Politecnico di Torino

ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Computer and Control Engineering (35^{th}cycle)

A Study on Data Imbalance: Using Metrics on Input Data to Foresee Bias and Fairness in Classification Outcomes

Mariachiara Mecati

******

Supervisors:

Prof. Marco Torchiano, Supervisor

Prof. Antonio Vetrò, Co-Supervisor

Doctoral Examination Committee:

Prof. Michael Littman, Referee, Brown University

Prof. Andrea Marrella, Referee, Sapienza Università di Roma

Prof. Letizia Tanca, Politecnico di Milano

Prof. Tania Cerquitelli, Politecnico di Torino

Prof. Fulvio Corno, Politecnico di Torino

Politecnico di Torino

2023

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Mariachiara Mecati
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

To my parents, who have always supported me, from my lovely tiny hometown in Romagna, throughout my journey at the Politecnico di Torino; they have always allowed me to pursue my dreams and have always believed in me.
To all those individuals who have been unjustifiably treated based on personal traits because of untransparent and unaware employment of automated decision-making and recommender systems, as well as biased software outputs, I wish that my studies can advance and foster research in order to prevent further discriminatory consequences.

Acknowledgements

This dissertation represents the ending step of my journey at the Politecnico di Torino, as well as the basis of my future path of personal and working growth. During these three years and a half of Ph.D. program, I could always rely on people that made my journey so special, first and foremost my supervisors Prof. Marco Torchiano and Prof. Antonio Vetrò, to whom I am extremely grateful for introducing me to the valuable and fascinating world of Data Ethics, for their help and continuous support, above all throughout the pandemic period, for always encouraging me to overcome obstacles, for their precious advice, for their effort in trying to make me an increasingly mature researcher, for sharing their working knowledge as well as personal values.
I would thank again my supervisor Prof. Marco Torchiano for giving me the opportunity to take part in-person in two international conferences, the first one in Seoul (South Korea, June 2022) and the second one in Osaka (Japan, December 2022), two wonderful experiences that I will always remember, but also for recruiting me as a teaching assistant, a gratifying experience that taught me so much in terms of communication and relational skills, both with students and course colleagues, Prof. Luca Ardito and Prof. Riccardo Coppola, to whom I am very grateful for the ongoing suggestions and exchanges since the first day of my Ph.D. journey.
A very special thank for all his support goes to my colleague and friend Simone Leonardi, with whom I shared each event, thoughts and moment of this journey, besides the difficulties of the pandemic period.
Last but not least, I would like to mention all my colleagues and friends from Lab1, in random order: Diego Monti, Isabeau Oliveri, Francesco Manigrasso, Tommaso Fulcini, Giacomo Garaccione, Edoardo Giusto, Pietro Chiavassa, Edoardo Battegazzorre, Francesco Strada.

Abstract

Data has become a fundamental element of our society in conjunction with the increasing adoption of automation software in a variety of organizational and production processes, and especially of automated decision-making (ADM) systems, which may affect multiple aspects of our lives. Indeed, when software makes decisions that allocate resources or opportunities, might disparately impact people based on personal traits (for example, gender, ethnic group, etc.) and thus might systematically (dis)advantage certain social groups; for these reasons, bias in software systems is a serious threat to human rights. One of the potential causes of unfairness lies in the quality of the data used to train ADM systems. In particular, bias in input data is a relevant socio-technical issue that emerged in recent years, and it still lacks a commonly accepted solution: the "bias in-bias out" problem is one of the most significant risks of discrimination, which encompasses technical fields, as well as ethical and social perspectives. Among the causes of bias, one of the most relevant issues is represented by data imbalance, that is, an unequal distribution of data between the classes of an attribute.

We enrich the current body of research on this topic by proposing a risk assessment approach based on the measurement of data imbalance, which is derived from the principles outlined in ISO standards for software quality and risk management. We look at data imbalance in a given dataset as a potential risk factor for detecting discrimination caused by ADM systems: specifically, we aim to evaluate whether it is possible to identify the risk of bias in a classification output by measuring the level of (im)balance of protected attributes in training data. After that, we investigate the issue of data imbalance more and more thoroughly: we define a methodology to identify imbalance thresholds in input data to achieve desired levels of algorithmic fairness; then, we study imbalance on intersectional protected attributes and on the combination of the target variable with protected attributes.

To conduct our studies, we selected a set of indexes of balance (Gini, Simpson, Shannon, Imbalance ratio) and we first assess their capability to detect (im)balance in synthetic attributes. Then, we tested their ability to identify unfair classification outcomes in large datasets belonging to different application domains, that is, their capacity to foresee a certain level of discrimination risk –which depends on the context, the dataset's domain, and the choice of the measures. Specifically, we applied the indexes of balance to protected attributes in the training sets, while we computed the unfairness by applying different fairness criteria to the same protected attributes in the test sets. In subsequent studies we tested our approach on a large number of data mutations with different classification tasks and on a variety of combinations of balance-unfairness-algorithm in order to identify specific imbalance thresholds. Lastly, we investigated whether measures of balance on intersectional attributes are helpful to detect unfairness in classification outcomes, and whether the computation of balance on the combination of a target variable with protected attributes improves the detection of unfairness.

The results show that our approach is suitable for the proposed goal, thus the balance measures can properly detect unfairness of software output. Indeed, a negative correlation holds between balance and unfairness measures, as low levels of balance in protected attributes are related to high levels of unfairness in the output; in addition, we found that measures of balance on intersectional protected attributes are helpful to detect unfairness in classification outcomes. However, the choice of the index has a relevant impact on the detection of discriminatory outcomes, and thus on the threshold to consider as risky. Overall, to increase the generalizability of our findings, it would be recommended to extend our studies on a wider number of datasets as well as indexes of balance, for instance by considering measures for non-categorical attributes. Given the different behaviors of the balance measures in detecting possible unfairness risks, we elaborated specific pragmatic recommendations for their application.

We believe that our approach for assessing the risk of discrimination should encourage to take more conscious and appropriate actions, as well as to prevent adverse effects caused by the "bias in-bias out" problem. Especially, we hope that our findings on data imbalance will improve the identification and assessment of discrimination risks in ADM systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The implementation and development of automated decision-making (ADM) systems have a significant impact on several aspects of our daily life [1]. Indeed, as a consequence of the general phenomenon of the digitization of organizational processes in our societies [2][3], the automation of decision processes is speedily expanding [4][5]. Initially, this trend was made possible by the computerization of our physical environment and the widespread use of internet connectivity, and more recently by the large availability of data and the development of technical tools for their analysis. The creation of predictive, classification, and ranking models, which form the basis of automated decision-making (ADM) systems, laid the groundwork for the rapid adoption of data-driven decision-making processes [6]. In our studies, we rely on the definition of Automated Decision-Making given by Algorithm Watch [4]:

> *Systems of automated decision-making (ADM systems) are always a combination of different 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).*

ADM systems are employed to categorize people and predict their behaviors based on patterns extracted from data gathered on them or other individuals. Thus, decisions can be based upon software-generated recommendations or even entirely automated: the most commonly adopted technical approaches range from simple tools –such as macros or scripts that compute and sort data according to predefined sets of rules or parameters [7]–, to more sophisticated neural networks [5].

ADM systems are used in a wide variety of tasks, including predicting debt repayment capacity [8], selecting the top job candidates [9], identifying social welfare frauds [10], and advising on which university to attend [11], to name but a few. On one hand, the benefits of using these systems range from the scalability of the operations and ensuing economic efficiency, to the removal of public servants' discretion [12] [13] [14]. On the other hand, the advantages only become apparent if the underlying data is of high quality, otherwise errors may result in significant additional costs [15] and also give rise to serious ethical issues. The problem of bias in information systems, although present in the scientific literature of software systems during the past quarter century –for example, see the pioneering work proposed in [16]– got wider attention only in the mid 2010s, in connection with the large investments in Artificial Intelligence (AI) / Machine Learning (ML), digital automation of organizational processes and, in general, automated decision-making (ADM) systems. Indeed, according to a substantial body of evidence found in both scientific literature [17] and journalistic essays [18] –especially the influential book by O'Neil (2017) [19]– ADM systems might replicate the same bias of our societies, systematically discriminating against the weakest people and escalating existing inequalities. The topic is so important that it has drawn in experts from both information technology and social sciences, and it has been acknowledged by the institutions [20] as well. Indeed, as affirmed by Margrethe Vestager (2020), who has been 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, automating decisions using historical data is a two-edged sword [21]:

> "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 view, biased data means *imbalanced* input dataset [22]: indeed, often the cause for the software-biased impact lies in the imbalance of training data [23]. Specifically, data imbalance is an unequal distribution of data between the classes of a given attribute (such as gender, education, country of origin, age category, etc.) [24], a condition that happens when there are large disparities in the number of data points between the classes. As acknowledged in the excerpt from Vestager's speech, imbalances can be caused by mistakes or constraints in the design and operations of the data collection process, or merely by inequities in the current reality that the data itself reproduce, where there might be imbalanced with respect to certain characteristics. Indeed, data imbalance skews the performances of classifiers, leading to varying accuracy among the classes of given attributes in the data. This consequence has been documented in a variety of domains and technologies, for example, male dominance in training data can perpetuate such bias in the output of automatic generation of images [25], while geographic imbalance in the content production that feeds recommender systems can generate (dis)advantage toward a specific group [26] or, again, gender imbalance in data about nurses is a rather common occurrence.
In particular, imbalance is defined as between-class when only two classes are considered and one class is over-represented relative to the other, or multiclass when imbalances occur between multiple classes. Herein we deal with the more general case, represented by the multiclass imbalance.

For a long time, imbalanced data has been recognized as a problematic matter in the field of machine learning [24] [27], and it is still relevant [22] [28], particularly because it can impair the performance of supervised learning algorithms in terms of very heterogeneous accuracy across the classes of data. When people are the subject of an automated decision, the algorithm's inconsistent outcomes actually represent a systematic form of discrimination that can be generally described as the following [16]:

> "unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category".

Thus, a biased software can be described as one that "systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others [by denying] an opportunity for a good or [assigning] an undesirable outcome to an individual or groups of individuals on grounds that are unreasonable or inappropriate" [16], or alternatively, as one that exposes a group (for example, a member of an ethnic minority or a certain class of worker) to unfair treatment [29], that is, as an algorithm that –often in order to achieve its optimization purposes– may discriminate and filter between people into account, with the result being a disparate impact on different population groups. Biased software is a significant socio-technical issue because it frequently arises when people are the target of predictions or classifications, casing a disparate impact on particular social groups as a result. A social group is recognized as a collection of people who have similar physical, cultural, or identitarian characteristics. When these traits are noted in datasets, those groups correspond to individuals who have the same value for a particular protected attribute. Specifically, we define as *protected attributes* –also said *sensitive attributes*– and thus, as social groups of category object of possible discrimination, those attributes identified by the characteristics provided in "Article 21 - Non- discrimination" of the EU Charter of Fundamental Rights [30]:

> *1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.*
>
> *2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.*

Following the line of reasoning exposed above, it is possible to detect the risk of bias in the classification output by measuring the level of (im)balance of the protected attributes in a dataset. Specifically, we propose a risk assessment approach based on quantitative measures to evaluate imbalance in the input datasets of automated decision-making systems, with a view to foreseeing a potential risk of discriminatory outcomes by revealing the presence of imbalances in training data.

Advancement with respect to the state of the art. Our proposal differentiates from the reference literature (exposed in Chapter 3) for two main reasons: in the first place, it is built upon a series of international standards that incorporate by design a multi-stakeholder perspective –we refer to Chapter 2 for an exhaustive explanation of the theoretical foundations of our approach. In the second place, 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 a very low probability of success. Indeed, given the socio-technical nature of the problem, we firmly believe that a risk approach is preferable because it keeps the ultimate responsibility in the realm of human agency.

1.1 Evidence of Discrimination by ADM Systems

Recently, both scientific literature and investigative reporting have gathered a significant body of evidence of discrimination by ADM systems. Hereafter, we highlight a few instances of automated discrimination brought on by ADM systems; we do not provide a comprehensive review of the literature on automated discrimination, but rather aim to emphasize how these systems affect people's lives.

The creation of a software system by Amazon [31] with a view to evaluating resumes of potential employees obtained from the internet is a well-known illustration of the aforementioned problem. The project, which had been started in 2014 with the intention of predicting successful future employees using word patterns extracted from CVs from the previous ten years, was discontinued in 2017 because, according to the news agency report, female profiles were systematically devalued. Given that men make up the majority of employees in the technology sector, the problem stemmed from the fact that training data was primarily composed of males.

Similar to the previous example, unfair treatment due to gender imbalance in the input data was shown by a scientific experiment on the search engine Common Crawl [32]. Through the comparison of three machine learning techniques for occupational classification with nearly 400.000 biographies, it was revealed that in each of the three cases, the rate of correct classifications reflected the existing gender imbalances of the occupational groups, even without explicitly adopting gender indicators.

Another case is provided by the “Black box Schufa” [33]. The most well-known credit bureau in Germany, Schufa, claims to have data on over 67 million consumers and to generate scores for each of them. These scores are used by retailers, telecom companies as well as banks to support them run their businesses, for instance, to choose which customers might be approved for a loan or which users are allowed to see a particular advertisement. Researchers reverse-engineered how Schufa functions and discovered that younger people are frequently evaluated worse than older people thanks to a crowdsourcing project that involved 2800 volunteers who requested their free personal credit reports. Similarly, males are ranked lower than females. Age and gender are legally but unfairly included in the score because the General Equal Treatment Act, which was created to shield consumers from discrimination based on gender and age, is ineffective with regard to credit bureaus.

According to a different study [34], Facebook job advertisements were significantly biased toward specific gender and ethnic group, which resulted in unequal job opportunities and ongoing discrimination throughout the duration of an advertisement. Contrary to what is stated in the Art. 21 of the EU Charter of human rights [30], people are denied opportunities based on personal characteristics as a result of such a conservative mechanism. The Department of Housing and Urban Development of the United States sued Facebook in March 2019 for violating the Fair Housing Act due to the discriminatory impact of its advertisements, which were disproportionately targeted with respect to different personal traits, such as gender and race [35].

Another reference example can be found in the research area of image classification, specifically facial recognition systems, which have drawn a lot of criticism for both the issue of discrimination and the technology itself. The case involves commercial gender classification: in [36], the authors showed how performance disparity in gender and race have an impact on automated facial image analysis: specifically, the gender classification on female faces performs noticeably worse than classification on male faces, and higher performances are reached on lighter skin tones than darker ones.

If these issues happen in the justice or medical fields, where the combined use of ADM systems and historical data is rapidly expanding, negative consequences could become even worse and life-altering for certain individuals. The most well-known case in the criminal justice system is represented by the investigation conducted by the no-profit organization Pro-Publica on the COMPAS algorithm (Correctional Offender Management Profiling for Alternative Sanctions) [37], adopted by judges to assess the probability of recidivism of defendants. Pro Publica revealed that the COMPAS algorithm was distorted against black defendants: indeed, black defendants who were not rearrested had nearly twice the likelihood of being incorrectly classified as higher risk (false positive) than white defendants. Contrarily, white defendants who actually got rearrested were nearly twice as likely to be misclassified as low-risk than black defendants. The root cause of this distorted effect was that there were significantly fewer records in the dataset pertaining to white defendants than there were to black defendants.

In a recent study of risk assessment for juvenile justice conducted in Catalonia [38], male defendants and members of a particular national group were more frequently marked as recidivists than other groups. Researchers developed a method to evaluate fairness and predictive performance of machine learning algorithms used to predict juvenile recidivism based on this problem of discrimination. Then, the results were compared to SAVRY (Structured Assessment of Violence Risk in Youth), a widely used risk assessment tool used to estimate the risk of violence in juvenile justice. Demographic parity and error rate balance were proposed by researchers as two metrics for measuring fairness. In practise, they found that machine learning algorithms become discriminatory when using SAVRY demographic features: indeed, foreigners were more likely to be classified as high risk and male defendants were more likely to be labelled as recidivists even though they were non-recidivists.

In the medical industry, a study conducted recently [39] reported the case of a widely-used commercial system for selecting which patients should be admitted to an intensive care program: practically, an algorithm was trained on historical information about medical spending and health-service utilization in order to generate risk scores subsequently used by medical doctors. According to research, white patients had a markedly higher likelihood than black patients of being assigned to the intensive care program in situations where their health status was equivalent. In this instance, as well, the system was impacted by ethnicity-based discrimination because the risk score reflected the expected treatment costs – which was heavily correlated with the patients' economic well-being – more than actual health conditions.

Finally, it is important to bring up the "Report of the Special rapporteur on extreme poverty and human rights" [40], which was published by the United Nations and expressly criticizes the way governments are automating welfare management. According to the evidence gathered, these systems routinely discriminate against the most vulnerable individuals of society and worsen already existing inequalities.

The examples mentioned above succinctly illustrate how an imbalance in data can spread and manifest itself in the output of ADM systems, becoming a socio-technical issue of paramount importance to public sector services, where high stakes decisions are always more frequently delegated to automated decision-making, in whole or in part, with grave consequences for people.

1.2 Logical Framework and Research Questions

We conducted several studies in order to delve into all the main aspects regarding data imbalance, specifically with a view to deeply understanding how to use metrics of balance on input data to forecast bias and fairness in classification outcomes. More formally, we answer six main research questions derived from the necessity to comprehend the abstract construct of *imbalance* and how we can identify it in data, followed by in-depth studies on the proposed risk assessment approach based on quantitative measures to assess imbalance in the input datasets of ADM systems in order to foresee a potential risk of discriminatory automated decisions.

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

To explain the abstract construct of *imbalance*, a number of measures have been put forth in the literature; in Chapter 4 we select a set of indexes of balance with the goal to understand how these indexes represent our subjective –and probably limited– perception of imbalance. The related study has been published in the journal article entitled "A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision-Making Systems" (2021) [41].

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

A substantial body of research from both scientific literature and journalistic essays demonstrates that biased data can cause discriminatory behavior when used to train an ADM system (see Section 1.1). The aim of the analysis presented in Chapter 5 is to determine whether the data imbalance, measured through the indexes analyzed in the previous study, may indicate a risk of discrimination in the ADM system. This research question has been addressed first in an exploratory study published in the conference paper "Identifying

Risks in Datasets for Automated Decision–Making" (2020) [42], and then in the more extensive journal article "A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision-Making Systems" (2021) [41].

RQ 3. *Is it possible to measure the risk of bias in a classification output by measuring the level of (im)balance in the protected attributes of the training set?*

The goal of the study reported in Chapter 6 is to understand how the balance characteristics of protected attributes in training data can be used to evaluate –and thus, to forecast– the risk of algorithmic unfairness in subsequent classification tasks. In this analysis we examine how well the balance measures (previously analyzed) applied to a specific protected attribute reflect a discrimination risk, separately in the case of a binary protected attribute and in the case of a multiclass one. Particularly, the characteristic element of this study is represented by the adoption of specific pre-processing methods as mutation techniques in order to vary the level of (im)balance of the protected attributes in analysis. Moreover, differently from the previous research question, we conducted two separate studies which are reported in the following publications:

- "Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data" (2021) [43], a conference paper in which we analyzed the behaviour of the balance measures when applied to *multiclass* protected attributes;
- "Detecting Risk of Biased Output with Balance Measures" (2022) [44], a journal article in which we focus on *binary* protected attributes.

RQ 4. *Is it possible to identify a threshold s (for balance measures) such that if the* balance *of the training set is greater than s, then the* unfairness *of the classification on the test set is expected to be less than a threshold f?*

In Chapter 7 we focus on the construction of risk thresholds for balance measures, and the respective thresholds for fairness criteria, in order to better understand how the balance of protected attributes in training data can be used to assess the risk of algorithmic unfairness; we propose a specific procedure for identifying risk thresholds and we test it on several datasets. The findings of this study are published in the conference paper titled "Identifying Imbalance

Thresholds in Input Data to Achieve Desired Levels of Algorithmic Fairness" (2022) [45].

RQ 5. *Is it possible to identify the risk of biased output by detecting the level of (im)balance in intersectional protected attributes?*

One more key point to investigate concerns intersectional classes due to the fundamental role they play in understanding risks of discrimination and inequalities that result from the intersection of particular social identities and are even amplified in correspondence of such intersections. To this end, we formulated two research questions that will drive the investigation reported in Chapter 8:

RQ 5.1. *How do intersectional attributes relate to the corresponding primary attributes, in terms of balance and fairness?*

First of all, it is of paramount importance to extend our knowledge on how the imbalance of the primary attributes (binary or multiclass) affects the imbalance of the intersectional attribute, in addition to understand how the fairness with respect to an intersectional attribute relates to the fairness with respect to the primary attributes.

RQ 5.2. *Can the measure of balance on intersectional attributes detect unfairness risks?*

From the studies conducted to answer the previous research questions, there is proof that when working at the level of protected primary attributes, the balance of the classes of a given attribute can detect the risk of unfair classifications with regard to such attribute; here we want to ascertain whether this capability also applies to intersectional attributes.

These research questions have been addressed in the journal article titled "Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications" (2023) [46].

RQ 6. *Does the combination of the target variable with protected attributes improve the detection of unfair classification risks?*

Finally, in Chapter 8 we investigate the contribution of the target variable to the unfairness detection: the level of (im)balance of a target variable can be examined by taking into account its combination with protected attributes (both primary and intersectional), and evaluating whether the risk of unfair classification can be identified through the measurement of the balance of the combined attribute. Keep in mind that attributes given by the combination of the target variable with protected attributes (primary or intersectional) are referred to as *combined attributes* hereinafter. This study has been published in the journal article entitled "Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications" (2023) [46].

All the aforementioned study have been conducted through RStudio – an integrated development environment for R, a programming language for statistical computing and graphics. The dissertation is organized as follows: first of all, in this Chapter we outlined motivations and research context of the Ph.D. program; in Chapter 2 we introduce a Data Imbalance-based Risk Assessment Approach and discuss the theoretical foundations of our proposal, while in Chapter 3 we position our work in relation to the body of literature by demonstrating how it is connected to several existing research strands. Then, starting from Chapter 4 through Chapter 8 we addressed all the research questions listed above. Particularly, in Chapter 4 we investigate how to identify imbalance in datasets using metrics of balance, and subsequently in Chapter 5 we analyze data imbalance as a risk indicator. In Chapter 6 we move our investigation forward by detecting the risk of bias in classification outcomes by specifically analyzing the use of the balance measures on binary and multiclass protected attributes, where the (im)balance of the attributes has been varied through the application of different mutation techniques. In Chapter 7 we set up a procedure to identify imbalance thresholds to achieve desired levels of algorithmic fairness and we apply the methodology to different datasets. In Chapter 8 we address a different aspect of the data imbalance issue by measuring imbalance on intersectional protected attributes and on target variables to foresee unfair classifications. Finally, in Chapter 9 we provide practical implications for the usage of the indexes of balance, as well as discuss open issues and possible future works, while in Chapter 10 we formulate the conclusions of our research activity.

Chapter 2

Background and Experimental Design Fundamentals

Building on the line of thought exposed in the Introduction, we present the fundamental concepts that guide our perspective and proposals: identifying data imbalance as a risk factor for systematic discrimination caused by ADM systems. The whole approach stems from the principles outlined in ISO standards for software quality and risk management, as analytically and thoroughly described in [47] that originated the series of studies presented in this dissertation.

2.1 A Risk Assessment Approach Based on the Measurement of Data Imbalance

The first foundational concept is the ISO/IEC 25000:2014 series of standards, referred to as "***Systems and Software engineering – Software product Quality Requirements and Evaluation (SQuaRE)***" [48]. SQuaRE outlines the quality modeling and assessment of data, software services and software products. Note that a *software product* is defined in ISO/IEC 12207:1998 as a "set of computer programs, procedures, and possibly associated documentation and data", and in SQuaRE standards, *software quality* stands for *software product quality*. It defines quality with a set of measurable characteristics and sub-characteristics, and models data quality specifically in ISO/IEC 25012:2008 with 15 quantifiable characteristics, such as recoverability, completeness and efficiency. These characteristics can be seen from two perspectives: "inherent", which depend solely on the data themselves, or "system dependent", influenced by the hardware or software used to store, analyze, and retrieve the data. Some characteristics can even fall under both viewpoints.

One example of a inherent characteristic is Completeness, which refers to the extent to which all necessary data has been entered and stored in the computer system. For instance, in a database of university students, all essential information about each student should be present to meet the needs of users. To evaluate Completeness, one metric used is "Record completeness" (Com-I-1), calculated as the ratio of data items with non-null values in a record to the total number of data items in that record. On the other hand, Availability is a system-dependent characteristic that refers to the ability of data to be always accessible. A measure of this characteristic is "Probability of data available" (Ava-D-2), which is the ratio of times that data is available to the number of times it is requested in a given time period. Finally, Efficiency is a characteristic that falls into both inherent and system-dependent categories: it is defined as the ability of data to be processed (accessed, acquired, updated, etc.) with appropriate levels of performance using the appropriate number and type of resources under certain conditions. Efficiency has different measures for inherent and system-dependent perspectives.

The ISO/IEC 25012:2008 standard does not mention data imbalance (nor its dual concept of data balance) as part of its definition of data quality. However, the SquaRE standard introduces a concept of paramount importance in our scenario, that is, the chain of effects and dependencies. This principle states that improving the quality of a product, service, or data will have a positive impact on the system's overall quality and ultimately benefit the software system's users. Note that this relationship holds also in the opposite way and among aspect pairs, for instance: the quality in use is impacted by the product quality, which then affects the data quality. The upper section of Figure 2.1 summarizes how this chain of effects is formalized in SQuaRE. In the field of data quality, a simplified version of this idea is the well-known GIGO principle, which stands for "garbage in, garbage out" and states that the outcome of a software will be unreliable if the input data are outdated, inaccurate, incomplete, or flawed.

Our argument is that the chain of effects is still applicable in the presence of data imbalance, meaning that imbalanced data might lead to imbalanced software outputs,

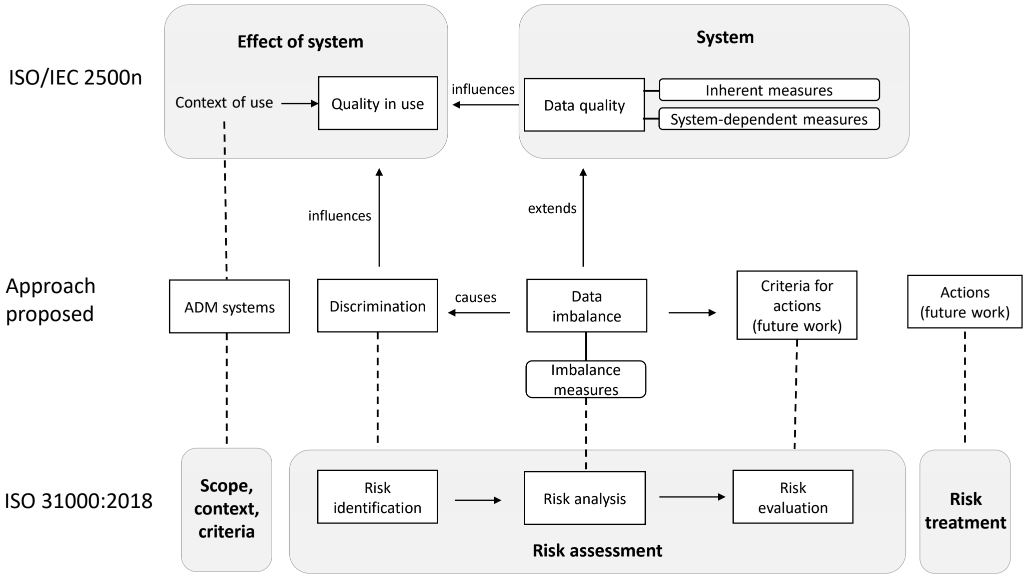


Fig. 2.1 The proposed approach in relation to ISO standards adopted as reference frameworks.

which in the context of ADM systems could result in differentiation of information, products and services based on personal characteristics. As noted in Chapter 1, this differentiation in areas such as employment, education, wages and social benefits can lead to unjustified unequal treatment and potentially unlawful discrimination. Hence, it is essential to consider data imbalance as a potential risk factor in all ADM systems that use historical data and automate decisions impacting individuals' rights and freedoms: a particularly relevant case is exactly represented by ADM systems utilized in public sector services.
In this context, data imbalance is viewed as a part of the data quality model outlined in ISO/IEC 25012:2008. It is considered an inherent characteristic that will be quantified using appropriate metrics, which are an extension of those defined in ISO/IEC 25024:2015.

The second fundamental concept of our methodology is based on the ISO 31000:2018 standard for ***Risk management*** [49]. This standard provides the fundamental principles for risk management and establishes a framework for incorporating it into organizational structures, as well as a process for managing risks at different levels, such as the strategic, operational, program, or project levels. In regards to our proposal, special attention is given to the risk management process, taking into account both data imbalance and potential discrimination from ADM systems.

Specifically, we propose to consider data imbalance as a risk factor in all those systems that –on the basis of historical data– automate decisions on important aspects of people's lives, impacting their rights and freedoms. A summary of the key elements of the ISO 31000:2018 risk management approach can be seen in the bottom section of Figure 2.1. In particular, we concentrate on the *risk assessment* stage, which encompasses the identification, analysis, and evaluation of risks, as described below in connection to our approach.

- *Risk identification.* It involves locating, recognizing, and describing potential risks within a specific context and scope, using predetermined criteria. In our scenario, making reference to the Article 21 "Non discrimination" of the Charter of Fundamental Rights of the European Union [30], this stage can be traced back to the statements of "unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category" [16], since ADM systems are adopted in areas affecting individuals' rights and freedoms.
- *Risk analysis.* The objective of risk analysis is to identify the features and potential extent of risk. This stage assumes metrics of data imbalance as indicators of discrimination risk because of the effects of the bias propagation discussed previously. Different balance measures will be presented in Section 2.2.2 and they will be then applied to real datasets –described in Section 2.2.1– throughout our studies, with the aim of detecting discrimination risk in case an ADM system is trained using such data.
- *Risk evaluation.* This final stage involves determining the level of risk based on the outcome of the analysis: the findings are used to determine if additional analysis or risk management strategies are necessary, and who should implement them. In our case, the metrics of data imbalance should be examined within the framework of the algorithms that handle such data; additionally, the effects on end-users and the applicable legal regulations for the relevant field should be taken into account. However, these aspects are beyond the scope of this work and will be addressed later in Section 5.3.

In summary, making reference to the conceptual frameworks previously described, we propose a metric-based approach to evaluate imbalance in a given dataset in order to foresee risks of biased output from ADM systems. Figure 2.1 displays a comprehensive overview of the approach and its alignment with the internationally recognized ISO/IEC standards. The top portion showcases the key components of the SQuaRe series (2500n) that are relevant to our study. The bottom section displays the central components of the risk management process of ISO 31000. Our proposed approach is depicted in the center of the Figure, with emphasis placed on its links to both the SQuaRe series and ISO 31000 elements.

2.2 Experimental Design Fundamentals

Hereinafter we present the fundamental elements that characterized each of the experimental design that will be described in the following Chapters.

2.2.1 Datasets

In our studies, we looked for some variety in the data selection with a view to testing the versatility of our approach in various application domains of ADM systems. Thus, we analyzed eight datasets that belong to different fields of application: criminal justice (including the juvenile justice system), financial services, as well as personal earnings, drug consumption, medical diagnosis and education in social-related areas. A summary of the key features of the selected datasets can be found in Table 2.1, while the source links for each dataset are reported in Table 2.2.

COMPAS Recidivism racial bias dataset. This dataset includes variables used by the homonymous COMPAS algorithm in scoring criminal defendants in Broward County (Florida), in addition to their outputs within two years of the decision. The original data consist of 28 variables, among which the target variable is identified by *two_year_recid*, while the classifier is represented by *risk score*, which indicates a "recidivism degree" with scores ranging from 1 to 10 and being considered high risk if they are above 4, so as to be a binary classifier.
The COMPAS dataset has garnered attention in scientific communities for its potential biases, with a study by the U.S. non-profit organization ProPublica[1] revealing

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, last visited on June 1, 2023

disparities in risk scores for white and black defendants. The analysis[2] showed that black defendants were more likely to be misclassified as high risk compared to white defendants (false positive); conversely, white defendants who were actually rearrested were nearly twice as likely to be misclassified as low risk than black defendants (false negative). The main reason was that there were significantly more black defendants' records in the dataset than there were white defendants' records, and there were much more black recidivists than white recidivists as well.

Credit card default dataset. This dataset consists of credit card client information from Taiwan spanning from April 2005 to September 2005. It includes data on history of payment, credit data, bill statements, and demographic factors such as education and sex. Particularly, the dataset comprises 25 variables, with *default.payment.next.month* as target variable. This particular dataset was selected due to the significant impact of using ADM systems in this field. Moreover, at the time of our researches it was –and it is still ranked that way– the fourth-most highly rated credit card dataset on Kaggle[3].

Drug consumption. This dataset is provided by the UCI Machine Learning Repository and holds information for 1885 individuals, including their personality traits and demographic details. Participants were asked about their usage of 18 legal and illegal drugs, plus a fictitious drug (Semeron) was included to identify over-claimers, with the possibility to chose one of the following answers: never used, used over a decade ago, or used in the last decade, year, month, week, or day. This dataset has been used for two distinct classification tasks: predicting drug consumption based on personality data, and predicting personality traits based on drug consumption. In our study, these two scenarios are considered as separate datasets.

Heart disease. This data collection is retrieved from the UCI Machine Learning Repository and covers a variety of heart-related conditions, including but not only blood vessel illnesses (for example, coronary artery disease), heart rhythm issues, and congenital heart defects. It is made up of 303 instances and 14 variables, mostly consisting of medical information. Although the initial dataset consisted

[2] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, last visited on June 1, 2023

[3] https://www.kaggle.com/datasets?search=credit+card&sort=votes, last visited on June 1, 2023.

of 76 variables, only the subset that was examined in our study is indicated in all referenced reseach.

Income. These data, provided by the UCI Machine Learning Repository and also referred to as the "Adult" or "Census Income" dataset, were obtained by Barry Becker from the 1994 Census database and the aim was to predict a person's yearly income based on the records. The target variable, represented by *test.income*, can take on two values of either $<= 50K$ or $> 50K$ and a regression model was constructed to predict the respective score.

Juvenile justice. This collection of 4753 records presents the statistical details and recidivism rates of minors and young adults who participated in an educational program in Catalonia in 2010. The information provides an overview of their profiles and their involvement with the juvenile justice system. The recidivism rate, specific rates, profile of recidivists and recidivism are also included. Specifically, the SAVRY variables indicate the risk of recidivism and the areas of risk and need for these individuals, with the *SAVRY_total_score* serving as an indicator of the "total recidivism degree" on a scale from 1 to 100. In order to consider it as a binary score variable, we refer to the COMPAS dataset where the total percentage of moderate and high recidivism risk is around 45%, thus we select the same percentage of data in the Juvenile dataset as moderate-high risk: we considered a score –of the variable *SAVRY_total_score*– above 15 as an affirmative (estimated) risk of recidivism. In addition, the target variable *reincidencia_2013* represents the recidivity by the end of 2013.

Statlog. This dataset, obtained from the UCI Machine Learning Repository, was contributed by German professor Hans Hofmann as part of a collection of datasets from an European project "Statlog"[50]. The data are a stratified sample of 1000 credits, 700 of which are considered good and 300 bad, and were gathered from a large regional bank with about 500 branches, both urban and rural ones, in Southern Germany between 1973 and 1975. The bad credits were oversampled to provide sufficient information for differentiation from good credits [51]. Each record in the dataset represents an individual who applied for a credit by a bank and is characterized by 20 categorical attributes. The goal is to determine the credit risk, classified as good or bad, based on these attributes. As specified with the Statlog data (whose source link is provided in Table 2.2), one may analyze the misclassification cost: indeed, the cost of misclassifying a bad risk as good is assumed to be five times higher than the cost of misclassifying a good risk as bad [51], thus the *cost_matrix* variable, with a value of either 0 or 1, is used as target variable.

Student. These two collection of data were built with school reports and questionnaires gathered in 2014 and provide information on the academic performance of secondary school students in two Portuguese schools. Two datasets were retrieved from the UCI Machine Learning Repository and they include the students' grades, as well as various demographic, social, and school-related characteristics. In particular, the data relate to the performance of different students in Mathematics and Portuguese language. In our studies, we analyze both the datasets and used the variable *G3* as the target variable, which represents the final year grade ranging from 1 to 20, and is considered positive if it is above 9, or negative otherwise[52].

Table 2.1 Complete list of the *datasets* and their main characteristics.

Dataset	**Domain**	**Size**	**Target variable**	**Score**
COMPAS	Justice	6172×13	recidivism risk	`COMPAS_risk_score`
Default of credit cards clients (Dccc)	Financial	30000×25	default payment next month	*missing*
Drug (Cannabis)	Welfare	1885×32	cannabis consumption	*missing*
Drug (Impulsive)	Welfare	1885×32	impulsiveness	*missing*
Heart disease	Welfare	303×14	diagnosis	*missing*
Income	Welfare	32561×15	income bracket	*missing*
Juvenile justice	Justice	4753×132	recidivism risk	`SAVRY_total_score`
Statlog	Financial	1000×21	creditworthiness	*missing*
Student (Mathematics)	Welfare	395×33	final grade	*missing*
Student (Portuguese)	Welfare	649×33	final grade	*missing*

Table 2.2 Complete list of the *datasets* and their *source links*.

Dataset	**Source link**
COMPAS	https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv
Default of credit card clients (Dccc)	https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
Drug (Cannabis and Impulsive)	https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
Heart disease	https://archive.ics.uci.edu/ml/datasets/heart+disease
Income	https://archive.ics.uci.edu/ml/datasets/adult
Juvenile justice	http://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html
Statlog	https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)
Student (Mathematics and Portuguese)	https://archive.ics.uci.edu/ml/datasets/Student+Performance

2.2.2 Balance Measures

In our studies we focused on *categorical* attributes and chose four indexes of data balance that are commonly referenced in the literature of various fields of research. Note that to assess the balance in data, we will refer to *balance measures*, *metrics of balance* as well as *indexes of balance* interchangeably.

The measures were normalized to satisfy two standards:

i) range in the interval $[0,1]$;

ii) share the same interpretation: the closer the metric is to 1 and the more balanced the distribution of categories is, meaning that the frequencies of each category are similar; on the contrary, values closer to 0 show that there is a higher concentration of frequencies in a smaller number of categories, resulting in an imbalanced distribution.

Gini index. This index, which is described in the literature as a measure of heterogeneity, is widely used across many fields of study and referred to with various terms such as political polarization, market competition, ecological diversity, and racial discrimination. The index reflects the degree of diversity, or the extent to which different types (such as protected groups) are represented. In statistical analysis, the heterogeneity of a discrete random variable, which has m categories and frequencies f_i (with $i = 1,\ldots,m$), can range from a minimum value (degenerate case) to a maximum value (equiprobable case), where all categories are equally represented. In other words, for a given m, heterogeneity increases as the probabilities become more equal, meaning that the different protected groups are represented in similar numbers. The Gini index is calculated as follows:

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

Shannon index. Diversity indexes are a valuable method for assessing imbalances in a community by considering the proportion of various species or classes. One of the most commonly used indexes in biology, ecology, and phylogenetics is the Shannon Index. This index measures the diversity of species in a community by determining the relative abundance of each species. First, the proportion of species i, represented as f_i and relative to the total number of species, is considered; then, the natural logarithm of this proportion ($\ln f_i$) is taken and multiplied by f_i. This calculation is performed for each species and the results are summed and multiplied by -1. The final formula for the Shannon Index is the following:

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

Simpson index. It is a metric used to quantify diversity by determining the chance of two randomly selected individuals belonging to the same species or category. It is utilized in fields such as social and economic sciences for evaluating wealth, uniformity, and equity, as well as in ecology for examining the diversity of living organisms in a specific area. The metric is based on a discrete random variable, which has m categories and frequency f_i where $i = 1, \ldots, m$. This frequency represents the proportion of species i out of the total number of species. The calculation of the Simpson index is performed as follows:

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

Imbalance Ratio index. The IR index is a frequently utilized indicator that is computed as the ratio of the highest frequency to the lowest frequency. To standardize it to a range of $[0,1]$ and make it a metrics of balance comparable to the previous measures, the inverse is taken. The formula applied is as follows:

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

2.2.3 Fairness Criteria

We evaluated the *unfairness* of automated classification outcomes by means of three criteria formalized by [53] in Chapter 3 “Classification”. Bear in mind that hereinafter we will refer to *fairness criteria* and *unfairness measures* analogously, as we assume the following Fairness criteria as indicators to evaluate the unfairness of a classification output.

First of all, to assess the unfairness we take into account a categorical protected attribute A that can assume various values $\{a_1, a_2, ...\}$, a binary target variable Y (that is, $Y = 0$ or $Y = 1$) and a predicted class (or score) R; being Y binary, R is binary as well (that is, $R = 0$ or $R = 1$).

Practically, we aim at measuring to what extent an ADM system behaved fairly with respect to the different values of a protected attribute when determining a predicted class.

Independence criterion. According to this criterion, in order to assess whether the acceptance rate is the same across all groups, we adopt the concept of statistical parity or demographic parity, which requires the probability of acceptance (that is, $R = 1$) to be equivalent for all groups. In other words, the independence criterion is satisfied if all groups have equal rate of selection. In terms of probability, it is expressed through the following condition:

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

If A is binary (that is, $A = a_1$ or a_2), then we can compute the Independence criterion as:

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

Separation criterion. In plain words, when the protected characteristic is associated with the target variable –as it happens in various circumstances– the Separation criterion permits a correlation between the score and the sensitive attribute as long as it is justified by the target variable. Indeed, this criterion is also referred to as equalized odds, equality of opportunity, or even conditional procedure accuracy. This means that for each level of the protected attribute being evaluated, the true positive rate and false positive rate must match in order to satisfy the separation criterion:

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

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

Hence, if A is binary we can calculate two Separation unfairness measures ($\mathfrak{U}$) in the following ways:

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

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

Sufficiency criterion. This criterion assumes the calibration of the model for the various categories of a given protected attribute, that is, Parity of Positive/Negative Predictive Values (respectively R=1 or 0) for each level of the protected attribute taken into consideration:

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

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

As previously indicated, if A is binary we can compute two Sufficiency unfairness measures ($\mathfrak{U}$) as follows:

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

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

When dealing with non-binary attributes, that is $m > 2$ (where m indicates the number of categories of a given attribute), all the definitions above can be extended by considering the mean of indexes can be computed by taking all the possible pairs of levels in A:

All of the definitions above can be expanded when dealing with multiclass attributes, that is $m > 2$, by considering the mean of the measures calculated by taking into account all the possible pairs of levels in A.

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

Finally, we remind that all the fairness criteria previously introduced range in the interval $[0,1]$: the higher the values, the greater the unfairness; therefore, values equal to zero denote an entirely fair classification, whereas values close to 1 indicate unfair behavior.

Chapter 3

Related Work

In recent years, substantial effort has been made to enhance and introduce new methods and strategies to promote fairness in ADM systems. The main corpus of researches has centered on identifying and mitigating systematic discrimination through different definitions of unfairness. Among several notable works that provide a comprehensive overview on fairness in machine learning, we mention the ongoing study by Barocas et al. (2019) [53], which served as the basis for the unfairness measures used our studies, the survey on bias and fairness in machine learning by Mehrabi et al. (2019) [54], as well as the review of discrimination measures for algorithm decision making by Žliobaitė (2017) [55]. A significant challenge in defining software outputs as fair or not lies in the mathematical difficulty of meeting multiple definitions of fairness simultaneously [56] [57], giving rise to an ontological limitation: indeed, in order to define a "fair impact" it would be necessary –and essential– to include several political, economic and cultural aspects [58], therefore a universally accepted notion of fairness can not exist. The swiftly achieved relevance of this issue contributed to the birth of a new field of research, whose main forum today is the ACM Conference on Fairness, Accountability, and Transparency [1], which has been designed and promoted not only for computer scientists working in this field, but also for scholars and practitioners from *"law, social sciences, and humanities to investigate and tackle issues in this emerging area"*. The topic has become relevant for policymakers as well: for example, *avoidance of unfair bias* is one of the key requirements listed in the Ethics Guidelines for Trustworthy AI [59],

[1] ACM FAccT, https://facctconference.org, previously named ACM FAT, founded in 2018 https://facctconference.org/2018/index.html

a foundational document for the European efforts to regulate AI, currently going through the last steps of the European legislative process. In the meantime, the major institutions for technology standardization are also devoting special attention to the topic: in 2022, the US National Institute of Standards and Technology has published the draft of the future Standard for Identifying and Managing Bias in Artificial Intelligence [60]. This initiative follows the publication of the Technical Report “Bias in AI systems and AI aided decision making" by the International Standard Organization [61]. The potential danger to fundamental human rights that is posed by AI and, more in general, by data-driven technologies –as highlighted by several jurists such as in [62] and[63]– is the main driver of many initiatives in the field of ethics and governance of AI, of which those mentioned above are only a small fraction. As a matter of fact, consider the dozens of principles and guidelines for ethical artificial intelligence (AI) issued by private companies, research institutions and public sector organizations [64].

Our approach fits within the realm of interdisciplinary discussions and adds to the existing body of research on algorithmic bias and fairness. Rather than just examining the results of automated decision-making systems, we shift the focus to their inputs, whose investigation is necessary according to several recent studies [65]: *“There is a need to consider social-minded measures along the whole data pipeline”* [66] and *“Returning to the idea of unfairness suggests several new areas of inquiry [. . .] a shift in focus from outcomes to inputs and processes“* [67].

Our proposal differentiates from the reference literature for two additional properties: i) it is built upon a series of international standards, which incorporate by design a multi-stakeholder perspective; 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. Indeed, given the socio-technical nature of the problem, we firmly believe that a risk approach is preferable because it keeps the ultimate responsibility in the realm of human agency. Therefore, the proposed methodology address the need to better document the AI pipeline, particularly relevant in the algorithmic fairness community as shown in the exhaustive work of Fabris et al. [68]. Reporting imbalances in a synthetic and meaningful way is part of the necessary further efforts of the AI/ML community in devoting more attention to the dataset documentation, as acknowledged by Königstorfer and Thalmann: *“one should also record whether there were imbalances* *in the training data with regards to the target categories or how these imbalances were corrected"* [69].

An approach similar to ours but wider in scope is the work of Takashi Matsumoto and Arisa Ema (2020) [70], 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 was 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, passing through the service operation management and aspects related to the code of conduct of the service provider as well as the communication with users. The authors consider both data quality and data imbalance as risk factors –without indicating specific measures– and they stress the importance of visualizing the relations between risk factors for the purpose of a better planned 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. Furthermore, given that data quality metrics are well-established in SQuaRE, we specify that we did not address data quality as a risk factor. However, we recognize that specific studies should be conducted for selecting the most suitable and accurate measures for data quality in the management of ADM system risks.

Other approaches which can be related to ours are in the direction of labeling datasets. "The Dataset Nutrition Label Project" [2] [71] 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 for exploring the previously mentioned aspects. The ultimate goal is to prevent flawed, incomplete, skewed or problematic data from having a negative impact on automated decision-making systems and to foster the creation of more inclusive algorithms.

A similar goal was proposed in the work "Ethically and socially-aware labeling" (EASAL) [72], where authors propose a conceptual and operational framework to label datasets and identify possible risks of discriminatory output when used in decision making or decision support systems. Thus, it aims to plan and develop datasets metadata to help software engineers to be aware of the risks of discrimination.

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

Particularly, the authors 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 with imbalanced data. Indeed, the same authors then published a data annotation and visualization schema based on Bayesian statistical inference [73], always for the purpose of warning about the risk of discriminatory outcomes of a given dataset.
Yet another labelling approach is proposed by Gebru et al., "Datasheets for Datasets" [74]: with respect to the previous proposals, this research work consists of more discursive technical sheets for the purpose of encouraging an increasingly clear and comprehensive communication between users of a dataset and its creators.
Another noteworthy work is "DataTags – Share Sensitive Data with Confidence" [75], a project conducted by members of the Privacy Tools project in collaboration with the IQSS Dataverse team. The goal of DataTags is to support researchers who are not legal or technical experts in investigating considerations about proper handling of human subjects data, and make informed decisions when collecting, storing, and sharing sensitive data.

Finally, it is important to mention the development of tools: in the recent years researches both in the profit sector and in universities developed toolkits for bias detection and mitigation [76]. For example:

- the AI Fairness 360 Open Source Toolkit [77], an open source library developed by IBM and designed to examine and mitigate bias in the output of machine learning models; it provides several metrics to analyze the unfairness of the models, as well as pre-processing algorithms to transform the dataset;
- the What-If Tool [78] developed by Google, which can be used to analyze the characteristics of a dataset and of the models derived from it; the unfairness of these models can be analyzed with respect to various measures, and an interactive graphical user interface let the user perform a sensitivity analysis by shifting classification thresholds for the selected features;
- Aequitas is an open source bias audit toolkit [79] designed by the Center for Data Science and Public Policy at the University of Chicago: it allows to generate a bias report that includes multiple unfairness measures based on the user's selection of reference groups;

- the Themis software [80] developed by the University of Massachusetts Amherst differentiates from the previous tools because it is based on the concept of causal discrimination: a test suite captures the causal relationships between inputs and outputs, providing a causal discrimination score for a particular set of characteristics.

- The FairMask algorithm proposed in [81] is a model-based extrapolation method that is capable of both mitigating bias and explaining the cause; the authors aim to offset the biased predictions of the classification model by rebalancing the distribution of protected attributes, with a view to better detecting and mitigating algorithmic discrimination in machine learning software problems.

These and other fairness toolkits do not consider measures of balance in the input datasets: again, we emphasize the complementarity of our work with existing approaches and potential future integration.

In summary, in this dissertation we study how an imbalance in the input data of an ADM system can be used as an indicator of potential unfair software output, combining concepts from data quality measurement and risk management. The proposed approach has been tested first in [41] on a few hypothetical exemplar distributions and then on several real datasets. Then, we ran more exhaustive studies by applying two different mutation techniques to generate a number of derived synthetic datasets having different levels of balance, in one case to multiclass attributes [43] and in the other case to binary attributes [44]. After that, we define a methodology for identifying thresholds of balance to forecast a defined level of algorithmic fairness [45]. Finally, we move our investigation on data imbalance forward by analyzing intersectionality among the classes of protected attributes, and the impact of an imbalanced distribution in the target variable [46].
Particularly, there are two fundamental recurring elements between these studies:

- the experimental procedure, as we will see in the following chapters, the method that we adopted to collect synthetic data remains substantially unchanged;

- the computation of the relationship between balance and unfairness measures, in accordance with the usage of balance measures as indicators of the risk of systematic discrimination.

Chapter 4

Identifying Imbalance in Datasets with Balance Measures

Several measures have been proposed in the literature in order to explain the abstract construct of *imbalance*; in this Chapter we aim at understanding how these indexes reflect our subjective, and probably limited, perception of imbalance. Particularly, we formulated the following research question:

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

Hereinafter we report the very first study conducted during my Ph.D. journey and published as "A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision-Making Systems" (2021) [41], a journal article in which we analyzed both Research Question 1 and Research Question 2 proposed in this dissertation.

4.1 Experimental Design

In order to address the first research question, we assess a set of measures that are able to measure balance in the data –and consequently its absence, that is, imbalance. As shown in Figure 4.1, we defined a set of synthetic attributes with a known and simple exemplar distribution whose balance can be judged; then we asses the values

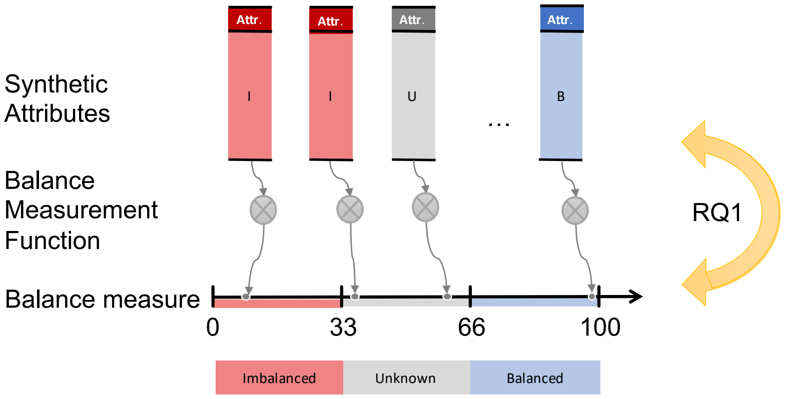


Fig. 4.1 Investigative approach for RQ1.

of the balance measures against the human assessment. More in detail, we followed this procedure:

- we defined a set of *synthetic attributes* with a simple description of the distribution between the classes and our expectation in broad terms;
- we attributed to each synthetic attribute of a specific balance judgement;
- we applied the balance measures described in Section 2.2.2 on this set of synthetic attributes;
- we compared measures versus expectations in order to assess the performance of each index.

In simple terms, we examined the behaviour of the balance measures by analyzing different cases of distribution of the occurrences between the classes of certain synthetic attributes.

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 with respect to max balance and the other half is 10% lower, we expect overall high value measures;

4. *One off*: occurrences are equally distributed between all the classes except for one, which is empty;

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, that is, 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

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

Furthermore, note that in this experiment we deal with *empty classes*, that is, classes that *exist* (potentially there could be occurrences) but are *not* represented in the given synthetic attribute: indeed, trying to simulate a real scenario, a dataset that contains no instance of a given class –for example, all males and zero females– is imbalanced.

Figure 4.2 summarizes all the exemplars of synthetic distributions. Overall we defined 24 distributions (6 cases of distribution × 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%,

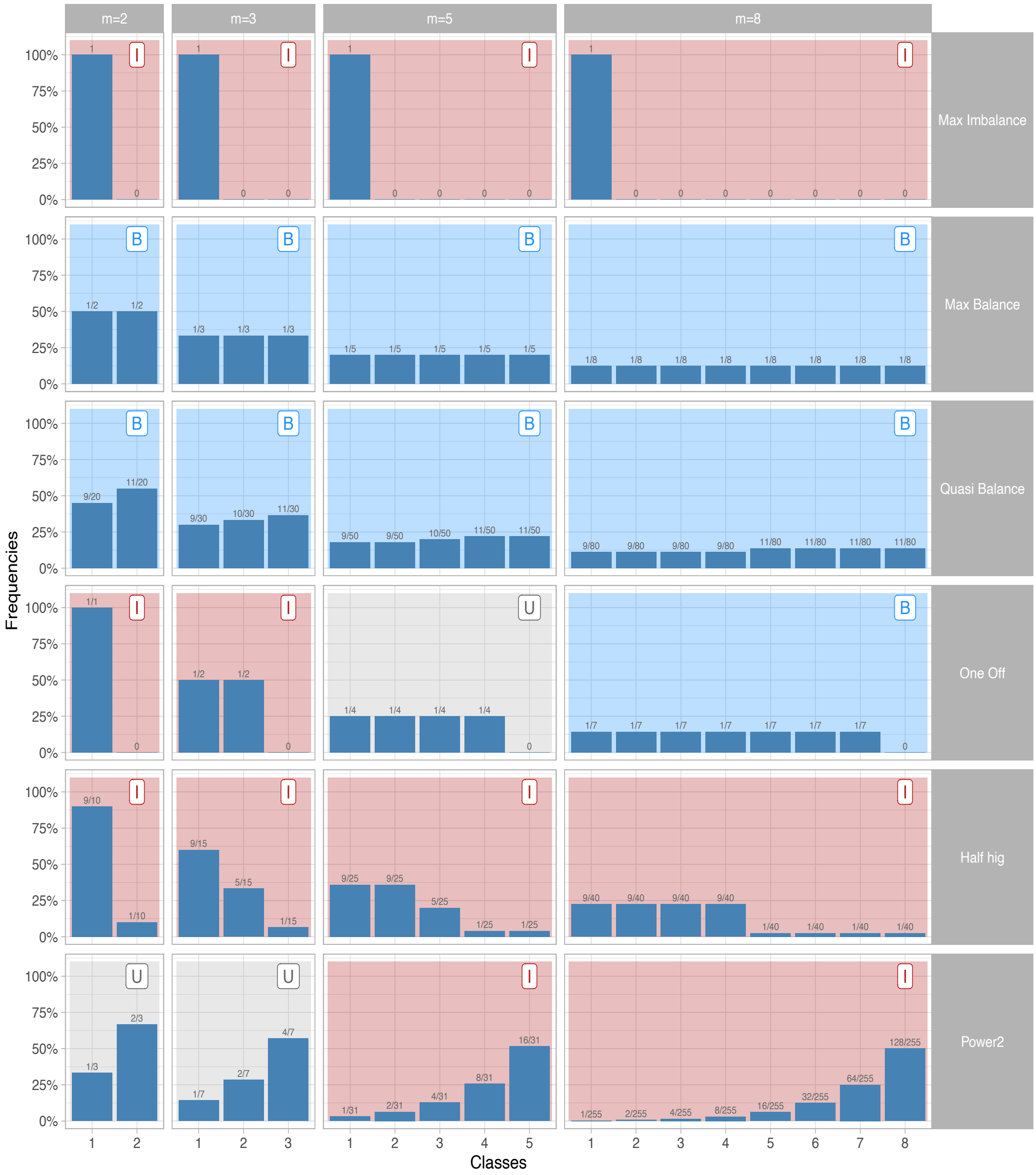


Fig. 4.2 Summary of the synthetic exemplar distributions with the relative balance level expected by the authors, where I indicates imbalance, B stands for balance, and U for uncertain.

- B = balanced if we expect the measure to be greater than 67%,
- 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 –that is, the 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 4.2 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 the different balance measures described in Section 2.2.2. To compare the performance of the different balance measures we compute the accuracy in predicting imbalanced distribution. We assess the accuracy in terms of Accuracy and F-score as they are common metrics used in classifier evaluation.

4.2 Results and Discussion

As detailed in Section 4.1, we applied the indexes of balance to the attributes created ad-hoc in order to test their behaviour in presence of different distributions of the occurrences between the classes of a certain attribute.

Results are reported in Figure 4.3, which reports 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 the balance 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 9 classes. The same results can be read from the perspective of these three latter distributions (that is, those in the second row of Figure 4.3): One Off is the distribution where the indexes performed better, with 8 correct cases out of 16 (mostly in correspondence of $m = 2$ and $m = 8$), two correct cases for each index; in Half high, we observe 5 correct cases out of 16, of which 4 are from the Imbalance Ratio index, and the last one being Simpson index with $m = 2$; in Power 2 distributions, only 4 cases out of 16 were correctly detected by the indexes: 3 for Imbalance Ratio index in correspondence of $m = 2; 5; 8$ and 1 for Simpson index with $m = 8$.
Also, we observe that 13 out of the possible 32 cases of imbalance were detected (40%), 3 out of 4 of balance, and only 1 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, 7 correct classifications are in correspondence of $m = 2$, 6 correct classifications are in correspondence of $m = 8$, 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.

Finally, the capability of the indexes to detect imbalance can be summarized in terms of the overall accuracy of the classification: the values corresponding to the four balance measures are reported in Figure 4.4.

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

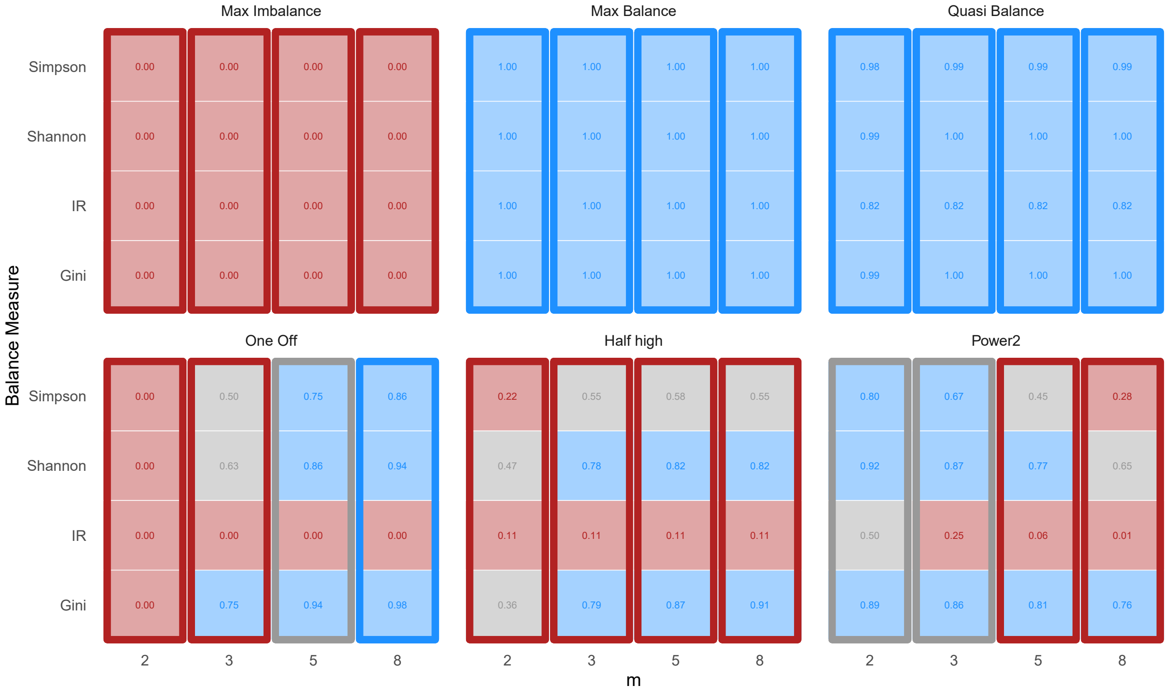


Fig. 4.3 Classification of the synthetic attributes based on the balance measures, where the colored tile along with the value of the measure indicates the level of balance (blue means balanced, red imbalanced, and gray uncertain), while the border color represents the expected class using the same encoding above.

4.3 Limitations

The results are highly dependent on the judgements of the authors (construct validity) and on the choice of the exemplar distributions (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. However, a higher number of notable distributions and a larger pool of judgements (for instance, via crowd-voting) would be necessary conditions to increase the validity of the findings.

Finally, a sensitivity analyses on the thresholds used to define the three classes Imbalanced/Balanced/Uncertain should improve the reliability of the results.

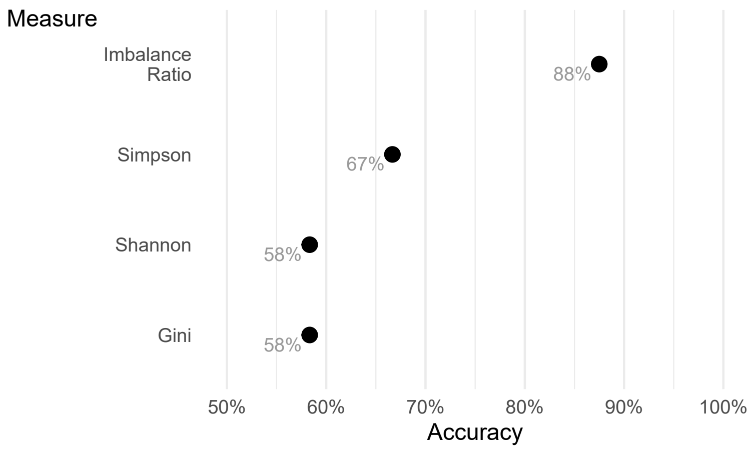


Fig. 4.4 Ability of the balance measures to detect imbalance.

4.4 Research Outcomes

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: indeed, when just one class is empty, the index drops to 0. An intermediate result is achieved by the Simpson index, while the Gini and Shannon indexes exhibit the lowest performances, as they generally present higher values than the former two for the same distributions. Thus, it may be necessary to study how the indexes behave with the application of different thresholds.
Moreover, we did not observe any specific trend associated with the number of classes.

As a further observation, we highlight that we decided to analyze a synthetic attribute with a power law with base 2 distribution since power law distributions are very common in several real cases (for example, income): we noted that the indexes achieved the worst performances exactly in this case, therefore a more in-depth analysis should be conducted in order to understand how to adequately deal with this family of distributions.

In general, these findings are encouraging enough to continue the exploration of data imbalance identified through the application of balance measures.

Chapter 5

Balance Measures as Risk Indicators

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 1.1). The ultimate focus of this study consists in assessing whether the imbalance in data, measured by means of the selected indexes of balance, may signal a discrimination risk in the ADM system. Specifically, we aim at answering the following research question:

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

To answer RQ2, we first conducted an exploratory analysis of data imbalance on two datasets with only two balance measures, published as a conference paper titled "Identifying Risks in Datasets for Automated Decision–Making" (2020) [42]. This first study laid the basis of our proposal –that is, a risk assessment approach based on quantitative measures to evaluate imbalance in the input datasets of ADM systems–, which was subsequently tested in detail in the journal article entitled "A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision-Making Systems" (2021) [41], where we extended the quantitative analysis with more balance measures and datasets. Hereinafter we report the research work published in these articles.

5.1 Experimental Design

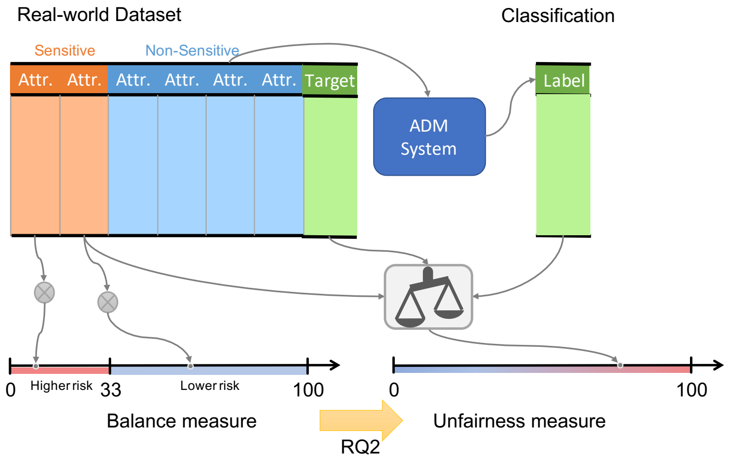


Fig. 5.1 Method of analysis used for RQ2.

The approach we adopted to investigate the research question above is summarized in Figure 5.1: given a dataset, its non-sensitive attributes and a target variable can be used to train an ADM system that performs a classification task. The unfairness of the classification with respect to a certain sensitive attribute can be evaluated considering the target class and the predicted class. The balance of the sensitive attributes can be quantified by applying any of the indexes described in Section 2.2.2, in order to understand the ability of such balance measures to reveal a potential discrimination risk –that is, *un-fairness*. Specifically, we followed this procedure:

- we selected seven large *datasets* from different domains (see Subsection 5.1.1); some datasets include both the *target variable* and the *score variable* that correspond to the actual ADM-predicted class; other datasets include the *target variable* but not the result of an ADM classification, so we trained a simple ADM system and complemented the original datasets with the *score variable*;

- we identified the *protected attributes* in the selected datasets; the identification was performed taking as reference the definition provided in “Article 21 - Non-discrimination” of the EU Charter of Fundamental Rights[30], already reported in Chapter 1;

- we evaluated the *balance* of the protected attributes using the measures described in 2.2.2; in order to assess the risk more in detail, we classified the protected attributes into *Higher risk* and *Lower risk* based on the value of the balance measures, using the threshold of 33%;

- we assessed the *unfairness* of the predictions with respect to the sensitive attributes; we computed the unfairness measures ($\mathfrak{U}$) related to the *Separation* and the *Independence* criteria described in Section 2.2.3;

- we analyzed the relationship between the balance measures and the fairness criteria: we observed the unfairness versus the balance induced risk. Specifically, we compare the values of the unfairness measures related to the protected attributes classified as *Higher risk* versus 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*.

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, but rather we aim to check if a looser – rank-based – relation holds.

Finally, in Algorithm 1 we present the pseudocode for the measurements of both balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$ for each protected attribute in the selected datasets.

5.1.1 Datasets

In Table 5.1 we report both the datasets analyzed in this study, whose complete description and characteristics are provided in Section 2.2.1, and the related protected attributes.

Note that for the datasets that do *not* contain a pre-computed classification, 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*.

Furthermore, bearing in mind the indications provided in Chapter 1, it is important

Algorithm 1: Measurements of balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$ for RQ2.

```
Require: categorical protected attributes $A_i$, each with number of categories $m > 0$;
    balance measures $\mathfrak{B} =$ Gini, Shannon, Simpson or IR;
    unfairness measures $\mathfrak{U} = \mathfrak{U}_{Sep\_TP}, \mathfrak{U}_{Sep\_FP}$ or $\mathfrak{U}_{Ind}$.
    Input  : Dataset $D_j$ with $j = 1, ..., 7$
    Output : Balance measures $\mathfrak{B}$, Unfairness measures $\mathfrak{U}$
for all $D_j \in \{D_1, \ldots, D_7\}$ do
    identification of protected attributes $A_{j,i}$
    if the score variable exists then
        $\mathfrak{B}_{j,i} \leftarrow \mathfrak{B}(A_{j,i})$
        $\mathfrak{U}_{j,i} \leftarrow \mathfrak{U}(A_{j,i})$
    else if the score variable is missing then
        randomly data splitting of 70%-30% into training-test sets
        prediction of the score variable with a classification model
        $\mathfrak{B}_{j,i} \leftarrow \mathfrak{B}(A_{j,i}) \in$ training set
        $\mathfrak{U}_{j,i} \leftarrow \mathfrak{U}(A_{j,i}) \in$ test set
    end if
end for
return $\mathfrak{B}_{j,i}$, $\mathfrak{U}_{j,i}$
```

to underline that protected attributes are never taken into account in classification models.

In this study we deal with empty classes, that is, classes that *exist* (potentially there could be occurrences) but are *not* represented in the dataset. Indeed, in our view a dataset that contains no instance of a given class –for example, all males or all whites– is imbalanced with respect to that protected attribute; therefore, we decided to take into consideration *all* the classes of each protected attribute identified in the datasets, including classes with zero occurrences.

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

Table 5.1 Complete list of the *datasets* with the analyzed *attributes*.

Dataset	**Domain**	**Protected attributes**
COMPAS	Justice	*ethnicity, sex, age category*
Default of credit cards clients (Dccc)	Financial	*sex, education*
Income	Welfare	*education, race, sex, native country*
Juvenile justice	Justice	*sex, stranger, country of origin, area of origin, age category, age*
Statlog	Financial	*status, sex, foreign worker*
Student (Mathematics)	Welfare	*sex, age, mother's job, father's job, mother's education, father's education*
Student (Portuguese)	Welfare	*sex, age, mother's job, father's job, mother's education, father's education*

5.2 Results and Discussion

We report results of applying balance measures on the protected attributes of the datasets indicated in Section 5.1.1, as well as the unfairness measures computed with respect to such attributes.

For each combination of balance and unfairness measures, Figure 5.2 reports a boxplot that shows the distribution of the unfairness values for higher risk attributes versus lower risk attributes: we remind from Section 5.1 that we adopted a threshold of 33% such that "imbalanced" corresponds to higher risk, while "unknown" + "balanced" corresponds to lower risk. The more a boxplot leans to the right, the more unfair the treatment of those attributes, and vice-versa: the more a boxplot is close to the left (that is zero) the more the relative attributes are treated fairly. As a general rule, if the boxes of the riskier attributes (colored red) and the one for the less riskier (colored yellow) are not overlapping, then the imbalance-based approach to risk identification is able to discriminate between actually fair and unfair classifications. We observe that:

- Gini index has 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 has a good discrimination ability for both the indicators of the Separation criterion, and a limited ability for the Independence criterion;

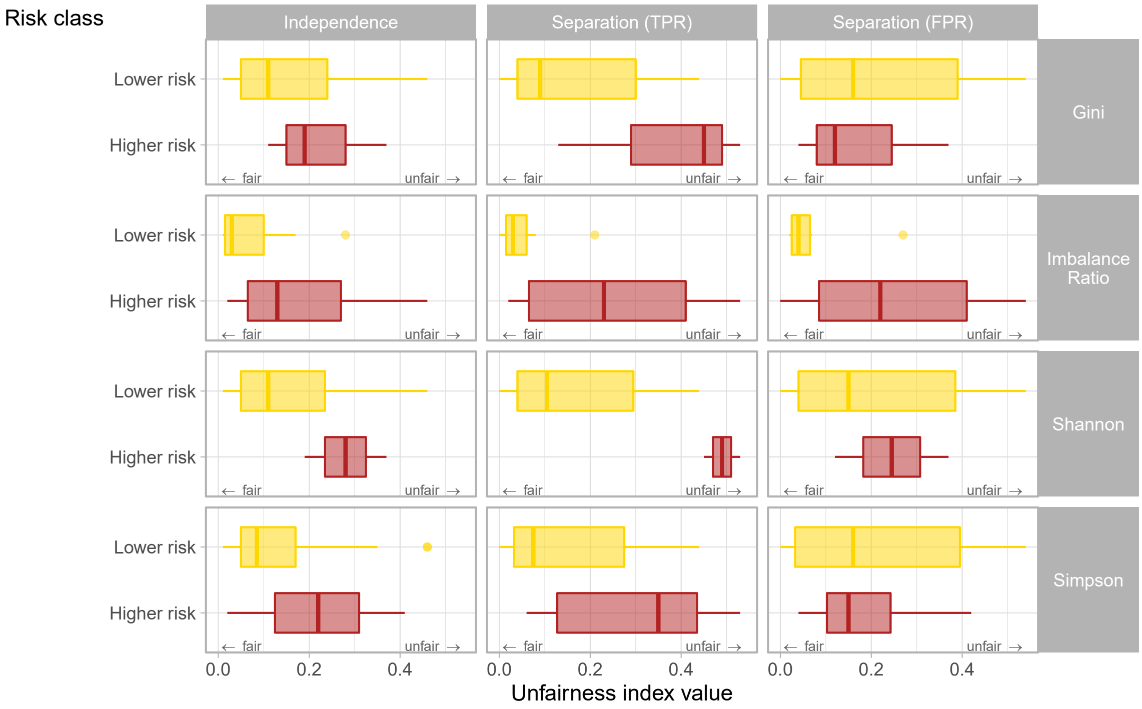


Fig. 5.2 Boxplot of unfairness measures versus balance classification, for different balance measures: the more a boxplot leans to the right, the more unfair the treatment of those attributes; vice-versa, the more a boxplot is close to the left, the more the relative attributes are treated fairly.

- Shannon index has 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 of the Separation criterion;
- the Simpson index has a limited capability on the 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 terms of substantial difference of false positives.

The values that have been summarized in Figure 5.2 are reported in detail in Table 5.2: for each dataset we show all the protected attributes with the related balance measurements –of the Gini, Shannon, Simpson and Imbalance Ratio indexes–

and the corresponding unfairness values relating to the Independence and Separation criteria.

Following the line of reasoning explained above, for high level of unfairness we expect low-value balance indexes, which reveal imbalanced data. Looking at the single attributes starting from the COMPAS dataset, previous studies [37] showed that the data are 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 observation, 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 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 a negative correlation with unfairness measures also in this case. For example, 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. But this trend does not held for all the attributes: for instance, 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.

Therefore, we integrate the analysis with the computation of the Spearman correlation coefficient between *balance* and *unfairness* measures, as reported in Table 5.3. Specifically, we expect the coefficient to be negative: the stronger the negative correlation, the stronger is the relationship between unfairness and balance measures, as we expect the measures to be high (meaning low imbalance) if the unfairness values are low (indicating higher fairness). Hence, in terms of correlation,

Table 5.2 Values of balance measures and unfairness measures.

								Separation	
Dataset	Attribute	m	Gini	Shannon	Simpson	IR	Independence	(TPR)	(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

the best balance measure is the Imbalance Ratio index as it always presents a strong negative correlation –meaning that the higher the indexes, the lower the unfairness measures– followed by the Simpson and the Shannon indexes respectively. The less accurate measure is the Gini index, with two negative correlation values, but weaker than the correlation values of the previous indexes, and a weak positive correlation for the FP rates of the Separation criterion. The correlation analysis confirm that false positive differences are the most difficult to detect with the four balance indexes, while results are encouraging for the Independence criterion and for the discrimination with respect to true positives rates.

Table 5.3 Correlation between balance measures and unfairness measures.

Balance Measures / Fairness criteria	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

5.3 Limitations

As limitations of our approach, first of all we remind that for the datasets where a classification score was present, we have no knowledge about the type of classification model used on those 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. However, it has been widely acknowledged in the literature that the imbalance in the input data plays a significant role in the observed discrimination in the controversial COMPAS case.

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. Further possible extensions regard the usage of other unfairness measures, for example the sufficiency criterion [53].
In addition, as already stated in the Limitations 4.3 of RQ1, an in-depth sensitivity analyses on the thresholds used for the balance and unfairness measures should improve the reliability and generalizability of the overall results.

Overall, it is important to stress again that our study focused on the level of risk analysis, while risk evaluation (that is, 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 [82] [83] is a useful resource to select and test imbalance mitigation techniques, 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 methods mitigate bias on the already predicted scores (data level). It should be observed that these data-engineering aspects are still object of research because of inconsistent and conflicting results [82], and 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.

5.4 Research Outcomes

In this Chapter we proposed and tested a metric-based approach to evaluate imbalance in a given dataset as a potential risk factor for discriminatory outcomes of ADM systems. We analyzed four widely used indexes of balance (Gini, Simpson, Shannon, Imbalance Ratio) and tested their ability to detect discrimination occurring in the classification outcomes of ADM systems trained with seven large datasets. We observed that the balance measures performed differently with respect to different fairness criteria: as a general consideration, evidence suggests that a combined usage is preferable to detect possible discrimination risks, since there is no single balance measure providing the basis for an ideal risk identification across all datasets analyzed. Similarly to the previous research question (whose research outcomes are reported in Section 4.4), the Imbalance Ratio foresees discrimination better than other indexes, although the correlation analysis showed that all indexes are able to detect both the Independence criterion and the Separation criterion with respect to the true positive rate. Instead, discrimination due to the Separation criterion for the false positive rate is much more difficult to be detected, especially for the Gini index.

It is important to emphasize that we performed our risk classification using a unique 33% threshold on all the measures; however, as we observed in the analysis of RQ 1 reported in Section 4.2, the balance measures taken into account can assume quite different values depending on the level of data imbalance.

Chapter 6

Detecting the Risk of Bias in Classification Outcomes with Balance Measures

In this Chapter we investigate to which extent it is possible to assess the risk of unfairness in classification outcomes by measuring the imbalance of protected attributes in training data. Differently from the previous study, in this Chapter we examine how well the balance measures applied to a specific protected attribute reflect a discrimination risk, where the (im)balance of the attributes has been varied through different mutation techniques. This study has been conducted in detail separately for *binary* and *multiclass* protected attributes. In particular, we formulated the following research question:

RQ 3. Is it possible to measure the risk of bias in a classification output by measuring the level of (im)balance in the protected attributes of the training set?

The cornerstone of our approach remains unchanged: by measuring the level of (im)balance of specific attributes in a dataset, it is possible to detect the risk of bias in the classification output from ADM systems. However, as mentioned above, in this study we introduce a mutation technique to generate a number of derived synthetic datasets having different levels of balance; moreover, we add the computation of the

Sufficiency criterion of fairness, in addition to Independence and Separation already taken into account in Chapter 5.

Particularly, to answer RQ 3, we conducted two different studies, one for multiclass attributes and one specific for binary attributes. Hereinafter we will present and discuss these works in chronological order of publication:

- "Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data" (2021) [43], a conference paper in which we analyzed the behaviour of the balance measures when applied to *multiclass* protected attributes;
- "Detecting Risk of Biased Output with Balance Measures" (2022) [44], a journal article in which we focus on *binary* protected attributes.

6.1 Experimental Design

For the purpose of understanding how the balance of protected attributes in training data can be used to assess the risk of algorithmic unfairness in subsequent classification tasks, we selected a set of indexes that are able to measure balance in the data –and thus its absence, that is imbalance–, and we assessed how well such balance measures applied to a given dataset reflect a discrimination risk. Specifically, we followed the following procedure, separately for *multiclass* and *binary* protected attributes:

1. we selected one large dataset in the *multiclass* case, and a multiclass protected attribute with cardinality "m"; whereas, we chose five large datasets in the *binary* case, each with the binary protected attribute "sex" (see Section 6.1.1);
2. using two distinct mutation techniques, one for *multiclass* attributes and one for *binary* attributes, we generated a number of derived synthetic datasets having different levels of balance (see Section 6.1.2); specifically, we adopted a pre-processing method as mutation technique and we mutate the distribution of the occurrences between the classes of a certain attribute by adjusting a certain parameter: `C.perc` in the *multiclass* case and `p` in the *binary* case;

3. we implemented a *binomial logistic regression* model in order to predict the *score variable* for each synthetic dataset; particularly, we trained a binary classifier on a training set composed by the 70% (randomly selected) of the data and we ran it on the remaining 30%, which represents the test set;

4. we measured the level of (im)balance of the protected attribute in the training set through four different widely used *balance measures* (described in Section 2.2.2);

5. we applied two distinct fairness criteria in the *multiclass* case (precisely, the Independence and Separation criteria) to the protected attribute in the test set –that is, to the classifications obtained from the model– for a total of three unfairness measures on each output, whereas we applied three fairness criteria in the *binary* case (that is, Independence, Separation and Sufficiency criteria) for a total of five unfairness measures on each output (see Section 2.2.3);

6. we assessed the relationship between balance measures and fairness criteria by checking whether a negative correlation holds, that is, whether a lower level of balance corresponds to a higher level of unfairness, and vice-versa.

Finally, note that in the Algorithms 3 and 2 we present the pseudocode for the measurements of balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$, separately in the case of *binary* and *multiclass* protected attributes.

6.1.1 Datasets

First of all, note that both in the *multiclass* and in the *binary* case, we decided to include missing values in the analysis by treating them as a separate category "NA" given that in real datasets we can often find missing values.

Moreover, since the selected datasets do *not* contain a pre-computed classification, for each of them 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.

Algorithm 2: Measurements of balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$ for RQ3 in the case of *multiclass* protected attributes.

```
Require: categorical protected attribute $A$, with number of categories $m > 2$;
    mutation technique $M =$ SmoteClassif (for multiclass attributes);
    exemplar distributions $E = \{Balance, QuasiBalance, OneOff, HalfHigh, Power2\}$;
    balance measures $\mathfrak{B} = \{Gini, Shannon, Simpson, IR\}$;
    unfairness measures $\mathfrak{U} = \{\mathfrak{U}_{Sep\_TP}, \mathfrak{U}_{Sep\_FP}, \mathfrak{U}_{Ind}\}$.
    Input  : Dataset $D$
    Output : Balance measures $\mathfrak{B}$, Unfairness measures $\mathfrak{U}$
identification of a multiclass protected attribute $A$ in $D$
for all $e \in E$ do
    application of the mutation technique SmoteClassif $M_e \leftarrow M(A)$
    randomly data splitting of 70%-30% into training-test sets
    prediction of the score variable with a classification model
    $\mathfrak{B}_e \leftarrow \mathfrak{B}(M_e) \in$ training set
    $\mathfrak{U}_e \leftarrow \mathfrak{U}(M_e) \in$ test set
end for
return $\mathfrak{B}_e, \mathfrak{U}_e$
```

Multiclass case. With a view to exploring the potential of our approach in one of the prominent application domain of ADM systems, we examined a dataset belonging to the field of financial services: *Default of Credit Card Clients* (Dccc), whose properties have been summarized in Table 2.1.
Among the protected attributes present in Dccc, we chose the *multiclass* attribute "education" with number of categories $m = 6$; indeed, as a consequence of our decision to include missing values in the analysis and count them as a separate category, this attribute consists of six classes: "NA", "graduate school", "university", "high school", "others", "unknown".

Binary case. We examined five datasets belonging to two different application domains, welfare and financial; the descriptions of the datasets are provided in Section 2.2.1 and their main properties are summarized in Table 2.1, while here we present only the list of the selected datasets with the related domain:

- Financial: **Default of credit cards clients (Dccc)** and **Statlog**;
- Welfare: **Income**, **Student-Mathematics** and **Student-Portuguese**.

Algorithm 3: Measurements of balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$ for RQ3 in the case of *binary* protected attributes.

```
Require: categorical protected attribute $A$ with number of categories $m = 2$;
    mutation technique $M =$ ovun.sample (for binary attributes);
    mutation parameter $P = \{0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$;
    balance measures $\mathfrak{B} = \{Gini, Shannon, Simpson, IR\}$;
    unfairness measures $\mathfrak{U} = \{\mathfrak{U}_{Sep\_TP}, \mathfrak{U}_{Sep\_FP}, \mathfrak{U}_{Suf\_PP}, \mathfrak{U}_{Suf\_PN}, \mathfrak{U}_{Ind}\}$.
    Input  : Dataset $D_j$ with $j = 1, ..., 5$
    Output : Balance measures $\mathfrak{B}$, Unfairness measures $\mathfrak{U}$
for all $D_j \in \{D_1, \ldots, D_5\}$ do
    identification of a binary protected attribute $A_j$
    for all $p \in P$ do
        application of the mutation technique ovun.sample $M_{j,p} \leftarrow M(A_{j,p})$
        randomly data splitting of 70%-30% into training-test sets
        prediction of the score variable with a classification model
        $\mathfrak{B}_{j,p} \leftarrow \mathfrak{B}(M_{j,p}) \in$ training set
        $\mathfrak{U}_{j,p} \leftarrow \mathfrak{U}(M_{j,p}) \in$ test set
    end for
end for
return $\mathfrak{B}_{j,p}$, $\mathfrak{U}_{j,p}$
```

For all these datasets, we decided to examine the *binary* protected attribute “sex” –with the two classes “Male” and “Female”–, as it is one of the most common sources of imbalance and consequent discrimination [19].

6.1.2 Mutation Techniques

We adopted two specific *pre-processing* methods as mutation techniques in order to generate a large number of variations of the distribution of the occurrences between the classes of a given protected attribute.

Multiclass case. We used the `UBL-package` provided by RDocumentation [1]:

> *“The package provides a diversity of pre-processing functions to deal with both classification (binary and multi-class) and regression problems that encompass non-uniform costs and/or benefits.”*

[1] https://rdocumentation.org/packages/UBL/versions/0.0.6/topics/UBL-package, last visited on June 1, 2023

In particular, we chose the `SmoteClassif` function [2] as mutation technique:

"*This function handles unbalanced classification problems using the SMOTE method. Namely, it can generate a new 'SMOTEd' data set that addresses the class unbalance problem.*"

This method has been applied with the following settings:

- "`education∼`" is the multi-class protected attribute chosen as formula.
- "`C.perc`" is a list containing the percentages of under-sampling or/and over-sampling to apply to each class of the protected attribute in the formula: an over-sampling percentage is a number above 1, while an under-sampling percentage should be a number below 1; in particular, a class remains unchanged if the number 1 is provided for that class; note that there exists an infinite number of possible combinations of the percentages of the classes. Alternatively, `C.perc` may be set to the two values "balance" (the default) or "extreme", cases where the sampling percentages are automatically estimated either to balance the examples between the minority and majority classes, or to invert the distribution of examples across the existing classes transforming the majority classes into the minority, and vice-versa.
- "`repl=FALSE`" is a boolean value controlling the possibility of having (or not, as in this case) repetition of examples when performing under-sampling by selecting among the majority class(es) examples.

In our study we decided to examine five different cases for the parameter `C.perc`: first, we set the parameter to the pre-established value "balance" –which is the perfect uniform distribution, with all the occurrences equally distributed between the classes–, then we assigned four different lists of percentages for the classes of the protected attribute, corresponding to the exemplar distributions "Power2", "HalfHigh", "OneOff" and "QuasiBalance" already analyzed in Chapter 4 (where we discussed the experiment published in [41]). Briefly, these exemplar distributions are described as follows:

- *Power 2*: occurrences are distributed according to a power-law with base 2, that is, distributions among the classes increase like the powers of 2;

[2] https://www.rdocumentation.org/packages/UBL/versions/0.0.6/topics/SmoteClassif, last visited on June 1, 2023

- *Half High*: occurrences are distributed mostly among half of the classes while the remaining have a very low frequency –specifically, a ratio of 1:9 has been chosen for the frequencies of the two halves;
- *One Off*: occurrences are distributed among all classes but one;
- *Quasi Balance*: half of the classes are 10% higher with respect to max balance and the other half is 10% lower.

In addition, for each exemplar distribution we considered 6 permutations of the values of the percentages assigned to the different classes of the protected attribute. This means that, for instance, in the *One Off* configuration the four different permutations have each a different class with zero occurrences.
Finally, in order to increase the variability –and thus the reliability– of our method, we decided to vary a `seed` (an integer recommended for reproducibility purposes to keep track of the samples) by setting 100 randomly sampled values between 1 and 1000.

Therefore, for the discussion of the results in the *multiclass* case we kept track of the outcomes for each value of the `seed` in the case of the mutation with `C.perc`=“balance", for a total of 1×100=100 values for each measurement –both *balance measures* and *fairness criteria*–; whereas in the case of the mutations corresponding to the four different lists of percentages, we collected a total of
4 (exemplar distributions) × 6 (permutations) × 100 (seed) = 2400 values for each measurement, leading to a grand total of 100+2400=2500 values for each *balance measure* and 2500 values for each *unfairness measure*.

Binary case. In order to generate a variant of an original dataset with respect to the protected attribute “sex”, we adopted the `ROSE-package`[3] [84] provided by RDocumentation:

> “*Functions to deal with binary classification problems in the presence of imbalanced classes. Synthetic balanced samples are generated according to ROSE (Menardi and Torelli, 2014).*”

[3] https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ROSE-package, last visited on June 1, 2023

Specifically, we applied the `ovun.sample` function [4] as mutation technique:

"*Creates possibly balanced samples by random over-sampling minority examples, under-sampling majority examples or combination of over- and under-sampling.*"

This technique was implemented with the following settings:

- "`sex~`" is the binary protected attribute chosen as formula, since it is one of the most common sources of imbalance and consequent discrimination;
- "`both`" as method, which indicates a combination of over-sampling minority examples and under-sampling majority examples to perform the random sampling;
- "`N`" equal to the same number of rows of the dataset under analysis as desired sample size of the resulting dataset;
- "`p`" represents "*the probability of resampling from the rare class*" and it has been set to 9 different values in order to vary as much as possible the distribution of the occurrences between the two categories of the attribute "sex": 0.01 (corresponding to the case of minimum balance), 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5 (maximum balance). When the value of `p` is set to 0.5, it means aiming for the most balance distribution between the two classes, whereas lower values correspond to less balanced distribution;
- "`seed`" is "*a single value, interpreted as an integer, recommended to specify seeds and keep track of the sample*", therefore we decided to vary such value by randomly selecting 50 values between 1 and 1000, in order to enhance the variability and consequently the reliability of our approach.

In order to increase the variability and the reliability of our method, given the random nature of the resampling, we generated 100 different mutations (for each value of `p`) using distinct seeds. Overall we applied this technique to the five datasets listed in Table 6.1.1:
5 datasets × 9 levels of `p` × 100 seeds = 4500 synthetically mutated datasets.

[4] https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ovun.sample, last visited on June 1, 2023

Finally, note that in both cases –multiclass attributes and binary attributes– the generated mutated datasets have the same number of rows as the original ones, while the distribution of the other variables in the dataset remains unchanged.

6.2 Results and Discussion

First of all, note that hereinafter the values of both balance measures and fairness criteria are multiplied by 100, so that all measures range in the interval [0,100], in order to simplify the readability of the results. For the interpretability of the measures, we remind that:

- in the case of *balance measures*, values close to 0 indicate a high imbalance, vice-versa the closer the measure to 100 and the higher the balance;
- in the case of *unfairness measures*, values close to 0 reveal a fair classification, on the contrary, high values indicate unfair behavior.

6.2.1 Multiclass case

Before addressing the research question, we observe the behavior of both balance measures and fairness criteria as the permutation of a specific mutation varies –for each of the four mutations corresponding to the exemplar distributions.
Given a certain mutation, we note that the values of the balance measures remain substantially unchanged for all six permutations, suggesting that permutations have a very weak effect or no effect at all on the balance measures.
On the contrary, regarding the fairness criteria, we observe an irregular behavior particularly in the case of mutations that lead to more imbalanced distributions –Power2, HalfHigh and OneOff–, while the values tend to be more stable for QuasiBalance; thus, permutations result to have some effect on the measures of unfairness.

After this preliminary check, we analyzed first the behavior of both balance and unfairness measures in response to mutations, and then we examined the relationship between balance measures and fairness criteria by checking whether a negative correlation holds, that is, whether a lower level of balance corresponds to a higher level of unfairness, and vice-versa.

a) Analysis of the balance measures in response to mutations

With a view to analyzing more in-depth the behavior of the indexes, we report in Figure 6.1 the box plots of the whole distributions for each balance measure with respect to mutations. Specifically, we expect balance measures to increase as the mutation tends to be increasingly balanced: keeping in mind the description of the exemplar distributions in Section 6.1.2, the most imbalanced distribution is represented by Power2, followed by HalfHigh (which is slightly more balanced with respect to Power2), OneOff (slightly more balanced again), QuasiBalance and Balance –which is the best case, with all the occurrences equally distributed between the classes. Indeed, we note an overall absence of variance and we observe that balance measures increase as the mutations become increasingly balanced, with the lowest values in correspondence of the case Power2, respectively followed by HalfHigh (which presents higher values with respect to the previous, indicating a more balanced distribution) and OneOff (with even higher values); then, we observe the highest outcomes corresponding to the cases QuasiBalance and Balance, confirming our general expectations.

Looking at the individual measures, Gini and Shannon indexes present a similar behavior, with values in the range between 75 and 100, and apparently no difference in detecting QuasiBalance and Balance, both with values close to 100. The Simpson index covers a larger range, about 38-100, with a slight difference between the cases QuasiBalance and Balance. Finally, the IR index appears to be spanned over the whole range [0,100], with well distinct values for the two most balanced cases, and the uncommon presence of zero values in correspondence of the mutation OneOff: indeed, by definition of IR, in the special case of one or more *empty* classes[5] the value of the IR index results to be zero, that is the reason for which we observe null values in the case of OneOff.

b) Analysis of the fairness criteria in response to mutations

An analogous analysis has been performed for the unfairness measures and is reported in Figure 6.2, which presents the box plots of the whole distributions for each fairness

[5] We remind from Chapter 4 that we defined as *empty class* a class with null frequency as there are no occurrences, that is, a class that *exists* because potentially there could be occurrences, but is *not* represented in the dataset.

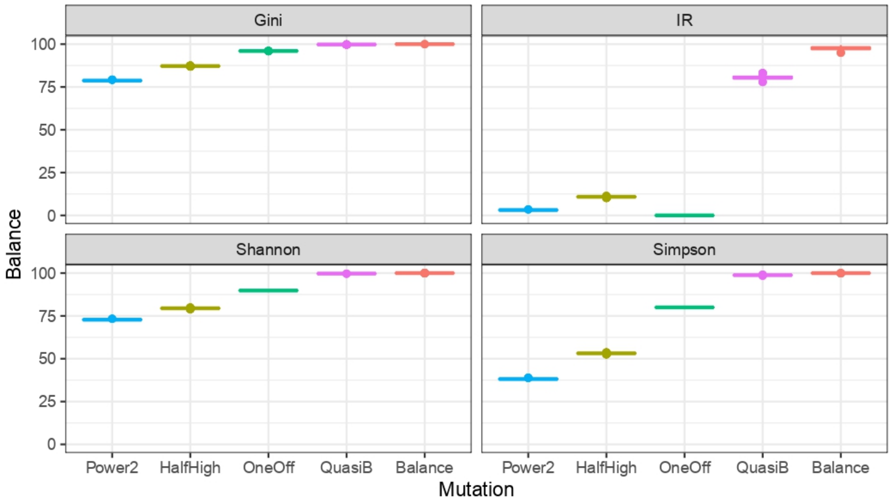


Fig. 6.1 Distributions of the *balance measures* with respect to mutations.

criterion in correspondence of the five mutations. First of all, we observe that the variance decreases as the mutations tend to be more and more balanced, with a very large variance in the cases Power2 and HalfHigh; then, the variance tends to drop in the intermediate case OneOff, and becomes much smaller for QuasiBalance and Balance. Such variance trend is substantially the same for all the unfairness measures, but looking at the individual measures, we observe that the Separation criterion in the case of TP rate assumes values in the range [0,23], while it assumes values between 0 and 4 in the case of FP rate; finally, we observe that the Independence criterion ranges in the interval [0,7].

Despite the different ranges of values, we note that all the unfairness measures present overall very similar distributions with respect to mutations: we observe the highest values in correspondence of Power2 (thus indicating the most unfair classification output), followed by HalfHigh, OneOff, QuasiBalance and Balance which all present lower values compared to the case Power2, revealing a fairer classification output, but substantially no difference between the mean values.

c) Analysis of the fairness criteria in response to the balance measures

In the subsequent analysis we examine the trends of the *fairness criteria* in response to the *balance measures* with respect to the different mutations, by considering (for each mutation) first the mean value of *Unfairness*, as reported in Figure 6.3 (a), and

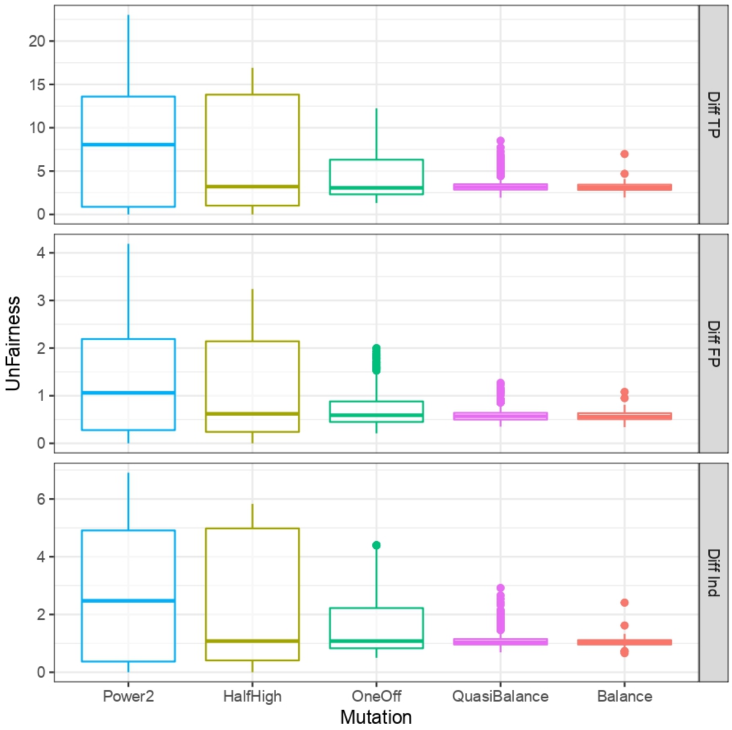


Fig. 6.2 Distributions of the *unfairness measures* with respect to mutations.

then the maximum value, which is represented in Figure 6.3 (b). Particularly, we aggregate data for mutation and we plot the distributions of the three fairness criteria (Y axis) with respect to the increase of the balance measures (X axis); therefore, the dashed lines trace the trend of the unfairness measures as the balance measures increase. We also specify that regarding the balance measures we always consider the mean values for each mutation (as in the previous analysis of Figure 6.1 we observed an absence of variance); whereas, concerning the unfairness measures, after aggregating data for mutation, we first compute the mean values (“Mean case") and then we take the maximum values (“Worst case", which corresponds to the most unfair output for that given mutation), since we previously observed in Figure 6.2 a large variance, above all in correspondence of highly imbalanced distributions. Overall, we observe a decrease in the unfairness measures as the balance measures increase.

This trend is confirmed in both the Mean and the Worst cases, but looking at the individual indexes of balance we observe an irregular behavior for the IR index: indeed, we already explained the special case of one or more classes with null frequency that make the IR index drop to zero, therefore in correspondence of the mutation OneOff the IR index results to be zero, while the unfairness level assumes an intermediate value between the mutation HalfHigh and QuasiBalance, thus reflecting the same order of the *unfairness levels* in response to the other balance measures. In turn, we observe that the unfairness measures decrease –thus indicating an increasingly fair classification– as the mutations tend to be increasingly balanced, with the highest values in the case of Power2, respectively followed by HalfHigh (which presents lower values compared to the previous, revealing a fairer classification output) and OneOff (with even lower values); then, the lowest values are obtained in the cases QuasiBalance and Balance, thus indicating the fairest output.

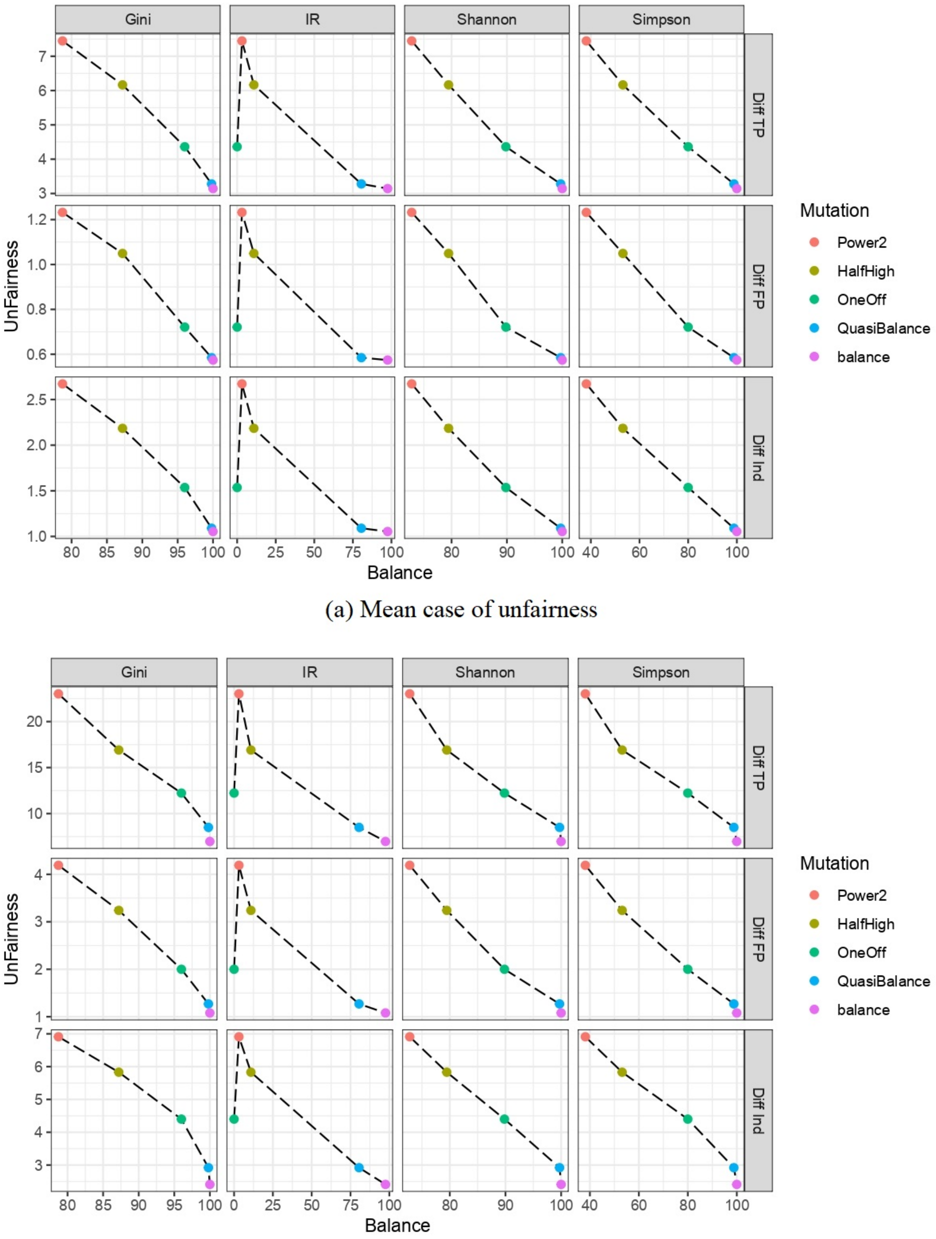
To analyze results more in depth, we integrate our study with the computation of the Spearman correlation coefficient between *balance* and *unfairness* measures. Specifically, we expect the coefficient to be negative, as we expect the balance measures to be high (meaning low imbalance) if the unfairness values are low (indicating higher fairness). Thus, the stronger the negative correlation, the stronger is the relationship between balance and unfairness measures.
As we can observe from Table 6.1, all the balance measures present a negative correlation with the fairness criteria, meaning that the higher the indexes of balance, the lower the unfairness measures; in addition, the computations reveal that such values are all significant (p-value<0.05) except for the IR index in correspondence of Separation TPR. More in detail, we notice that the Imbalance Ratio index always presents a weaker negative correlation (between -0.018 and -0.049) with respect to the other three balance measures, which seem to reflect very similarly the different unfairness measures; specifically, the more accurate balance measure is the Shannon index, followed by Gini and Simpson indexes respectively, each one with correlation values between -0.08 and -0.1.
From the perspective of the unfairness measures, the Separation criterion in the case of True Positive rate results to be the most difficult to detect (with correlation values around -0.08, and even -0.018 in correspondence of the IR index), followed by the Independence criterion –which presents a slightly stronger negative correlation–, while the Separation criterion in the case of False Positive rate appears to be the best to detect, showing a stronger negative correlation above all with the Gini, Shannon and Simpson indexes (with correlation values around -0.1).
Although correlations are weak, they are always negative and significant, that is

p-value<0.05 (except for IR with respect to Separation TPR), thus the correlation analysis do not reject the hypothesis that the balance measures are capable of revealing unfairness of software output, with some variation among the balance measures (for example, we observed halved correlation values in correspondence of the IR index, which is highly sensitive to extreme values of balance and imbalance).

Table 6.1 Correlation between balance measures and unfairness measures.

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	-0.088	-0.089	-0.087	-0.046
Separation (TPR)	-0.083	-0.084	-0.082	-0.018
(FPR)	-0.102	-0.103	-0.100	-0.049

In conclusion, on the basis of all the highlighted observations and within the limits of this study, we positively answer our initial research question: it is possible to identify the risk of unfairness in a classification output by detecting the level of (im)balance in the input data.



(a) Mean case of unfairness

(b) Worst case of unfairness

Fig. 6.3 Trends of the *fairness criteria* in response to the *balance measures* with respect to the different mutations, by considering for each mutation (a) the Mean value of *unfairness*, and (b) the maximum value, that is, the Worst case of unfairness corresponding to the most *unfair* output.

6.2.2 Binary case

Before addressing the main research question, we performed a sanity check by observing the behavior of the balance measures as the mutation parameter p varies. Figure 6.4 reports the average values for different balance measures and datasets. We observe an increasing trend of all the balance measures with respect to increasing p, in all training sets and test sets. More in detail, Gini and Shannon indexes have a super-linear increase; Simpson index is closer to a linear trend; finally, IR index has a sub-linear increase until $2/3$ of the course and then it turns to have a slight super-linear increase. This observation confirms the ability of the mutation approach to generate synthetic datasets that spread the whole range of conventional balance measures.

Figure 6.5 reports the variation of the five fairness criteria (Y axis) with respect to the increase of balance measures (X axis). The lines are smoothed regression of the individual mutations. For sake of legibility, we omitted Gini since it is very similar to Shannon. We can observe from the curves that very low levels of balance –roughly in the range $[0\ ,15]$ and up to 50 in a few cases– correspond to higher levels of unfairness. As shown in the preliminary results, the indexes react slightly differently to different levels of balance: as a consequence, the distinct unfairness criteria reflect different levels of balance in a slightly different way. By looking at the single fairness criteria, as well as at the specific trend lines in figure 6.5, we observe that:

- the trend of unfairness with respect to IR is often *not* monotonic: Independence, Separation-TP and Sufficiency-PP, after an initial decreasing phase, they slightly increase within the range $[15\ ,25]$ before stabilizing; Separation-FP slightly increases in the range $[50\ ,100]$ for Student_port; Sufficiency-PN is

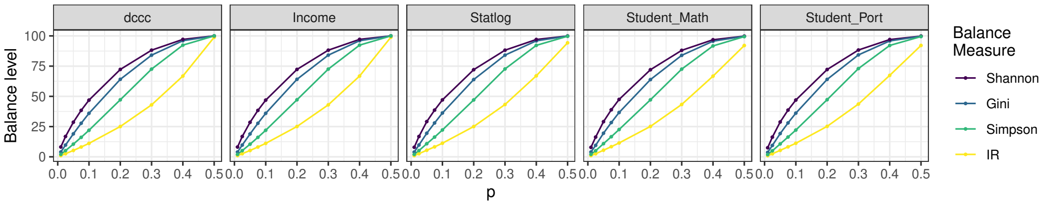


Fig. 6.4 Values of balance measures versus mutation parameter p.

much less regular among datasets, and the correlation between high unfairness and low balance holds only partially;

- modest final surges in correspondence of maximum levels of the balance – around the range [90 , 100] – are observable above all for Separation-FP, Sufficiency-PP and Sufficiency-PN;
- overall, the datasets Dccc and Income have lower levels of unfairness even with an extremely low balance, therefore the correlation high unfairness–low balance is much less pronounced for Separation-TP and Sufficiency-PP, and absent for Indepencence, Separation-FP and Sufficiency-PN;
- in general, Sufficiency-PN presents the most irregular trends especially in the dataset Student_port: it increases within [0 , 20], then it decreases till around 80 and it surges again in the final range; a similar behavior can be observed for Sufficiency-PN in Student_math. However, a follow-up analysis on Sufficiency-PN with respect to p showed that Sufficiency-PN tends to slightly decrease as p increases (that is, as balance increases): the reason for such irregular behavior should be further investigated and we cannot rely on the current results of Sufficiency-PN.

6.3 Limitations

As limitations of our approach, first of all we highlight that the binomial logistic regression used for the classification task assumes linearity between the dependent variable and the independent variables, and limited or no multi-collinearity between independent variables. These requirements were not taken into account and verified in our analyses.

In addition, applying more classification algorithms (each with different parameters) would improve the external validity of the relationship we found between balance and unfairness in the classification output, and would help to identify how the different types of classification algorithms propagate the imbalance from the training set to the output.

Eventually, other types of mutation techniques should be taken into account, for instance by adopting different pre-processing methods to reproduce several distributions of the occurrences between the classes of the protected attributes.

Concerning the mutations of the datasets used, the set of values of parameter p could be enriched with further entries, to track the relationship with unfairness in a more granular way.
Finally, as the choice of the balance measure has a relevant impact on the threshold to consider as risky, a thorough sensitivity analyses on the thresholds to be used should improve the reliability of the findings exposed in this Chapter.

6.4 Research Outcomes

In this Chapter we assess the (im)balance in a given dataset as a potential risk factor for detecting discrimination occurring in subsequent classification tasks, by measuring the level of imbalance of specific protected attributes. We performed the study separately for *multiclass* and *binary* protected attributes.

Multiclass case. Overall, the results reveal that our approach is suitable for the proposed goal, however the choice of the balance measure has a relevant impact on the detection of discriminatory output from ADM systems.

In particular, as regards the analysis of the balance measures in response to mutations, we can confirm the ability of the mutation approach to generate synthetic datasets that spread the whole range of conventional balance measures, which increase as the mutations become increasingly balanced.

Concerning the analysis of the fairness criteria in response to mutations, overall we note that the unfairness measures present very similar distributions with respect to mutations, with higher values of unfairness (indicating unfair classification outcomes) in correspondence of the mutations that presented higher values of imbalance, and vice-versa. However, as the imbalance of the mutations decrease, we observe substantially no difference between the mean values of the unfairness measures. Thus, only on condition of considering the highest extreme values of unfairness for each mutation, the general trend of the unfairness measures seems to decrease (thus indicating an increasingly fair classification) as the mutations tend to be increasingly balanced.

Then, regarding the analysis of the fairness criteria in response to the balance measures, overall the unfairness measures decrease as the balance measures increase.

Indeed, the correlation analysis confirmed that the balance measures are capable of predicting unfairness of software output, with some variation among the balance measures as they assume different behaviors in reflecting the level of (im)balance: particularly, the IR index is highly sensitive to extreme values of balance/imbalance.

Binary case. Also in this case, on the basis of the discussion provided in Section 6.2.2 we positively answer our initial research question. In addition, we can identify tentative thresholds of balance measures and formulate the following practical recommendation:

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

In conclusion, overall the results showed that our approach is suitable for the proposed goal, however the choice of the balance measure has a relevant impact on the threshold to consider as risky.

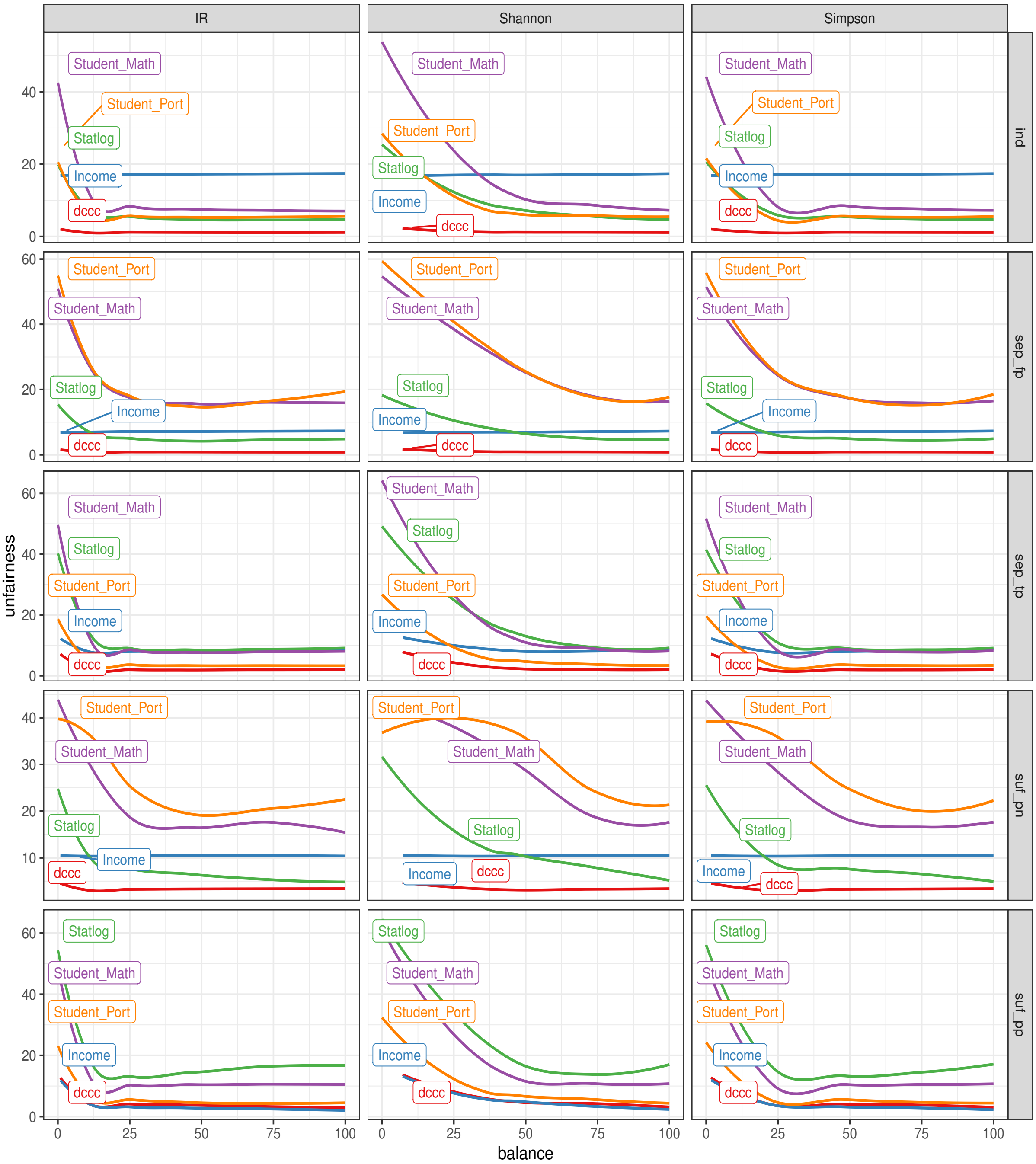


Fig. 6.5 Trends of the *fairness criteria* as a response to the *balance measures*.

Chapter 7

Identifying Imbalance Thresholds in Input Data to Achieve Desired Levels of Algorithmic Fairness

Given that the choice of the balance measure has a relevant impact on the threshold to consider as risky –as concluded in the previous Chapter– herein we focus on the construction of risk thresholds for balance measures in order to achieve desired levels of algorithmic unfairness. Specifically, we formulated the following research question:

> **RQ 4.** Is it possible to identify a threshold s (for balance measures) such that if the *balance* of the training set is greater than s, then the *unfairness* of the classification on the test set is expected to be less than a threshold f?

Since in our previous studies we successfully tested the reliability of the balance measures as risk indicators, in this Chapter we move forward by defining specific risk thresholds for balance measures and for fairness criteria, such that if the balance of the training set is greater than a threshold s, then the unfairness of the classification on the test set is expected to be less than a threshold f. Other novelties of this work are given by the analysis of much more datasets and protected attributes (both binary and multiclass), as well as the application of two mutation techniques simultaneously on different protected attributes of a given dataset; moreover, we adopted four different algorithms to simulate different classification tasks in order to increase the variability of the output and the generalizability of the results.

The findings of this study are published in the conference paper titled "Identifying Imbalance Thresholds in Input Data to Achieve Desired Levels of Algorithmic Fairness" (2022) [45].

7.1 Experimental Design

To answer the above research question, we set up the following procedure that was applied to the binary case and the multiclass case separately:

1. we collected seven different datasets and, for each dataset, we selected both a binary and a multiclass protected attribute;
2. using two specific mutation techniques (one for the binary case and one for the multiclass case) we generated a large number of synthetic datasets with different levels of balance;
3. we assumed four classification algorithms, then, for each algorithm and for each synthetic dataset, we performed a classification with training-test sets randomly split of 70%-30%;
4. we computed the balance of the protected attributes in the training set with the four *balance measures* analyzed so far;
5. we applied three different *unfairness measures* (the Independence, Separation and Sufficiency criteria) to the protected attributes in the test set –that is, to the classifications obtained from the model– for a total of five unfairness measures on each protected attribute;
6. we built the thresholds s (for balance measures) and f (for unfairness measures) using the first collection of data by following the procedure specified below;
7. we generated a new collection of data by repeating steps 2 and 3;
8. using the second collection of data, we assessed and analyzed the performances of the thresholds previously defined through different evaluation metrics.

In Algorithm 4 we present the pseudocode for the measurements of both balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$ for each protected attribute in the selected datasets.

Algorithm 4: Measurements of balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$ for RQ4.

```
Require: categorical protected attributes $A_i$, each with number of categories $m > 0$;
    mutation technique $M = \cdot$ ovun.sample (for binary attributes) or
                              $\cdot$ SmoteClassif (for multiclass attributes);
    exemplar distributions $E = \{Balance, QuasiBalance, OneOff, HalfHigh, Power2\}$;
    mutation parameter $P = \{0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$;
    classifier $C = \{logit, svm, rF, K\text{-}nn\}$;
    balance measures $\mathfrak{B} = \{Gini, Shannon, Simpson, IR\}$;
    unfairness measures $\mathfrak{U} = \{\mathfrak{U}_{Sep\_TP}, \mathfrak{U}_{Sep\_FP}, \mathfrak{U}_{Suf\_PP}, \mathfrak{U}_{Suf\_PN}, \mathfrak{U}_{Ind}\}$.
    Input  : Dataset $D_j$ with $j = 1, ..., 8$
    Output : Balance measures $\mathfrak{B}$, Unfairness measures $\mathfrak{U}$
 1: for all $D_j \in \{D_1, \ldots, D_8\}$ do
 2:     identification of a multiclass protected attribute $A_{m,j}$ and a binary protected
        attribute $A_{b,j}$
 3:     for all $e \in E$: application of the mutation technique SmoteClassif:
        $M_{m,j} \leftarrow M_{SmoteClassif}(A_{m,j})$
 4:     for all $p \in P$: application of the mutation technique ovun.sample:
        $M_{b,j} \leftarrow M_{ovun.sample}(A_{b,j})$
 5:     randomly data splitting of 70%-30% into training-test sets
 6:     for all $c \in C$ do
 7:         prediction of the score variable with classifier $c$
 8:         $\mathfrak{B}_{b,j,c} \leftarrow \mathfrak{B}(M_{b,j}) \in$ training set
 9:         $\mathfrak{B}_{m,j,c} \leftarrow \mathfrak{B}(M_{m,j}) \in$ training set
10:         $\mathfrak{U}_{b,j,c} \leftarrow \mathfrak{U}(M_{b,j}) \in$ test set
11:         $\mathfrak{U}_{m,j,c} \leftarrow \mathfrak{U}(M_{m,j}) \in$ test set
12:     end for
13: end for
14: return $\mathfrak{B}_{b,j,c}, \mathfrak{B}_{m,j,c}, \mathfrak{U}_{b,j,c}, \mathfrak{U}_{m,j,c}$
```

The method for identifying risk thresholds (at step 5 of the procedure above) relies on the first collection of data to identify the thresholds f and s: it is necessary to empirically observe the distribution of the unfairness to understand where the unfairness thresholds could be reasonably placed. We built five different configurations[1] in which f is placed differently relative to the distribution of the unfairness

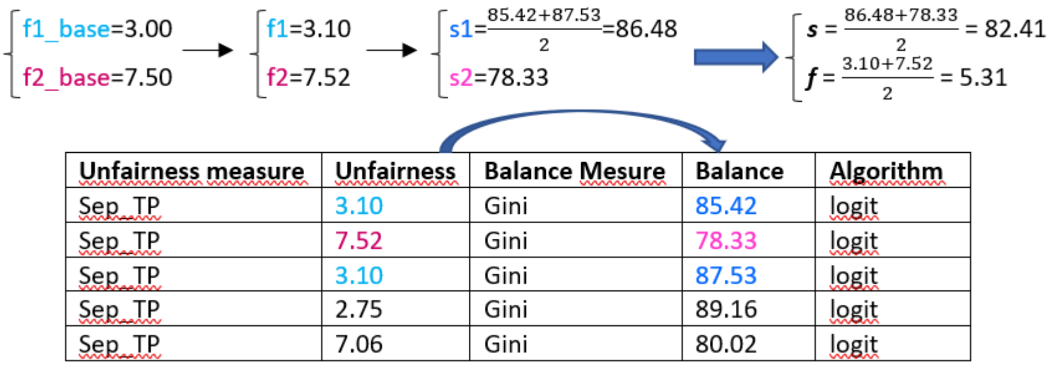
[1] The five configurations with all the specifications on their construction can be found in the Appendix A.1

and, associating to each f the corresponding thresholds of balance s, we got five potential sets of thresholds; among them, we select the one that presented the highest accuracy.

We followed this procedure for each combination of balance measures, unfairness measures, and algorithms, basically filtering the collection of data with respect to these factors. In Figure 7.1 we report a numerical example of this procedure, which is described as follows:

1. we define 2 theoretical values of unfairness thresholds, f1_base and f2_base, which identify the following brackets (where u = unfairness):
   - $u \leq f1\_base \longrightarrow$ low unfairness
   - $f1\_base < u \leq f2\_base \longrightarrow$ medium unfairness
   - $u > f2\_base \longrightarrow$ high unfairness
2. in the first collection of data, we select the values of unfairness that are nearest to f1_base and f2_base, and define them as f1 and f2;
3. as for each unfairness value there exists a corresponding value of balance, and vice versa, we identify the two values of balance corresponding to f1 and f2 –that is, the values in correspondence of f1 and f2 in the data– and define them as s1 and s2. If more than one balance value is found corresponding to f1 or f2, we take their mean (for example, if we find 2 values equal to f1, we can find two different values for the corresponding s1, thus we assume as s1 the mean of the two values);
4. we define the threshold of unfairness f as the mean between f1 and f2, and the threshold of balance s as the mean between s1 and s2.

A possible variation of this procedure consists in defining only one value of the unfairness f_base at step 1, instead of two different values. In this case, there is only one value for f and s at step 2-3, and it is possible to define the two thresholds at step 4 without computing a mean. The reason for these choices is to distribute the values of f evenly in the desired range, which is –based on the initial observation of the distribution of the unfairness– where we observed the highest concentration of unfairness values, approximately between the minimum and the mean of the distribution.



Unfairness measure	Unfairness	Balance Mesure	Balance	Algorithm
Sep_TP	3.10	Gini	85.42	logit
Sep_TP	7.52	Gini	78.33	logit
Sep_TP	3.10	Gini	87.53	logit
Sep_TP	2.75	Gini	89.16	logit
Sep_TP	7.06	Gini	80.02	logit

Fig. 7.1 Numerical example of the procedure for the identification of the thresholds s and f, for the combination Gini-Sep_TP-logit.

Note that for generating each of the two collections of data we varied a `seed` by setting 50 randomly sampled values between 1 and 1000, in order to keep track of both the samples and the mutations for reproducibility purposes, and with a view to increasing the variability –and thus the reliability– of our method.

7.1.1 Datasets

To conduct this study we selected a range of datasets reported in Table 7.1; exhaustive descriptions and specific characteristics of all datasets are provided in Section 2.2.1.

Note that each of them includes a binomial target variable that we predict through different classification tasks: specifically, we trained each classifier on a training set composed of 70% of the original dataset (randomly selected) and we used the remaining 30% as the test set.

To be consistent with our previous works, also in this study we treated empty classes, that is, classes that *exist* (potentially there could be occurrences) but are *not* represented in the dataset, as in our view a dataset that contains no instance of a given class –for example, all males or all whites– is imbalanced with respect to that protected attribute; thus, we included missing values in the analysis by considering them as a separate category “NA”.

Finally, for each selected dataset we identified both a *multiclass* protected attribute (with a number of categories $m > 2$) and a *binary* protected attribute ($m = 2$). For the choice of the multiclass attribute we looked for some variety among the different domains and datasets, while we set “sex” as a binary protected attribute for all the datasets, as it is a highly common source of imbalance and consequent discrimination [19].

Table 7.1 List of the *datasets* with the analyzed *attributes*.

Dataset	**Domain**	**Binary Protected attribute**	**Multiclass Protected attribute**	**m**
Default of credit cards clients (Dccc)	Financial	*sex*	*education*	6
Drug (Cannabis)	Social	*sex*	*ethnicity*	7
Drug (Impulsive)	Social	*sex*	*ethnicity*	7
Heart disease	Welfare	*sex*	*age category*	5
Income	Welfare	*sex*	*race*	5
Statlog	Financial	*sex*	*age category*	5
Student (Mathematics)	Welfare	*sex*	*father's education*	5
Student (Portuguese)	Welfare	*sex*	*father's job*	5

7.1.2 Mutation techniques

We used two specific *pre-processing* methods as mutation techniques, one for multiclass attributes and one for binary attributes, in order to generate synthetic datasets with different levels of balance. In both cases, the generated mutated datasets have the same number of rows as the original ones and, as the mutation technique applies to a single attribute at a time, the distribution of the other variables in the dataset remains unchanged. Particularly, we applied the two methods previously adopted in Chapter 6, which we briefly describe again.

Multiclass case. For multiclass attributes we used the function `SmoteClassif`[2] from the R `UBL-package`. The relevant parameter is *C.perc*, a list containing the percentages of under-sampling or/and over-sampling to apply to each class of the sensitive attribute. We examined five different configurations for this parameter: the default configuration "balance" (namely, the perfect uniform distribution, with all the occurrences equally distributed among the different classes), and four additional configurations corresponding to the four exemplar distributions "Power2", "HalfHigh",

[2] https://www.rdocumentation.org/packages/UBL/versions/0.0.6/topics/SmoteClassif, last visited on June 1, 2023

"OneOff" and "QuasiBalance" already used to test this mutation technique in Chapter 6. For the exhaustive description of the exemplar distribution, we refer to Section 4.1.1 of Chapter 4.

In this study, for each exemplar distribution we considered 4 permutations of the percentages assigned to the different classes. For example, in the *One Off* configuration the four different permutations have each a different class with zero occurrences.
Moreover, in order to increase variability and reliability of our results, we varied a `seed` (an integer recommended for reproducibility purposes to keep track of the samples) by setting 50 randomly sampled values between 1 and 1000.
If we multiply by 8 datasets, we obtain $8 \times 50 \times (4 \times 4 + 1) = 6800$ synthetic datasets. Considering that each dataset is processed by 4 different algorithms, we reach a total of $6800 \times 4 = 27200$ classifications.

Binary case. For binary attributes we applied the function `ovun.sample`[3] from the R `ROSE-package`. The relevant parameter of this mutation is p, which determines the probability of resampling from the minority class. We set 9 values for p, ranging from 0.01 (high imbalance) to 0.5 (perfect balance):
p = {0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5}.
If we multiply by 8 datasets and 50 seeds, we obtain $8 \times 50 \times 9 = 3600$ synthetic datasets. Since each dataset is processed by 4 different algorithms, we have a total of $3600 \times 4 = 14400$ classifications.

7.1.3 Algorithms

In our analysis, we adopted four different algorithms in order to simulate different classification tasks. We highlight that we searched for possible significant differences when establishing the thresholds with respect to the different algorithms in order to generalize our findings, but we were not interested in the specific performance of each algorithm; for this reason, we did not perform hyper-parameters tuning and we kept the default parameters. Specifically, we selected the following algorithms:

[3] https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ovun.sample, last visited on June 1, 2023

- *logistic regression* (logit): function *glm*, with argument `family`=binomial(link="logit"), from the package `stat` [4];
- *support vector machine* (svm): function *svm* from the package `e1071` [5];
- *random forest* (rF): function *randomForest* from the package `randomForest` [6];
- *K-nearest neighbors* (K-nn): function *knn* from the package `class` [7].

7.1.4 Evaluation Metrics

We assumed different evaluation metrics to assess the reliability of the thresholds. First, we remind that the first collection of data has been used to build the thresholds, whereas the second has been used to evaluate them: in simple words, we evaluate whether a classification (obtained with the second collection of data) respects or not the conditions on balance and unfairness measures defined through the first collection. Given the two thresholds s and f (for balance measures and unfairness measures respectively), when the balance of the training set is over s, we expect the unfairness of the classification to be under f; if this happens, we have a positive instance, otherwise we have a negative instance. Hence, we define the following instances related to the confusion matrix in Figure 7.2:

- *if* balance$< s$ & unfairness$> f \longrightarrow$ True Positive (TP)
- *if* balance$< s$ & unfairness$< f \longrightarrow$ False Positive (FP)
- *if* balance$> s$ & unfairness$< f \longrightarrow$ True Negative (TN)
- *if* balance$> s$ & unfairness$> f \longrightarrow$ False Negative (FN)

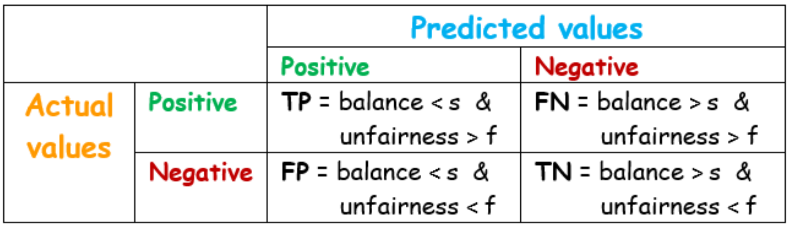
In particular, we adopted the five *evaluation metrics* reported in Table 7.2, whose values range in the interval [0, 1]: Accuracy evaluates the percentage of correctly classified values, while Precision (also called Positive Predictive Value) represents the fraction of positive instances correctly identified with respect to all the positive

[4] https://www.rdocumentation.org/packages/stats/versions/3.6.2, last visited on June 1, 2023

[5] https://www.rdocumentation.org/packages/e1071/versions/1.7-13, last visited on June 1, 2023

[6] https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.1, last visited on June 1, 2023

[7] https://www.rdocumentation.org/packages/class/versions/7.3-22, last visited on June 1, 2023



		Predicted values	
		Positive	Negative
Actual values	Positive	TP = balance < s & unfairness > f	FN = balance > s & unfairness > f
	Negative	FP = balance < s & unfairness < f	TN = balance > s & unfairness < f

Fig. 7.2 Confusion matrix of the statistical classification based on the different levels of balance/unfairness.

predicted instances; Sensitivity, also called Recall,indicates how many positive instances are correctly detected (TP) among those that actually present the condition; instead, Specificity represents how many negative instances are correctly identified (TN) among all those that do not present the condition; finally, F1-score is the harmonic mean of precision and sensitivity.

Table 7.2 The *evaluation metrics* with the respective formula.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
F1-score	$\frac{2TP}{2TP+FP+FN}$

7.2 Results and Discussion

Before addressing the research question of this study, we report a preliminary analysis of the correlation between fairness criteria and balance measures; after that, we show the overall results for thresholds and evaluation metrics, and finally we assess the goodness of our findings by aggregating with respect to balance measures, fairness criteria and algorithms, in order to better understand how the different factors affect the goodness of the outcomes.

a) Analysis of the correlation between balance measures and fairness criteria

First of all, we assessed the correlation between balance and unfairness measures in order to verify whether the negative correlation holds (the higher the balance, the lower the unfairness). Indeed, compared to our previous works [41][43][44] reported in Chapters 5 and 6, in this analysis we introduced much more datasets and algorithms to classify data, so as to increasingly thoroughly assess the balance measures as risk indicators.

To understand the values, we remind from the findings of our previous studies that we expect the correlation between balance measures and fairness criteria to be negative, as we expect the balance measures to be high (indicating low imbalance) if the unfairness values are low (that is, higher fairness). Thus, the stronger the negative correlation, the stronger the relationship between balance and unfairness measures.

Table 7.3 Correlation between balance measures and fairness criteria for the *binary* case.

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**IR**
Independence	0.008	0.008	0.008	0.006
Separation – TP	-0.393	-0.396	-0.379	-0.341
Separation – FP	-0.073	-0.074	-0.071	-0.063
Sufficiency – PP	-0.400	-0.407	-0.382	-0.345
Sufficiency – PN	-0.115	-0.116	-0.110	-0.097

Table 7.4 Correlation between balance measures and fairness criteria for the *multiclass* case.

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**IR**
Independence	0.133	0.111	0.100	0.081
Separation – TP	-0.028	-0.022	-0.032	-0.008
Separation – FP	0.032	0.014	0.012	-0.011
Sufficiency – PP	-0.039	-0.036	-0.045	-0.016
Sufficiency – PN	-0.019	-0.032	-0.034	-0.017

As we can observe from Table 7.3 and 7.4, most of the balance measures present a moderate or low negative correlation with the fairness criteria, above all in the case of binary attributes, meaning that the higher the indexes of balance, the lower the unfairness measures. Note that the computations reveal that such values are all significant, with a p-value$<$0.05.

More in detail, for the **binary** case we observe correlation values between -0.063 and -0.407 for all the fairness criteria except for the Independence criterion, which presents no correlation (around 0.008) indicating that this criterion is the most difficult to detect; whereas the Sep_TP and the Suf_PP criteria present the stronger negative correlation values, between -0.341 and -0.407.

Specifically for **multiclass** attributes, instead, we note a weak negative correlation in correspondence of the Sep_TP, Suf_PP and Suf_PN criteria, between -0.008 and -0.045, and a weak positive correlation in the range 0.012–0.133 for the Independence and the Sep_FP criteria, meaning that overall the level of unfairness in the case of multiclass attributes is more difficult to detect.

In general, we observe that the balance measures respond very similarly to the different fairness criteria, therefore we deduce that the negative correlation depends mostly on the unfairness measures, rather than on the specific balance measure.

b) Assessment of the thresholds through evaluation metrics

To define the thresholds s (for balance measures) and f (for fairness criteria), for each combination of balance-unfairness-algorithm we selected the configuration that has the highest accuracy. We chose the accuracy as a discriminant for the identification of the thresholds s and f because it showed the smallest interquartile range (or IQR, which graphically corresponds to the height of the box) indicating that the accuracy index is the one with the lowest variability among the selected evaluation metrics (see Figure 7.3 and Figure 7.4). The complete results are reported in Appendix A.2 in separate Tables for the binary and the multiclass cases, ordered by balance measure (Gini, Shannon, Simpson, and IR indexes), for each combination of balance-unfairness-algorithm. For sake of legibility, we report values for the thresholds of both fairness criteria and balance measures multiplied by 100, that is on a scale $[0, 100]$. Hereinafter, we show the aggregated and overall results for thresholds and evaluation metrics. We remind that the aim of this study is to define two thresholds s (for balance measures) and f (for unfairness measures) such that if the balance of the training set is greater than s, then the unfairness of the classification on the test set is expected to be less than *f*. As before, we examine results separately for binary and multiclass attributes.

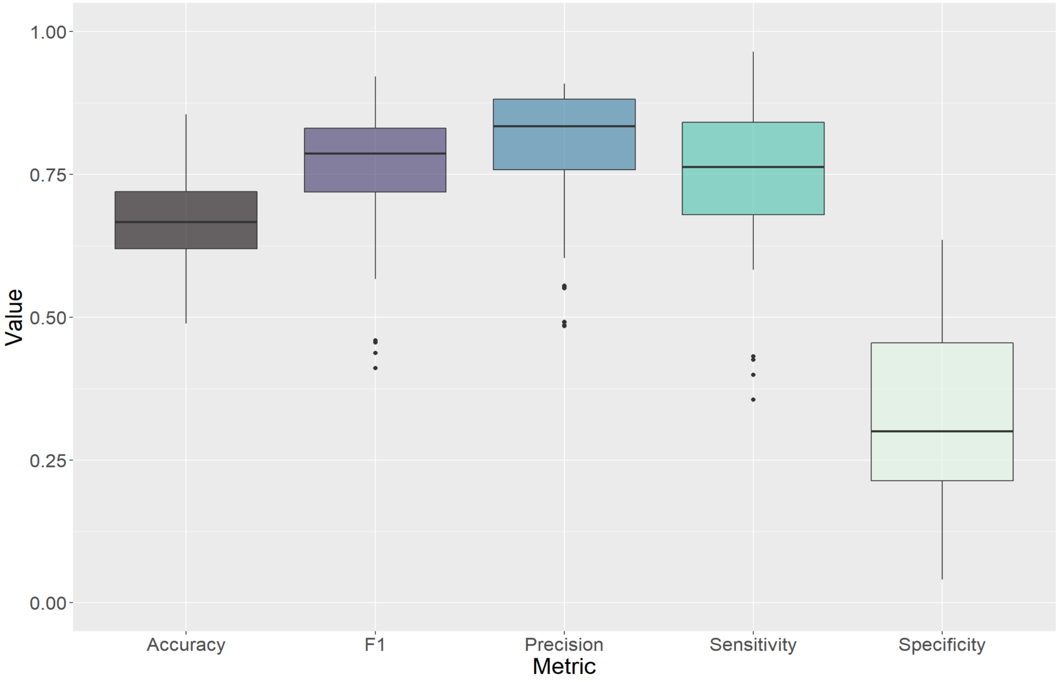


Fig. 7.3 Boxplot of the evaluation metrics in the binary case.

Binary case. Overall the thresholds assume values close to the extremes of the range, with the thresholds for the fairness criteria being between 0 and 10, and the thresholds for the balance measures being between 80 and 100 except for the IR index, which presents lower balance threshold values, around 60 (we can retrieve such data from Appendix A.2). Looking at Figure 7.3, we observe that the Accuracy is on average 0.7, but among all the evaluation metrics the Precision index is the one that presents the highest values, around 0.85 on average, indicating a high fraction of positive instances correctly identified with respect to all the positive predicted instances. Instead, the Sensitivity –or Recall– is on average around 0.75, meaning that the number of instances misclassified as negatives (FN) is higher than the number of instances misclassified as positives (FP); in terms of thresholds, it means that the number of instances in which values of balance are over s (indicating high balance) and the unfairness is over f (indicating high unfairness) differently from the expectation to find low unfairness, is *higher* than the number of instances in which values of balance are under s (indicating low balance) and the unfairness

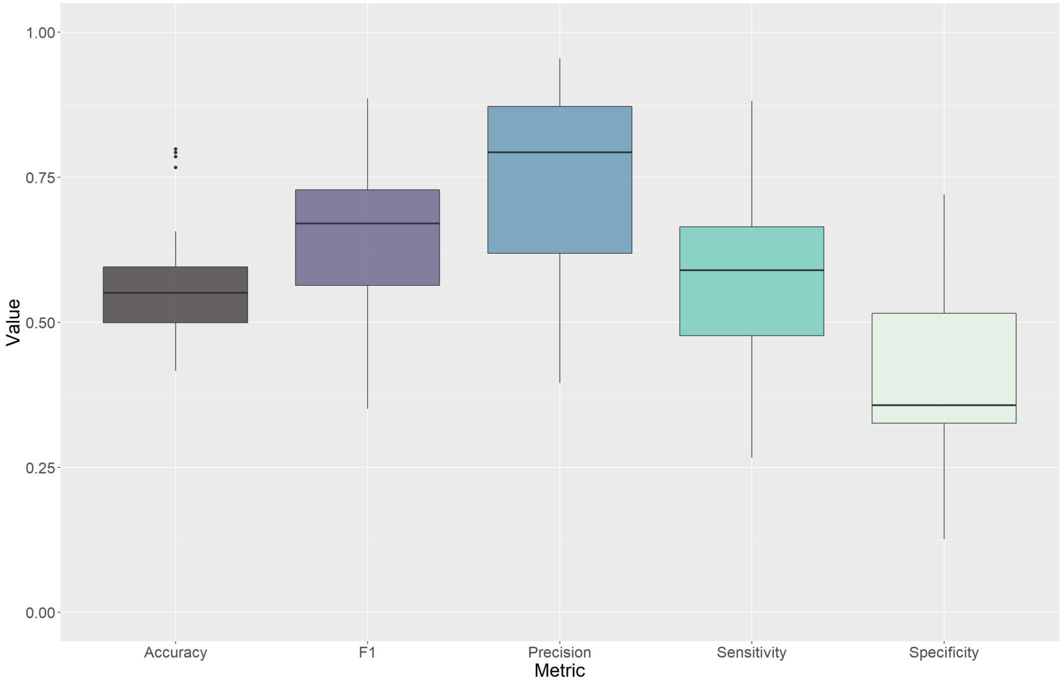


Fig. 7.4 Boxplot of the evaluation metrics in the multiclass case.

is under f (indicating low unfairness) differently from the expectation to find high unfairness. Considering the F1-score, which represents the harmonic mean between Precision and Sensitivity, it assumes values around 0.8 on average. Finally, the worst performances are identified through the Specificity index, with significantly lower values (around 0.3 on average) with respect to the other indexes, indicating that the number of true negative instances is very small with respect to the number of false positives, that is, when evaluating the thresholds we obtain a high number of instances in which values of balance are under s (indicating low balance) but the unfairness is under f (indicating low unfairness) differently from the expectation to find high unfairness.

Multiclass case. In Appendix A.2 we can observe that overall the thresholds for the balance measures are between 70 and 100 except for the IR index, which presents much lower balance threshold values of around 30, while the thresholds for the fairness criteria are in the range 0 and 15. Looking at Figure 7.4 we note that overall the evaluation metrics assume lower values with respect to the binary case: Accuracy decrease to around 0.55, Precision is around 0.8 and Sensitivity is around 0.6 on average, with F1-score around 0.7; on the contrary, Specificity slightly increase to around 0.35 on average. Thus, according to the evaluation metrics taken into account, the identified thresholds perform better in the binary case than in the multiclass case.

Overall, we deduce that the thresholds are responsive to risk, that is, values of balance under s indicate levels of unfairness over f, but they even tend to overestimate the risk (as we can infer from the low Specificity caused by the high number of false positives).

c) Assessments of the thresholds' goodness with respect to balance measures, fairness criteria and algorithms

As we defined the thresholds for each combination of balance-unfairness-algorithm, hereinafter we assess the thresholds' accuracy with respect to the different balance measures, fairness criteria and algorithms involved in the study, in order to understand how and to which extent each factor affects the performances of the thresholds.

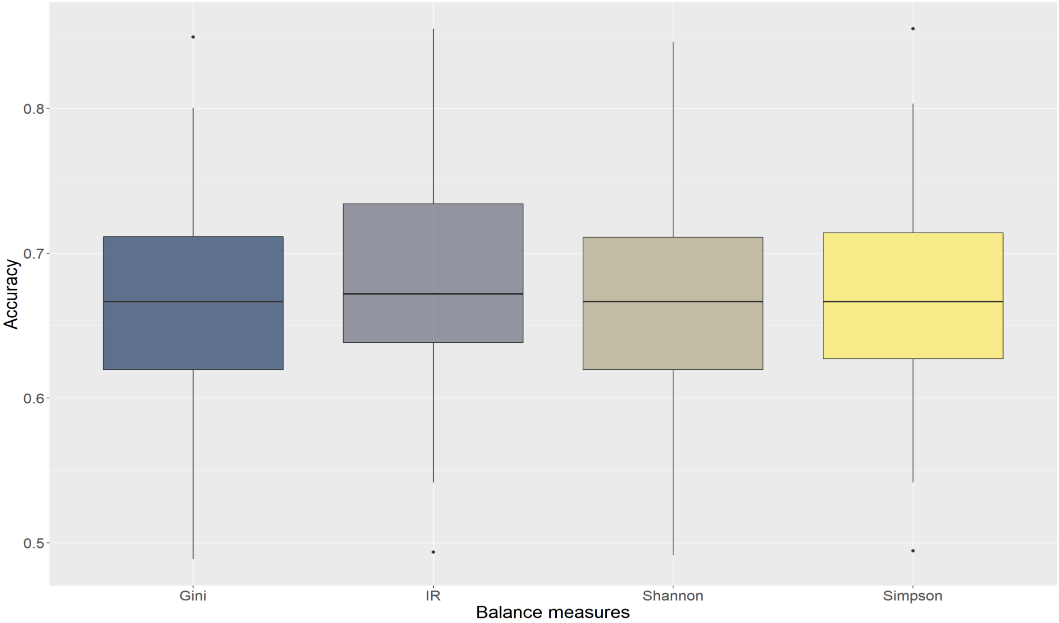


Fig. 7.5 Boxplot of the thresholds' Accuracy with respect to the *balance measures* in the binary case.

Binary case. Looking at Figure 7.5 we observe that the thresholds' accuracy with respect to the four *balance measures* is around 0.67 overall, with the IR index

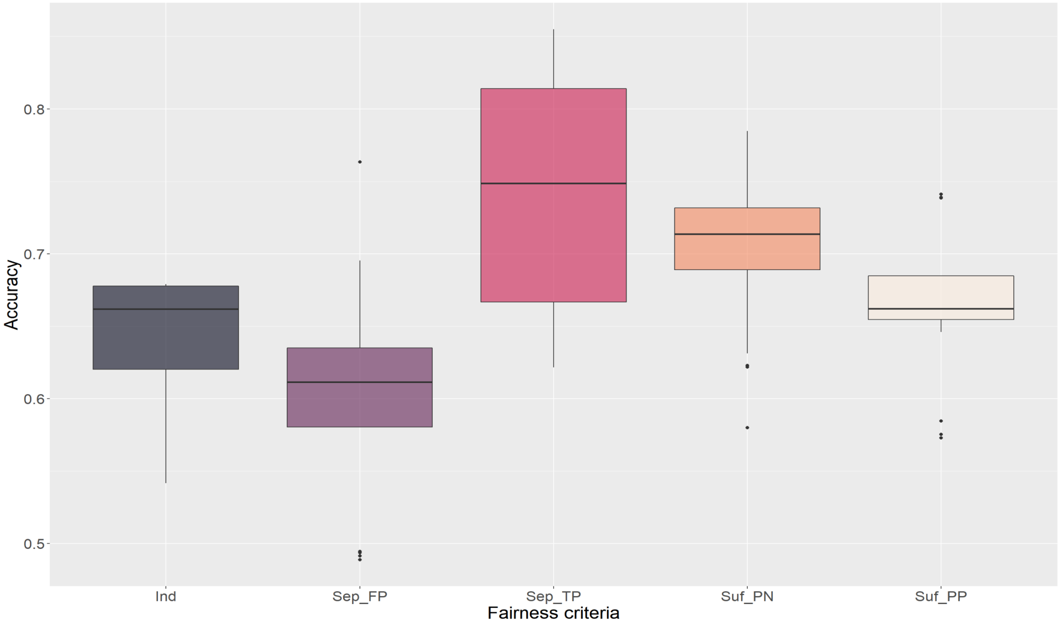


Fig. 7.6 Boxplot of the thresholds' Accuracy with respect to the *fairness criteria* in the binary case.

slightly higher than the other measures, but with no significant differences between the indexes.

Conversely, from Figure 7.6 we note that the thresholds perform very differently with respect to the different *fairness criteria*: particularly, the Sep_TP criterion presents the highest values of thresholds' accuracy, around 0.75 on average, with the largest interquartile range (or IQR, which graphically corresponds to the height of the box) indicating that the Sep_TP measure is the one with the largest variability of the accuracy values; then we find the two Sufficiency conditions, with thresholds' accuracy respectively around 0.72 for Suf_PN and 0.66 for Suf_PP, and the Independence criterion, always with an accuracy around 0.66 on average; the lowest values of the thresholds' accuracy are found in correspondence of Sep_FP, around 0.62.

Looking at Figure 7.7 on the thresholds' accuracy with respect to the *algorithms*, we note that the best performances are reached with the K-nn classifier, with accuracy values around 0.70 on average; the logit and the svm algorithms present similar accuracy values around 0.67 on average, while the worst performances of the thresholds correspond to the random forest classifier with values around 0.67 on average; the random forest also presents the largest variability, with the highest values close to the ones of K-nn.

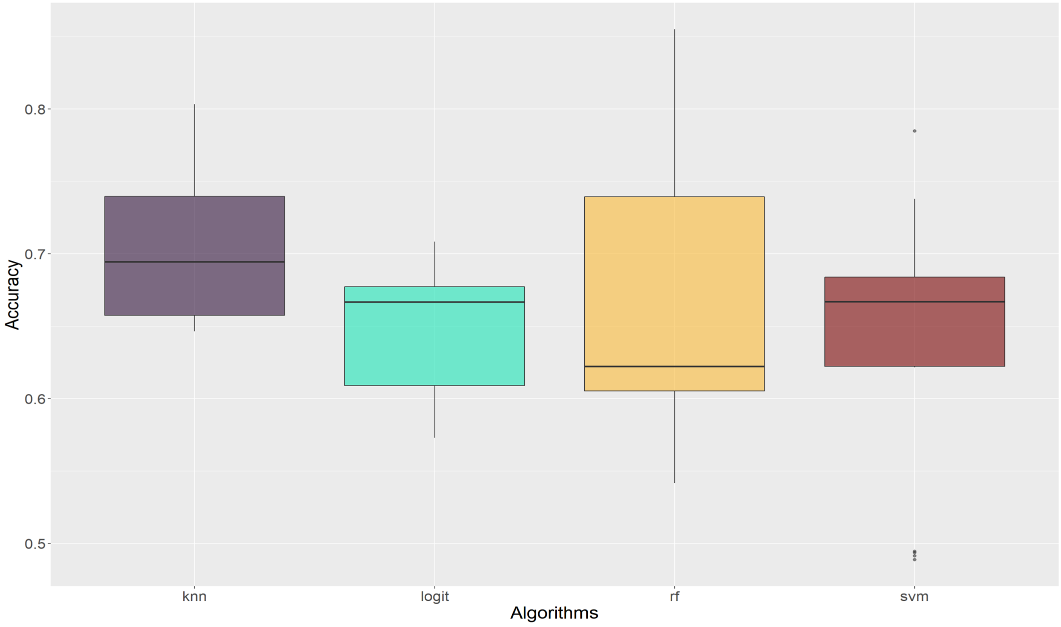


Fig. 7.7 Boxplot of the thresholds' Accuracy with respect to the *algorithms* in the binary case.

Multiclass case. From Figure 7.8 we note that the thresholds perform differently with respect to the different *balance measures* –contrary to the binary case–. Specifically, the IR index presents the highest accuracy values, around 0.63 on average, while the values decrease to around 0.55, 0.54 and 0.5 for the Shannon, Simpson and Gini indexes respectively.

Then, looking at Figure 7.9 on the thresholds' accuracy with respect to the *fairness criteria*, we observe the same pattern as in the binary case, but with lower values and greater variability for all the measures: the highest accuracy values are in correspondence of the Sep_TP condition and decrease to around 0.6 on average, followed by Suf_PN and Suf_PP (around 0.57), and by Independence and Sep_FP around 0.52 on average.

Finally, about the thresholds' accuracy with respect to the *algorithms* represented in Figure 7.10, we note a completely different pattern with respect to the binary case: the accuracy in correspondence of the random forest remains stable at around 0.6 with a wide variability and it presents the best accuracy value among the other algorithms (contrary to the binary case); indeed, the thresholds' accuracy decrease on average to around 0.58, 0.54 and 0.51 respectively for the svm, K-nn and the logit classifiers.

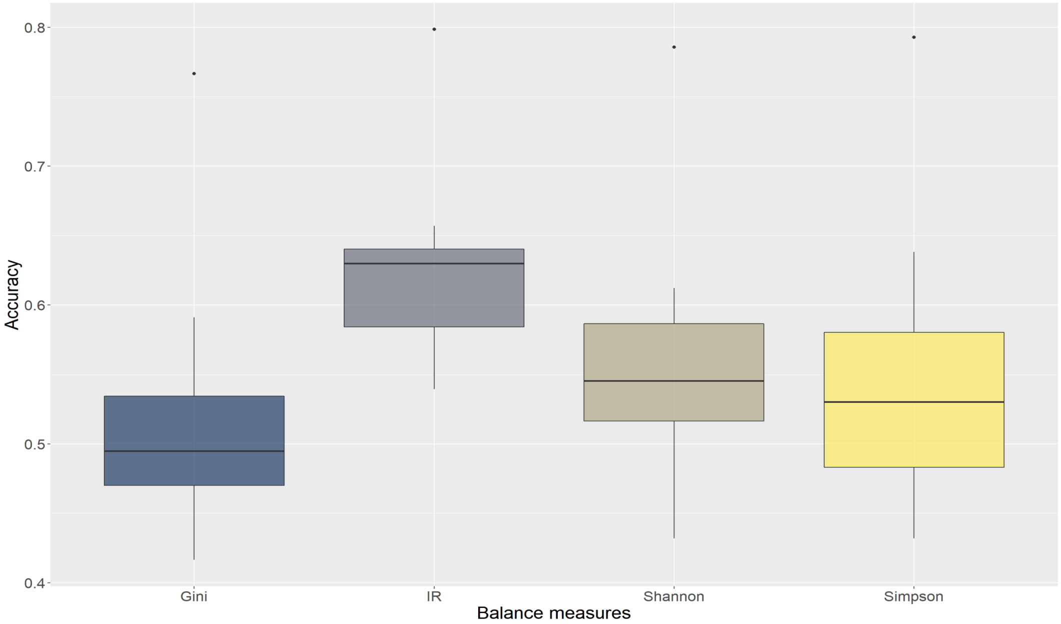


Fig. 7.8 Boxplot of the thresholds' Accuracy with respect to the *balance measures* in the multiclass case.

7.3 Limitations

As limitations of this study, first of all we highlight that we did not perform the hyper-parameters tuning of the algorithms involved in our study as we were not interested in a specific algorithms performance analysis, but rather in varying the classifier in order to increase the variability of the output and the generalizability of the results; nevertheless, a better fitting of the data could reveal more meaningful differences among the different algorithms.

We also remark that we chose the accuracy as a discriminant for the identification of the thresholds, but an analogous study can be conducted by considering a different evaluation metric as a reference and identifying the thresholds according to the performances based on such metric.

Finally, other kinds of mutation techniques could be considered by adopting different pre-processing methods in order to extend the variability and reliability of our results.

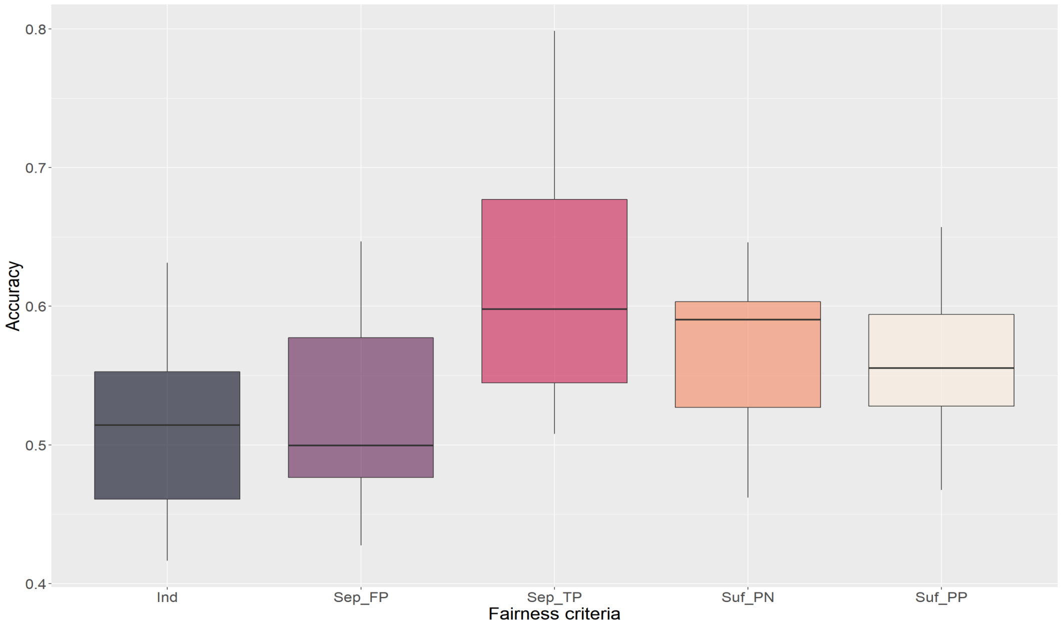


Fig. 7.9 Boxplot of the thresholds' Accuracy with respect to the *fairness criteria* in the multiclass case.

7.4 Research Outcomes

In this study we defined and tested a methodology to identify thresholds of balance such that the unfairness of the classification is expected to be less than a certain desired level. To conduct this analysis, we adopted a previously defined metric-based approach to assess imbalance in a given dataset as a risk indicator of discriminatory classification outcomes of automated decision-making systems[41], and we built the thresholds s and f –for *balance* and *unfairness* measures respectively– by following a specific procedure, separately for *binary* and *multiclass* attributes. Specifically, for each combination of balance-unfairness-algorithm we selected the configuration of thresholds that presented the highest accuracy.

As regards the initial analysis of the correlation between balance and unfairness measures, overall we observed that the balance measures respond very similarly to the different fairness criteria, thus we deduce that the negative correlation depends mostly on the unfairness measures, rather than on the specific balance measure; indeed, given a certain fairness criterion, all the balance measures behave similarly with respect to such criterion.

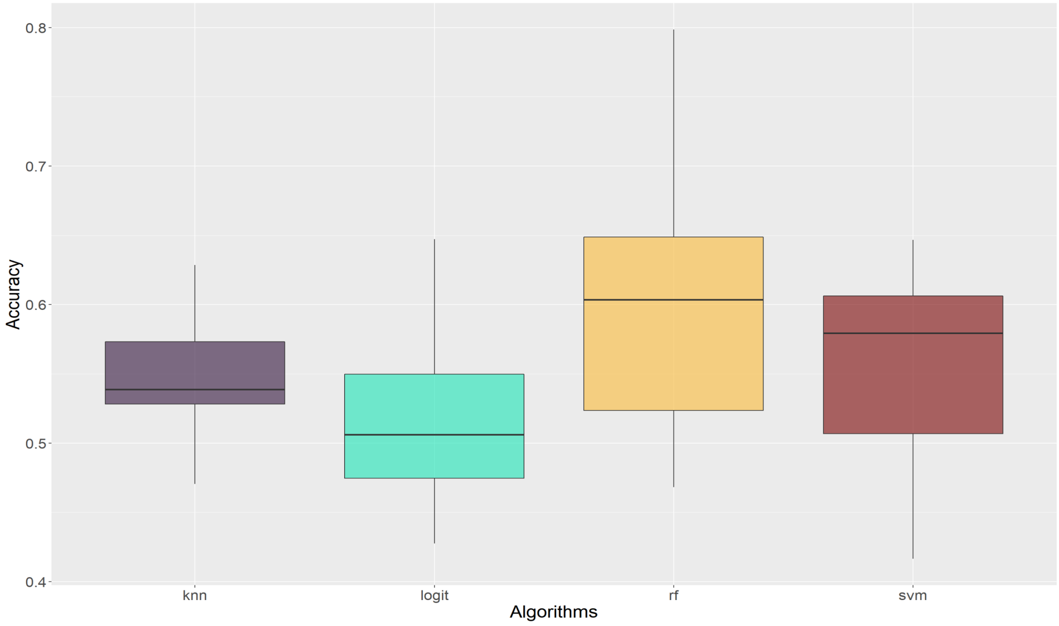


Fig. 7.10 Boxplot of the thresholds' Accuracy with respect to the *algorithms* in the multiclass case.

Then, the assessment of the thresholds through the different evaluation metrics revealed that the values of balance under *s* indicate levels of unfairness over *f*, but they even tend to overestimate the risk. In both the binary and the multiclass cases, the Precision index –which indicates the Positive Predictive value– is the one that presents the highest values among all the evaluation metrics, while the worst performances are given by the Specificity index (due to a high number of false positives), suggesting that the thresholds tend to overestimate the risk. Overall, we also noted that the identified thresholds perform better in the binary case than in the multiclass case.

Lastly, also regarding the evaluation of the thresholds' accuracy with respect to balance-unfairness-algorithm, the thresholds perform better in the binary case than in the multiclass case, with higher accuracy values overall.
Specifically, the *balance measures* seem to have a slight impact only in the multiclass case; indeed, overall there is no significant difference between the different indexes of balance, except for the IR index, which presents the highest accuracy values in both the binary and the multiclass cases.
Conversely, the thresholds perform very differently with respect to the different *fairness criteria*, and overall, for the multiclass case we observe the same pattern as

in the binary case, but with lower values and greater variability for all the fairness criteria.

For the thresholds' accuracy with respect to the different *algorithms*, we found completely different values between the binary and the multiclass case, thus the algorithms have an impact on the performances of the thresholds, but in this study we could not identify a specific pattern.

Chapter 8

Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications

Briefly summarizing, our previous studies tested the reliability of the balance measures as risk indicators only when applied to single protected attributes: in Chapters 4 and 5 we first tested the measures on a few hypothetical exemplar distributions and then on a variety of protected attributes selected from datasets belonging to several application domains of automated decision-making systems [41]. After that, in Chapter 6 we conducted more exhaustive analyses by applying mutation techniques to generate a number of derived synthetic datasets having different levels of balance, in one case to multiclass attributes [43] and in the other case to binary attributes [44], while in Chapter 7 we identified imbalance thresholds to foresee unfair classification outcomes (separately for binary and multiclass protected attributes) [45].
The novelties of the study reported in this Chapter are given by:

i) the integration of the concept of intersectionality among the classes of two or more single protected attributes;

ii) the impact of an imbalanced distribution of the target variable as well as the contribution of the target variable to the unfairness detection.

Particularly, we put forward the research questions below.

RQ 5. Is it possible to identify the risk of biased output by detecting the level of (im)balance in intersectional protected attributes?

Indeed, intersectional classes play a crucial role in understanding the risks of discrimination and inequalities that are even exacerbated in correspondence of the intersection of certain social identities. Therefore, we believe that it is fundamental to analyze the nature of intersectional classes; to this end, we formulated two more specific research questions:

RQ 5.1. *How do intersectional attributes relate to the corresponding primary attributes, in terms of balance and fairness?*

It is of paramount importance to understand to what extent the imbalance of the primary attributes (binary or multiclass) affects the imbalance of the intersectional attribute, as well as how the fairness with respect to an intersectional attribute is linked to the fairness with respect to the primary attributes.

RQ 5.2. *Can the measure of balance on intersectional attributes detect unfairness risks?*

From the studies conducted to answer the previous research questions, there is evidence that working at the level of protected primary attributes, the balance of their classes can reveal the risk of classification unfairness with respect to such attributes; our goal here is to understand whether this capability also extends to intersectional attributes.

Finally, in this Chapter we investigate the contribution of the target variable to the unfairness detection: specifically, the imbalanced distribution of target classes can be taken into consideration by looking at their combination with protected attributes (both primary and intersectional) and assessing whether the measurement of the balance of the combined attribute can detect the risk of unfair classification. In particular, we aim at answering the following research question:

RQ 6. Does the combination of the target variable with protected attributes improve the detection of unfair classification risks?

Note that in the following we call *combined attributes* those attributes given by the combination of the target variable with protected attributes (primary or intersectional).

The work carried out to analyze the research questions formulated above with all the related results is reported in the journal article entitled "Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications" (2023) [46].

8.1 Background

As stated above, our previous studies showed that lower levels of balance in protected attributes are related to higher levels of unfairness in the output [41, 43–45]. In this work, we investigate other two relevant aspects for the assessment of the balance measures as risk indicators of systematic discrimination.
First, the intersectionality among the classes of protected attributes, which is of paramount importance since social identities and inequality are interdependent for groups –for instance, *black women*– and not mutually exclusive [85].
Second, the contribution of the target variable to the unfairness detection, which is recognized as a challenge in a variety of domains, for example, fraud detection, network intrusion detection, medical diagnostics, and a number of other fields [86]. Indeed, positively labeled instances are often lower than negatively labeled instances, but the former are associated with the most significant events for end users (for example, a fraud).

To our knowledge, none of the current approaches to intersectionality and to the target variable combined with protected attributes (and to their effects on classifications) use specific metrics for measuring data (im)balance. This integration and related analysis represent the main contribution to the state of the art.

8.2 Related Work

Intersectionality was introduced in the late 80s in the Black Feminist literature in relation to the intersection of gender and race [87] and it has been successively extended to embrace other traits such as disability status, socioeconomic class, sexual orientation, etc. The concept has recently appeared in the context of fairness and machine learning, related to issues of intersectional discrimination in different domains: Buolamwini and Gebru (2018) studied the impact of the intersection of gender and skin color on computer vision performance [36]; Holman et al. (2020) explored intersectionality in the medical field [88]; whereas Subramanian et al. (2021) advocated for the use of intersectional groups in the validation of NLP models to better intercepts the social and cultural biases reflected in the corpus of training data [89]; other works present attempts of introducing intersectionality in fairness measures [90] and in causal models [91]. However, up to our knowledge, none of these studies and others in the AI/ML fairness literature constructed and applied synthetic measures of (im)balance to intersectional protected attributes.

Concerning the imbalance of the target variable, a comprehensive survey has been conducted by Branco et al. (2016), who collected existing techniques for handling the problem for both classification and regression tasks [86]. The same authors examined more in-depth the context of regression tasks [92], where the target variable is continuous: they presented three new pre-processing approaches to tackle the problem of forecasting rare values of a continuous target variable. Other works concern the mitigation of the imbalance issue of the target variable, and they have been developed with the aim of improving the predictive accuracy of rare cases in forecasting tasks through the adoption of different resampling methods –for instance, see [93, 94].

The closest work to ours is the one by Thabtah et al. (2020) [95], who studied the impact of varying class imbalance ratios on classifier accuracy: they identify nine different imbalance ratios (from 10%:90% to 90%:10%, with steps of 10% increase/decrease) and compute their effect on standard measures of classifier performance (error rate, predictive accuracy, recall and precision). Thus, they focus on the nature of the relationship between the degree of class imbalance and the corresponding classifier performance, but they neither use specific and synthetic measures of balance nor consider multilevel attributes. The same consideration applies to the other studies mentioned above.

8.3 Experimental Design

With the goal of investigating the research questions outlined above, we generated a number of synthetic datasets by aggregating sensitive attributes (also with the target variable) and mutating the distributions of the occurrences between their classes. Then, we evaluated data (im)balance through the four balance measures analyzed so far, as well as the fairness occurring in classification outcomes by means of three fairness criteria. After that, we assessed whether the balance/unfairness of intersectional attributes can be inferred from the balance/unfairness of the primary attributes, and finally, whether the combination of a protected attribute with the target variable improves identifying the unfairness.

First of all, we define an intersectional attribute as a multiclass attribute whose classes are given by the combination –in all the possible ways– of the classes of (single) primary attributes that can be either binary or multiclass. Similarly to the definition of imbalance already stated in Chapter 1, intersectionality is between-attributes when only two attributes are taken into consideration, or multiattribute when the intersectionality involves multiple attributes. In this work, we will explore the concept of multi attributes intersectionality in greater detail.

Secondly, as regards the analysis of the target variable, in general data are imbalanced with respect to the target if at least one of its values has a significantly smaller number of instances when compared to the other values. In this study we focus on binomial target variables, namely, target variables that can assume two possible values, positive (equal to 1) or negative (associated with 0).

Specifically, we set up the following procedure:

1. we chose a sizable *dataset* (as described in Section 8.3.1) that includes two *protected attributes*: the multiclass attribute "education" with cardinality *m*=4, and the binary attribute "sex" (with *m*=2);
2. we generated several derived synthetic datasets with different levels of balance by means of two suitable *mutation techniques*: specifically, we adopted two processing methods, one specific for multiclass attributes and one for binary attributes; we adjusted the parameters of the two methods to alter the distribution of occurrences among the classes –and consequently the balance– of the two protected attributes under analysis (see Section 8.3.2);

Table 8.1 Balance measurements with the respective unfairness measurements for each protected attribute.

Balance measurement	**Unfairness measurement**
$\mathfrak{B}$(sex)	$\mathfrak{U}$(sex)
$\mathfrak{B}$(education)	$\mathfrak{U}$(education)
$\mathfrak{B}$(sex_education)	$\mathfrak{U}$(sex_education)
$\mathfrak{B}$(sex_target)	$\mathfrak{U}$(sex)
$\mathfrak{B}$(education_target)	$\mathfrak{U}$(education)
$\mathfrak{B}$(sex_education_target)	$\mathfrak{U}$(sex_education)

3. we aggregate the two primary protected attributes in one intersectional attribute “sex_education” by combining the classes in all the possible ways, thus creating an intersectional multiclass attribute of cardinality m equal to 8 (= 2 “sex” × 4 “education”); likewise we aggregate the three previous attributes with the target variable, obtaining three combined multiclass attributes, namely “sex_target” (m=4), “education_target” (m=8) and “sex_education_target” (m=16);

4. we used four different *balance measures* $\mathfrak{B}$ (described in Section 2.2.2) to compute the level of (im)balance of both the primary protected attributes and the intersectional attribute in the training set;

5. we built a *binomial logistic regression* model in order to forecast the *score variable* for each synthetic dataset: we trained a binary classifier on a training set composed of the 70% (chosen randomly) of the data, and then tested it on the remaining 30% (which represents the test set);

6. we applied three different*fairness criteria* $\mathfrak{U}$ (see Section 2.2.3) to both the primary protected attributes and the intersectional attribute in the test set –that is, to the classifications obtained from the model– for a total of five distinct unfairness measures, following the pattern described in Table 8.1; note that for the protected attributes combined with the target variable we compute the unfairness on the corresponding protected attribute without target in the test set;

7. we analyzed the collected results in order to answer the research questions.

In Algorithm 5 we present the pseudocode for the generation of the synthetic attributes –first mutated, and subsequently intersected and/or combined with the target variable– and for measuring the balance $\mathfrak{B}$ and the unfairness $\mathfrak{U}$ of each protected attribute under analysis.

8.3.1 Dataset

We selected a dataset from the financial services context, as it is one of the most considerable application domains of ADM systems: **Default of Credit Card Clients** (Dccc), whose main properties are summarized in Table 2.1, while a complete description can be found in Section 2.2.1.
In particular, we took into account two protected attributes: the first one is "education", which is composed of six classes in the original dataset, but two of the classes –*NA* and *unknown*– do not represent an actual category of individuals, therefore we exclude such unknown and missing values (NA) from the analysis; thus, the resulting dataset is composed of 29655 rows, where the classes of the protected attribute "education" are composed as follows: 10585 *graduate school*, 14030 *university*, 4917 *high school* and 123 *others*.
The second protected attribute is the binary attribute "sex", which is composed of 11760 instances of the class *male* and 17895 instances of the class *female*.

In addition, note that this dataset does *not* contain a pre-computed classification, thus we implemented a *binomial logistic regression* model in order to foresee the *score* variable: specifically, we trained a binary classifier on a *training* set represented by the 70% (randomly selected) of the original dataset and we ran it on the *test* set, composed of the remaining 30% of the data. Finally, we highlight again that we choose to keep out missing values (NA) from the analysis as we were interested in examining existing intersectional classes of protected attributes.

8.3.2 Mutation Techniques

Two distinct *pre-processing* methods were employed as mutation techniques, one for multiclass attributes and one for binary attributes, in order to generate multiple variations of the distribution of the occurrences between the classes of the protected attributes taken into account. Specifically, we applied the two methods previously adopted both in Chapter 6 and in Chapter 7, which we briefly describe in the following.

Multiclass case. To mutate the classes of the multiclass protected attribute "education", we employed the function `SmoteClassif`[1] from the R `UBL-package`. The relevant parameter is *C.perc*, a list that holds the percentages of under-sampling or/and over-sampling to apply to each class of the protected attribute. We analyzed five different configurations for this parameter: the default configuration "balance" (which represents the case where the sampling percentages are automatically estimated to balance the samples between the minority and majority classes, that is, the perfect uniform distribution) plus four additional configurations corresponding to the four exemplar distributions "QuasiBalance", "OneOff", "HalfHigh" and "Power2" already used to test this mutation technique in Chapter 6. For the exhaustive description of the exemplar distribution, we refer to Section 4.1.1 of Chapter 4.
For each exemplar distribution we looked at 4 different permutations of the values of the percentages assigned to the various classes of the protected attribute. For instance, in the *One Off* configuration the four different permutations have each a different class with zero occurrences.

Binary case. For binary attributes we applied the function `ovun.sample`[2] from the R `ROSE-package`. The relevant parameter of this mutation is p, which represents the probability of resampling from the rare class and it has been set to 17 different values in order to vary as much as possible the distribution of the occurrences between the two categories of the attribute "sex": 0.01 (corresponding to the case of minimum balance), 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5 (maximum balance), 0.6, 0.7, 0.8, 0.9, 0.925, 0.95, 0.975, 0.99. When the value of `p` is set to 0.5, it indicates aiming for the most even distribution between the two classes; lower values of `p` will result in a less balanced distribution, while increasing the value from 0.5 to 1 will lead to a more balanced distribution, but with inverted proportions.

[1] https://www.rdocumentation.org/packages/UBL/versions/0.0.6/topics/SmoteClassif, last visited on June 1, 2023

[2] https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ovun.sample, last visited on June 1, 2023

Note that in both cases –multiclass attributes and binary attributes– the generated mutated datasets have the same number of rows as the original ones, and the distribution of the other variables in the dataset remains unchanged.

Finally, with a view to increasing the variability and the reliability of our approach, given the random nature of the resampling we decided to vary a `seed` (an integer value used to ensure reproducibility and keep track of the samples) by randomly generating 50 different values between 1 and 1000.
This means that for the analysis and discussion of the results we always kept track of the outputs for each value of the `seed` and for each value of the parameter `p`, then: for the mutations obtained by setting `C.perc`=“balance” we collected a total of
6 (attributes) × 50 (seed) × 17 (levels of `p`) = 5100 values for the *balance measures* and 5100 values for the *fairness criteria*.
Instead, for the mutations obtained through the four different lists of percentages, we gathered a total of 6 (attributes) × 50 (seed) × 17 (levels of `p`) × 4 (exemplar distributions) × 4 (permutations) = 81600 values for both the *balance measures* and the *fairness criteria*.
The sum of all these elements adds up to 5100 + 81600 = 86700 values for the *balance measures* and 86700 values for the *unfairness measures*. Note that the 6 attributes considered in these computations, are those summarized in the left column of Table 8.1.

8.4 Results and Discussion

In this Section we examine and discuss the results of our investigation, according to the research questions stated at the beginning of this Chapter.

RQ 5.1 - Intersectional versus Primary attributes

Method

In order to investigate the relationship between intersectional and primary attributes, we observe the results of an ANOVA on two linear regression models, one for balance measures and one for fairness criteria:

$$\mathfrak{B}(sex\_education) = c_{sex} \cdot \mathfrak{B}(sex) + c_{education} \cdot \mathfrak{B}(education) + c_0$$

$$\mathfrak{U}(sex\_education) = c_{sex} \cdot \mathfrak{U}(sex) + c_{education} \cdot \mathfrak{U}(education) + c_0$$

The first model was applied with all the four distinct balance measures reported in Section 2.2.2, while the second model was evaluated using all the five unfairness measures described in Section 2.2.3. To answer RQ 5.1, we look at two results from the analysis: adjusted R^2 and p-value. The adjusted R^2 is a goodness-of-fit measure for linear models and it is an indicator of the model accuracy, as it identifies the percentage of variance in the output variable that is explained by the input variables. In fact, R^2 tends to optimistically estimate the fit of the linear regression: a value of 1 indicates a model that perfectly predicts dependent values, whereas a value closer to 0 means that the model has no predictive capability. Thus, in our specific case, values of R^2 close to 1 mean that the measure related to the intersectional attribute can be explained by those related to the primary attributes. Smaller values indicate that the intersectional attribute cannot be explained by primary attributes alone. To assess the statistical significance of the results, we observe the p-value and consider significant the relationships whose p-value is lower than 5%. In addition, looking at the coefficients c_{sex} and $c_{education}$, we evaluate whether the two primary attributes provide an equal contribution.

a) Balance

The results of the regression for the balance measures are reported in Table 8.2. We observe that in the cases of the Gini index the R^2 is very close to 1 (0.941), and for the Shannon and Simpson indexes the R^2 is around 0.86; while it is much smaller (0.540) for the Imbalance Ratio index. For all the cases the p-value is $< 2.2 \cdot 10^{-16}$, indicating statistically significant results. This means that in three cases out of four, the balance measures related to the intersectional attribute can be explained by those related to the primary attributes, thus we can accurately infer the balance of the multiclass intersectional attribute from the balance of the primary attributes which compose the intersectional attribute itself; in the case of IR we have a smaller correlation, probably due to the fact that for many data points the IR assumes values close to zero more frequently than the other measures.

In addition, we computed the regression coefficients, reported in the three rightmost columns of Table 8.2. Indeed, looking at the coefficients c_{sex} and $c_{education}$, we observed that overall the balance measurements of the primary attributes have a high positive correlation with the intersectional attribute: in particular, such a positive correlation is higher in correspondence of the primary multiclass attribute "education", except for the IR index, which presents a coefficient $c_{education}$ much smaller with respect to the other coefficients.

Table 8.2 Balance measures: evaluation of the linear regression model $\mathfrak{B}(sex\_education) = c_{sex} \cdot \mathfrak{B}(sex) + c_{education} \cdot \mathfrak{B}(education) + c_0$.

Balance measure	***Adjusted R^2***	***p-value***	**Coefficients**		
			c_0	c_{sex}	$c_{education}$
Gini	0.941	$< 2.2 \cdot 10^{-16}$	17.308	0.195	0.671
IR	0.540	$< 2.2 \cdot 10^{-16}$	-2.908	0.475	0.107
Shannon	0.877	$< 2.2 \cdot 10^{-16}$	13.296	0.345	0.533
Simpson	0.850	$< 2.2 \cdot 10^{-16}$	-6.652	0.476	0.564

b) Unfairness

The results of the regression for the unfairness measures are reported in Table 8.3. Differently from the balance measures, for the fairness criteria we observe overall lower values of the adjusted R^2: in particular, we found values of R^2 around 0.6 for the independence, separation-TP and sufficiency-PN criteria, and even lower values – around 0.4 – in the case of the separation-FP and sufficiency-PP criteria. For all the cases the p-value is <2.2$\cdot$10^{-16}, indicating statistically significant results. In addition, we computed the regression coefficients reported in the three rightmost columns of Table 8.3. As before for the balance measures, overall the unfairness measurements of the primary attributes have a positive correlation with the intersectional attribute: specifically, the coefficient $c_{education}$ –which is between 0.430 and 0.622– assumes higher values than the coefficient c_{sex} for all the fairness criteria except for the separation-FP criterion, indicating overall a higher positive correlation in correspondence of the primary multiclass attribute "education". Overall –in four cases out of five– there exists a higher positive correlation between the unfairness measurements of the intersectional attribute and those of the primary attribute "education", with respect to the correlation between the intersectional attribute and the primary attribute "sex".

Table 8.3 Unfairness measures: evaluation of the linear regression model $\mathfrak{U}(sex\_education) = c_{sex} \cdot \mathfrak{U}(sex) + c_{education} \cdot \mathfrak{U}(education) + c_0$.

Unfairness measure	***Adjusted*** R^2	***p-value***	**Coefficients** c_0	c_{sex}	$c_{education}$
Independence	0.624	$< 2.2{\cdot}10^{-16}$	0.858	0.395	0.567
Separation – TP	0.625	$< 2.2{\cdot}10^{-16}$	1.971	0.472	0.622
Separation – FP	0.395	$< 2.2{\cdot}10^{-16}$	0.525	0.610	0.547
Sufficiency – PP	0.393	$< 2.2{\cdot}10^{-16}$	5.432	0.285	0.430
Sufficiency – PN	0.549	$< 2.2{\cdot}10^{-16}$	1.896	0.249	0.646

RQ 5.2 - Balance as intersectional unfairness predictor

Method

In order to answer RQ 5.2, we analyze the relationships between balance measures and fairness criteria for the intersectional multiclass attribute sex_education. We compute the correlation between the balance and the unfairness measures, for each index of balance and each fairness criterion. We use the Spearman correlation coefficient since we do not expect a linear relationship. A negative and statistically significant correlation coefficient –that is, low balance corresponding to high unfairness– suggests a positive answer to the research question.

We remind from our previous studies on primary protected attributes [41, 43–45] that the balance measures properly detect unfairness of software output, but their effectiveness in identifying unfairness is dependent on the selected metric of balance, which has a relevant impact on the threshold to consider as risky, and thus on the detection of discriminatory outcomes. As we are investigating the balance measures as unfairness predictors when specifically applied to intersectional attributes, we plot LOESS curve to better understand the relationship between balance and unfairness in the case of intersectional protected attributes.

a) Correlation

The correlation coefficients are reported in Table 8.4. We observe that they are all negative and the corresponding p-values are all smaller than $2.2{\cdot}10^{-16}$. Thus, we can answer the research question positively.

Table 8.4 Correlation between balance and unfairness for the intersectionl attribute *sex_education*: $\mathfrak{B}$(sex_education) $\sim$ $\mathfrak{U}$(sex_education).

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	-0.1614	-0.1652	-0.1687	-0.1706
Separation – TP	-0.2583	-0.2799	-0.2696	-0.2902
Separation – FP	-0.2130	-0.2244	-0.2203	-0.2340
Sufficiency – PP	-0.1836	-0.1842	-0.1905	-0.1862
Sufficiency – PN	-0.1487	-0.1654	-0.1425	-0.1631

b) Relationship

Figure 8.1 reports the trend lines –as smoothed regression– of the five fairness criteria (along the Y-axis) with respect to the increase in balance measures (along the X-axis), in percentage values. It has to be noted that maximum levels of unfairness are higher for Sufficiency (PP and PN) and Separation-TP (more than 10% in correspondence of the lowest values of balance) and less than 5% in the other cases. Overall we observe decreasing trends, in accordance with negative correlation values, however often *not* monotonic, which explains why correlation values were not high. In general the trends are consistent with our previous studies on primary protected attributes, with most irregular patterns related to Sufficiency. Since the specific unfairness criteria reflect different levels of balance in slightly different ways, we recommend choosing distinct thresholds of risks for the four balance measures: the specific application context might suggest using more sensitive balance measures – IR and Simpson – for cases where unfairness tolerance is low, and the less sensitive Gini and Shannon when higher levels of unfairness can be socially accepted.

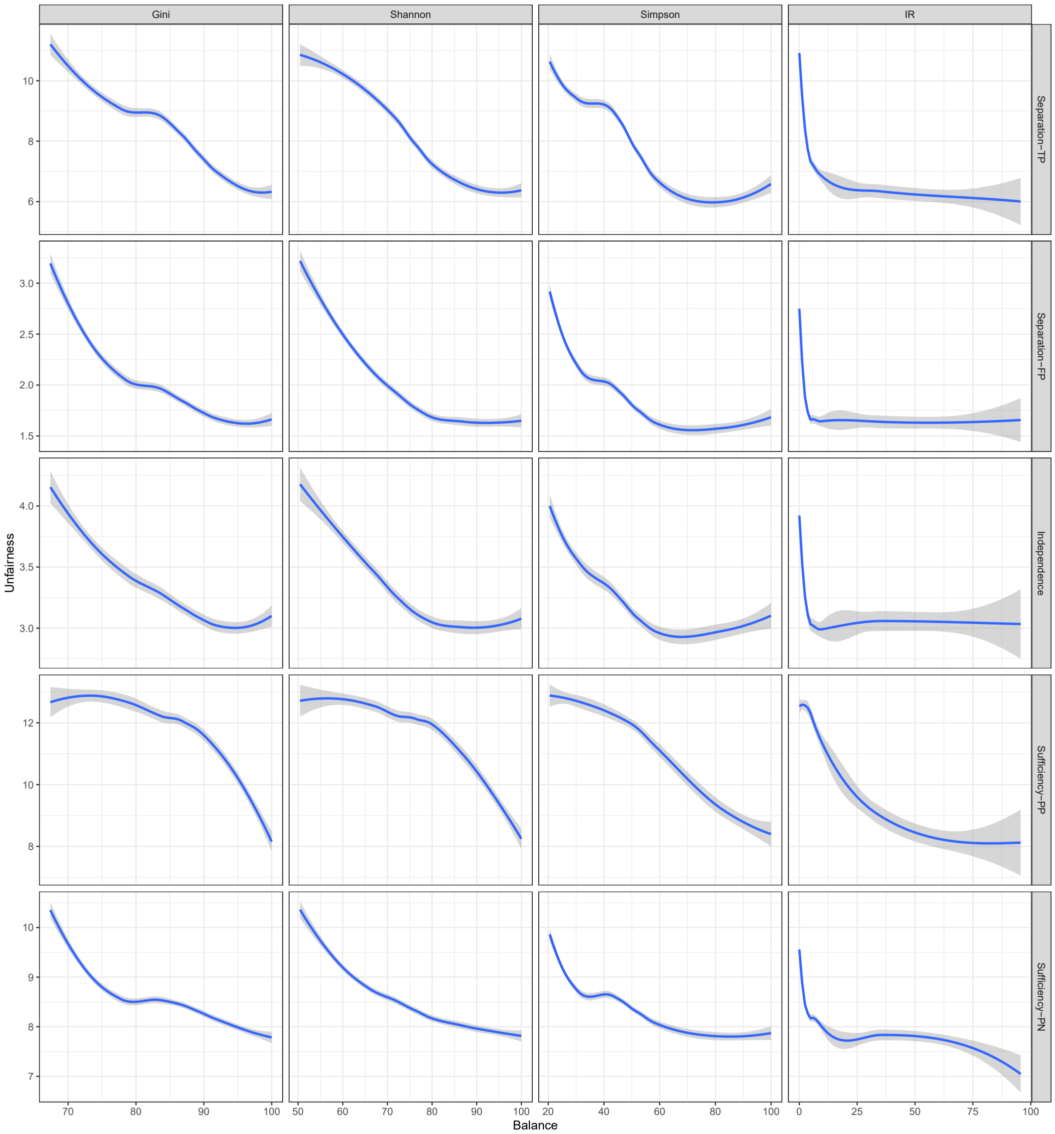


Fig. 8.1 Trends of the *fairness criteria* as a response to the *balance measures* for the intersectional protected attribute *sex_education*.

RQ 6 - Contribution of combined target

Method

Before looking into the contribution of the target variable combined with protected attributes to the detection of unfairness, we consider the relationship between the balance values of the protected attributes (primary or intersectional) by themselves and when considered in combination with the target variable.

To answer the research question 6, we computed the Spearman correlation coefficients of unfairness measures versus balance measures, and compared the coefficients for the attributes with and without the combination with the target variable, with a view to investigating whether the combination of a protected attribute with the target variable improves the detection of the unfairness. Then, to examine in-detail our findings, we computed the difference between the correlation of a protected attribute (primary or intersectional) and the correlation of the same attribute combined with the target, for three different cases:

- $\text{diff}_{sex} = \text{cor}(sex) - \text{cor}(sex\_target)$
- $\text{diff}_{education} = \text{cor}(education) - \text{cor}(education\_target)$
- $\text{diff}_{sex\_education} = \text{cor}(sex\_education) - \text{cor}(sex\_education\_target)$

where the expression "cor(*protected attribute*)" indicates the correlation between balance and unfairness measures for a given *protected attribute* (primary or intersectional). We remind that we expect the correlations to be negative, which would mean that high balance is associated with low unfairness values, and vice-versa.

a) Combination with target variable

Figure 8.2 reports the scatter plot of the corresponding values with a smoothed interpolation curve. We can observe very different patterns. The "sex" primary attribute shows a relationship to its combination with the target that is close to linear for all the balance measures. As far as the "education" attribute is concerned, we observe an irregular relationship that changes among the different balance measures. The intersectional attribute encompassing both the former attributes exhibits a close to linear relationship for three balance measures except for the IR index, which presents a different more irregular pattern.

b) Differences in correlation

We report all the numerical results in Appendix B, while here we provide only a synthetic and more readable overview in Figure 8.3, where we report the correlation values for all combinations of balance and unfairness measures divided by attribute.

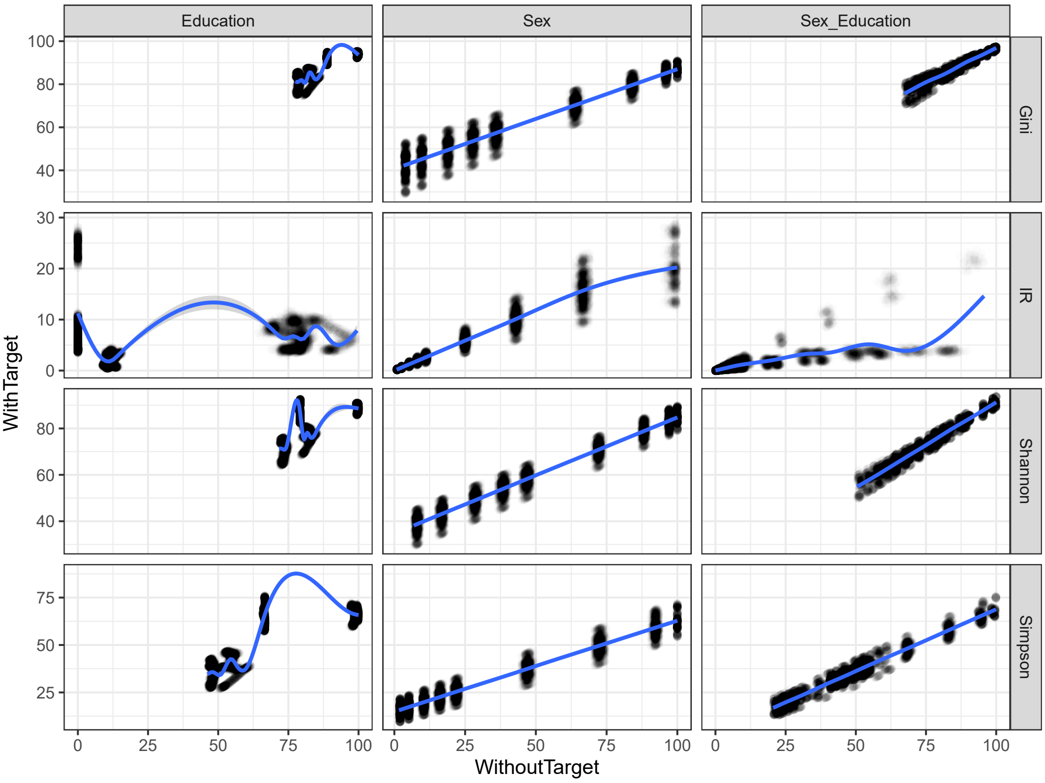


Fig. 8.2 Balance measures of protected attributes combined *with* the target variable versus protected attributes *without* target.

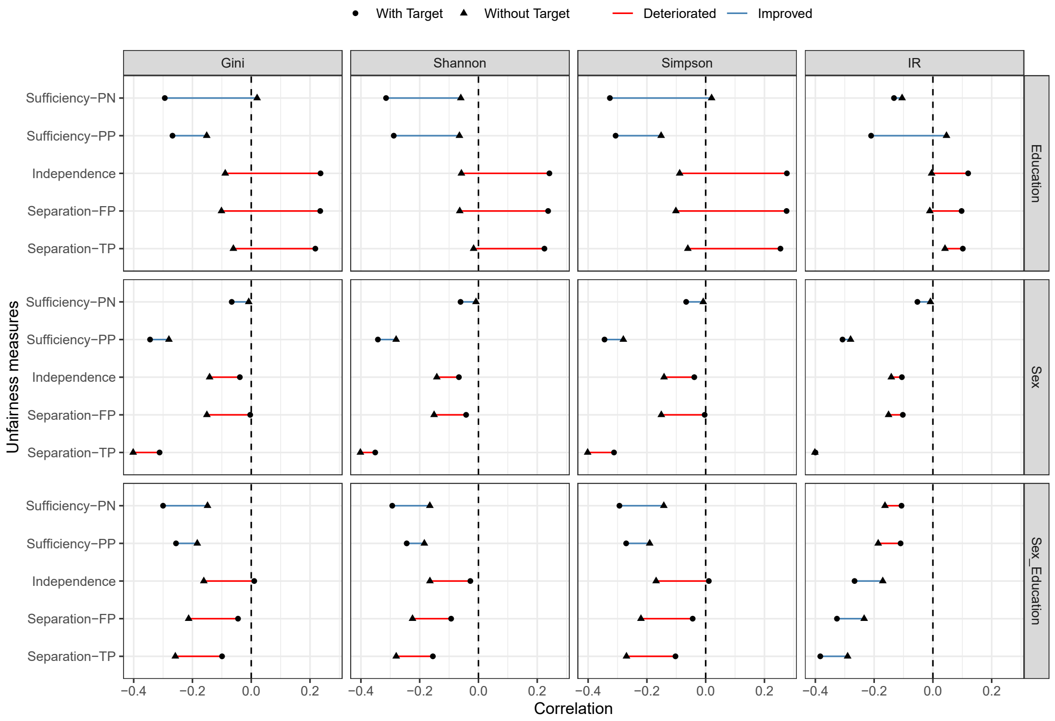


Fig. 8.3 Correlation Balance-Unfairness for protected attributes combined *with* the target variable and *without* target, for different fairness criteria and balance measures. The farther left to the zero the points, the better they are; if the circle marker is left of the triangle, then the combination *with* the target variable improves the unfairness detection.

The diagram can be interpreted as follows: the farther left to the zero (represented by the dashed black line) the points, the better they are; if the circle marker is left of the triangle then the combination with the target variable improves the correlation, that is the capability of detecting unfairness risk.

As concerns the binary attribute “sex”, we observe that all the correlation values are negative (the data points are to the left of the dashed line), but we note a small improvement of the correlation only for the sufficiency criterion –both for Parity of Positives and Parity of Negatives–, whereas we observe a worsening in correspondence of the independence and separation criteria.

A similar pattern can be observed for the multiclass attribute “education”, although with much larger differences. In particular, the deterioration of the independence and the separation criteria is so significant that the combination with the target variable makes all correlations positive, indicating that the combination of the target with multiclass protected attributes worsens the unfairness detection even more than binary attributes. Conversely, in the case of the sufficiency criterion, the combined attributes improve the identification of the unfairness to a greater extent with respect to the previous case of the binary attribute.

Finally, for the intersectional protected attribute "sex_education" we notice the same strengthening/weakening pattern, but with the notable exception of the Imbalance Ratio index for which we find exactly the opposite pattern for all the fairness criteria. Hence, combining the target variable with the intersectional protected attribute improves the identification of the unfairness assessed through the sufficiency criterion with respect to the Gini, Shannon and Simpson indexes, but not to the IR index –which by contrast worsens in correspondence of the sufficiency criterion, and improves according to the independence and separation criteria.

As a further observation, we note that given the nature of the fairness criteria –whose definitions are based on the target variable, score and protected attributes– and the two mutation techniques that we applied –which leave the distribution of the other variables unchanged, and thus also the distribution of the target variable remains unchanged– the level of balance of the target variable in the original dataset certainly plays a role in the final interpretation of our results. Indeed in our dataset, where the frequency of the positive target is much lower than the negative target (in the original dataset, only 6636 instances out of 30.000 belong to the positive class, which corresponds to 22% of the occurrences), the combination of protected attributes with the target variable improves identifying a discrimination risk when we apply the sufficiency criterion (except for the Imbalance Ratio index applied to intersectional protected attributes); in fact, the sufficiency criterion implies the calibration of the model for the different groups as it requires the conditional probability of the target variable to be equal to 1.

8.5 Limitations

As concerns the limitations of our study, we highlight that an investigation of more datasets and protected attributes would extend the analysis to a wider range of intersectional classes and combinations with target variables belonging to different domains of interest in the large landscape of automated decision-making systems. Moreover, we remark that in our dataset the predicted class was not present, therefore we ran a binomial logistic regression in order to build a classification label, but all

the limitations of the algorithm hold: most notably, the assumption of limited or no multi-collinearity between independent variables, as well as the assumption of linearity between the dependent variable and the independent variables.
Applying more classification algorithms (each with different parameters) would increase the generalizability of the results, by helping to identify how the different types of classification algorithms propagate the imbalance from the training set to the output. In addition, other kinds of mutation techniques could be considered by adopting different pre-processing methods in order to extend the variability and reliability of our results.
Finally, a more in-depth analysis is also necessary to better understand the relation between the level of imbalance of the target variable in the dataset and the application of a specific fairness criterion, for instance by applying a mutation technique to the target variable and examining the behavior of the fairness criteria as a response to the balance measures, in order to better understand which fairness criterion to choose to assess the risk of discrimination; it is important to highlight that the choice of the fairness criterion should also depend on the domain and context of use.

8.6 Research Outcomes

In this work we move forward on the assessment of balance measures as risk indicators of systematic discrimination by including two more aspects: i) intersectionality among the classes of protected attributes, and ii) the impact of imbalanced distributions in target variables.

RQ 5.1. First of all, we investigated how intersectional attributes relate to the corresponding primary attributes in terms of balance and fairness. Concerning the investigation of the balance measures, we found that the measures of the Gini, Shannon and Simpson indexes related to the intersectional attribute can be explained by those related to the primary attributes; while the measure of the IR index related to the intersectional attribute is explained by the measures of the IR index related to primary attributes alone to a smaller extent with respect to other indexes.
As regards the analysis of the fairness criteria, we made the following observations:

- there exists a correlation between the unfairness measurements of the intersectional attribute and the primary attributes, but the former is only partly determined by the latter;
- the unfairness measures related to the intersectional attribute can be explained by those related to the primary attributes, but to a lower extent with respect to the balance measures.

RQ 5.2. Secondly, we analyzed whether the measure of balance on intersectional attributes can detect unfairness risks: we observe a moderate negative correlation between balance measures and fairness criteria, indicating that intersectional protected attributes can be taken into account to identify unfairness risks.
Indeed, the behavior of the fairness criteria in response to the balance measures presents a decreasing trend, even though the distinct fairness criteria reflect different levels of balance in slightly different ways.

RQ 6. Finally, we investigated whether the combination of the target variable with protected attributes improves the detection of the risk of unfair classification outcomes. From our findings we can conclude as follows:

- the combination of *primary* protected attributes (binary or multiclass) with the target variable improves the detection of the unfairness measured through the sufficiency criterion (both Parity of Positives and Parity of Negatives), but worsens the detection of the unfairness measured through the independence or the separation criteria;
- the combination of *intersectional* protected attributes with the target variable improves the identification of the unfairness measured through the sufficiency criterion in the cases of the Gini, Shannon and Simpson indexes, but not in the case of the Imbalance Ratio index, for which the detection of the unfairness is improved when measured through the independence or the separation criteria.

Algorithm 5: Measurements of balance $\mathfrak{B}$ and unfairness $\mathfrak{U}$ for RQ5.

```
Require: categorical protected attributes $A_i$, each with number of categories $m > 0$;
    • mutation technique $M =$ · ovun.sample (for binary attributes) or
                                · SmoteClassif (for multiclass attributes);
    • exemplar distributions $E = \{Balance, QuasiBalance, OneOff, HalfHigh, Power2\}$;
    • mutation parameter $P = \{0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.925, 0.95, 0.975, 0.99\}$;
    • balance measures $\mathfrak{B} = \{Gini, Shannon, Simpson, IR\}$;
    • unfairness measures $\mathfrak{U} = \{\mathfrak{U}_{Sep\_TP}, \mathfrak{U}_{Sep\_FP}, \mathfrak{U}_{Suf\_PP}, \mathfrak{U}_{Suf\_PN}, \mathfrak{U}_{Ind}\}$.
    Input  : Dataset $D$
    Output : Balance measures $\mathfrak{B}$, Unfairness measures $\mathfrak{U}$
identification of a multiclass protected attribute $A_m$ and a binary protected attribute $A_b$ in $D$
for all $e \in E$ do
    application of the mutation technique SmoteClassif: $M_{m,e} \leftarrow M(A_m)$
    for all $p \in P$ do
        application of the mutation technique ovun.sample: $M_{b,p} \leftarrow M(A_b)$
        intersection of mutated protected attributes and combination with target variable:
        · intersection of $M_{b,p}$ and $M_{m,e}$: $M_{bm,p,e} = M_{b,p} + M_{m,e}$
        · combination of $M_{b,p}$ and target variable: $M_{b,p,target} = M_{b,p} + target$
        · combination of $M_{m,e}$ and target variable: $M_{m,e,target} = M_{m,e} + target$
        · combination of $M_{bm,p,e}$ and target variable:
        $M_{bm,p,e,target} = M_{bm,p,e} + target$
        randomly data splitting of 70%-30% into training-test sets
        prediction of the score variable with a classification model
        application of balance and unfairness measures according to Table 8.1:
        $\mathfrak{B}_{b,p} \leftarrow \mathfrak{B}(M_{b,p}) \in$ training set
        $\mathfrak{B}_{m,e} \leftarrow \mathfrak{B}(M_{m,e}) \in$ training set
        $\mathfrak{B}_{bm,p,e} \leftarrow \mathfrak{B}(M_{bm,p,e}) \in$ training set
        $\mathfrak{B}_{b,p,target} \leftarrow \mathfrak{B}(M_{b,p,target}) \in$ training set
        $\mathfrak{B}_{m,e,target} \leftarrow \mathfrak{B}(M_{m,e,target}) \in$ training set
        $\mathfrak{B}_{bm,p,e,target} \leftarrow \mathfrak{B}(M_{bm,p,e,target}) \in$ training set
        $\mathfrak{U}_{b,p} \leftarrow \mathfrak{U}(M_{b,p}) \in$ test set
        $\mathfrak{U}_{m,e} \leftarrow \mathfrak{U}(M_{m,e}) \in$ test set
        $\mathfrak{U}_{bm,p,e} \leftarrow \mathfrak{U}(M_{bm,p,e}) \in$ test set
    end for
end for
return $\mathfrak{B}_{b,p}, \mathfrak{B}_{m,e}, \mathfrak{B}_{bm,p,e}, \mathfrak{B}_{b,p,target}, \mathfrak{B}_{m,e,target}, \mathfrak{B}_{bm,p,e,target}, \mathfrak{U}_{b,p}, \mathfrak{U}_{m,e}, \mathfrak{U}_{bm,p,e}$
```

Chapter 9

Practical Implications and Future Work

We propose a risk assessment approach based on quantitative measures to evaluate imbalance in the input datasets of automated decision-making systems, and we aim to highlight a potential risk of discriminatory outcomes by revealing these imbalances in training data. For this purpose, we built an open and extensible benchmark of balance measures, so as to attract further contributions in this direction. Specifically, we measured the imbalance of the protected attributes in the input data while we evaluated the unfairness of classification outcomes through different fairness criteria.

9.1 Practical Implications

On the basis of the studies conducted in this dissertation, we provide several recommendations for the usage of the balance measures as indicators of potential unfairness in the output. For this purpose, we will make use of the following two examples (formalized by keeping in mind the different behaviors of the indexes of balance that we observed overall in our studies).

Given a certain protected attribute in a dataset, suppose to obtain a value around 0.3 for the IR and Simpson indexes and a value around 0.5 for the Gini and Shannon indexes. If we define a single threshold of balance valid for all indexes such that a value below the threshold identifies imbalance, whereas a value above the threshold

indicates balance, and we set this threshold at 0.4, then the IR and Simpson indexes reveal imbalance and, consequently, a higher risk of unfairness, while the Gini and Shannon indexes –being above the threshold– indicate the absence of imbalance and therefore a lower risk of unfairness. Now, suppose to have a threshold of unfairness equal to 0.2 such that a value higher than the threshold identifies the presence of unfairness, then, according to the IR and Simpson indexes we expect a value of unfairness greater than 0.2; on the contrary, according to the Gini and Shannon indexes we expect an unfairness value lower than 0.2 (that is, the absence of unfairness). If we have an actual case of unfairness (that is, the unfairness is *higher* than 0.2), then we have True Positives in the case of the IR and Simpson indexes (as they correctly anticipated the presence of unfairness), while we have False Negatives in the case of the Gini and Shannon indexes (as they revealed a balance situation, and consequently the absence of unfairness). Thus, we should increase the threshold of balance, for instance at 0.6, in order to have all the indexes correctly revealing unfairness. The presence of False Negatives is a significant issue, for instance in the medical field, where it can lead to missed healthcare and to the loss of effective treatments.
Now suppose a different situation in which we have a single threshold of balance set at 0.8 and the actual unfairness is *lower* than 0.2 (which means, an actual case of fairness). Suppose to obtain a value around 0.9 for the Gini/Shannon indexes and a value around 0.7 for the IR/Simpson indexes: the first two indexes correctly predict a low risk of unfairness; the second two indexes, on the other hand, although assuming a rather high value, indicate imbalance and consequently a high risk of unfairness in the output, thus representing a False Positive. In this case, we should decrease the threshold of balance, for instance at 0.6, in order to have all the indexes correctly indicating a low risk of unfairness, thus avoiding the presence of False Positives.

As it clearly emerges from these two simple examples, the choice of the metric has a relevant impact on the threshold to consider as risky, therefore, it is advisable to consider different thresholds (instead of a single one as in the examples above) depending on the context and dataset's domain, as well as on the choice of the fairness criterion.

Specifically, we recommend taking the following aspects into account when using the indexes of balance to foresee unfairness risks measured through the *Independence*, *Separation* or *Sufficiency* criteria:

- the **Gini** and **Shannon** indexes are able to detect discrimination, but overall they tend to assume higher values with respect to the Simpson and Imbalance Ratio indexes; for this reason, we recommend using the Gini and Shannon indexes with higher thresholds (with respect to those adopted for the other two indexes given a specific context) in order to avoid losing relevant cases of imbalance and, consequently, of unfairness;

- the **Imbalance Ratio** index has a good capacity to detect imbalance in datasets and to reveal discrimination risks, but it shall not be used in presence of empty classes, as it drops to 0 in presence of at least one class with zero occurrences, and tend to assumes very low values when very few classes compared to the total number of classes are empty;

- the **Simpson** index has a good ability to detect imbalance and identify discrimination risks; we recommend using 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 presents correlations values comparable to those of the Imbalance Ratio index.

Furthermore, we provide specific recommendations derived from our studies on intersectional attributes and combinations of protected attributes with the target variable. From the investigation on *intersectional attributes*, we found that balance measures are suitable for identifying the unfairness risk in classification outcomes, however, due to some variance in the observed trends, we strongly recommend first selecting a single fairness criterion of interest, and then choosing the balance measure that is more appropriate to the application case:

- the **Simpson** and **Imbalance Ratio** indexes are recommended for rapid reaction to unfairness, in all those cases in which slight deviations of fairness correspond to severe damages to people's lives;

- the **Gini** and **Shannon** indexes are suitable in all other cases because they have a smoother response to unfairness risks and the cost of a wrong fairness detection can be minimized.

Concerning the combination of the *target variable* with protected attributes, we recommend using the **Gini**, **Shannon** or **Simpson** indexes when the sufficiency criterion of fairness is preferable, otherwise the **Imbalance Ratio** index.

9.2 Future Work

In order to increase the generalizability of our findings and to better assess the reliability of the balance measures as risk indicators for a potential emergence of discriminatory behavior by automated decision-making systems, first of all it would be recommended to extend our studies on a wider number of datasets with all the concerning information. We are confident that investigating on a wider amount of data could help to interpret more profoundly the suitability of the adopted indexes.

In the second place, for the purpose of improving our approach based on quantitative measures, it would be advantageous to take into account other kinds of metrics, examining and comparing their performance with the Gini/Shannon/Simpson/IR indexes. In particular, this research could be expanded by including balance measures for non-categorical attributes and additional unfairness measures.

In addition, the application of different mutation techniques could further enrich this research, as well as the study of a larger number of exemplar distributions, both in the initial phase of our analysis and when they are taken into account for the application of the mutation techniques.
Moreover, in order to better generalize our investigation on risk thresholds and to thoroughly assess their reliability, an analogous study could be conducted by considering other evaluation metrics (different from the accuracy) as a discriminant to define the best thresholds for each combination of balance-unfairness-algorithm, and also performing the hyper-parameters tuning for each classifier.

Further work shall be devoted to testing in more detail our approach through the adoption of other classification and prediction algorithms. In particular, the generalizability of our results would certainly benefit from the confirmation of the study through the application of more complex deep learning systems. Given that we deeply demonstrated the strength of our results by applying the proposed approach to a huge amount of data having different levels of balance, including extreme cases (for example, the presence of categories with zero occurrences), we expect the results to hold even for more sophisticated systems, as their complexity due to a large number of parameters does not minimize the problem of data imbalance, which indeed represents an important socio-technical issue that needs to be addressed –we remind of the examples exposed in Chapter 1, where we illustrated how an imbalance in data can spread and manifest itself in the output of different ADM systems. As a further strength of our approach, we highlight the following point: although the application of deep learning methods could certainly enrich this research work from an exploratory point of view, we should keep in mind that these methods require a much larger amount of data than those used in our study, therefore they will not necessarily lead to an improvement in terms of details in the research.

Finally, potential mitigation strategies and actions should be examined, that encompass 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.

Chapter 10

Conclusions

In this study we proposed and tested a metric-based approach to evaluate imbalance in a given dataset as a potential risk factor for discriminatory outcomes of automated decision-making systems.
The rationale of our approach is founded on two conceptual frameworks: the first one is the ISO/IEC 25000:2014 series of standards [48], also known as SQuadRE, in which a chain of quality effects is described: in the field of data quality, a simplified version of this standard series is the well-known GIGO principle, which stands for "garbage in, garbage out" and states that a software outcome will be unreliable if the input data are outdated, inaccurate, incomplete, or flawed. Following this line of reasoning, our hypothesis is that bias on input data will probably cause biased output data: in terms of automated decision systems, this would lead to potential discriminatory outputs. The second fundamental concept of our methodology is based on the ISO 31000:2018 standard for Risk management [49]; taking into account both data imbalance and potential discrimination from ADM systems, we focus on the risk assessment stage, and particularly on the first two phases: i) risk identification, in order to recognize and describe potential risks within a specific context and scope, and ii) risk analysis, which consists in identifying the features and potential extent of risk, and assumes metrics of data imbalance as indicators of discrimination risk.

In the following we summarize our findings for each research question.

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

In our very first investigation, we selected four widely used indexes of data balance (the Gini, Shannon, Simpson and Imbalance Ratio indexes), normalized them to share the same range of values and semantics, and tested their ability to detect different levels of imbalance in synthetic attributes. According to the exemplar distributions chosen, the best result is achieved by the Imbalance Ratio index, even though it is very sensitive in presence of classes with 0 occurrences: for its definition, the index drops to 0 when at least one class is empty. An intermediate result is achieved using the Simpson index, while the Gini and Shannon indexes showed the lowest performances as they constantly assume higher values (suggesting balanced data), therefore they should be further investigated, for example using different thresholds. As a further observation, we did not find any specific trend associated with the number of classes.

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

In this study, we aim at assessing the reliability of the balance measures as risk indicators of biased decisions or distorted recommendations in the context of automated decision-making systems. Thus, we tested the capabilities of the balance measures to reveal discrimination occurring in the classification output of ADM systems trained with several large datasets belonging to different application domains. Overall, the results indicate that the approach is suitable for the proposed goal, even though the balance measures performed differently with respect to different fairness criteria and can be ranked similarly as for the previous study: the IR index achieved the best performances, followed by the Simpson index and the Gini/Shannon indexes. As a general indication, evidence suggests that a combined usage of the indexes of balance is preferable in order to detect possible discrimination risks –as indicated in the pragmatic recommendations provided in the previous Chapter.

RQ3. *Is it possible to measure the risk of bias in a classification output by measuring the level of (im)balance in the protected attributes of the training set?*

With the third research question we evaluated how the balance of protected attributes in training data can be used to assess the risk of algorithmic unfairness in subsequent classification tasks, by using the set of balance measures previously analyzed. Differently from the previous study, we conducted the analysis separately for *multiclass* and *binary* protected attributes, and we adopted a pre-processing method as mutation technique in order to mutate the distribution of the occurrences between the classes of a certain attribute so as to generate a number of derived synthetic datasets having different levels of balance. Then, we assessed whether imbalanced distributions of protected attributes in the training data can lead to discriminatory output of ADM systems.
Both in the case of *multiclass* and *binary* protected attributes, our investigations showed that overall the unfairness measures decrease as the balance measures increase. Indeed, the correlation analysis confirmed that the balance measures are capable of predicting unfairness of software output, with some differences between the indexes: specifically, the IR index is highly sensitive to extreme values of balance/imbalance, while the Gini and Shannon indexes tend to assume higher values overall. Thus, the choice of the balance measure has a relevant impact on the threshold to consider as risky, and thus on the detection of discriminatory outcomes of ADM systems.

RQ4. *Is it possible to identify a threshold s (for balance measures) such that if the* balance *of the training set is greater than s, then the* unfairness *of the classification on the test set is expected to be less than a threshold f?*

In this study we defined and tested a methodology to identify thresholds of balance such that the unfairness of the classification is expected to be less than a desired levels. To conduct this analysis, we adopted the metric-based approach previously defined in order to evaluate imbalance in a given dataset as a risk indicator of discriminatory classification outcomes. Thus, we generated a large number of synthetic datasets and measured the different levels of imbalance in the training sets by means of the balance measures previously analyzed, while we assessed the discrimination occurring in the classification output through a set of fairness criteria. After that, we built the thresholds s and f by following a specific procedure. Specifically, for each combination of balance-unfairness-algorithm we selected the configuration of thresholds that presented the highest accuracy. We conducted the experiment and analyzed the results separately for binary and multiclass attributes.

By assessing the thresholds through the evaluation metrics (Accuracy, Precision, Sensitivity, Specificity and F1-score), we observed that the values of balance under s indicate levels of unfairness over f, but they even tend to overestimate the risk (as we can infer from the low values assumed by the Specificity index, which is caused by a high number of false positives). We also noted that the identified thresholds perform better in the binary case than in the multiclass case, with higher accuracy values overall. Then, from the assessment of the thresholds' accuracy with respect to balance-unfairness-algorithm, we found that the balance measures seem to have an impact only in the multiclass case, whereas the fairness criteria affect the thresholds' accuracy in both cases following the same pattern; also the algorithms have an impact –as they present different accuracy values– but without a precise trend.

RQ5. *Is it possible to identify the risk of biased output by detecting the level of (im)balance in intersectional protected attributes?*

In this work we studied to which extent it is possible to rely on balance measures as risk indicators of systematic discrimination when dealing with the intersection of protected attributes. We conducted an empirical study to test whether: i) it is possible to infer the balance of intersectional attributes from the balance of the primary attributes, ii) measures of balance on intersectional protected attributes are helpful to detect unfairness in classification outcomes.
Again, we selected four indexes of balance (Gini, Simpson, Shannon, Imbalance Ratio), we generated a large number of synthetic datasets and measured different levels of imbalance in the training sets, whereas we evaluated the discrimination occurring in the classification outcome on the test sets. Overall, the results on intersectional attributes show that balance measures are suitable for identifying unfairness risks in a classification output. Indeed, we observe a moderate negative correlation between balance measures and fairness criteria, indicating that intersectional protected attributes can be taken into account to identify unfairness risks, even though the distinct fairness criteria reflect different levels of balance in slightly different ways.

RQ6. *Does the combination of the target variable with protected attributes improve the detection of unfair classification risks?*

In the last study we investigate the contribution of the target variable to the unfairness detection. From our findings we conclude that the combination of the target variable with protected attributes, both in the case of *primary* and *intersectional* attributes, improves the detection of the unfairness depending on the choice of the fairness criterion, and also on the selected index of balance in the case of *intersectional* protected attributes.
Particularly, the combination of *primary* protected attributes (binary or multiclass) with the target variable improves the detection of the unfairness depending on the choice of the fairness criterion: the identification of discrimination risks improves if the sufficiency criterion is chosen, while it is worsens through the independence or the separation criteria.
As regards the combination of *intersectional* protected attributes with the target variable, the detection of the unfairness is improved through the sufficiency criterion only in the cases of the Gini, Shannon and Simpson indexes, but not in the case of the Imbalance Ratio index, for which the identification of the unfairness is improved when measured through the independence or the separation criteria.

We remark that we looked at data imbalance as a risk factor and not as a technical fix, in order to create space for active human considerations and interventions, thus entrusting the ultimate responsibility to human decisions: we strongly recommend keeping in mind this important premise when applying our approach to real cases or further scientific researches.
Indeed, we strongly believe that our results will encourage to take more aware and appropriate actions, as well as to prevent adverse effects caused by the "bias in-bias out" problem: we suggest getting the assistance of domain experts, professionals from human and social science, and impacted stakeholders in selecting the most appropriate balance measures and fairness criteria for the case at hand, in order to fully understand and best address the socio-technical nature of the bias problem in software systems and, particularly, in automated decision-making systems.

References

[1] Lorenz Matzat. Atlas of Automation. https://atlas.algorithmwatch.org/wp-content/uploads/2019/04/Atlas_of_Automation_by_AlgorithmWatch.pdf, 2019.

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

[3] Byron Reese. *The fourth age: Smart robots, conscious computers, and the future of humanity*. Simon and Schuster, April 2018.

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

[5] Bernd W. Wirtz, Jan C. Weyerer, and Carolin Geyer. Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, 42(7):596–615, May 2019. Publisher: Routledge_eprint.

[6] Erik Brynjolfsson and Kristina McElheran. The rapid adoption of data-driven decision-making. *American Economic Review*, 106(5):133–39, May 2016.

[7] L. Willcocks, M. Lacity, and A. Craig. Robotic process automation: Strategic transformation lever for global business services? *Journal of Information Technology Teaching Cases*, 7(1):17–28, 2017.

[8] Ira I. Makrygianni and Angelos 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, page 124–128, New York, NY, USA, 2016. Association for Computing Machinery.

[9] Z. Siting, H. Wenxing, Z. Ning, and Y. Fan. Job recommender systems: A survey. In *2012 7th International Conference on Computer Science Education (ICCSE)*, pages 920–924, 2012.

[10] Doaa Abu Elyounes. 'Computer Says No!': The Impact of Automation on the Discretionary Power of Public Officers. https://papers.ssrn.com/abstract=3692792, September 2020.

[11] Sumitkumar Kanoje, Debajyoti Mukhopadhyay, and Sheetal Girase. User Profiling for University Recommender System Using Automatic Information Retrieval. *Procedia Computer Science*, 78:5–12, January 2016.

[12] Antonio Cordella and Niccolò Tempini. E-government and organizational change: Reappraising the role of ICT and bureaucracy in public service delivery. *Government Information Quarterly*, 32(3):279–286, July 2015.

[13] Jeffrey B. Wenger and Vicky 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):313–333, 02 2008.

[14] Peter André 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*, pages 1–10, January 2017.

[15] Thomas C. Redman. Seizing Opportunity in Data Quality.

[16] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, July 1996.

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

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

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

[20] Ed Pilkington. Digital dystopia: how algorithms punish the poor. https://www.theguardian.com/technology/2019/oct/14/automating-poverty-algorithms-punish-poor, October 2019.

[21] Margrethe 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.

[22] Goce Ristanoski, Wei Liu, and James Bailey. Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1529–1532, New York, NY, USA, 2013. Association for Computing Machinery.

[23] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, page 429–440, New York, NY, USA, 2021. Association for Computing Machinery.

[24] Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009.

[25] Eliza Strickland, Charles Q. Choi, Samuel K. Moore, and Prachi Patel. Dall-e2's failures reveal the limits of ai - openai's text-to-image generator struggles with text, science, and bias. *IEEE Spectrum*, 59(8):5–12, 2022.

[26] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Provider fairness across continents in collaborative recommender systems. *Information Processing & Management*, 59(1):102719, 2022.

[27] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, October 2002.

[28] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, November 2016.

[29] Michael Rovatsos, Brent Mittelstadt, and Ansgar Koene. Landscape summary: Bias in algorithmic decision-making. Technical report, Centre for Data Ethics and Innovation (CDEI), 2019.

[30] European Union Agency 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.

[31] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. https://reut.rs/2Od9fPr, October 2018.

[32] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman 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, page 120–128, New York, NY, USA, 2019. Association for Computing Machinery.

[33] Marcel Pauly. Black box Schufa - Data Journalism Awards. https://datajournalismawards.org/projects/black-box-schufa/, 2019.

[34] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

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

[36] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[37] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine Bias—ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[38] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proc. of the 17th Int. Conf. on Artificial Intelligence and Law*, ICAIL '19, pages 83–92, New York, NY, USA, 2019. Association for Computing Machinery.

[39] Ziad Obermeyer and Sendhil 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, page 89, New York, NY, USA, 2019. Association for Computing Machinery.

[40] Philip Alston. Report of the special rapporteur on extreme poverty and human rights. https://undocs.org/A/74/493, October 2019.

[41] Antonio Vetrò, Marco Torchiano, and Mariachiara Mecati. A data quality approach to the identification of discrimination risk in automated decision making systems. *Government Information Quarterly*, 38(4), 2021.

[42] Mariachiara Mecati, Flavio Emanuele Cannavò, Antonio Vetrò, and Marco Torchiano. Identifying Risks in Datasets for Automated Decision–Making. In Gabriela Viale Pereira, Marijn Janssen, Habin Lee, Ida Lindgren, Manuel Pedro Rodríguez Bolívar, Hans Jochen Scholl, and Anneke Zuiderwijk, editors, *Electronic Government*, Lecture Notes in Computer Science, pages 332–344, Cham, 2020. Springer International Publishing.

[43] Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. Detecting discrimination risk in automated decision-making systems with balance measures on input data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4287–4296. IEEE, 2021.

[44] Mariachiara Mecati, Antonio Vetrò, and Marco Torchiano. Detecting risk of biased output with balance measures. *J. Data and Information Quality*, 14(4), nov 2022.

[45] Mariachiara Mecati, Andrea Adrignola, Antonio Vetrò, and Marco Torchiano. Identifying imbalance thresholds in input data to achieve desired levels of algorithmic fairness. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4700–4709, 2022.

[46] Mariachiara Mecati, Marco Torchiano, Antonio Vetrò, and Juan Carlos De Martin. Measuring imbalance on intersectional protected attributes and on target variable to forecast unfair classifications. *IEEE Access*, 11:26996–27011, 2023.

[47] Antonio Vetrò. Imbalanced data as risk factor of discriminating automated decisions: a measurement-based approach. *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law*, 12(4):272–288, 2021.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Paulo Cortez and Alice Silva. Using Data Mining to Predict Secondary School Student Performance. In *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pages 5–12, Porto, Portugal, April 2008.

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

[54] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. http://arxiv.org/abs/1908.09635, September 2019.

[55] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, July 2017.

[56] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. http://arxiv.org/abs/1609.07236, September 2016.

[57] Jon Kleinberg. Inherent Trade-Offs in Algorithmic Fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '18, page 40, New York, NY, USA, June 2018. Association for Computing Machinery.

[58] Elena Beretta, Antonio Santangelo, Bruno Lepri, Antonio Vetrò, and Juan Carlos De Martin. The Invisible Power of Fairness. How Machine Learning Shapes Democracy. In Marie-Jean Meurs and Frank Rudzicz, editors, *Advances in Artificial Intelligence*, pages 238–250, Cham, 2019. Springer International Publishing.

[59] European Commission and Directorate-General for Communications Networks, Content and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019.

[60] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. Towards a standard for identifying and managing bias in artificial intelligence, 2022-03-15 04:03:00 2022.

[61] ISO. Iso/iec tr 24027:2021 information technology — artificial intelligence (ai) — bias in ai systems and ai aided decision making. https://www.iso.org/standard/77607.html, 2021.

[62] Alessandro Mantelero. *Beyond data: human rights, ethical and social impact assessment in AI*. Number volume 36 in Information technology and law series. Asser Press, Berlin Heidelberg, 2022.

[63] Filippo A Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. Artificial intelligence & human rights: Opportunities & risks. *Berkman Klein Center Research Publication*, (2018-6), 2018.

[64] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, September 2019.

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

[66] Evaggelia Pitoura. Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias. *Journal of Data and Information Quality*, 12(3):12:1–12:8, July 2020.

[67] Ben Hutchinson and Margaret Mitchell. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 49–58, New York, NY, USA, January 2019. Association for Computing Machinery.

[68] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, September 2022.

[69] Florian Königstorfer and Stefan Thalmann. Software documentation is not enough! requirements for the documentation of ai. *Digital Policy, Regulation and Governance*, 2021.

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

[71] Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence, 2022.

[72] Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. Ethical and Socially-Aware Data Labels. In Juan Antonio Lossio-Ventura, Denisse Muñante, and Hugo Alatrista-Salas, editors, *Information Management and Big Data*, pages 320–327, Cham, 2019. Springer International Publishing.

[73] Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. Detecting discriminatory risk through data annotation based on bayesian inferences, 2020.

[74] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[75] M. Bar-Sinai, L. Sweeney, and M. Crosas. Datatags, data handling policy spaces and the tags language. In *2016 IEEE Security and Privacy Workshops (SPW)*, pages 1–8, May 2016.

[76] Michelle Seng Ah Lee and Jatinder Singh. The landscape and gaps in open source fairness toolkits, 2020.

[77] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep 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):4–1, 2019.

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

[79] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, pages 98–108, 2018.

[80] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510, 2017.

[81] Kewen Peng, Joymallya Chakraborty, and Tim Menzies. Fairmask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering*, pages 1–14, 2022.

[82] Joffrey L. Leevy, Taghi M. Khoshgoftaar, Richard A. Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. *J. Big Data*, 5(1):42–30, Nov 2018.

[83] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani,

Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

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

[85] Lisa Bowleg. When black+lesbian+woman≠black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. *Sex roles*, 59(5):312–325, 2008.

[86] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[87] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]. In *Feminist legal theory*, pages 57–80. Routledge, 2018.

[88] Daniel Holman, Sarah Salway, and Andrew Bell. Mapping intersectional inequalities in biomarkers of healthy ageing and chronic disease in older english adults. *Scientific reports*, 10(1):1–12, 2020.

[89] Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[90] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[91] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal Intersectionality and Fair Ranking. In Katrina Ligett and Swati Gupta, editors, *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 7:1–7:20, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[92] Paula Branco, Luis Torgo, and Rita P Ribeiro. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343:76–99, 2019.

[93] Nuno Moniz, Paula Branco, and Luís Torgo. Resampling strategies for imbalanced time series forecasting. *International Journal of Data Science and Analytics*, 3(3):161–181, 2017.

[94] Xiao Yu, Jin Liu, Zijiang Yang, Xiangyang Jia, Qi Ling, and Sizhe Ye. Learning from imbalanced data for predicting the number of software defects. In *2017 IEEE 28th international symposium on software reliability engineering (ISSRE)*, pages 78–89. IEEE, 2017.

[95] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

Appendix A

Thresholds of Balance to Forecast Algorithmic Fairness

A.1 Configurations of the Thresholds

In this Appendix we provide the five configurations of thresholds that we defined during the procedure of identification of risk thresholds provided in Chapter 7. The configurations have been built so as to distribute the values of f evenly in the range with the highest concentration of unfairness values, which is approximately between the minimum and the mean of the distribution (around the first quartile). Hence, for each configuration we specify the two theoretical values of unfairness thresholds that we chose a priori, f1_base and f2_base, or f_base if we are in the case of only one threshold defined a priori. In the five figures, we report a violin plot that represents the compact display of the (continuous) distribution of the values of the Separation criterion in the case of the True Positive rate; indeed, this kind of plot allows to show the probability density of the data at different values. Thus, we took as reference the Sep_TP criterion as it is the one with the largest range of values with respect to the other fairness criteria (which presented very high probability density in correspondence of very low values). In the figures, we color f1_base and f2_base with gray and highlight their mean f in red color (or f_base in the case of the definition of only one threshold), in a way such that the red line is progressively moved to the left (where we observe the highest concentration of unfairness values). In particular: configurations 1, 2 and 4 belong to the general case of the two thresholds f1_base and f2_base defined a priori, whereas configurations 3 and 5 belong to the case with only one f_base threshold.

Configuration 1

- f1_base: *1st quartile*
- f2_base: *mean*

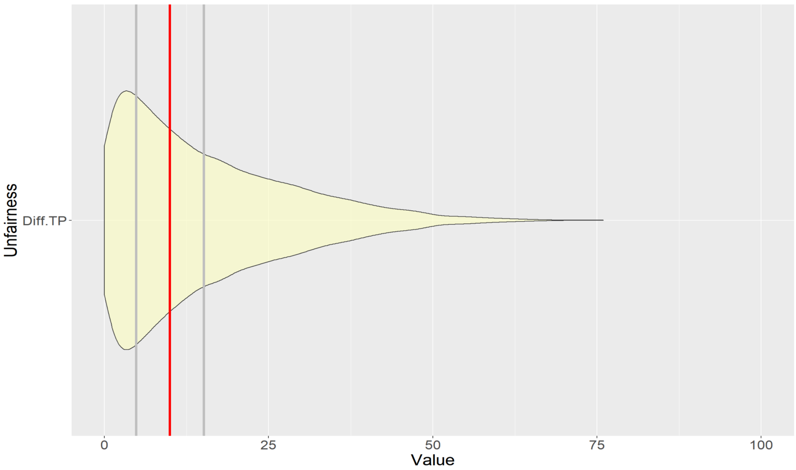


Fig. A.1 Thresholds Configuration 1.

Configuration 2

- f1_base: mean between *minimum* and *1st quartile*
- f2_base: mean between *1st quartile* and *mean*

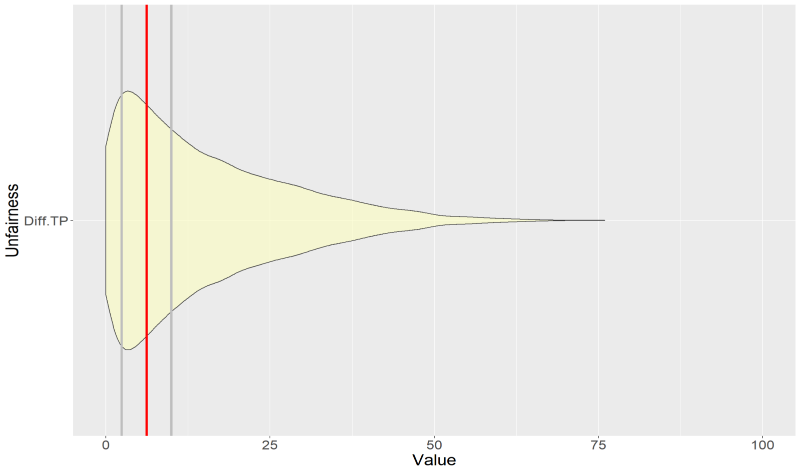


Fig. A.2 Thresholds Configuration 2.

Configuration 3

- f_base: *1st quartile*

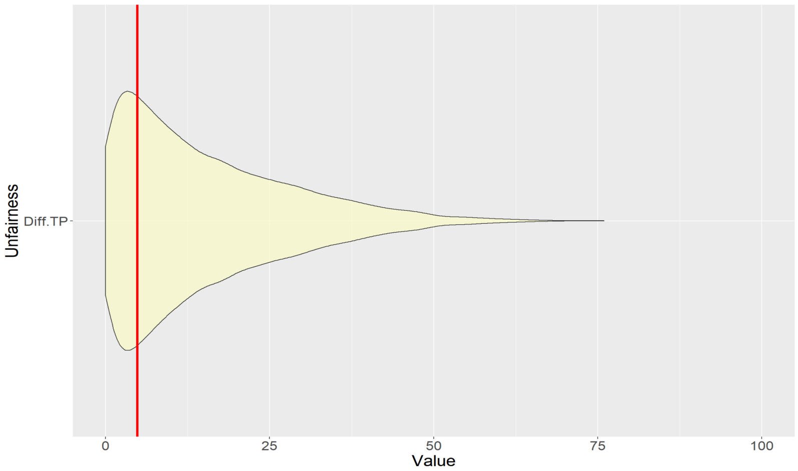


Fig. A.3 Thresholds Configuration 3.

Configuration 4

- f1_base: mean between *minimum* and *1st quartile*
- f2_base: *1st quartile*

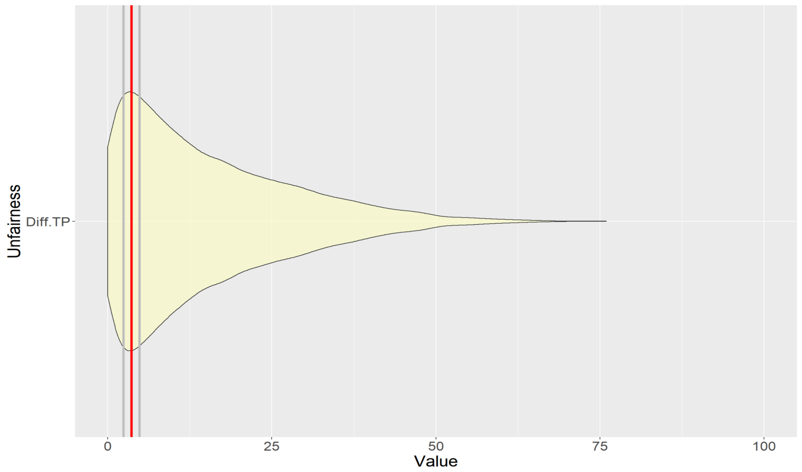


Fig. A.4 Thresholds Configuration 4.

Configuration 5

- f_base: mean between *minimum* and *1st quartile*

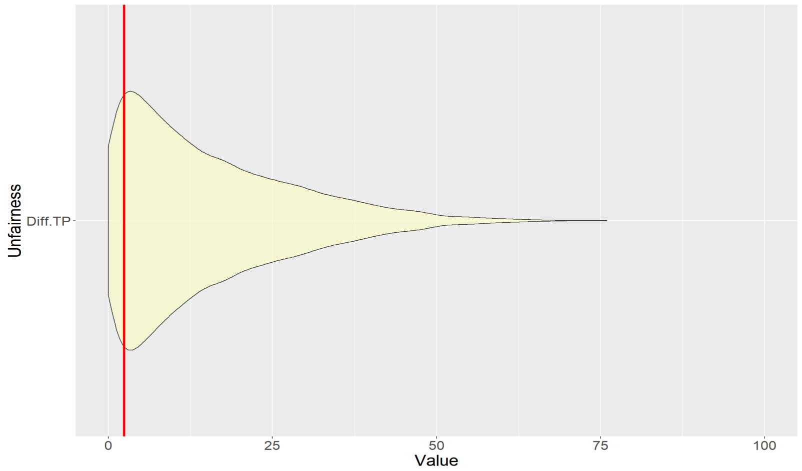


Fig. A.5 Thresholds Configuration 5.

A.2 Final Thresholds and Evaluation Metrics

In this Appendix, we report a set of tables concerning the research outcomes obtained in Chapter 7. Particularly, for each combination of balance-unfairness-algorithm we report the best thresholds selected by accuracy, the configuration they correspond to (among the 5 options described in Appendix A.1), and all the evaluation metrics related to those thresholds. For sake of legibility, we report values for the thresholds of both fairness criteria and balance measures multiplied by 100, i.e. on a scale $[0, 100]$. Results are presented in separate tables for the binary and the multiclass cases, grouped by balance measure (Gini, Shannon, Simpson, and IR indexes), and ordered by unfairness measures (Independence, Separation_TP, Separation_FP, Sufficiency_PP, and Sufficiency_PN criteria); then, in each table results vary according to the algorithm used in the classification task (logistic regression, support vector machine, random forest, k-nearest neighbors). Finally, we remind that the aim of this study was to define two thresholds s (for balance measures) and f (for unfairness measures) such that if the balance of the training set is greater than s, then the unfairness of the classification on the test set is expected to be less than f.

Binary attributes

Gini index

Balance	Unfairness	Algorithm	Configuration	s	f	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Independence	logit	3	97,62	3,20	0,68	0,75	0,21	0,84	0,79
Gini	Independence	svm	5	95,49	1,66	0,68	0,86	0,30	0,75	0,80
Gini	Independence	rf	3	80,59	4,01	0,54	0,76	0,41	0,58	0,66
Gini	Independence	knn	3	96,18	2,33	0,65	0,77	0,21	0,78	0,77

Table A.1 Thresholds and evaluation metrics for the combination Gini-Independence in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	s	f	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Sep_TP	logit	3	94,55	4,72	0,67	0,80	0,33	0,76	0,78
Gini	Sep_TP	svm	3	79,29	3,99	0,63	0,86	0,56	0,64	0,73
Gini	Sep_TP	rf	5	99,99	3,31	0,85	0,88	0,07	0,96	0,92
Gini	Sep_TP	knn	5	99,96	2,06	0,80	0,85	0,11	0,93	0,89

Table A.2 Thresholds and evaluation metrics for the combination Gini-Sep_TP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Sep_FP	logit	3	84,68	2,89	0,61	0,76	0,42	0,68	0,72
Gini	Sep_FP	svm	1	38,44	6,77	0,49	0,48	0,58	0,40	0,44
Gini	Sep_FP	rf	1	54,58	7,875	0,61	0,55	0,64	0,59	0,57
Gini	Sep_FP	knn	5	91,50	0,91	0,90	0,74	0,35	0,69	0,81

Table A.3 Thresholds and evaluation metrics for the combination Gini-Sep_FP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Suf_PP	logit	5	72,02	2,50	0,58	0,89	0,50	0,59	0,71
Gini	Suf_PP	svm	5	85,00	2,16	0,67	0,91	0,46	0,70	0,79
Gini	Suf_PP	rf	4	96,53	4,46	0,74	0,82	0,29	0,86	0,84
Gini	Suf_PP	knn	3	96,27	6,10	0,66	0,71	0,26	0,85	0,77

Table A.4 Thresholds and evaluation metrics for the combination Gini-Suf_PP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Suf_PN	logit	5	95,77	2,17	0,71	0,88	0,21	0,78	0,82
Gini	Suf_PN	svm	5	95,80	1,92	0,72	0,88	0,25	0,79	0,83
Gini	Suf_PN	rf	1	67,60	7,84	0,62	0,60	0,53	0,71	0,65
Gini	Suf_PN	knn	5	96,41	2,47	0,73	0,85	0,25	0,82	0,84

Table A.5 Thresholds and evaluation metrics for the combination Gini-Suf_PN in the case of binary attributes.

Shannon index

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Independence	logit	3	98,28	3,20	0,68	0,75	0,21	0,84	0,79
Shannon	Independence	svm	5	96,72	1,66	0,68	0,86	0,30	0,75	0,80
Shannon	Independence	rf	3	85,51	4,01	0,54	0,76	0,41	0,58	0,66
Shannon	Independence	knn	3	97,232,33	2,33	0,65	0,77	0,21	0,78	0,77

Table A.6 Thresholds and evaluation metrics for the combination Shannon-Independence in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Sep_TP	logit	3	96,03	4,72	0,67	0,80	0,33	0,76	0,78
Shannon	Sep_TP	svm	3	83,03	3,99	0,62	0,87	0,58	0,63	0,73
Shannon	Sep_TP	rf	5	99,99	99,99	0,85	0,88	0,08	0,95	0,92
Shannon	Sep_TP	knn	5	99,97	2,06	0,80	0,85	0,12	0,93	0,89

Table A.7 Thresholds and evaluation metrics for the combination Shannon-Sep_TP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Sep_FP	logit	3	88,65	2,89	0,61	0,76	0,41	0,68	0,72
Shannon	Sep_FP	svm	1	46,58	6,77	0,49	0,49	0,63	0,36	0,41
Shannon	Sep_FP	rf	1	63,19	7,88	0,61	0,55	0,64	0,58	0,57
Shannon	Sep_FP	knn	5	93,68	0,91	0,69	0,90	0,35	0,74	0,81

Table A.8 Thresholds and evaluation metrics for the combination Shannon-Sep_FP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Suf_PP	logit	5	77,06	2,50	0,57	0,89	0,50	0,58	0,70
Shannon	Suf_PP	svm	5	88,89	2,16	0,67	0,91	0,45	0,910,69	0,79
Shannon	Suf_PP	rf	4	97,48	4,46	0,74	0,82	0,29	0,86	0,84
Shannon	Suf_PP	knn	3	97,30	6,10	0,66	0,71	0,26	0,84	0,77

Table A.9 Thresholds and evaluation metrics for the combination Shannon-Suf_PP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Suf_PN	logit	5	96,93	2,17	0,71	0,88	0,21	0,78	0,82
Shannon	Suf_PN	svm	5	96,90	1,92	0,72	0,88	0,25	0,78	0,83
Shannon	Suf_PN	rf	1	73,07	7,84	0,62	0,61	0,56	0,68	0,64
Shannon	Suf_PN	knn	5	97,39	2,47	0,73	0,85	0,25	0,82	0,84

Table A.10 Thresholds and evaluation metrics for the combination Shannon-Suf_PN in the case of binary attributes.

Simpson index

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Independence	logit	3	95,35	3,20	0,68	0,75	0,21	0,84	0,79
Simpson	Independence	svm	5	91,37	1,66	0,68	0,86	0,30	0,75	0,80
Simpson	Independence	rf	3	67,50	4,01	0,54	0,76	0,41	0,58	0,66
Simpson	Independence	knn	3	92,65	2,33	0,65	0,77	0,21	0,770,78	0,77

Table A.11 Thresholds and evaluation metrics for the combination Simpson-Independence in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Sep_TP	logit	3	89,66	4,72	0,67	0,80	0,33	0,76	0,78
Simpson	Sep_TP	svm	3	73,11	3,99	0,65	0,85	0,47	0,70	0,77
Simpson	Sep_TP	rf	5	99,99	3,31	0,85	0,88	0,04	0,96	0,92
Simpson	Sep_TP	knn	5	99,93	2,06	0,80	0,85	0,11	0,93	0,89

Table A.12 Thresholds and evaluation metrics for the combination Simpson-Sep_TP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Sep_FP	logit	3	73,43	2,89	0,61	0,76	0,41	0,68	0,72
Simpson	Sep_FP	svm	1	29,43	6,77	0,49	0,49	0,56	0,43	0,46
Simpson	Sep_FP	rf	1	40,32	7,88	0,61	0,55	0,63	0,59	0,57
Simpson	Sep_FP	knn	5	85,47	0,91	0,70	0,90	0,35	0,74	0,81

Table A.13 Thresholds and evaluation metrics for the combination Simpson-Sep_FP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Suf_PP	logit	5	63,82	2,50	0,58	0,89	0,48	0,60	0,72
Simpson	Suf_PP	svm	5	73,91	2,16	0,67	0,91	0,45	0,69	0,79
Simpson	Suf_PP	rf	4	93,36	4,46	0,74	0,82	0,29	0,86	0,84
Simpson	Suf_PP	knn	3	92,82	6,10	0,66	0,71	0,26	0,84	0,77

Table A.14 Thresholds and evaluation metrics for the combination Simpson-Suf_PP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Suf_PN	logit	5	91,88	2,17	0,71	0,88	0,21	0,78	0,82
Simpson	Suf_PN	svm	5	92,54	1,92	0,74	0,88	0,23	0,81	0,84
Simpson	Suf_PN	rf	1	60,61	7,84	0,63	0,61	0,52	0,74	0,67
Simpson	Suf_PN	knn	5	93,07	2,47	0,73	0,85	0,25	0,82	0,84

Table A.15 Thresholds and evaluation metrics for the combination Simpson-Suf_PN in the case of binary attributes.

Imbalance Ratio index

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Independence	logit	3	73,27	3,20	0,68	0,75	0,21	0,84	0,79
IR	Independence	svm	5	64,96	1,66	0,68	0,86	0,30	0,75	0,80
IR	Independence	rf	3	38,84	4,01	0,54	0,76	0,41	0,58	0,66
IR	Independence	knn	3	67,32	2,33	0,65	0,77	0,21	0,78	0,77

Table A.16 Thresholds and evaluation metrics for the combination IR-Independence in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Sep_TP	logit	3	62,14	4,72	0,67	0,80	0,33	0,76	0,78
IR	Sep_TP	svm	4	51,20	3,00	0,70	0,89	0,43	0,74	0,81
IR	Sep_TP	rf	5	98,29	3,31	0,85	0,88	0,04	0,96	0,92
IR	Sep_TP	knn	5	96,28	2,06	0,80	0,85	0,11	0,93	0,89

Table A.17 Thresholds and evaluation metrics for the combination IR-Sep_TP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Sep_FP	logit	3	43,74	2,89	0,61	0,76	0,41	0,68	0,72
IR	Sep_FP	svm	1	19,31	6,77	0,49	0,49	0,56	0,43	0,46
IR	Sep_FP	rf	1	22,13	7,88	0,62	0,55	0,63	0,60	0,57
IR	Sep_FP	knn	5	70,24	0,91	0,76	0,89	0,19	0,84	0,86

Table A.18 Thresholds and evaluation metrics for the combination IR-Sep_FP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Suf_PP	logit	5	43,87	2,50	0,65	0,89	0,43	0,68	0,77
IR	Suf_PP	svm	5	44,16	2,16	0,67	0,91	0,45	0,69	0,79
IR	Suf_PP	rf	4	70,25	4,46	0,74	0,82	0,26	0,87	0,84
IR	Suf_PP	knn	3	67,64	6,10	0,66	0,71	0,26	0,84	0,77

Table A.19 Thresholds and evaluation metrics for the combination IR-Suf_PP in the case of binary attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Suf_PN	logit	5	65,88	2,17	0,71	0,88	0,21	0,78	0,82
IR	Suf_PN	svm	5	82,40	1,92	0,78	0,88	0,16	0,87	0,88
IR	Suf_PN	rf	1	52,13	7,84	0,58	0,55	0,35	0,81	0,66
IR	Suf_PN	knn	5	8,13	2,47	0,73	0,85	0,25	0,82	0,84

Table A.20 Thresholds and evaluation metrics for the combination IR-Suf_PN in the case of binary attributes.

Multiclass attributes

Gini index

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Independence	logit	1	88,54	8,45	0,44	0,43	0,55	0,33	0,38
Gini	Independence	svm	1	93,73	10,08	0,42	0,40	0,37	0,47	0,43
Gini	Independence	rf	5	91,09	3,61	0,47	0,92	0,50	0,47	0,62
Gini	Independence	knn	5	93,71	1,69	0,52	0,86	0,38	0,55	0,67

Table A.21 Thresholds and evaluation metrics for the combination Gini-Independence in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Sep_TP	logit	1	92,23	10,04	0,51	0,65	0,54	0,49	0,56
Gini	Sep_TP	svm	2	94,64	8,21	0,56	0,80	0,44	0,59	0,68
Gini	Sep_TP	rf	5	99,83	4,55	0,77	0,89	0,14	0,84	0,87
Gini	Sep_TP	knn	1	93,50	9,98	0,53	0,64	0,55	0,52	0,57

Table A.22 Thresholds and evaluation metrics for the combination Gini-Sep_TP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Sep_FP	logit	1,00	87,58	8,32	0,43	0,53	0,56	0,34	0,41
Gini	Sep_FP	svm	5,00	92,02	1,46	0,48	0,96	0,51	0,48	0,63
Gini	Sep_FP	rf	1,00	91,96	9,55	0,48	0,55	0,50	0,47	0,51
Gini	Sep_FP	knn	1,00	94,10	8,79	0,47	0,49	0,35	0,58	0,53

Table A.23 Thresholds and evaluation metrics for the combination Gini-Sep_FP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Suf_PP	logit	1	87,96	13,29	0,48	0,56	0,59	0,40	0,46
Gini	Suf_PP	svm	1	89,61	11,99	0,47	0,57	0,56	0,41	0,47
Gini	Suf_PP	rf	5	95,16	4,85	0,59	0,93	0,34	0,61	0,73
Gini	Suf_PP	knn	1	86,83	12,88	0,55	0,60	0,72	0,39	0,47

Table A.24 Thresholds and evaluation metrics for the combination Gini-Suf_PP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Gini	Suf_PN	logit	1	86,67	12,04	0,48	0,53	0,69	0,30	0,38
Gini	Suf_PN	svm	3	98,92	7,60	0,59	0,80	0,31	0,66	0,72
Gini	Suf_PN	rf	2	94,68	7,72	0,52	0,76	0,35	0,57	0,65
Gini	Suf_PN	knn	2	93,05	8,58	0,53	0,87	0,62	0,51	0,64

Table A.25 Thresholds and evaluation metrics for the combination Gini-Suf_PN in the case of multiclass attributes.

Shannon index

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Independence	logit	3	84,89	5,19	0,48	0,76	0,54	0,46	0,57
Shannon	Independence	svm	3	85,52	6,54	0,43	0,68	0,42	0,44	0,53
Shannon	Independence	rf	5	87,27	3,61	0,55	0,90	0,27	0,58	0,70
Shannon	Independence	knn	5	86,10	1,69	0,53	0,86	0,36	0,55	0,67

Table A.26 Thresholds and evaluation metrics for the combination Shannon-Independence in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Sep_TP	logit	4	87,50	5,14	0,57	0,82	0,41	0,60	0,69
Shannon	Sep_TP	svm	2	93,80	8,21	0,60	0,79	0,35	0,66	0,72
Shannon	Sep_TP	rf	5	99,79	4,55	0,79	0,89	0,14	0,87	0,88
Shannon	Sep_TP	knn	1	92,80	9,98	0,54	0,61	0,33	0,67	0,64

Table A.27 Thresholds and evaluation metrics for the combination Shannon-Sep_TP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Sep_FP	logit	1	83,97	8,32	0,47	0,57	0,51	0,44	0,49
Shannon	Sep_FP	svm	5	89,82	1,46	0,58	0,95	0,23	0,60	0,73
Shannon	Sep_FP	rf	3	86,44	4,36	0,52	0,74	0,33	0,58	0,65
Shannon	Sep_FP	knn	3	87,54	4,23	0,52	0,71	0,31	0,59	0,64

Table A.28 Thresholds and evaluation metrics for the combination Shannon-Sep_FP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Suf_PP	logit	1	84,93	13,29	0,53	0,61	0,57	0,50	0,55
Shannon	Suf_PP	svm	2	84,67	8,20	0,52	0,81	0,58	0,50	0,62
Shannon	Suf_PP	rf	2	93,24	8,72	0,61	0,80	0,34	0,68	0,74
Shannon	Suf_PP	knn	1	82,46	12,88	0,56	0,60	0,65	0,48	0,53

Table A.29 Thresholds and evaluation metrics for the combination Shannon-Suf_PP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Shannon	Suf_PN	logit	1	81,04	12,04	0,46	0,50	0,64	0,31	0,38
Shannon	Suf_PN	svm	3	97,16	7,60	0,59	0,80	0,31	0,66	0,72
Shannon	Suf_PN	rf	2	94,02	7,72	0,59	0,79	0,34	0,66	0,72
Shannon	Suf_PN	knn	2	89,19	8,58	0,59	0,85	0,46	0,61	0,71

Table A.30 Thresholds and evaluation metrics for the combination Shannon-Suf_PN in the case of multiclass attributes.

Simpson index

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Independence	logit	1	62,19	8,45	0,50	0,51	0,54	0,47	0,49
Simpson	Independence	svm	3	66,97	6,54	0,43	0,68	0,42	0,44	0,53
Simpson	Independence	rf	4	66,00	5,42	0,47	0,85	0,50	0,47	0,60
Simpson	Independence	knn	5	74,88	1,69	0,53	0,86	0,36	0,55	0,67

Table A.31 Thresholds and evaluation metrics for the combination Simpson-Independence in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	s	f	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Sep_TP	logit	3	77,24	6,85	0,55	0,76	0,39	0,60	0,67
Simpson	Sep_TP	svm	2	84,92	8,21	0,60	0,79	0,35	0,66	0,72
Simpson	Sep_TP	rf	5	99,14	4,55	0,79	0,89	0,15	0,87	0,88
Simpson	Sep_TP	knn	1	83,05	9,98	0,54	0,62	0,40	0,63	0,62

Table A.32 Thresholds and evaluation metrics for the combination Simpson-Sep_TP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	s	f	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Sep_FP	logit	1	63,30	8,32	0,48	0,58	0,51	0,47	0,51
Simpson	Sep_FP	svm	5	76,82	1,46	0,58	0,95	0,23	0,59	0,73
Simpson	Sep_FP	rf	3	69,14	4,36	0,48	0,76	0,50	0,48	0,59
Simpson	Sep_FP	knn	3	73,71	4,23	0,48	0,73	0,48	0,48	0,58

Table A.33 Thresholds and evaluation metrics for the combination Simpson-Sep_FP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	s	f	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Suf_PP	logit	1	64,35	13,29	0,53	0,61	0,57	0,51	0,55
Simpson	Suf_PP	svm	3	66,23	8,84	0,52	0,78	0,57	0,50	0,61
Simpson	Suf_PP	rf	2	87,44	8,72	0,61	0,80	0,34	0,68	0,74
Simpson	Suf_PP	knn	1	60,97	12,88	0,56	0,58	0,57	0,55	0,56

Table A.34 Thresholds and evaluation metrics for the combination Simpson-Suf_PP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	s	f	Accuracy	Precision	Specificity	Sensitivity	F1-score
Simpson	Suf_PN	logit	1	53,71	12,04	0,47	0,51	0,71	0,27	0,35
Simpson	Suf_PN	svm	4	84,02	5,70	0,64	0,94	0,28	0,66	0,77
Simpson	Suf_PN	rf	2	85,30	7,72	0,59	0,79	0,34	0,66	0,72
Simpson	Suf_PN	knn	3	68,87	9,54	0,53	0,81	0,59	0,51	0,63

Table A.35 Thresholds and evaluation metrics for the combination Simpson-Suf_PN in the case of multiclass attributes.

Imbalance Ratio index

Balance	Unfairness	Algorithm	Configuration	s	f	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Independence		3	18,61	5,19	0,59	0,76	0,35	0,67	0,71
IR	Independence	svm	4	10,92	4,91	0,56	0,83	0,27	0,61	0,70
IR	Independence	rf	5	32,69	3,61	0,63	0,91	0,27	0,66	0,77
IR	Independence	knn	4	12,08	2,54	0,57	0,77	0,30	0,64	0,70

Table A.36 Thresholds and evaluation metrics for the combination IR-Independence in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Sep_TP	logit	5	24,12	3,42	0,65	0,89	0,41	0,68	0,77
IR	Sep_TP	svm	5	42,71	4,44	0,64	0,91	0,36	0,66	0,77
IR	Sep_TP	rf	5	80,85	4,55	0,80	0,89	0,13	0,88	0,89
IR	Sep_TP	knn	4	22,47	4,86	0,61	0,82	0,34	0,67	0,74

Table A.37 Thresholds and evaluation metrics for the combination IR-Sep_TP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Sep_FP	logit	2	16,07	4,83	0,56	0,68	0,32	0,67	0,67
IR	Sep_FP	svm	5	56,18	1,46	0,65	0,95	0,23	0,67	0,78
IR	Sep_FP	rf	4	20,81	3,27	0,64	0,88	0,34	0,68	0,77
IR	Sep_FP	knn	4	18,33	3,17	0,59	0,79	0,28	0,88	0,72

Table A.38 Thresholds and evaluation metrics for the combination IR-Sep_FP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Suf_PP	logit	1	30,65	13,29	0,54	0,58	0,35	0,68	0,63
IR	Suf_PP	svm	4	15,57	6,63	0,63	0,86	0,35	0,68	0,76
IR	Suf_PP	rf	5	18,36	4,85	0,66	0,94	0,34	0,68	0,79
IR	Suf_PP	knn	3	16,39	9,40	0,58	0,72	0,34	0,68	0,70

Table A.39 Thresholds and evaluation metrics for the combination IR-Suf_PP in the case of multiclass attributes.

Balance	Unfairness	Algorithm	Configuration	*s*	*f*	Accuracy	Precision	Specificity	Sensitivity	F1-score
IR	Suf_PN	logit	2	17,71	8,11	0,59	0,79	0,30	0,66	0,72
IR	Suf_PN	svm	4	25,90	5,70	0,64	0,94	0,28	0,66	0,77
IR	Suf_PN	rf	5	46,74	4,11	0,65	0,95	0,31	0,66	0,78
IR	Suf_PN	knn	4	19,22	7,16	0,63	0,88	0,35	0,67	0,76

Table A.40 Thresholds and evaluation metrics for the combination IR-Suf_PN in the case of multiclass attributes.

Appendix B

Imbalance of Intersectional Protected Attributes and Target Variable

As a complement to the discussion of the research question RQ 6 presented in Section 8.4 of Chapter 8, in this Appendix we report the correlations between balance and unfairness measures for both protected attributes combined with the target variable and protected attributes without target (in Tables B.1, B.2, B.4, B.5, B.7, B.8). We also report the differences between the aforementioned correlations, for the protected attributes sex, education and sex_education (in Tables B.3, B.6, B.9). For the sake of better interpretability of the numerical values in the tables, we make the following specification: as we expect the correlation between balance measures and fairness criteria to be negative for a given protected attribute, we assess the difference between the correlation of a protected attribute (primary or intersectional) and the correlation of the same attribute combined with the target. If this difference is positive, it means that the correlation between balance measures and fairness criteria for that protected attribute combined *with the target variable* is stronger than the correlation obtained *without* combining the protected attribute with the target variable, thus adding the target will improve the unfairness detection. On the contrary, if the difference is negative, the combination of the protected attribute with the target variable does *not* improve identifying the unfairness.

Indeed, the numerical results reflect all the observations about the plot in Figure 8.3: as regards tables reporting the differences in correlation, we have a positive value if combining the target with protected attributes improves detecting the unfairness, while we find a negative value if the combination of the target variable with protected attributes worsens the identification of the unfairness.

Table B.1 Correlation between balance and unfairness measures for the primary attribute *sex*: $\mathfrak{B}(\text{sex}) \sim \mathfrak{U}(\text{sex})$.

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.1417	-0.1417	-0.1417	-0.1416
Separation – TP	-0.4017	-0.4018	-0.4017	-0.4017
Separation – FP	-0.1509	-0.1510	-0.1509	-0.1509
Sufficiency – PP	-0.2801	-0.2802	-0.2802	-0.2801
Sufficiency – PN	-0.0085	-0.0085	-0.0084	-0.0084

Table B.2 Correlation between balance and unfairness measures for the attribute *sex_target*: $\mathfrak{B}(\text{sex\_target}) \sim \mathfrak{U}(\text{sex})$.

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.0388	-0.0665	-0.0388	-0.1056
Separation – TP	-0.3119	-0.3513	-0.3119	-0.3999
Separation – FP	-0.0035	-0.0420	-0.0035	-0.1024
Sufficiency – PP	-0.3443	-0.3422	-0.3443	-0.3075
Sufficiency – PN	-0.0664	-0.0609	-0.0664	-0.0527

Table B.3 Difference between the correlation tables B.1 and B.2:
$\text{diff}_{sex} = \text{cor}(sex) - \text{cor}(sex\_target)$.

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.1029	-0.0752	-0.1029	-0.0360
Separation – TP	-0.0898	-0.0504	-0.0898	-0.0018
Separation – FP	-0.1474	-0.1089	-0.1474	-0.0485
Sufficiency – PP	0.0641	0.0620	0.0641	0.0274
Sufficiency – PN	0.0579	0.0525	0.0580	0.0444

Table B.4 Correlation between balance and unfairness measures for the primary attribute *education*: $\mathfrak{B}$(education) $\sim$ $\mathfrak{U}$(education).

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	-0.0885	-0.0581	-0.0885	-0.0044
Separation – TP	-0.0608	-0.0165	-0.0608	0.0411
Separation – FP	-0.1013	-0.0637	-0.1014	-0.0106
Sufficiency – PP	-0.1515	-0.0647	-0.1516	0.0462
Sufficiency – PN	0.0200	-0.0601	0.0201	-0.1049

Table B.5 Correlation between balance and unfairness measures for the attribute *education_target*: $\mathfrak{B}$(education_target) $\sim$ $\mathfrak{U}$(education).

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	0.2358	0.2413	0.2762	0.1199
Separation – TP	0.2181	0.2244	0.2544	0.1018
Separation – FP	0.2348	0.2370	0.2753	0.0976
Sufficiency – PP	-0.2679	-0.2882	-0.3065	-0.2103
Sufficiency – PN	-0.2941	-0.3145	-0.3260	-0.1325

Table B.6 Difference between the correlation tables B.4 and B.5:
$\text{diff}_{education}$ = cor(*education*) – cor(*education_target*).

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	-0.3243	-0.2994	-0.3647	-0.1243
Separation – TP	-0.2789	-0.2409	-0.3152	-0.0607
Separation – FP	-0.3361	-0.3008	-0.3767	-0.1082
Sufficiency – PP	0.1164	0.2235	0.1549	0.2565
Sufficiency – PN	0.3141	0.2544	0.3461	0.0276

Table B.7 Correlation between balance and unfairness measures for the intersectionl attribute *sex_education*: $\mathfrak{B}$(sex_education) $\sim$ $\mathfrak{U}$(sex_education).

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	-0.1614	-0.1652	-0.1687	-0.1706
Separation – TP	-0.2583	-0.2799	-0.2696	-0.2902
Separation – FP	-0.2130	-0.2244	-0.2203	-0.2340
Sufficiency – PP	-0.1836	-0.1842	-0.1905	-0.1862
Sufficiency – PN	-0.1487	-0.1654	-0.1425	-0.1631

Table B.8 Correlation between balance and unfairness measures for the attribute *sex_education_target*: $\mathfrak{B}$(sex_education_target) $\sim$ $\mathfrak{U}$(sex_education).

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	0.0103	-0.0275	0.0107	-0.2666
Separation – TP	-0.0993	-0.1552	-0.1028	-0.3832
Separation – FP	-0.0450	-0.0930	-0.0443	-0.3266
Sufficiency – PP	-0.2559	-0.2441	-0.2706	-0.1100
Sufficiency – PN	-0.3000	-0.2934	-0.2935	-0.1069

Table B.9 Difference between the correlation tables B.7 and B.8:
$\text{diff}_{sex\_education}$ = cor(*sex_education*) – cor(*sex_education_target*).

Balance Measures **Fairness criteria**	**Gini**	**Shannon**	**Simpson**	**Imbalance Ratio**
Independence	-0.1717	-0.1376	-0.1794	0.0960
Separation – TP	-0.1590	-0.1247	-0.1667	0.0930
Separation – FP	-0.1680	-0.1314	-0.1760	0.0926
Sufficiency – PP	0.0723	0.0600	0.0800	-0.0762
Sufficiency – PN	0.1513	0.1279	0.1510	-0.0562

Appendix C

Publication List

The studies discussed in this dissertation have been published in the following conference proceedings or journals, starting from the most recent in descending date order.

- Mecati, M., Torchiano, M., Vetrò, A., De Martin, J. C. (2023). *Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications*. IEEE Access.
https://dx.doi.org/10.1109/ACCESS.2023.3252370

- Mecati, M., Adrignola, A., Vetrò, A., Torchiano, M. (2022). *Identifying Imbalance Thresholds in Input Data to Achieve Desired Levels of Algorithmic Fairness*. In: 2022 IEEE International Conference on Big Data (BigData).
https://doi.org/10.1109/BigData55660.2022.10021078

- Mecati, M., Adrignola, A., Vetrò, A., Torchiano, M. (2022). *Appendix for "Identifying Imbalance Thresholds in Input Data to Achieve Desired Levels of Algorithmic Fairness"*. Zenodo – general-purpose open repository.
https://dx.doi.org/10.5281/ZENODO.7350599

- Mecati, M., Vetrò, A., Torchiano, M. (2022). *Detecting Risk of Biased Output with Balance Measures*. ACM Journal of Data and Information Quality (JDIQ).
https://dl.acm.org/doi/10.1145/3530787

- Mecati, M., Vetrò, A., Torchiano, M. (2021). *Detecting Discrimination Risk in Automated Decision-Making Systems with Balance Measures on Input Data*.

In: 2021 IEEE International Conference on Big Data (BigData). https://doi.org/10.1109/BigData52589.2021.9671443

- Vetrò, A., Torchiano, M., Mecati, M. (2021). *A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision Making Systems*. Government Information Quarterly (GIQ). https://doi.org/10.1016/j.giq.2021.101619

- Vetrò, A., Torchiano, M., Mecati, M. (2021). *Reproducibility package for "A Data Quality Approach to the Identification of Discrimination Risk in Automated Decision Making Systems"*. Code Ocean – reproducibility platform. https://doi.org/10.24433/CO.0509748.v3

- Mecati, M., Cannavò, F. E., Vetrò, A., Torchiano, M. (2020). *Identifying Risks in Datasets for Automated Decision–Making*. In: Electronic Government conference (EGOV2020): Lecture Notes in Computer Science. Best Paper Award. https://doi.org/10.1007/978-3-030-57599-1_25

- Mecati, M., Cannavò, F. E., Vetrò, A., Torchiano, M. (2020). *Reproducibility package for "Identifying Risks in Datasets for Automated Decision–Making"*. Code Ocean – reproducibility platform. https://doi.org/10.24433/CO.5067135.v2

- Labate, D., Pahari, B. R., Hoteit, S., Mecati, M. (2020). *Quantitative Methods in Ocular Fundus Imaging: Analysis of Retinal Microvasculature*. In: Landscapes of Time-Frequency Analysis. Applied and Numerical Harmonic Analysis (ATFA19). https://doi.org/10.1007/978-3-030-56005-8_9

- Mecati, M. (2020). *Different shades of "FAIRNESS" in automated decision-making*. Blog article in the webpage of the "Institute for Internet & the Just Society", Berlin, GERMANY. https://www.internetjustsociety.org/different-shades-of-fairness-in-automated-decision-making