A data quality approach to the identification of discrimination risk in automated decision making systems

This is the authors' post-print version
The editorial version is available at
Elsevier Government Information Quarterly website:
https://doi.org/10.1016/j.giq.2021.101619

# A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision Making Systems (rev.1)

Antonio Vetrò[a], Marco Torchiano[a], Mariachiara Mecati[a]

[a]*Politecnico di Torino, Turin, Italy*

**Abstract**

Automated decision-making (ADM) systems may affect multiple aspects of our lives. In particular, they can result in systematic discrimination of specific population groups, in violation of the EU Charter of Fundamental Rights. One of the potential causes of discriminative behavior, i.e. unfairness, lies in the quality of the data used to train such ADM systems.

Using a data quality measurement approach combined with risk management, both defined in ISO standards, we focus on *balance* characteristics and we aim to understand how balance indexes (Gini, Simpson, Shannon, Imbalance ratio) identify discrimination risk in six large datasets containing the classification output of ADM systems.

The best result is achieved using the Imbalance Ratio index. Gini and Shannon indexes tend to assume high values and for this reason they have modest results in both aspects: further experimentation with different thresholds is needed.

In terms of policies, the risk-based approach is a core element of the EU approach to regulate algorithmic systems: in this context, balance measures can be easily assumed as risk indicators of propagation – or even amplification – of bias in the input data of ADM systems.

*Keywords:* Automated decision making, Data ethics, Data quality, Data bias, Algorithm fairness

*Email addresses:* antonio.vetro@polito.it (Antonio Vetrò), marco.torchiano@polito.it (Marco Torchiano), mariachiara.mecati@polito.it (Mariachiara Mecati)

## 1. Introduction

The automation of decision processes is rapidly expanding [1][2] as a result of the more general phenomenon of digitization of organizational processes in our societies [3][4]. Such trend was enabled at first by the computerization of our physical environments and the large diffusion of internet connectivity, and more recently by the large availability of data and the emergence of technical means for their analysis. The foundations of this rapid adoption of data driven decision-making [5] lay on the development of predictive, classification, and ranking models that are at the core of automated decision-making (ADM) systems[1]. Decisions are either based on software-generated recommendations or even completely automated: the adopted technical approaches range from sophisticated neural networks [2] to simple tools such as macros or scripts that compute and sort data according to predefined sets of rules [6].

The tasks delegated to or supported by ADM systems range from predicting debt repayment capability [7] to identifying the best candidates for a job position [8], from detecting social welfare frauds [9] to suggesting which university to attend [10], just to mention a few cases. Advantages for using these systems concern not only scalability of the operations and consequent economic efficiency, but they are also supposed to remove discretion of public service workers [11] [12] [13]. However, a large amount of evidence both in scientific literature [14] and journalistic essays [15] [16] shows that ADM systems may perpetuate the same bias of our societies, systematically discriminating the weakest people and exacerbating existing inequalities. The issue is so relevant to involve not only specialists from information technology as well as social sciences, but it has been recognized by the institutions [17]. As stated

[1]In our writing we adopt the definition of Automated Decision Making provided by Algorithm Watch [1]: *Systems of automated decision-making (ADM systems) are always a combination of the following social and technological parts: i) a decision-making model; ii) algorithms that make this model applicable in the form of software code; iii) data sets that are entered into this software, be it for the purpose of training via Machine learning or for analysis by the software; iv) the whole of the political and economic ecosystems that ADM systems are embedded in (elements of these ecosystems include: the development of ADM systems by public authorities or commercial actors, the procurement of ADM systems, and their specific use).*

by Margrethe Vestager[2], automating decisions using historical data is a double-edged sword [18]:

> If they're trained on biased data then they can learn to repeat those same biases. Sadly, our societies have such a history of prejudice that you need to work very hard to get that bias out. And if we don't know how they're making their decisions, we can't be sure that those choices aren't based on harmful stereotypes – and to challenge those decisions, if they're unfair.

From a data engineering perspective, biased data means *imbalanced* input dataset [19]: data imbalance is an unequal distribution of data between the classes – e.g. gender, country, etc. – of a given attribute [20]. Causes of imbalance can be errors or limitations in the data collection design and operations, or no other reason than disproportions in the current reality that the data itself reproduce, as acknowledged by the excerpt of Vestager's speech. Specifically, imbalance is between-class when only two classes are taken into consideration and one class is over-represented with respect to the other, or multiclass when imbalances exist between multiple classes. Herein we focus on the more general case, i.e., multiclass imbalance.

Imbalanced data is known since long time to be a problematic aspect in the machine learning domain [20] [21] – and it is still relevant [19] [22]– especially because it can corrupt the performance of supervised learning algorithms in terms of very heterogeneous accuracy across the classes of data. When the objects of automated decision are individuals, such disparate performance of the algorithm means in practice to

> "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 out- come to an individual or groups of individuals on grounds that are unreasonable or inappropriate"[23].

[2]Margrethe Vestager is the Executive Vice President of the European Commission for A Europe Fit for the Digital Age since December 2019 and European Commissioner for Competition since 2014.

A well-know example of this issue is the development of a software system by Amazon for the purpose of evaluating the CVs of potential employees collected from the web [24]. The project started in 2014 with the aim of predicting successful future employees using word patterns extracted from CVs of the previous 10 years, but it was stopped in 2017 as female profiles were systematically downgraded. The problem came from the fact that training data consisted mostly of men, since the majority of employees in the technology sector is male.

Similarly, a scientific experiment on the search engine Common Crawl [25] revealed an unequal treatment due to gender imbalance in the input data (almost 400.000 biographies): authors compared three techniques of machine learning for occupational classification and showed that in each case the rate of correct classifications followed the existing gender imbalances of the occupational groups, even without explicitly using gender indicators.

Another study [26] reported that Facebook advertisements for employment opportunities were significantly distorted towards gender and ethnic group, leading to unequal job opportunities and persistent discriminatory treatments for all the lifetime of an advert. Due to such a conservative mechanism, people are deprived of opportunities based on personal characteristics, contrary to the statements in the Art. 21 of the EU Charter of human rights [27]. For this reason, in the United States, the Department of Housing and Urban Development sued Facebook in March 2019 for violating the Fair Housing Act because of the discriminatory effect of its advertisements, as housing ads were disproportionately targeted with respect to race, gender, and other personal traits [28].

These negative consequences could become even worse and life-altering if they occur in the medical or in the justice fields, where the combined use of ADM systems and historical data is rapidly increasing. The most famous case in the criminal justice system is represented by the investigation on COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) , an algorithm used by judges to assess the probability of recidivism of defendants. The no-profit organization Pro Publica showed that the algorithm was distorted in favor of white defendants [29]: in fact, those of them who got rearrested, they were nearly twice as likely to be misclassified as low risk than black defendants. On the contrary, black defendants who did not get rearrested were nearly twice as likely to be misclassified as higher risk (false positive) than white defendants. The main reason of this distorted effect was that the number

of records related to white defendants was much smaller than the number of records of black defendants in the dataset.

Concerning the medical field, a recent study [30] reported the case of a widely-used commercial system for deciding which patients should get into an intensive care program. Medical doctors applied risk scores generated by an algorithm trained on historical data about medical expenditure and use of health services. It has been found that in cases of equivalent health status, white patients were significantly more likely than black patients to be assigned to the intensive care program. Also in this case, the system was affected by ethnicity-based discrimination, as the risk score reflected more the expected cost of treatment than the actual health conditions, with the former being highly correlated with the economic wealth of the patients.

The examples briefly reported above, explicitly show how imbalance in data can propagate and be reflected in the output of ADM systems, becoming a socio-technical issue particularly important to public sector services, where high stake decisions are increasingly delegated to such systems. With a view to facing this important issue, we propose a risk assessment approach based on quantitative measures to evaluate imbalance in the input datasets of ADM systems. Specifically, by revealing imbalance in input data we aim to highlight a potential risk of discriminatory automated decisions : we believe that this approach should encourage to take appropriate actions and to prevent adverse effects.

The rest of this paper is organized as follows. In Section 2 we discuss the theoretical foundations of our proposal. In Section 3, we show how our work is related to several research strands, building a landscape picture of the research context. In Section 4 we describe the design of the exploratory work, which includes the methodology, the datasets and the four balance measures we employed. We report the analysis of results in Section 5 and discuss them in Section 6, along with an analysis of the relations of our approach to current European Policy efforts (Section 6.1) and an overview of the limitations to be addressed in future work (Section 6.2). In the end, we highlight conclusions and future line of research in Section 7.

## 2. Background

Along the line of though adopted above, and motivated by the evidence that supports it, herein we outline the founding concepts underpinning our approach: data imbalance as a risk factor for systematic discrimination caused by ADM systems. The approach originates from software quality and risk management ISO standards.

The first conceptual pillar is the series of standards ISO/IEC 25000:2014 : "Systems and Software engineering — Software product Quality Requirements and Evaluation (SQuaRE)" [31]. SQuaRe defines quality modelling and measures of software products[3], data, and software services. The quality of these three elements is modeled with a set of measurable characteristics and sub-characteristics. In particular, data quality is modeled in ISO/IEC 25012:2008 with 15 characteristics (e.g., completeness, efficiency, recoverability), each of which is quantifiable through measures of quality-related properties defined in ISO/IEC 25024:2015. The characteristics belong to the "inherent" point of view if they depend on the data themselves, or to the "system dependent" perspective if they are influenced by the computer systems hardware or software used to store, analyze, retrieve, etc. them. Some characteristics can belong to both points of view. An example of inherent characteristic is Completeness, defined as *the extent to which all necessary values have been assigned and stored in the computer system*: for instance, in a dataset on university's students, all the necessary information on students should be present to satisfy the needs of the users. One of the measures of Completeness is "Record completeness" (Com-I-1), defined as the ratio between the number of data items with not null associated value in a record and the number of data items of the record itself. An example of system dependent characteristic is Availability, defined as *the capability of data to be always retrievable*; one of its measure is the "Probability of data available" (Ava-D-2), i.e., the ratio between the number of times that data items are available in a given time and the number of times that data items are requested during that same time. Finally, an example of characteristic belonging to both points of view is

[3]A software product is a *"set of computer programs, procedures, and possibly associated documentation and data"* as defined in ISO/IEC 12207:1998. In SQuaRE standards, *software quality* stands for *software product quality*.

Efficiency, defined as *the capability of data to be processed (accessed, acquired, updated, etc) and to provide appropriate levels of performance using the appropriate amounts and types of resources under stated conditions*. Efficiency has distinct measures for the inherent and the system dependent point of views.

Although neither data imbalance nor its dual concept data balance are part of data quality in ISO/IEC 25012:2008, the SquaRE standard puts forward a concept that appears extremely relevant in our context, that is, the chain of effects and dependencies. According to this principle, improving product, service or data quality will have a positive effect on the system quality in use and will eventually benefit the users of a software system[4]. The top portion of figure 1 reports synthesis of how this chain of effects is formalized in SQuaRE. In the realm of data quality, a simplified dual concept is the well-known GIGO – i.e. "garbage in, garbage out" – principle : outdated, inaccurate, incomplete or flawed input data, make the output of the software unreliable.

We maintain that the chain of effects holds even for data imbalance, in the sense that imbalanced datasets may lead to imbalanced software outputs, which means – in the context of ADM systems – differentiation of products, information and services based on personal characteristics. As mentioned in the introduction of the manuscript, in specific applications such as wages, education, working positions, social benefits, etc. such differentiation can lead to unjustified unequal treatment and even unlawful discrimination. For this reason, data imbalance shall be considered as a risk factor in all those ADM systems that rely on historical data and that automate decision on aspects that concern the exercise of rights and freedoms: AMD systems developed and deployed in public sector services are certainly one of these cases. In this specific context, we treat data imbalance as an extension of the data quality model formalized in ISO/IEC 25012:2008: more precisely, it can be considered an inherent characteristic, which will be quantified by proper measures, extending those already defined in ISO/IEC 25024:2015.

The second pillar behind our approach is represented by the ISO 31000:2018 standard on

[4]The relationship holds also in the opposite direction and between pairs of aspects, e.g.: the quality in use depends on the product quality, which in turn influences the data quality

risk management [32]. This standard provides the guiding principles for risk management, a framework for integrating it into organizational contexts, and a process for managing risks at *"strategic, operational, program or project levels"*. In the context of our proposal, data imbalance as well as ADM systems discrimination shall be explicitly taken into account within the risk management process. The main elements of the ISO 31000:2018 approach to risk is summarized in the bottom portion of figure 1. In particular we focus on the *risk assessment* phase: it consists of risk identification, analysis and evaluation, briefly described below in relation to our approach.

- *Risk identification.* It refers to finding, recognizing and describing risks within a certain context and scope, and with respect to specific criteria established prior to risk assessment. In our case, this phase can be traced back to the discrimination of individuals based on their membership in a certain group or category [23] and to the Article 21 "Non discrimination" of the Charter of Fundamental Rights of the European Union [27], as ADM systems operate in contexts relevant to the rights and freedoms of individuals.

- *Risk analysis.* The goal of this activity is to understand the characteristics of the risk and -when possible- its levels. This is the phase where measures of data imbalance are used as indicators of the risk of discrimination and it is the focus of this paper.

- *Risk evaluation.* In this last step, the results of the analysis are taken into consideration to decide whether the level of risks requires additional analyses, treatments or other types of actions. In our case, the measures of data imbalance should be analyzed in the context of the specific algorithms which deal with such data, the severity of the impact on the users and the specific legal requirements for a given domain. This aspect is out of the scope of the current work, and it will further discussed in Section 6.2.

Overall, figure 1 summarizes the approach and the connections with the international ISO/IEC standards adopted as reference frameworks. In the upper layer, we represent the elements of the SQuaRe series (2500n) which are most relevant for our scope. In the bottom, we report the main elements of the risk management process of ISO 31000. Our proposal is depicted in the middle of the figure, with all the relations to SQuaRe and ISO 31000.

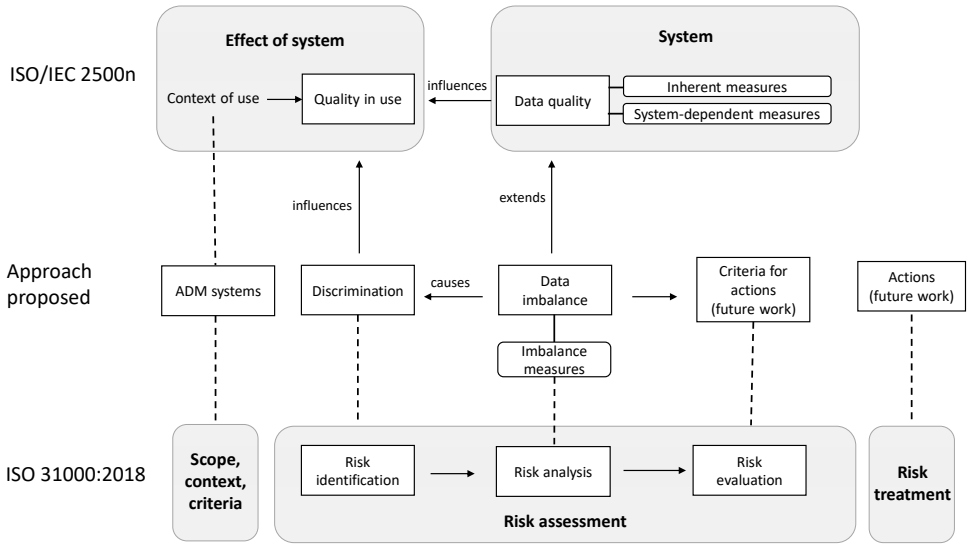


Figure 1: The proposed approach in relation to ISO standards adopted as reference frameworks.

## 3. Related Work

In the last few years a remarkable research work has been developed to make outcomes of ADM systems more equitable. The main studies have been focused on techniques to detect and mitigate systematic discrimination according to different definitions of unfairness. Among the most comprehensive works we remind the book by Barocas et al. [33] (from which we derived the unfairness measures used here), the survey on bias and fairness in machine learning by Mehrabi et al. [34] as well as the review of discrimination measures for algorithm decision making by Žliobaitė [35]. An important limitation of defining a software output as fair or not consists in the formal impossibility of concurrently satisfying different mathematical notations of fairness [36] [37]. This is an ontological limitation: in fact, a universally acceptable notion of fairness can not exist, as to define a "fair impact" it is necessary to include several political, economic and cultural aspects [38]. The ACM Conference on Fairness, Accountability, and Transparency [5] has recognized this issue and has been designed and promoted not only for computer scientists working in the area, but also for scholars and practitioners from *"law, social sciences and humanities to investigate and tackle issues in this emerging area"*.

[5]ACM FAccT, https://facctconference.org

Our approach can be placed in this space of inter-disciplinary discussion. It contributes to the main corpus of researches on algorithmic bias and fairness by moving the focus from the outcomes of ADM systems to their inputs, as indicated as necessary in [39] (*"There is a need to consider social-minded measures along the whole data pipeline"*) and in [40] (*"Returning to the idea of unfairness suggests several new areas of inquiry [...] a shift in focus from outcomes to inputs and processes"*). Our proposal differentiates from the reference literature for two additional properties: i) it is built upon a series of international standards, which incorporate a multi-stakeholder perspective by design; ii) we look at data imbalance as a risk factor and not as a technical fix: we believe that a risk approach creates space for active human considerations and interventions, rather than delegating the mitigation of the problem to yet another algorithm, with very low probability of success. In fact, given the socio-technical nature of the issue, we think it is preferable to keep the ultimate responsibility in the realm of human agency.

An approach similar to ours and with a wider scope is the work of Takashi Matsumoto and Arisa Ema [41], who proposed a risk chain model for risk reduction in Artificial Intelligence (AI) services, named RCM. By applying RCM in a given risk scenario, it can be proven that a propagation occurs from the technical components of AI systems (data and model) up to the user's understanding, behavior, and usage environment. The authors consider both data quality and data imbalance as risk factors, whereby they stress the importance of visualizing the relations between risk factors for better risk control. While our work is smaller in scope, we think that it can be easily plugged into the RCM framework, due to the fact that we offer a quantitative way to measure balance, backed by a structural relation to the ISO/IEC standards on software quality requirements and risk management.

Other approaches related to ours are in the direction of labeling datasets: "The Dataset Nutrition Label Project" [6] has been an inspiring work for us. Similar to nutrition labels on food, this initiative aims to identify the "key ingredients" in a dataset such as provenance, population, missing data. The label takes the form of an interactive visualization that allows

[6] It is the result of a joint initiative of MIT Media Lab and Berkman Klein Center at Harvard University: https://datanutrition.org/

for exploring the previously mentioned aspects. The ultimate goal is to avoid the fact that flawed, incomplete, skewed or problematic data would have a negative impact on automated decision systems, and to drive to the creation of more inclusive algorithms. A similar goal was declared by authors of the "Ethically and socially-aware labeling" (EASAL) [42], who identified three types of data input properties that could lead to downstream potential risks of discrimination: data quality, correlations and collinearity, and disproportions in datasets. The last property coincides to imbalanced data. The same authors lately published a data annotation and visualization schema based on Bayesian statistical inference [43], always for the purpose of warning about the risk of discriminatory outcomes of a given dataset.

Finally, it is important to mention the development of tools: in the recent years researchers both in the profit sector and in universities developed toolkits for bias detection and mitigation [44]. For example:

- the AI Fairness 360 Open Source Toolkit [45], an open source library developed by IBM, designed to examine and mitigate bias in the output of machine learning models. It provides several metrics to analyze the unfairness of the models and pre-processing algorithms to transform the dataset;

- the What-If Tool [46], by Google, which can be used to analyze the characteristics of a dataset and of the models derived from it. These models can also be examined for their unfairness w.r.t. various measures, and an interactive graphical user interface let the user perform a sensitivity analysis by moving classification thresholds for the selected features;

- Aequitas is an open source bias audit toolkit [47] developed by the Center for Data Science and Public Policy at the University of Chicago. It allows to generate a bias report which includes multiple unfairness measures based on the user's selection of reference groups;

- the Themis software [48] by the University of Massachusetts Amherst, is different from the previous ones because it is based on the concept of causal discrimination: a test suite captures the relationships between inputs and outputs, providing a causal discrimination score for a particular set of characteristics.

- the Amazon SageMaker Clarify tool [49] provides eight measures of pre-training bias [7], seven of which are focused on the outcomes distribution and only one is a measure of balance, i.e. "Class Imbalance CI" which is applicable to only two classes [8], while our measures can work with any number of classes.

Also in this case, we highlight the complementarity of our work with existing approaches and the potentiality for future integration.

## 4. Exploratory study design

The goal of our study is to anticipate possible discrimination risks on the basis of the balance features of the training data employed by ADM systems. To this end we formulated two research questions (RQ) and two corresponding methodologies (M) that will drive our investigation:

RQ 1. *How are existing measures able to detect imbalance among the classes of a given attribute in a dataset?*

Several measures have been proposed in the literature, trying to capture the abstract construct *imbalance*. We aim at understanding how the indexes reflect our – probably limited and subjective – understanding of imbalance.

M 1. We defined a set of synthetic attributes with a known and simple exemplar distribution whose balance can be judged; then we assessed the values of the balance measures against the human judgement. Details about RQ 1 methodology are available in section 4.1

RQ 2. *Are existing measures able to reveal a discrimination risk when an ADM system is trained with such data?*

A large corpus of scientific and journalistic evidence show that imbalanced data, when used to train an ADM system, may trigger a discriminatory behavior (see section 1). The ultimate focus of the paper consists in assessing whether the imbalance in data,

[7] `https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html`

[8] Called "facets", see `https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-bias-metric-class-imbalance.html`

measured by means of the selected indexes, may signal a discrimination risk in the ADM system.

M 2. We check the correlation between balance measures computed on six large datasets with discrimination measures on the classification outcome of ADM systems trained with those data. Details about RQ 2 methodology are available in section 4.2.

### *4.1.* ***RQ1 - Imbalance detection***

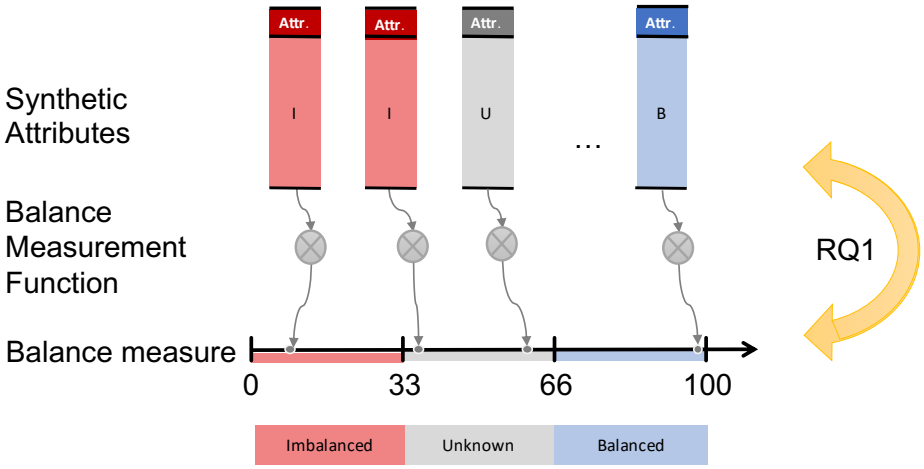


Figure 2: Investigation approach for RQ1 – *How are existing measures able to detect imbalance among the classes of a given attribute in a dataset?*

In order to address the first research question, we assess a set of measures that are able to measure balance in the data –and thus its absence, i.e. imbalance–. Figure 2 synthesizes the following procedure:

- definition of a set of *synthetic attributes* with a simple description of the distribution between the classes and our expectation in broad terms, specified by a balance judgement (*see section 4.1.1* );
- measurement on the datasets (*see section 4.1.2*);
- comparison of measures vs. expectations in order to assess the performance of the index (*see section 5.1*).

#### *4.1.1.* ***Synthetic attributes***

We identified six synthetic attributes, each with a certain exemplar distribution of the occurrences between the classes:

1. *Max Balance*: the perfect uniform distribution, we expect the measures to indicate the highest level of balance;
2. *Max Imbalance*: all classes are empty (zero occurrences) but one, we expect the measures to indicate the highest level of imbalance;
3. *Quasi Balance*: half of the classes are 10% higher w.r.t. max balance and the other half is 10% lower, we expect overall high value measures;
4. *One off*: occurrences are distributed among all classes but one;
5. *Half high*: occurrences are distributed mostly among half of the classes while the remaining have a very low frequency, we opted for a ratio of 1:9 for the frequencies of the two halves;
6. *Power 2*: occurrences are distributed according to a power law with base 2, i.e., distributions among the classes increase like the powers of 2.

For each of the above seven cases of distribution, we built different synthetic datasets with number of classes $m = 2, 3, 5, 8$. The cardinalities of the classes have been chosen according to the Fibonacci series to have enough diversity. For instance, in the *One off* case for $m = 5$ we have classes with frequencies

$$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0\right)$$

Figure 3 summarizes all the exemplars of synthetic distributions. Overall we defined 24 distributions (6 cases of distribution $\times$ 4 cases of number of classes), but the cases Max Imbalance and One Off for $m = 2$ are identical, leading to 23 unique distributions.

To formalize our expectations and to better judge the balance of the synthetic attributes, we defined three classes and the related associated thresholds:

- $I$ = imbalanced if we expect the measure to be lower than 33%,
- $B$ = balanced if we expect the measure to be greater than 67%,


Figure 3: Summary of synthetic exemplar distributions

- $U$ = uncertain if we expect a value between the two above thresholds.

We chose those thresholds because all the measures are defined to range in the interval $[0, 1]$ where 1 corresponds to the perfect balance and 0 means most extreme imbalance. In terms of understandability – i.e. capability for a human reader to look at a measure and understand its meaning – we assume the lowest values correspond to imbalance, the highest values to balance and the intermediate ones to an uncertain region.

The classification was performed collectively by the authors: each author proposed a class and a convergence to a common class was achieved after internal discussion. The final results of this process are reported in figure 3 as colored background – red for imbalance, gray for undecided, and blue for balanced – and with a label with the initial of the assigned class.

The goal is to assess the performance in terms of consistency of balance prediction with human judgments for different balance measures described in section 4.1.2 below. To compare the performance of the different balance measures we compute the accuracy in predicting imbalanced distributions, as it is a common metric used in classifier evaluation.

#### *4.1.2.* ***Balance Measures***

In this study we limited our attention to *categorical* attributes.

We selected four indexes of data balance, adjusted in order to meet three criteria:

- range in the interval $[0, 1]$;
- share the same interpretation: the closer the measure to 1 and the higher the balance (i.e. categories have similar frequencies), and vice-versa values closer to 0 means more concentration of frequencies in few categories, thus an imbalanced distribution;
- deal with empty classes, i.e., classes that *exist* (potentially there could be occurrences) but are *not* represented in the given dataset: we decided to take into account *all* the classes of each selected sensitive attribute, including also the classes with zero occurrences.

  The motivation for this choice is that in our view a dataset that contains no instance of a given class – e.g. all males or all whites – is imbalanced.

Often, in real datasets, we can find missing values (NA). We decided *not* to exclude missing values (NA) from the analysis and to consider them as a separate "NA" category.

***Gini index.*** It is a measure of heterogeneity [50] used in many disciplines and often discussed with different designations: examples are political polarization, market competition, ecological diversity as well as racial discrimination. Heterogeneity reflects how many different types (such as protected groups) are represented. In statistics, the heterogeneity of a discrete random variable which assumes $m$ categories with frequency $f_i$ (with $i = 1, ..., m$) can vary between a degenerate case (= 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:

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

Where we added the multiplication factor $\frac{m}{m-1}$ in order to normalize the index between 0 and 1.

***Shannon index.*** Diversity indexes represent a useful tool to measure imbalance providing information about community composition taking the relative amounts of different species (classes) into account. A widely employed concept in biology, phylogenetics and ecology is the Shannon index, which is a measure of species diversity in a community. We computed the index as follows:

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

In order to normalize the index we divide by $\ln m$. In addition since $ln(0) = -\infty$, to deal with empty classes – i.e. when $f_i = 0$ – we resort to the notable limit:

$$\lim_{x \to 0} x \ln x = 0$$

***Simpson index.*** The Simpson index is another indicator of diversity: it measures the probability that two individuals randomly selected from a sample belong to the same species (i.e., the same class or category). It is employed in social and economic sciences for measuring wealth, uniformity and equity, as well as in ecology for measuring the diversity of living beings in a given location. As before, we consider a discrete random variable which assumes $m$ categories with frequency $f_i$ where $i = 1, ..., m$ (that is, the proportion $f_i$ of the species $i$ with respect to the total number of species):

$$D = \frac{1}{m-1} \cdot \left( \frac{1}{\sum_{i=1}^{m} f_i^2} - 1 \right) \tag{3}$$

***Imbalance Ratio.*** The Imbalance Ratio (IR) is a widely used measure made of the ratio between the highest and the lowest frequency. We take the inverse in order to normalize it in the range $[0, 1]$ and to render it a balance measure :

$$IR = \frac{\min(\{f_{1..m}\})}{\max(\{f_{1..m}\})}$$

### *4.2.* ***RQ2 - Discrimination risk***

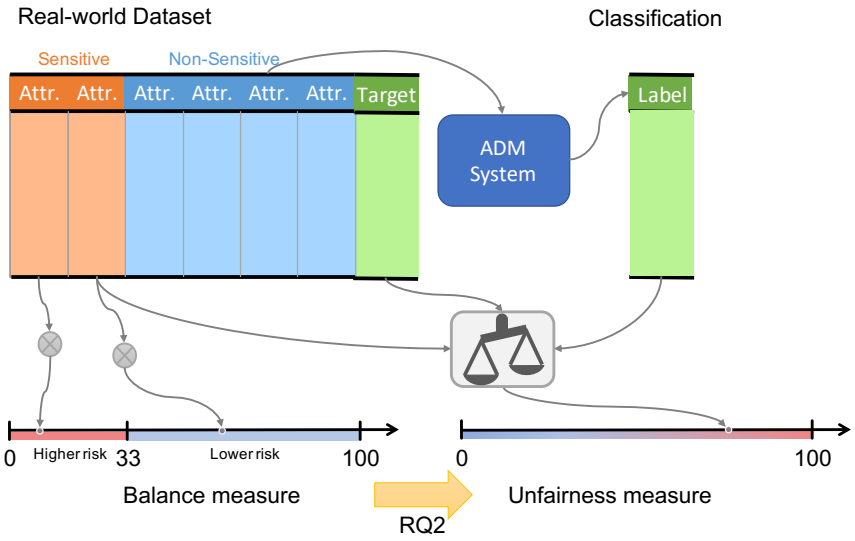


Figure 4: Analysis method for RQ2 – *Are existing measures able to reveal a discrimination risk when an ADM system is trained with such data?*

In the second stage of our analysis, we aim at assessing the reliability of the balance measures as risk indicators of biased decisions. The approach used is summarized in figure

4: given a dataset, its non-sensitive attributes and a target attribute can be used to train an ADM system that performs a classification task. The unfairness of the classification w.r.t. a sensitive attribute can be evaluated considering the target class and the predicted class. The balance of the sensitive attributes can be quantified applying any of the indexes described in the previous section, in order to understand the ability of such balance measures to reveal a potential discrimination risk –*un-fairness*. Specifically, we followed this procedure:

- selection of six large *datasets* from different domains and identification of the *protected attributes* [9] in the datasets (*see section 4.2.2*);
- assessment of the *unfairness* of the predictions w.r.t. the sensitive attributes present in the datasets; we computed the unfairness measures ($\mathfrak{U}$) related to the *Separation* and the *Independence* criteria (*see section 4.2.1*);
- evaluation of the *balance* of the protected attributes into *Higher risk* and *Lower risk*, using the threshold of 33%, which corresponds to judgements "imbalanced" for higher risk and to judgements "unknown" + "balanced" for lower risk (*see section 5.2)*;
- analysis of the relationship between the balance measures and the unfairness: we compare the values of the unfairness measures related to the protected attributes classified as *Higher risk* vs. the values related to those classified as *Lower risk*. The risk induced by imbalance in the protected attributes can be confirmed if we can observe higher unfairness relative to those classified as *Higher risk* (*see section 5.2*).

As a further assessment step we computed the correlation coefficient between balance and unfairness measures. For this purpose we selected the Spearman correlation coefficient since we are not expecting anything like a simple linear correlation, we aim to check if a looser – rank-based – relation exists (results details always in section5.2).

[9]The identification was performed taking as reference the attributes defined in "Article 21 - Non-discrimination" of the EU Charter of Fundamental Rights [27]: 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.

*4.2.1.* ***Fairness assessment***

We assessed the *unfairness* of automated classifications relying on two criteria formalized in [33]. In general, to evaluate unfairness we consider a sensitive categorical attribute $A$ that can assume values $(a_1, a_2, ...)$, a target variable $Y$, and a predicted class $R$. In practice we aim to check whether the ADM system, which assigned a predicted class, behaved fairly w.r.t. the different values of a sensitive attribute.

The ***Separation*** criterion, where R is binary (i.e., R=0 or R=1 and thus Y=0 or Y=1), requires the equivalence of true positive rate and false positive rate for each level of the protected attributed under analysis:

$$P(R = 1 \mid Y = 1, A = a_1) = P(R = 1 \mid Y = 1, A = a_2) = ...$$

$$P(R = 1 \mid Y = 0, A = a_1) = P(R = 1 \mid Y = 0, A = a_2) = ...$$

If A is binary (that is, $A = a_1$ or $a_2$), then we can compute two Separation *unfairness* measures ($\mathfrak{U}$) as:

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

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

The ***Independence*** criterion requires the acceptance rate to be the same in all groups, where acceptance correspond to the event R=1. In term of probability, this condition correspond to the following constraint:

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) = ...$$

As before, if $A$ is binary, we can compute the Independence *unfairness* measure as:

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

The definitions above can be easily extended to the case of non-binary attributes – i.e. $m > 2$ – by taking the mean of indexes computed considering all the possible pairs of levels in $A$:

$$\mathfrak{U}(a_1, ..., a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathfrak{U}(a_i, a_j)$$

The unfairness measures defined above range in the interval $[0, 1]$. They assume values equal to zero for a perfectly fair classification and higher values for unfair behavior.

Note that when the dataset under analysis contained the score variable, we assessed the unfairness and computed the balance measures on the full dataset; in case the score variable was not included in the original dataset we defined a logistic regression model, we assessed the unfairness on the test set of the model whereas we computed measures to the training set of the model.

### *4.2.2.* ***Datasets***

We examine six datasets belonging to three different application domains: criminal justice systems (also juvenile), financial services, and social-related topics – personal earnings and education–. We sought some variety among popular datasets in order to explore the potential of our approach in several fields of application of ADM systems. All datasets were retrieved from popular machine learning websites, such as `kaggle` [10] and the UCI Machine Learning Repository [11]. The main features of the selected datasets are summarized in table 1.

Table 1: Complete list of the datasets with the analyzed attributes.

| **Dataset** | **Domain** | **Protected attributes** | **Target** | **Score** |
|---|---|---|---|---|
| **COMPAS** | Justice | *ethnicity, sex, age category* | recidivism risk | `COMPAS_risk_score` |
| **Juvenile justice** | Justice | *sex, stranger, country of origin, area of origin, age category, age* | recidivism risk | `SAVRY_total_score` |
| **Default of credit cards clients** | Financial | *sex, education* | default payment next month | *missing* |
| **Statlog** | Financial | *status, sex, foreign worker* | creditworthiness | *missing* |
| **Income** | Social | *education, race, sex, native country* | income bracket | *missing* |
| **Student** | Social | *sex, age, mother's job, father's job, mother's education, father's education* | final grade (separate for Mathematics and Portuguese) | *missing* |

[10] `https://www.kaggle.com/datasets`

[11] `https://archive.ics.uci.edu/ml/datasets.php`

***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, among which we took *sex*, *race* and *age category* into account as sensitive attributes, while we assumed *two_year_recid* as target variable and the *risk score* as classifier R, which indicates a "recidivism degree" between 1 and 10, and can be interpreted as estimated recidivism risk if above 4, so that it represents a binary classifier [51].

We chose the COMPAS dataset because it is well-known in 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 was distorted in favor of white individuals, whereby those who were rearrested were nearly twice as likely to be misclassified as low risk than black defendants [12]. Furthermore, the black defendants who did not get rearrested were nearly twice as likely to be misclassified as higher risk (false positive) than white defendants. The major cause was that the number of records in the dataset related to black defendants was much higher than the number of records of white defendants, as well as the number of black recidivists compared to white recidivists.

***Juvenile justice***. This dataset consists of 4753 data and presents the statistical descriptive variables, as well as the recidivism of children and young people who completed an educational program in 2010 in Catalonia, between the date of completion of the program and the end of 2013 or the end of 2015 [52]. The dataset describes the profile of youths and also minors who had contact with juvenile justice in relation to the program done. Additional provided data are: the juvenile recidivism rate, the specific rates and the profile of the recidivist and recidivism according to the program. In particular, the SAVRY variables present the risks of recidivism among young people, as well as their specific areas of risk and needs; among them, $SAVRY\_total\_score$ indicates a "total recidivism degree" between 1 and 100. In order to assume it as binary score variable, we make reference to the COMPAS dataset, where the total percentage of moderate and high recidivism risk is around 45%, so

[12] `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm` , last visited on Feb 4, 2021

we decide to consider the same percentage of data in the Juvenile dataset as moderate-high risk: following this line of reasoning, $SAVRY\_total\_score$ can be interpreted as affirmative (estimated) recidivism risk if above 15. Then, we assumed $reincidencia\_2013$ as target variable, which represents the recidivity by the end of 2013, and we examined the protected attributes *sex*, *stranger*, *country of origin*, *area of origin*, *age category* and *age*.

Differently than the first two, the following datasets do *not* contain a pre-computed classification, so we built a *binomial logistic regression* model in order to predict the score variable: in particular, we trained a binary classifier on a *training* set composed by the 70% (randomly selected) of the original dataset and we ran it on the remaining 30%, which represents the *test* set.

***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 [53]. The dataset is composed by 25 variables and some of them have demographic character; in particular, we consider as protected attributes *sex* and *education*, while we assumed $default.payment.next.month$ as target variable and $default\_pred$ as classifier $R$ (obtained through the binomial logistic regression). The credit card default dataset was chosen because of the high impact of using ADM systems 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 third most voted dataset on credit cards on Kaggle[13] 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.

***Statlog***. This widely used German credit dataset from the UCI Machine Learning Repository [54] has been provided by the German professor Hans Hofmann as part of a collection of datasets from an European project called "Statlog" [55] and will be simply called Statlog in the following. The data are a stratified sample of 1000 credits (700 good ones and 300 bad

[13] `https://www.kaggle.com/datasets?search=credit+card&sort=votes`, last visited on February 4, 2021.

ones) and have been collected between 1973 and 1975 from a large regional bank in southern Germany, which had about 500 branches, both urban and rural ones. Bad credits have been heavily oversampled, with a view to acquiring sufficient information for discriminating them from good ones [56]. Specifically, the dataset contains 1000 entries with 20 categorical attributes: each entry represents a person who takes a credit by a bank and is classified as good or bad credit risks according to the set of attributes, among which we considered *status*, *sex* and $foreign\ worker$ as sensitive attributes. As indicated with the Statlog data [54], one might examine misclassification cost: it has been suggested to allocate the cost for misclassifying a bad risk as good to be five times as high than the cost for misclassifying a good risk as bad [56], therefore we assumed $cost\_matrix$ as target variable (equal to 0 or 1) and we built the predictions through a binomial logistic regression model.

***Income***. The extraction of these data was realized by Barry Becker from the 1994 Census database; the prediction task is to determine whether a person makes over $\$50,000$ a year based on that set of reasonably clean records, also known as "Census Income" dataset [57]. Thus, *test.income* represents the target variable, which can assume the two values $<= 50K$ or $> 50K$, whereas we built a regression model to predict the corresponding score. In our analysis we took into account the protected attributes *education*, *race*, *sex* and *native country*.

***Student***. These two datasets contain information on student achievement in secondary education of two Portuguese schools and they have been built by using school reports and questionnaires in 2014. The attributes include student grades, as well as demographic, social and school related features. Two datasets are provided from the UCI Machine Learning Repository [58] regarding the performance of students (not necessarily the same students) in two distinct subjects: Mathematics and Portuguese language. In our study we took into account both the datasets and we considered the sensitive attributes *sex*, $mother's\ job$, $father's\ job$, *age*, $mother's\ education$ and $father's\ education$ for each of the two datasets, whereas we assumed the variable $G3$ as target variable, which indicates the final year grade (issued at the end of the school year) between 1 and 20, corresponding to a positive grade if above 9, or negative if lower [59]. Finally, we built the corresponding binary predictions

through the regression model.

## 5. Analysis of Results

In this section we examine the results of the analysis for each research question. The whole analysis environment (data, code, tools, settings) is openly available at `https://codeocean.com/capsule/3628819/tree/v3`, where it is possible to reproduce the results with a single click.

### *5.1.* ***RQ1 - Imbalance detection***

As detailed in 4.1, we tested the balance measures in presence of different distributions of the occurrences between the classes of a certain attribute. Results are reported in figure 5. The figure shows each synthetic attribute as a rectangle whose border color encodes the expected class: blue means balanced, red imbalanced, and gray undecided. For each combination of synthetic attribute and balance measure, the figure reports the result of the classification as a colored tile using the above encoding, along with the value of the measure.

We can observe that for the first three groups of synthetic distributions – Max Imbalance, Max Balance, and Quasi Balance – all measures provide an accurate identification of the class. Concerning the remaining cases:

- Gini and Shannon provided the right class in just 2 cases out of 12, Simpson detected correctly 4 classes, Imbalance Ratio was accurate in detecting 8 classes.
- The same results can be read from the perspective of these three latter distributions (i.e., those in the second row of figure 2):
  - One Off is the distribution where the indexes performed best, with 8 correct cases out of 16 (mostly in correspondence of $m = 2$ and $m = 8$), two correct cases for each one;
  - in Half high, we observe 5 correct cases out of 16, 4 or which are from the Imbalance Ratio index, and the last one being Simpson index with $m = 2$;

  - in Power 2 distributions, only 3 cases out of 16 were correctly detected by the indexes: 2 for Imbalance Ratio index ($m = 5$ and $m = 8$) and 1 for Simpson index ($m = 8$).

- Also, we observe that 13 out of the possible 32 cases of imbalance were detected (40%), 3 out of 4 of balance, and none of the 12 classes of the undecided category.

From the point of view of the number of classes, we cannot derive a clear tendency: in fact, 6 correct classifications are in correspondence of both $m = 2$ and $m = 8$, and 2 classifications are in correspondence of both $m = 3$ and $m = 5$.

By looking more in details at the values of the measures, we can observe that Imbalance Ratio has the lowest values: this can be explained by looking at its definition that takes the ratio of the two extreme frequencies. The highest values are those computed using the Gini and Shannon indexes, while the Simpson index has intermediate values.

The capability to detect imbalance can be summarized in terms of the overall accuracy of the classification reported in figure 6.

In general, we observe that all indexes have some drawbacks. We ought to emphasize that this is an exploratory study based on a limited number of synthetic attributes whose goal is to provide a basic understanding of the balance measures.

### *5.2.* ***RQ2 - Discrimination risk***

Figure 7 reports, for each combination of balance and unfairness measures, a boxplot with the distribution of unfairness measure values for higher risk vs. lower risk attributes: we remind from Section 4 that we used the threshold 33%, corresponding to "imbalanced" for higher risk, and "unknown" + "balanced" for lower risk. The more a boxplot leans to the right the more unfair the treatment of those attributes are treated. And vice-versa: the more a boxplot is close to left (i.e. zero) the more the relative attributes are treated fairly. As a general rule, when the boxes of the riskier attributes (colored in red) and the one for the less riskier (yellow) are not overlapping, it means that the imbalance-based approach to risk identification is able to discriminate between actually fair and unfair classifications. We observe that:

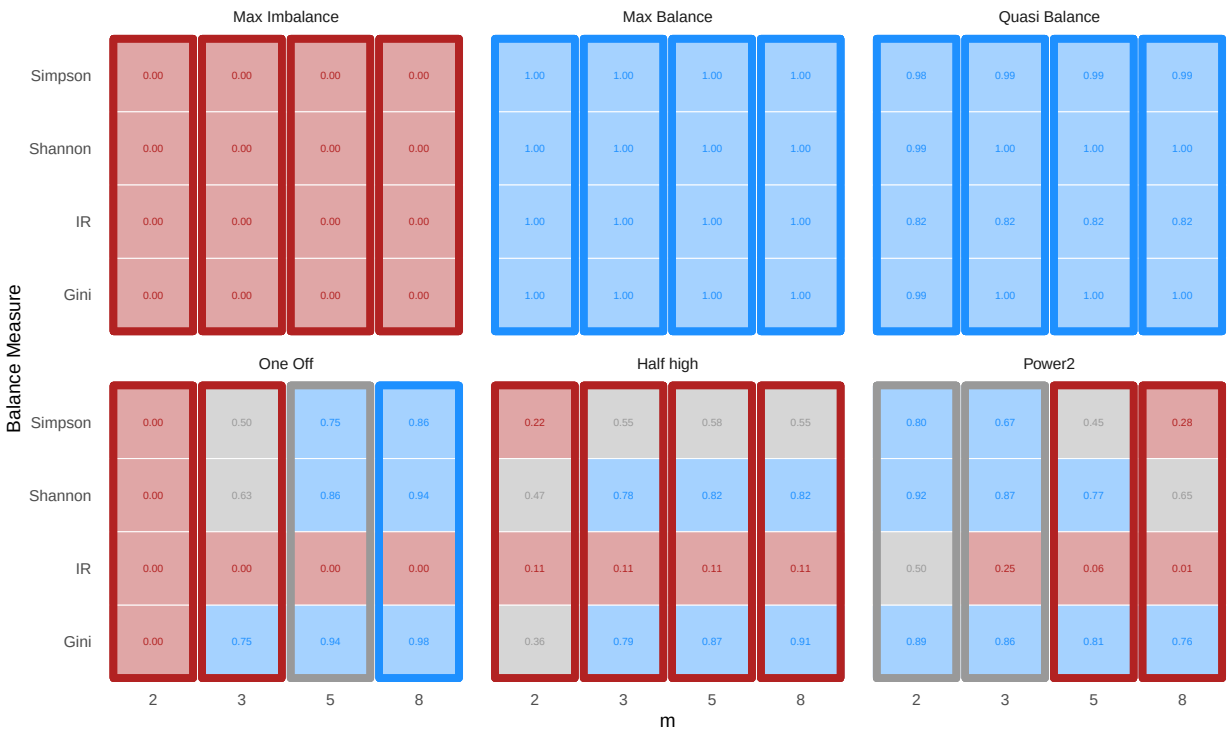


Figure 5: Classification of synthetic attributes based on the balance measures.

- Gini index had a good discrimination ability for the true positive rates of the Separation criterion and essentially no discrimination for the other two unfairness measures;
- Imbalance Ratio had a good discrimination ability for both the indicators of the Separation criterion, and a limited ability for the Independence criterion;
- Shannon index had a good discrimination for the Independence criterion, excellent for the true positives rates of the Separation criterion, and no discrimination for the false positive rates in the Separation criterion;
- the Simpson index had a limited capability on Independence criterion, and no discrimination for the Separation criterion.

According to this analysis, we can summarise that all indexes but Simpson were able to detect discrimination in terms of substantial difference of true positives; the indexes are moderately able to detect discrimination in terms of different acceptance rates; all indexes, with the notable exception of Imbalance Ratio, are not able to anticipate discrimination in

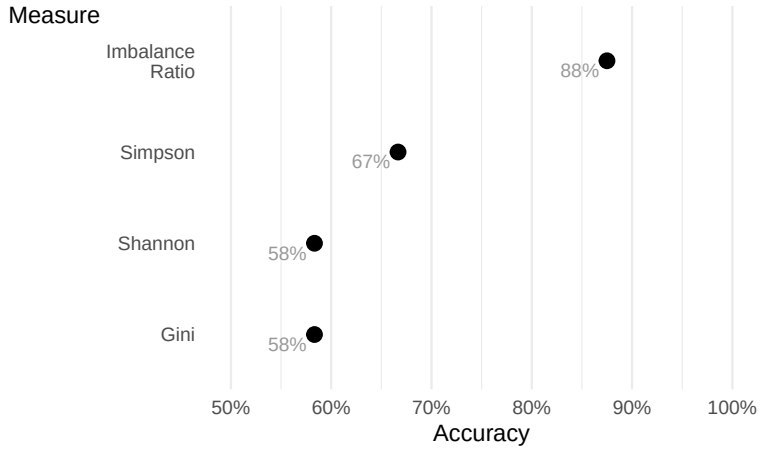


Figure 6: Balance measures ability to detect imbalance.

terms of substantial difference of false positives.

The values that have been summarized in figure 7 are reported in detail in table 2, where we show for each dataset (and target) all the protected attributes, and for each one the balance measures. The three rightmost columns contain the mean value of the unfairness measures related to Independence and Separation: following the line of reasoning explained above, for high level of unfairness we expect low-value indexes –that reveal imbalance in data. Looking at the single attributes :

- starting from the COMPAS dataset, previous studies [29] have shown that the data is imbalanced in favor of white people, as the highest levels of reoffending are observed in black individuals. Indeed, as regards "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. In confirmation of this, both the fairness criteria reveal high level of unfairness; at the same time, the balance measures confirm the presence of data imbalance, with low and medium values for the Imbalance Ratio and Simpson indexes, and just a relatively high value for the Gini and Shannon indexes.

- A similar relation between balance and unfairness measures can be observed, for instance, for the attribute "Country of origin" in the Juvenile justice dataset, but also

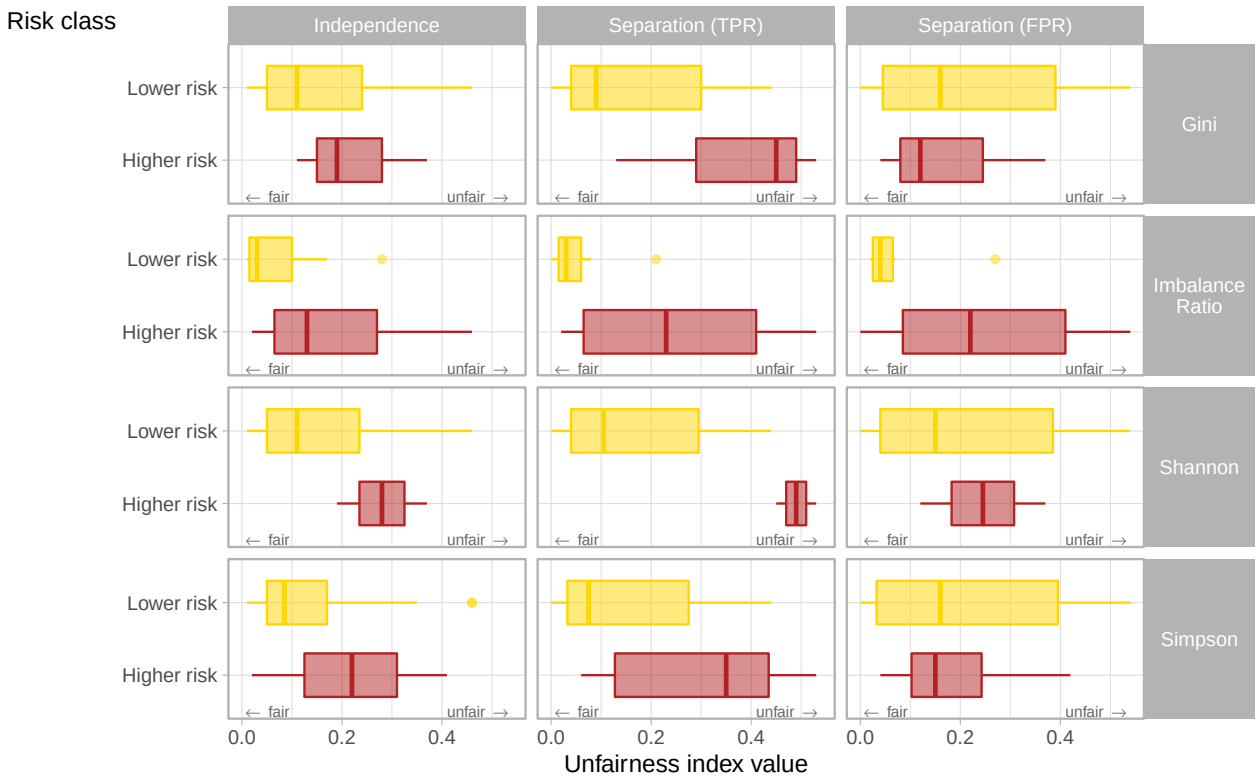


Figure 7: Boxplot of unfairness measures vs. balance classification, for different balance measures.

for "Foreign worker" in Statlog or "Native country" in the Income dataset.

- Vice versa, correspondingly to low unfairness values we note overall high balance indexes, denoting also in this case a negative correlation with unfairness measures. For instance, for the sensitive attributes "Stranger" in the Juvenile justice dataset, "Sex" in the Credit card default dataset, "Sex" in Statlog, "Sex" in both the Student-Mathematics and Student-Portuguese datasets, we found low unfairness levels, which are reflected by very high and similar balance measures –the Gini, Shannon and Simpson indexes above all.

- The previous trend does not held for all the attributes: for example, with respect to "Age category" in COMPAS, the fairness tests reveal high level of unfairness, but the balance measures tend to be higher than expected, with values between 0.36 and 0.89.

- Also for "Status" in Statlog we note medium and high unfairness values in correspondence of high balance measures, as well as for the attributes "Education" and "Sex"

in Income, "Age" and "Mother's education" in Student-Mathematics, or "Age" in Student-Portuguese.

We integrate the analysis with the computation of the Spearman correlation coefficient between *balance* and *unfairness* measures, as reported in table 3. Specifically, we expect the coefficient to be negative: the higher the balances indexes, the lower the unfairness measures. Hence, in term of correlation, the best balance measure is the Imbalance Ratio index as it always presents a strong negative correlation followed by the Simpson and the Shannon indexes respectively. The less accurate measure appear to be the Gini index . The correlation analysis confirms that false positive differences are the most difficult to detect with the four indexes, while results are encouraging for the discrimination with respect to both Independence and the TP rates of the Separation criterion.

## 6. Discussion

*RQ1. How are existing measures able to detect imbalance among the classes of a given attribute in a dataset?*

Overall we can conclude that the Imbalance Ratio index is the most precise measure for detecting imbalance among the classes of a given attribute, according to the exemplar distributions chosen. However, the index is very sensitive when classes have 0 occurrences: when just a class is empty, the index goes to 0. An intermediate result is achieved by Simpson index, while Gini and Shannon indexes exhibit the lowest performances: since, in general, they have higher values than the former two for the same distributions. We might need to study whether different thresholds should be applied. We did not observe any trend associated to the number of classes. We also observed that the worst performances occurred in the case of a Power law with base 2 distribution: since power law distributions are very common in a variety of real cases (e.g., income), we shall extend the analysis to understand how to adequately deal with this family of distributions. More in general, results are encouraging enough to continue the exploration with a more extensive catalogue of distributions.

*RQ2. Are existing measures able to reveal a discrimination risk when an ADM system is trained with such data?*

Table 2: Values of balance measures and unfairness measures.

| Dataset | Attribute | m | Gini | Shannon | Simpson | IR | Independence | Separation (TPR) | Separation (FPR) |
|---|---|---|---|---|---|---|---|---|---|
| **COMPAS** | | | | | | | | | |
| | Ethnicity | 6 | 0.73 | 0.62 | 0.31 | 0 | 0.25 | 0.29 | 0.20 |
| | Sex | 2 | 0.61 | 0.70 | 0.44 | 0.05 | 0.23 | 0.02 | 0.00 |
| | Age category | 3 | 0.87 | 0.89 | 0.69 | 0.36 | 0.28 | 0.21 | 0.27 |
| **Juvenile justice** | | | | | | | | | |
| | Sex | 2 | 0.44 | 0.54 | 0.28 | 0.14 | 0.02 | 0.12 | 0.05 |
| | Stranger | 2 | 0.94 | 0.96 | 0.90 | 0.63 | 0.03 | 0.04 | 0.03 |
| | Country of origin | 35 | 0.61 | 0.44 | 0.04 | 0 | 0.41 | 0.43 | 0.42 |
| | Area of origin | 5 | 0.70 | 0.67 | 0.32 | 0.02 | 0.13 | 0.06 | 0.14 |
| | Age category | 3 | 0.66 | 0.59 | 0.39 | 0 | 0.06 | 0.41 | 0.02 |
| | Age | 5 | 0.89 | 0.83 | 0.63 | 0.01 | 0.05 | 0.31 | 0.07 |
| **Credit card default** | | | | | | | | | |
| | Sex | 2 | 0.95 | 0.96 | 0.91 | 0.65 | 0.02 | 0.01 | 0.02 |
| | Education | 6 | 0.75 | 0.60 | 0.33 | 0 | 0.06 | 0.16 | 0.03 |
| **Statlog** | | | | | | | | | |
| | Status | 4 | 0.93 | 0.91 | 0.77 | 0.18 | 0.15 | 0.40 | 0.10 |
| | Sex | 2 | 0.85 | 0.89 | 0.75 | 0.45 | 0.01 | 0.02 | 0.06 |
| | Foreign worker | 2 | 0.17 | 0.26 | 0.09 | 0.04 | 0.37 | 0.53 | 0.37 |
| **Income** | | | | | | | | | |
| | Education | 16 | 0.86 | 0.73 | 0.28 | 0 | 0.29 | 0.41 | 0.16 |
| | Race | 5 | 0.32 | 0.34 | 0.08 | 0 | 0.11 | 0.13 | 0.04 |
| | Sex | 2 | 0.88 | 0.91 | 0.79 | 0.49 | 0.17 | 0.08 | 0.07 |
| | Native country | 42 | 0.20 | 0.17 | 0 | 0 | 0.19 | 0.45 | 0.12 |
| **Student** - Mathematics target | | | | | | | | | |
| | Sex | 2 | 0.99 | 0.99 | 0.99 | 0.95 | 0.03 | 0 | 0.02 |
| | Age | 8 | 0.89 | 0.77 | 0.51 | 0.01 | 0.46 | 0.44 | 0.40 |
| | Mother's job | 5 | 0.94 | 0.93 | 0.77 | 0.23 | 0.10 | 0.07 | 0.22 |
| | Father's job | 5 | 0.78 | 0.74 | 0.42 | 0.07 | 0.23 | 0.23 | 0.33 |
| | Mother's education | 5 | 0.91 | 0.86 | 0.69 | 0.03 | 0.46 | 0.41 | 0.44 |
| | Father's education | 5 | 0.93 | 0.87 | 0.74 | 0.01 | 0.17 | 0.09 | 0.38 |
| **Student** - Portuguese target | | | | | | | | | |
| | Sex | 2 | 0.97 | 0.97 | 0.94 | 0.70 | 0.01 | 0.03 | 0.04 |
| | Age | 8 | 0.87 | 0.74 | 0.47 | 0 | 0.35 | 0.29 | 0.48 |
| | Mother's job | 5 | 0.93 | 0.92 | 0.74 | 0.21 | 0.11 | 0.05 | 0.54 |
| | Father's job | 5 | 0.75 | 0.72 | 0.38 | 0.06 | 0.06 | 0.02 | 0.53 |
| | Mother's education | 5 | 0.93 | 0.86 | 0.72 | 0.02 | 0.11 | 0.04 | 0.51 |
| | Father's education | 5 | 0.93 | 0.86 | 0.72 | 0.02 | 0.07 | 0.04 | 0.32 |

Table 3: Correlation between balance measures and unfairness measures.

| Fairness criteria \ Balance Measures | Gini | Shannon | Simpson | Imbalance Index |
|---|---|---|---|---|
| **Independence** | -0.278 | -0.352 | -0.435 | -0.514 |
| **Separation (TPR)** | -0.474 | -0.575 | -0.604 | -0.667 |
| **(FPR)** | 0.012 | -0.085 | -0.181 | -0.288 |

As a general consideration we notice that there is no single balance measure providing the basis for an ideal risk identification across all datasets analysed. Similarly to the previous research question, the Imbalance Ratio anticipates discrimination better than other indexes, although the correlation analysis showed that all indexes are able to detect – each one with its own strengths and weaknesses – a substantial difference of acceptance rates and true positive rates. Discrimination due to false positive rates, instead, is much more difficult to be detected, especially for the Gini index.

*Recommendations for the usage of the indexes*

On the basis of the exploratory analysis conducted here, we can take into consideration different indexes to identify potential unfairness risks. We recommend to consider the following aspects when using the indexes:

- the Imbalance Ratio index has a good capacity to detect both imbalance in the exemplar distributions and discrimination in real cases, but it shall not be used if very few classes are empty or close to zero;

- Gini and Shannon indexes are moderately able to detect discrimination, but they have a tendency to assume high values (in fact, they had the worst performances in detecting imbalance with the exemplar distributions): for these reasons, we recommend to test them with lower thresholds to avoid missing relevant cases of imbalance;

- Simpson index has a good capability of detecting imbalance, according to the exemplar distributions used, but a limited capability to detect discrimination insofar it is used with the current thresholds: hence, we recommend to use it in combination with

Imbalance Ratio for a preliminary analysis of the possible cases of discrimination, since it is not affected by the presence of empty classes and it still has correlations comparable to those of Imbalance Ratio.

### *6.1.* ***Relation to Policy***

We extensively reported on how and why imbalance in data used to build ADM systems challenges a founding element of the rule of law in our democratic societies: the principle of non-discrimination [27]. The "Recommendation of the Committee of Ministers to member states on the human rights impacts of algorithmic systems" [60], published by the Council of Europe (CoE) on April 8, 2020, emphasizes the impact of algorithmic systems on human rights and the need for additional normative protections. Although the CoE cannot issue binding laws, it is the main organization for safeguarding human rights in Europe, and for this reason the recommendation is of particular interest for our purposes. The document defines "high risk" in correspondence with *"the use of algorithmic systems in processes or decisions that can produce serious consequences for individuals or in situations where the lack of alternatives prompts a particularly high probability of infringement of human rights, including by introducing or amplifying distributive injustice"* (p.5). In these situations, *"risk-management processes should detect and prevent the detrimental use of algorithmic systems and their negative impacts"* (p.6). The recommended obligations for the states include a continuous review of algorithmic systems throughout their entire lifecycle. In terms of data management, bias in the data as a risk factor for systematic discrimination is explicitly cited: *"States should carefully assess what human rights and non-discrimination rules may be affected as a result of the quality of data that are being put into and extracted from an algorithmic system, as these often contain bias and may stand in as a proxy for classifiers such as gender, race, religion, political opinion or social origin"* (p.7). The document adds that bias and discriminatory outputs should be properly tested since the analysis and modeling phase and that system development should be even *"discontinued if testing or deployment involves the externalisation of risks or costs to specific individuals, groups, populations and their environments"* (p.8). Precautionary measures should include risk assessment procedures to evaluate potential risks and minimize adverse effects, in cooperation with all relevant stake-

holders. Similar obligations are recommended to the private sector, as part of their social responsibility: after referring to the UN Guiding Principles on Business and Human Rights [61] in the introduction (p.3), the document prescribes that: *"Private sector actors should be cognisant of risks relating to the quality, nature and origin of the data they are using for training their algorithmic systems, ensuring that errors, bias and potential discrimination in datasets and models are adequately responded to within the specific context"* (p.12). Independent expert review and oversight should take place also for private entities (p.14), who are demanded to adjust or discontinue the development of the systems if risks cannot be mitigated (p.13).

Looking at the Institutions of the European Union (EU), the problem of biased ADM systems is widely recognized, as acknowledged by the words of Margrethe Vestager that we reported in the Introduction. The words of M. Vestager should be considered in the context of the ongoing efforts of the EU to redefine the markets rules in response to the rapid technological advancements related to the emergence of automated decision making processes. As a matter of fact, we recall the "Resolution on automated decision making processes and consumer protection" [62] which was approved by the EU Parliament on February 6, 2020. The document is relevant because it comes from the highest legislative Institution in the EU and because therein, we find explicit references to the two foundational elements of our proposals. More precisely, the Parliament stresses:

- *"the need for a risk-based approach to regulation, in light of the varied nature and complexity of the challenges created by different types and applications of AI and automated decision-making systems"* (p. 4);

- *"the importance of using only high-quality and unbiased data sets in order to improve the output of algorithmic systems and boost consumer trust and acceptance"* (p.11-12).

Although the general context of the Resolution is market surveillance, it is still within the ambit of the EU Charter of Fundamental Rights, and in particular Article 38 on consumer protection [63]. It is worth reminding that the European Commission acknowledged the problem of biased ADM since the publication of its communication "Artificial Intelligence for Europe" [64] on April 25, 2018 by stipulating *"Whilst AI clearly generates new*

*opportunities, it also poses challenges and risks, for example [...] bias and discrimination"* (p.15). Notwithstanding the non-binding value of the document, this communication paved the way for several other policy documents, among which we mention the Coordinated Plan on Artificial Intelligence [65] and the Strategy for Artificial Intelligence [66]. It is worth observing that the communication "AI for Europe" was published one month before than the General Data Protection Regulation had effect (GDPR, in effect from May 25, 2018), hence it is coherent to the obligations for data controllers to safeguard the data subject's rights and freedoms and legitimate interests in automated individual decision-making (Art. 22).

In the given policy document examples, the term "risk management" recurred often and hitherto it is indicated as the more suitable approach for regulating algorithmic systems. We corroborate our statement with a selection of the documents issued by several bodies of the European Union or by experts groups appointed by them. Since the following policy documents are heterogeneous in terms of binding value and domains of application, we present them following a simple chronological order.

- The EU White Paper on Artificial Intelligence - A European approach to excellence and trust [67], published on February 19, 2020, promotes an AI regulation proportional to the impact of systems on citizen's lives. It is reported that in high-risk cases (e.g., the health domain) mandatory testing and certification of adopted algorithms should be put in place, while in all the other cases a voluntary quality labelling scheme can be adopted (critics highlighted that the definition of risk is too vague). Bias and discrimination are reported as *"practical impacts of the correlations or patterns that the system identifies in a large dataset"* (p.12) while on operation: indeed, it has been widely discussed that disparities in the representation of population groups can be present since the training of algorithms phase.

- Between April and July 2020, several documents from the High-Level Expert Group (HLEG) on AI were published [14]. In the first document, i.e., the "Ethics Guidelines for Trustworthy AI" [68] (April 8, 2019), one of the seven identified requirements is

[14]The High-Level Expert Group was formed in 2018 to support the implementation of the European

non-discrimination, which requires the removal of unfair bias in datasets since the data collection phase (p. 18, p.29, p.36). In the following two documents, i.e., "Policy and Investment Recommendations" [69] (June 26, 2019) and "Assessment List" [70] (July 17, 2020), the requirement of diversity, non-discrimination and fairness is further described, and auditing mechanisms are encouraged (by public enforcement authorities or by independent third-party auditors) within the general context of risk-based governance of AI.

- The EU Committee of Civil Liberties, Justice and Home Affairs published the report "Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters" [71] in June 2020. It *"stresses the potential for bias and discrimination arising from the use of machine learning and AI applications"* (p.6), specifying that biases can be present both in historical data used for training algorithms and in data generated by systems already functioning.

- One month later, the sector-based study of "Artificial Intelligence and Civil Liability" [72] was published by the EU Policy Department for Citizens' Rights and Constitutional Affairs: a whole section of the document (from p.96 to p.121) is devoted to specifying a *"risk-management approach to the regulation of civil-liability in AI-based applications"* at theoretical and methodological level, and with specific considerations on four case studies.

- Finally, the recent (December 15, 2020) Digital Services Act [73] of the European Commission paves the way for risk management obligations and independent audits but only for very large platforms (i.e., reaching more than 10% of the 450 million consumers in Europe), in order to better protect customers from unfair practices and to safeguard the competition in contestable and fair markets. In part 1 of its impact assessment [74] (p.63) *"biases potentially embedded in the notification systems by users*

strategy on AI in identifying the principles that should guarantee the development of a "trustworthy AI". From its formation and up to the time of writing this manuscript, the HLEG published four deliverables, three of which are relevant in this discussion.

*and third parties"* are reported to make specific groups disproportionately affected by restrictions and removal measures adopted by the very large platforms where most of both private communications and public discourses occur.

In addition to the documents issued at European level, we mention that a similar approach has been recently indicated also by national authorities (for example, in Germany by the Federal Anti-Discrimination Agency [75] and the Data Ethics Commission [76]), and on the other side of the Atlantic Ocean, where the US Congress Algorithm Accountability Act of 2019 [77] prescribed that automated decision systems that may contribute to inaccuracy, bias, or discrimination, shall undergo assessments of these risks and identification of the actions to minimize them. Also in this case, the obligations might apply only to very large companies.

This overview of the most recent efforts on regulating algorithmic systems, with a clear focus on Europe, defines the legislative context in which our proposal should be placed. We showed that the risk-based approach is a cornerstone element of the European regulation of algorithmic systems, which is currently under redefinition[15]: the prescriptions of the policy documents are not only at a general and declamatory level but they also act in specific matters. Our proposal can potentially cross this path, whereby imbalance measures can be suitable risk indicators of propagation of bias in the input data of ADM systems. In addition, they can be used for certification and labeling purposes, as our notes in Section 3 highlights.

### *6.2.* ***Threats to validity and limitations***

The results about RQ1 (imbalance detection) are highly dependent on the judgements of the authors (construct validity) and on the exemplar distributions chosen (conclusion validity). Given the exploratory nature of the work, we aimed at simplicity and not at an exhaustive test of the possible levels of imbalance, which are infinite from a prospective of marginal increments. Nevertheless, a higher number of notable distributions and a larger pool of judgements (e.g., via crowd-voting) are necessary to increase the validity of the

[15]The AI regulation proposal by the European Commission (21 April 2021) was intentionally not reported in the list because subject to numerous future negotiations (it still has to go through the Parliament and the Council).

findings. In addition, in-depth sensitivity analyses on the thresholds used for the balance and unfairness measures should improve the reliability of the overall results.

The same sensitivity analysis on the thresholds could be performed in followup analyses of RQ2 (discrimination risk) to improve the reliability and generalizability of the findings. As far as other limitations of RQ2 results are concerned, we remark that for the datasets where a classification label was present, we have no knowledge about the type of classification model used on such datasets, thus we do not know whether the observed relationship is exclusively connected to the imbalance in the data: confounding factors may be present and affect the internal validity[16]. On the contrary, we obtained much more control over the datasets for which we ran a classification model, the binomial logistic regression specifically. In all these cases the limitations of the algorithm hold, most notably the assumption of linearity between the dependent variable and the independent variables, as well as the assumption of limited or no multi-collinearity between independent variables.

Applying more classification algorithms (each with different parameters) would be necessary not only to improve the reliability of the relationship found between balance and unfairness, but also to increase the generalizability of the results (external validity): it will help to identify how the different types of classification algorithms propagate the imbalance. Other possible extensions regard the usage of further unfairness measures, e.g., sufficiency from [33] and the identification of balance measures for non-categorical data.

Overall, it is important to stress again the fact that our study focused on the level of risk analysis. Risk evaluation (i.e., which criteria should activate which actions) has been left out of the scope of our research. In order to understand how to manage the discrimination risk, the literature on machine learning and big data will be a useful resource to select and test imbalance mitigation techniques [78] [79], that are usually classified according to the different phases of the machine learning pipeline: pre-processing techniques aim at re-balancing the training data, thus mostly operating at data level; in-processing techniques are applied at the training phase, operating both at algorithm level and at data level; post-processing

[16]However, for the largely debated COMPAS case, the imbalance in the input data has been widely recognized in the literature as a relevant factor for the observed discrimination

methods mitigate bias on the already predicted labels (data level). It should be observed that these data-engineering aspects are still object of research because of inconsistent and conflicting results [78]; in addition, they should be combined with other perspectives that factor in the socio-technical nature of the problem: for example, both ethical considerations and legal requirements shall be included to find meaningful thresholds of risks in relation to the context of use and the severity of the impact on individuals.

## 7. Conclusions

In this article we proposed and tested a metric-based approach to evaluate imbalance in a given dataset as a potential risk factor for discriminatory output of ADM systems. The approach combines aspects of data quality and risk management from the ISO standards and it resonates with the most recent European policy proposals for regulating digital services, including ADM systems. We selected four widely used indexes (Gini, Simpson, Shannon, Imbalance Ratio), normalized them to share the same range of values, and tested their ability to detect (i) different levels of imbalance in synthetic attributes and (ii) discrimination occurring in the classification outcome of ADM systems trained with six large datasets. Concerning the detection of imbalance, the best result is achieved using the Imbalance Ratio and the Simpson index, while the Gini and Shannon indexes constantly assume higher values (suggesting balanced data), and for this reason they should be further investigated, e.g. using different thresholds. Regarding the ability to detect discrimination, the balance measures performed differently with respect to different fairness criteria and can be ranked similarly as for the previous goal. As a general indication, evidence suggests that a combined usage is preferable to detect possible discrimination risks: for this reason, after discussing the results, we elaborated a few pragmatic recommendations for their application.

Overall, the results indicate that the approach is suitable for the proposed goal. However further work is needed to better assess the reliability of the balance measures as risk indicators, for instance by considering different classification and prediction algorithms, a larger number of exemplar distributions, and a sensitivity analysis on the thresholds. The work could be also expanded by including balance measures applicable to continuous attributes, additional criteria of fairness. In addition potential mitigation actions should be examined,

that factor in both data engineering aspects and procedural or organizational aspects that reflect the social dimension of the problem: for example, the severity of the impact on disadvantaged users, in combination with the legal and ethical issues related to specific application domains.

We hope that this exploratory analysis could stimulate a community effort for more extensive analyses from other researchers and investigations on real systems from policy actors: we aim at building an open and extensible benchmark of balance measures and measurements to attract further contributions. We also recommend software companies and research teams to include such measures in the toolkits mentioned in the related work.

We conclude by remarking that this work (and most of those we cited) fall within the wide landscape of principles, methodologies and tools for a data governance that should serve our society to develop ADM systems in a trustworthy and more democratic way [80]. To achieve such a goal, it is of paramount importance to protect and promote rights and freedoms in both their individual and collective dimensions. Developing socially sustainable ADM systems, especially those employed in public sector services, represents one of the mandatory actions to pursue this path.

## References


[1] F. Chiusi, S. Fischer, N. Kayser-Bril, M. Spielkamp, Automating Society Report 2020, `https://automatingsociety.algorithmwatch.org` (Oct. 2020).

[2] B. W. Wirtz, J. C. Weyerer, C. Geyer, Artificial Intelligence and the Public Sector—Applications and Challenges, International Journal of Public Administration 42 (7) (2019) 596–615, publisher: Routledge_eprint. `doi:10.1080/01900692.2018.1498103`.

[3] E. Brynjolfsson, A. McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, reprint edition Edition, W. W. Norton & Company, New York London, 2016.

[4] B. Reese, The fourth age: Smart robots, conscious computers, and the future of humanity, Simon and Schuster, 2018.

[5] E. Brynjolfsson, K. McElheran, The rapid adoption of data-driven decision-making, American Economic Review 106 (5) (2016) 133–39. doi:10.1257/aer.p20161016.

[6] L. Willcocks, M. Lacity, A. Craig, Robotic process automation: Strategic transformation lever for global business services?, Journal of Information Technology Teaching Cases 7 (1) (2017) 17–28. doi:10.1057/s41266-016-0016-9.

[7] I. I. Makrygianni, A. P. Markopoulos, Loan evaluation applying artificial neural networks, in: Proceedings of the SouthEast European Design Automation, Computer Engineering, Computer Networks and Social Media Conference, SEEDA-CECNSM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 124–128. doi:10.1145/2984393.2984407.

[8] Z. Siting, H. Wenxing, Z. Ning, Y. Fan, Job recommender systems: A survey, in: 2012 7th International Conference on Computer Science Education (ICCSE), 2012, pp. 920–924. doi:10.1109/ICCSE.2012.6295216.

[9] D. Abu Elyounes, 'Computer Says No!': The Impact of Automation on the Discretionary Power of Public Officers, https://papers.ssrn.com/abstract=3692792 (Sep. 2020).

[10] S. Kanoje, D. Mukhopadhyay, S. Girase, User Profiling for University Recommender System Using Automatic Information Retrieval, Procedia Computer Science 78 (2016) 5–12. doi:10.1016/j.procs.2016.02.002.

[11] A. Cordella, N. Tempini, E-government and organizational change: Reappraising the role of ICT and bureaucracy in public service delivery, Government Information Quarterly 32 (3) (2015) 279–286. doi:10.1016/j.giq.2015.03.005.

[12] J. B. Wenger, V. M. Wilkins, At the Discretion of Rogue Agents: How Automation Improves Women's Outcomes in Unemployment Insurance, Journal of Public Administration Research and Theory 19 (2) (2008) 313–333. doi:10.1093/jopart/mum044.

[13] P. A. Busch, The Role of Contextual Factors in the Influence of ICT on Street-Level Discretion, in: Proceedings of the 50th Hawaii International Conference on System Sciences, 2017, pp. 1–10. doi:10.24251/HICSS.2017.358.

[14] S. Barocas, A. D. Selbst, Big Data's Disparate Impact, https://papers.ssrn.com/abstract=2477899 (2016).

[15] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, reprint edition Edition, Broadway Books, New York, 2017.

[16] V. Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, St. Martin's Press, New York, NY, 2018.

[17] E. Pilkington, Digital dystopia: how algorithms punish the poor, https://www.theguardian.com/technology/2019/oct/14/automating-poverty-algorithms-punish-poor (Oct. 2019).

[18] M. Vestager, Algorithms and democracy - AlgorithmWatch Online Policy Dialogue, https://ec.europa.eu/commission/commissioners/2019-2024/vestager/announcements/algorithms-and-democracy-algorithmwatch-online-policy-dialogue-30-october-2020_en (Oct. 2020).

[19] G. Ristanoski, W. Liu, J. Bailey, Discrimination aware classification for imbalanced datasets, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 1529–1532. doi:10.1145/2505515.2507836.

[20] H. He, E. A. Garcia, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284. doi:10.1109/TKDE.2008.239.

[21] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis 6 (5) (2002) 429–449. doi:10.3233/IDA-2002-6504.

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

[23] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Trans. Inf. Syst. 14 (3) (1996) 330–347. doi:10.1145/230538.230561.

[24] J. Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, https://reut.rs/2Od9fPr (Oct. 2018).

[25] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. Kalai, Bias in bios: A case study of semantic representation bias in a high-stakes setting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 120–128. doi:10.1145/3287560.3287572.

[26] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. Rieke, Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes, Proc. ACM Hum.-Comput. Interact. 3 (CSCW). doi:10.1145/3359301.

[27] E. U. A. for Fundamental Rights, EU Charter of Fundamental Rights - Article 21 - Non-discrimination, https://fra.europa.eu/en/eu-charter/article/21-non-discrimination (December 2007).

[28] T. Jan, E. Dwoskin, Facebook is sued by HUD for housing discrimination, https://www.washingtonpost.com/business/2019/03/28/hud-charges-facebook-with-housing-discrimination.

[29] S. M. Julia Angwin, Jeff Larson, L. Kirchner, Machine Bias—ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).

[30] Z. Obermeyer, S. Mullainathan, Dissecting racial bias in an algorithm that guides health decisions for 70 million people, in: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 89. doi:10.1145/3287560.3287593.

[31] International Organization for Standardization, 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 (2014).

[32] International Organization for Standardization, ISO 31000:2018 Risk management — Guidelines, https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65694.html (2018).

[33] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.

[34] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, http://arxiv.org/abs/1908.09635 (Sep. 2019).

[35] I. Žliobaitė, Measuring discrimination in algorithmic decision making, Data Mining and Knowledge Discovery 31 (4) (2017) 1060–1089. doi:10.1007/s10618-017-0506-1.

[36] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, On the (im)possibility of fairness, http://arxiv.org/abs/1609.07236 (Sep. 2016).

[37] J. Kleinberg, Inherent Trade-Offs in Algorithmic Fairness, in: Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 40. doi:10.1145/3219617.3219634.

[38] E. Beretta, A. Santangelo, B. Lepri, A. Vetrò, J. C. De Martin, The Invisible Power of Fairness. How Machine Learning Shapes Democracy, in: M.-J. Meurs, F. Rudzicz (Eds.), Advances in Artificial Intelligence, Springer International Publishing, Cham, 2019, pp. 238–250.

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

[40] B. Hutchinson, M. Mitchell, 50 Years of Test (Un)fairness: Lessons for Machine Learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 49–58. doi:10.1145/3287560.3287600.

[41] T. Matsumoto, A. Ema, RCModel, a Risk Chain Model for Risk Reduction in AI Services, http://arxiv.org/abs/2007.03215 (Jul. 2020).

[42] E. Beretta, A. Vetrò, B. Lepri, J. C. De Martin, Ethical and Socially-Aware Data Labels, in: J. A. Lossio-Ventura, D. Muñante, H. Alatrista-Salas (Eds.), Information Management and Big Data, Springer International Publishing, Cham, 2019, pp. 320–327.

[43] E. Beretta, A. Vetrò, B. Lepri, J. C. De Martin, Detecting discriminatory risk through data annotation based on bayesian inferences (2020).

[44] M. S. A. Lee, J. Singh, The landscape and gaps in open source fairness toolkits (2020).

[45] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (4/5) (2019) 4–1. doi:10.1147/JRD.2019.2942287.

[46] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics 26 (1) (2020) 56–65. doi:10.1109/TVCG.2019.2934619.

[47] S. Udeshi, P. Arora, S. Chattopadhyay, Automated directed fairness testing, in: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, 2018, pp. 98–108. doi:10.1145/3238147.3238165.

[48] S. Galhotra, Y. Brun, A. Meliou, Fairness testing: testing software for discrimination, in: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, 2017, pp. 498–510. doi:10.1145/3106237.3106277.

[49] Amazon Web Services, Amazon SageMaker Clarify – Bias Detection and Explainability – Amazon Web Services, https://aws.amazon.com/sagemaker/clarify/.

[50] S. Capecchi, M. Iannario, Gini heterogeneity index for detecting uncertainty in ordinal data surveys, Metron 74 (2) (2016) 223–232.

[51] J. Angwin, J. Larson, S. Mattu, L. Kirchner, COMPAS Recidivism Racial Bias, propublica/compas-analysis, https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv (2016).

[52] Recidivism in juvenile justice, http://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html (2016).

[53] U. M. Learning, Default of Credit Card Clients Dataset, https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset (2016).

[54] UCI Machine Learning Repository: Statlog (German Credit Data) Data Set, https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) (1994).

[55] Comparative Testing and Evaluation of Statistical and Logical Learning Algorithms for Large-Scale Applications in Classification, Prediction and Control | STATLOG Project | FP2 | CORDIS | European Commission, https://cordis.europa.eu/project/id/5170.

[56] U. Groemping, South german credit data: Correcting a widely used data set, http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf (2019).

[57] UCI Machine Learning Repository: Adult Data Set, https://archive.ics.uci.edu/ml/datasets/adult (1996).

[58] UCI Machine Learning Repository: Student Performance Data Set, https://archive.ics.uci.edu/ml/datasets/Student+Performance (2014).

[59] P. Cortez, A. Silva, Using Data Mining to Predict Secondary School Student Performance, in: Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), Porto, Portugal, 2008, pp. 5–12.

[60] Council of Europe, Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems,

https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016809e1154 (April 2020).

[61] United Nations Human Rights - Office of the High Commissioner, Guiding Principles on Business and Human Rights, https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf (June 2011).

[62] Petra De Sutter, Motion for a Resolution - on automated decision-making processes: ensuring consumer protection and free movement of goods and services, https://www.europarl.europa.eu/doceo/document/B-9-2020-0094_EN.pdf (February 2020).

[63] E. U. A. for Fundamental Rights, EU Charter of Fundamental Rights - Article 38 - consumer protection, https://fra.europa.eu/en/eu-charter/article/38-consumer-protection (December 2007).

[64] European Commission, Artificial Intelligence for Europe, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625 (April 2018).

[65] European Commission, Coordinated Plan on Artificial Intelligence, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56018 (December 2018).

[66] European Commission, A European strategy for data, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN (February 2020).

[67] European Commission, White paper on artificial intelligence - a european approach to excellence and trust, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (February 2020).

[68] European Commission - High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (April 2019).

[69] European Commission - High-Level Expert Group on Artificial Intelligence, Policy and Investment Recommendations for Trustworthy AI, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60343 (June 2019).

[70] European Commission - High-Level Expert Group on Artificial Intelligence, The Assessment List for Trustworthy Artificial Intelligence (ALTAI) *for self-assessment*, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342 (July 2020).

[71] European Parliament - Committee on Civil Liberties, Justice and Home Affairs, Draft Report on Artificial Intelligence in criminal law and its use by the police and judicial authorities in criminal matters, https://www.europarl.europa.eu/doceo/document/LIBE-PR-652625_EN.pdf (June 2020).

[72] European Parliament - Policy Department for Citizens' Rights and Constitutional Affairs, Artificial Intelligence and Civil Liability, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf (July 2020).

[73] European Commission, Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act), https://ec.europa.eu/info/sites/info/files/proposal-regulation-single-market-digital-services-digital-services-act_en.pdf (December 2020).

[74] European Commission, Impact Assessment - *Accompanying the document* PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=72160 (December 2020).

[75] Carsten Orwat, Risks of Discrimination through the Use of Algorithms, https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/EN/publikationen/Studie_en_Diskriminierungsrisiken_durch_Verwendung_von_Algorithmen.pdf?__blob=publicationFile&v=2 (July 2020).

[76] Data Ethics Commission of the Federal Government, Opinion of the Data Ethics Commission, https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3 (December 2019).

[77] 116th United States Congress, Algorithmic Accountability Act of 2019 H.R. 2231, https://www.congress.gov/116/bills/hr2231/BILLS-116hr2231ih.pdf (April 2019).

[78] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, N. Seliya, A survey on addressing high-class imbalance in big data, J. Big Data 5 (1) (2018) 42–30. doi:10.1186/s40537-018-0151-6.

[79] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven artificial intelligence systems—an introductory survey, WIREs Data Mining and Knowledge Discovery 10 (3) (2020) e1356. doi:10.1002/widm.1356.

[80] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, T. Janowski, Data governance: Organizing data for trustworthy Artificial Intelligence, Government Information Quarterly 37 (3) (2020) 101493. doi:10.1016/j.giq.2020.101493.