$\mathfrak{U}_{Suf\_PN}(a_1, a_2) = |P(Y = 1 \mid R = 0, A = a_1) - P(Y = 1 \mid R = 0, A = a_2)|$$

3 RESULTS AND DISCUSSION

Before addressing the main RQ, we performed a sanity check to observe the behavior of the balance measures as the mutation parameter p varies. Figure 1 reports the average values for different balance measures and datasets. We observe an increasing trend of all the balance measures w.r.t. increasing p, in all training sets and test sets. More in detail, Gini and Shannon indexes have a super-linear increase; Simpson index is closer to a linear trend; finally, IR index has a sub-linear increase until 2/3 of the course and then it turns to have a slight super-linear increase. This observation confirms the ability of the mutation approach to generate synthetic datasets that spread the whole range of conventional balance measures.

Figure 2 reports the variation of the five fairness criteria (Y axis) w.r.t. the increase of balance measures (X axis). The lines are smoothed regression of the individual mutations. For sake of legibility, we omitted Gini since it is very similar to Shannon. We can observe from the curves that very low levels of balance – roughly in the range $[0\ , 0.15]$ and up to 0.50 in a few cases – correspond to higher levels of unfairness. As shown in the preliminary results, the indexes react slightly differently to different levels of balance: as a consequence, the distinct unfairness criteria reflect different levels of balance in a slightly different way. By looking at the single fairness criteria, as well as at the specific trend lines in figure 2, we observe that:

- the trend of unfairness with respect to IR is often *not* monotonic: Independence, Separation-TP and Sufficiency-PP, after an initial decreasing phase, they slightly increase within the range $[0.15\ , 0.25]$ before stabilizing; Separation-FP slightly increases in the range $[0.5\ , 1]$ for Student_port; Sufficiency-PN is much less regular among datasets, and the correlation between high unfairness and low balance holds only partially;
- modest final surges in correspondence of maximum levels of the balance – around the range $[0.9\ , 1]$ – are observable above all for Separation-FP, Sufficiency-PP and Sufficiency-PN;

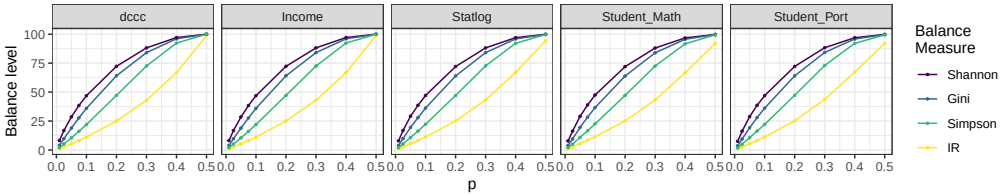

Fig. 1. Values of balance measures vs. mutation parameter p

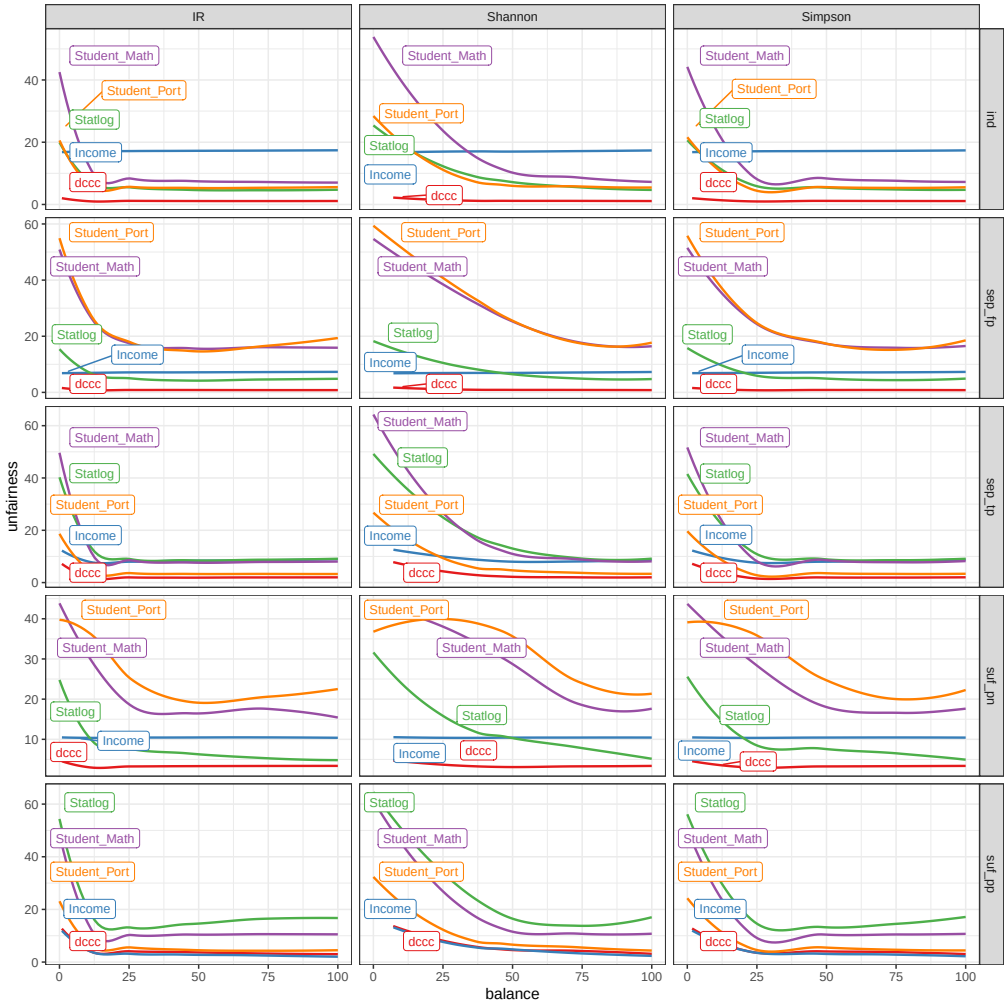

Fig. 2. Trends of the *fairness criteria* as a response to the *balance measures*.

- overall, the datasets Dccc and Income have lower levels of unfairness even with an extremely low balance, therefore the correlation high unfairness–low balance is much less pronounced for Separation-TP and Sufficiency-PP, and absent for Indepencence, Separation-FP and Sufficiency-PN.
- in general, Sufficiency-PN presents the most irregular trends especially in the dataset Student_port: for the indexes Gini and Shannon (and similarly for the indexes Simpson and IR) it increases within $[0\,, 0.2]$, then it decreases till around 0.8 and it surges again in the final range; a similar behavior can be observed for Sufficiency-PN in Student_math. However, a follow-up analysis on Sufficiency-PN w.r.t. p showed that Sufficiency-PN tends to slightly decrease as p increases (i.e., as balance increases): the reason for such irregular behavior should be further investigated and we cannot rely on the current results of Sufficiency-PN.

On the basis of these observations and within the limits of this proof concept, we positively answer our initial research question. Moreover we can identify tentative thresholds of balance measures and the following practical recommendation:

Values of indexes Shannon < 0.5, Gini < 0.4, Simpson < 0.3 and IR< 0.15 indicate a relevant risk of unfairness –which increases as the values of the balance measures decrease till 0– in terms of Independence, Separation and Sufficiency-PP.

4 RELATED WORK

Our contribution can be located in the main corpus of researches on algorithmic bias and fairness. While most of the literature focus on the outputs of ADM systems, we focus on the inputs and processes, following a direction suggested by several recent studies (e.g., [5], [17] and [8]). Our approach has its theoretical and methodological foundations in the ISO/IEC standards on data quality measurement [9] and on risk management [10]: for space reasons we can not analytically report on all the relations between our proposed approach and the two ISO/IEC standards, which can be found in [19]. This study expands the research reported in [20]: herein we introduced a mutation technique to generate a number of derived synthetic datasets having different levels of balance, instead of relying on a few exemplar distributions as done in the previous study. We applied a similar technique also in [14], but not specifically to binary attributes as done here. A further novelty in this paper is the computation of the Sufficiency criterion of fairness, in addition to Independence and Separation.

An approach similar to ours and with a wider scope is the work of Matsumoto and Ema [13], who proposed a risk chain model for risk reduction in Artificial Intelligence (AI) services, named RCM. The authors consider both data quality and data imbalance as risk factors. While our work is smaller in scope, we think that it can easily fit into the RCM framework, due to the fact that we offer a quantitative way to measure balance. Our work is also complementary to the existing toolkits for bias detection and mitigation [12], since the proposed measures of balance are not taken into consideration yet.

5 LIMITATIONS

The limited number of datasets that has been taken into account, as well as the set of balance measures constitute notable limitations to our study. More datasets and more metrics are necessary to generalize the findings of this exploratory work, also by including measures for non-categorical data. In addition, as the choice of the balance measure has a relevant impact on the threshold to consider as risky, in-depth sensitivity analyses on the thresholds should improve the reliability of the findings presented here.

Furthermore, as we ran the binomial logistic regression, all the limitations of this classification model hold, most notably the two assumptions of limited or no multi-collinearity between independent variables, and of linearity between the dependent variable and the independent variables. Applying more classification algorithms (each with different parameters) would improve the external validity of the relationship we found between balance and unfairness in the classification output, and would help to identify how the different types of classification algorithms propagate the imbalance in the training set.

Other kinds of mutation techniques should be also considered by adopting different pre-processing methods to create several variations of the distribution of the occurrences between the classes of a given protected attribute.

6 CONCLUSIONS AND FUTURE WORK

In this paper we evaluated whether imbalanced distributions of a binary protected attribute in the training data can lead to discriminatory output of ADM systems. We selected four balance

measures (the Gini, Simpson, Shannon, and Imbalance Ratio indexes, normalized to share the same range of values and the same meaning), applied them to training sets, and tested their ability to detect unfairness occurring in classification tasks. Overall the results showed that our approach is suitable for the proposed goal, however the choice of the balance measure has a relevant impact on the threshold to consider as risky. Hence, further work will be devoted to thorough and systematic test thresholds to be used, also in combination with different prediction models and mutation techniques.

REFERENCES

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[2] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. https://papers.ssrn.com/abstract=2477899.

[3] Erik Brynjolfsson and Andrew McAfee. 2016. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* (reprint edition ed.). W. W. Norton & Company, New York London.

[4] Fabio Chiusi, Sarah Fischer, Nicolas Kayser-Bril, and Matthias Spielkamp. 2020. Automating Society Report 2020. https://automatingsociety.algorithmwatch.org.

[5] Donatella Firmani, Letizia Tanca, and Riccardo Torlone. 2019. Ethical dimensions for data quality. *Journal of Data and Information Quality (JDIQ)* 12, 1 (2019), 1–5.

[6] European Union Agency for Fundamental Rights. 2007. EU Charter of Fundamental Rights - Article 21 - Non-discrimination. https://fra.europa.eu/en/eu-charter/article/21-non-discrimination.

[7] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. http://doi.acm.org/10.1145/230538.230561. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[8] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 49–58.

[9] ISO. 2014. ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE. https://www.iso.org/standard/64764.html.

[10] ISO. 2018. ISO 31000:2018 Risk management — Guidelines. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65694.html.

[11] Bartosz Krawczyk. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5, 4 (Nov. 2016), 221–232. https://doi.org/10.1007/s13748-016-0094-0

[12] Michelle Seng Ah Lee and Jatinder Singh. 2020. The Landscape and Gaps in Open Source Fairness Toolkits.

[13] Takashi Matsumoto and Arisa Ema. 2020. RCModel, a Risk Chain Model for Risk Reduction in AI Services. http://arxiv.org/abs/2007.03215.

[14] Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. 2021. Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 4287–4296. https://doi.org/10.1109/BigData52589.2021.9671443

[15] Giovanna Menardi and Nicola Torelli. 2014. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery* 28, 1 (2014), 92–122.

[16] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (reprint edition ed.). Broadway Books, New York.

[17] Evaggelia Pitoura. 2020. Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias. https://doi.org/10.1145/3404193. *Journal of Data and Information Quality* 12, 3 (July 2020), 12:1–12:8. https://doi.org/10.1145/3404193

[18] Thomas C. Redman. 2017. Seizing Opportunity in Data Quality. *MIT Sloan Management Review* 29 (2017). https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

[19] Antonio Vetrò. 2021. Imbalanced data as risk factor of discriminating automated decisions: a measurement-based approach. *JIPITEC* 12, 4 (2021), 272–288. https://doi.org/10.5281/zenodo.5795184

[20] Antonio Vetrò, Marco Torchiano, and Mariachiara Mecati. 2021. A data quality approach to the identification of discrimination risk in automated decision making systems. *Government Information Quarterly* 38, 4 (2021), 101619. https://doi.org/10.1016/j.giq.2021.101619