

Examining the Impact of Bias Mitigation Algorithms on the Sustainability of ML-enabled Systems: A Benchmark Study

Vincenzo De Martino, Gianmario Voria, Ciro Troiano, Gemma Catolino, Fabio Palomba

Software Engineering (SeSa) Lab — University of Salerno, Salerno (Italy)

Abstract

Context: As machine learning (ML) systems become increasingly prevalent across various industries, concerns regarding fairness have intensified. Bias mitigation algorithms—that aim to reduce bias in ML models—serve as solutions to mitigate this issue. However, these techniques can affect more than just social sustainability. They may alter the computational overhead and energy usage of ML systems, affecting their environmental sustainability. Similarly, they can influence businesses’ economic sustainability by shaping resource allocation and consumer trust. **Goal:** This work aims to provide a benchmark study of the implications of applying bias mitigation algorithms on the sustainability of ML solutions. We first corroborate previous findings by examining their effect on social sustainability metrics. Additionally, we complement existing studies by offering a comprehensive analysis of how bias mitigation affects environmental and economic sustainability, aiming to highlight trade-offs for practitioners designing ML solutions. **Method:** We evaluate six bias mitigation algorithms by conducting 3,360 experiments across multiple configurations of four ML algorithms and datasets. From these experiments, we compute metrics for social, environmental, and economic sustainability, evaluating them using statistical analysis. **Results:** Our quantitative findings show that all bias mitigation algorithms affect the three sustainability dimensions differently, indicating that applying these algorithms involves complex trade-offs. Furthermore, we expand our discussion with qualitative insights that arise from our results, also providing implications for both research and practice.

Conclusions: Our study emphasizes the need for a deeper investigation into the trade-offs bias mitigation algorithms introduce and how they impact various non-functional requirements of ML systems.

Keywords: Software Sustainability, Machine Learning-Enabled Systems, Software Engineering for Artificial Intelligence.

1. Introduction

The advancements in machine learning (ML) have stimulated transformative changes across various domains, including science, medicine, finance, and education [84]. However, alongside its widespread adoption, concerns about fairness and equity have emerged as pivotal issues in the design and development of ML-enabled systems [62]. Numerous real-world ML deployments have exhibited discriminatory behavior linked to protected attributes such as gender, race, and age [15, 12, 62, 28]. These challenges have prompted the research community to actively pursue development practices aimed at reducing bias and promoting fairness in ML-enabled systems, reflecting a commitment to ethical and responsible software engineering [15, 12, 62, 28]. Within software engineering (SE), the subfield of SE for Artificial Intelligence (SE4AI) recognizes fairness as a critical non-functional requirement,¹ particularly in ML-enabled systems, i.e., software systems that actively include at least one ML component [59, 42, 60]. Prior research has focused extensively on addressing bias and enhancing fairness through the development of bias mitigation algorithms. These algorithms often target protected attributes—demographic characteristics identified by

Email address: {vdemartino, gvorio, c.troiano17, gcatolino, fpalomba}@unisa.it (Vincenzo De Martino, Gianmario Voria, Ciro Troiano, Gemma Catolino, Fabio Palomba)

¹A non-functional requirement specifies criteria to judge a system’s operation rather than its specific behaviors [9].

legal, ethical, and societal frameworks to safeguard marginalized groups from systemic discrimination or inequitable treatment [2, 39]. Examples include pre-processing algorithms like FAIR-SMOTE [19], ensemble-based methods such as MAAT [25], and hybrid approaches like FAIRWAY [20]. These studies have primarily targeted the optimization of the trade-off between fairness and accuracy, demonstrating that bias mitigation can improve equity in ML systems while preserving the overall performance of the models.

Despite these advancements, much of the existing literature evaluates fairness as an *isolated* objective, focusing predominantly on metrics like demographic parity or equalized odds. However, even if these practices have been successful in mitigating discrimination, their broader implications, particularly how fairness techniques might influence non-functional attributes such as environmental and economic sustainability, remain unexplored. More specifically, bias mitigation methods may introduce computational overhead during training and deployment, which can increase energy consumption and carbon emissions, thereby contributing to the environmental footprint of ML models. Similarly, these methods may affect economic factors such as training time and storage requirements, potentially impacting the financial feasibility of deploying fairness-enhanced systems. These trade-offs, while critical to achieving responsible AI, remain poorly understood in the literature.

Our work addresses these gaps by *framing fairness within the broader theoretical construct of sustainability*, encompassing its social, environmental, and economic dimensions. Building on the conceptualization of sustainability as a stratified and multi-systemic construct [61], we operationalize fairness metrics as indicators of *social sustainability*, connecting them to societal goals such as equity, inclusiveness, and well-being. This framing allows us to evaluate bias mitigation algorithms not merely as tools for addressing fairness concerns but as components of sustainable software systems that balance equity with other non-functional requirements. Beyond fairness, we extend the evaluation of bias mitigation techniques to include two additional sustainability dimensions. On the one hand, we evaluate *environmental sustainability*, quantified through metrics such as energy consumption and carbon emissions. On the other hand, we expand the discussion on the trade-off between fairness and accuracy by exploring other *economic sustainability* measures such as training time and storage weight. Through the integration of these dimensions, we provide a more comprehensive understanding of the trade-offs involved in adopting bias mitigation techniques.

To the best of our knowledge, only a few recent studies have begun advocating for a broader view of fairness in ML, highlighting the importance of considering its trade-offs with other sustainability dimensions [16, 21], for instance by proposing multi-objective algorithms that attempt to optimize fairness alongside efficiency-related or environmental objectives. Compared to these earlier works, our study offers a large-scale, systematic benchmark that quantitatively examines the impact of bias mitigation algorithms on three key dimensions of sustainability—social, environmental, and economic—hence complementing prior efforts with empirical evidence and providing actionable insights into the practical trade-offs involved in deploying fairness-aware ML systems. We focus on analyzing multi-protected attributes, which involves ensuring fairness across multiple overlapping groups [40], as this type of evaluation represents a more realistic scenario for implementing fairness enhancements in practice [26]. More particularly, our work is routed around three main research questions that drive our exploration:

- **RQ₁.** *How do bias mitigation algorithms impact social sustainability regarding multiple protected attributes?*
- **RQ₂.** *How do bias mitigation algorithms impact environmental sustainability regarding multiple protected attributes?*
- **RQ₃.** *How do bias mitigation algorithms impact economic sustainability regarding multiple protected attributes?*

While **RQ₁** corroborates and builds upon existing studies by contextualizing fairness within the social sustainability dimension, **RQ₂** and **RQ₃** extend the current body of knowledge by assessing the environmental and economic implications of bias mitigation techniques. To achieve these objectives, we implemented a **quantitative benchmark study research approach** [73]. A benchmark study systematically evaluates and compares the performance of specific techniques, tools, or algorithms under controlled and reproducible conditions. This approach is particularly suited to our goals, as it enables an objective analysis of multiple metrics across different dimensions of sustainability. By applying six bias mitigation algorithms—*Reweighting* [52], *Meta Fair Classifier* [18], *Gerry Fair Classifier* [54], *Exponentiated Gradient Reduction* [3], *PrejudiceRemover* [53], and *Grid Search Reduction* [3]—to four datasets—*Adult* [4], *Statlog* [45], *Mep15* [43], and *Compas* [1]—, we conduct a total of 3,360 experimental runs, where we

systematically measure the impact of these bias mitigation algorithms on fairness, energy consumption, carbon emissions, accuracy, training time, and storage requirements. Our findings reveal multiple insights into the trade-offs across sustainability dimensions. From a social perspective (**RQ₁**), we corroborate previous findings in the field, showing that bias mitigation algorithms significantly improve fairness metrics, with Exponentiated Gradient Reduction (EG) emerging as the most effective technique. However, the results also indicate that certain algorithms, such as Gerry Fair Classifier (GF) and Grid Search Reduction (GS), may deteriorate fairness when applied to multiple protected attributes, emphasizing the complexity of achieving equitable outcomes in diverse scenarios. For environmental sustainability (**RQ₂**), the results demonstrate that algorithms like EG and GS consume significantly more energy and produce higher CO₂ emissions, raising concerns about their environmental footprint. In contrast, computationally efficient methods like GF and Meta Fair Classifier (MF) exhibit lower environmental impacts, highlighting the trade-offs between fairness improvements and ecological responsibility. From an economic sustainability perspective (**RQ₃**), the application of bias mitigation techniques generally increases training time and storage weight while reducing accuracy compared to baseline models. However, exceptions such as Prejudice Remover (PR) demonstrate competitive accuracy in specific datasets, offering valuable insights into the cost-effectiveness of bias mitigation strategies.

To sum up, the **contributions** of this article are the following:

1. We frame fairness within the broader theoretical construct of sustainability, moving beyond the traditional fairness-accuracy trade-offs to explore its intersections with social, environmental, and economic dimensions. This research approach not only highlights the multifaceted impacts of bias mitigation algorithms but also bridges the gap between fairness research and sustainability goals, providing a foundation for the development of more responsible and holistic ML system design;
2. Our findings reveal complex trade-offs between fairness improvements, energy consumption, carbon emissions, accuracy, training time, and model scalability. These insights not only confirm the need for multi-dimensional evaluation frameworks but also provide practitioners with a clearer understanding of the sustainability implications of their design decisions;
3. We provide a replication package [30] containing all raw data, scripts, datasets, and experimental setups of the quantitative benchmark study. This resource enables researchers to verify our results, extend our study, and explore new methodologies for assessing sustainability in bias mitigation.

Structure of the paper. Section 2.2 overviews the related work and emphasizes the novelty of our study, describing the terminology applied in our study. Section 2.2 describes the research questions and methods employed to address the objectives of our work, while Section 4 discusses the results achieved. In Section 5, we summarize the major findings obtained and outline the implications of our work. The limitations of the study are discussed in Section 6. Finally, Section 7 concludes the paper and discusses our future research agenda.


2. Theoretical Framework and Related work

In this section, we first report on the theoretical framework underpinning our study, emphasizing the basic constructs we rely on, particularly the dimensions of sustainability (social, environmental, and economic) and their relevance to fairness in ML. Secondly, we discuss the most closely related work, positioning our study within the context of existing research and highlighting the key distinctions and advancements it offers. Specifically, we illustrate how our work extends prior studies by incorporating sustainability considerations into the evaluation of bias mitigation algorithms, providing a more comprehensive perspective on their impacts.

2.1. Theoretical Framework

Sustainability, as originally defined by the Brundtland Report, refers to “*meeting the needs of the present without compromising the ability of future generations to meet their own needs*” [10]. While this definition emerged in the context of environmental and economic sustainability, its application has been extended to software engineering to ensure the long-term viability of software systems [5]. McGuire et al. [61] provide a multi-layered and stratified view of sustainability, conceptualizing it as a system’s ability to influence and endure at individual, team, organizational,

and societal levels. Within this framework, sustainability is typically classified into three core dimensions of the software process: *social*, *environmental*, and *economic* sustainability, each of which forms the foundation of our study. We argue that sustainability dimensions provide a holistic theoretical construct for understanding the impacts of ML-enabled systems, particularly when evaluating bias mitigation algorithms. Traditional fairness research has focused on the fairness-accuracy trade-off, isolating fairness from its broader consequences. By framing fairness within the sustainability framework, we aim to address previously underexplored interactions among social, environmental, and economic dimensions, providing a more comprehensive evaluation of bias mitigation techniques. A key element of this framing is given by the definition of the *protected attribute*, which we define as follows:

 **Protected Attribute:** Protected attributes are specific characteristics of individuals or groups, such as gender, race, age, or other legally or contextually defined attributes, that are recognized as critical to ensuring fairness in ML models. These attributes are typically defined based on legal statutes, ethical considerations, and societal norms, as they represent groups that have historically been subjected to systemic discrimination, unequal treatment, or exclusion [2, 39]. In the context of fairness-aware ML, these attributes serve as a basis for identifying and mitigating biases to ensure equitable outcomes across demographic groups.

Protected attributes play a pivotal role in defining privileged and unprivileged groups for fairness evaluation. For instance, in *social sustainability*, the focus is on ensuring that outcomes for unprivileged groups (e.g., based on gender or race) do not perpetuate inequities. For *environmental sustainability*, the way protected attributes are handled can influence resource efficiency, as iterative fairness methods may increase energy consumption. In *economic sustainability*, understanding how protected attributes impact storage or computational costs offers valuable insights into trade-offs between fairness and operational efficiency. Furthermore, applying bias mitigation algorithms that operate on these attributes in datasets may affect models’ predictive performances, hence deteriorating their values. By integrating protected attributes into the sustainability framework, our study has the ultimate goal of understanding their broader societal, environmental, and economic implications. In the context of our study, we focus on intersectional fairness, that is, the evaluation of bias in ML models considering more than one protected attribute simultaneously [40]. We decided to focus on this particular scenario for two reasons: on the one hand, this matter is still underexplored in fairness research [24]; on the other hand, it could provide a more realistic view of the trade-off between sustainability dimensions as it reflects better real-world contexts where more than a protected attribute has to be considered [26]. The selection of protected attributes in fairness evaluation, however, remains a complex issue. This is due to the context-dependent nature of ML fairness: for example, an automated hiring system should not make decisions considering the ‘sex’ attribute, whereas, in some complex medical applications, this information should be taken into account [36]. To handle this selection in our study, we relied on existing guidelines for each of the datasets we selected, as the datasets are drivers for the context of application [35]. In the following sections, we discuss each sustainability dimension, elaborating on their theoretical foundations and how they are operationalized in our study.

Social Sustainability: Social sustainability concerns the impact of systems on individuals, communities, and society. This dimension emphasizes equity, inclusiveness, and human well-being as fundamental principles for sustainable development [61]. In software engineering, socially sustainable systems should promote fairness, accessibility, and ethical behavior while minimizing harm, marginalization, and discrimination against individuals or groups. Fairness, as widely discussed in the ML fairness literature, directly aligns with these objectives, making it a natural foundation for evaluating social sustainability.

McGuire et al. [61] conceptualize social sustainability as influencing multiple levels: psychosocial well-being at the individual level, team cohesion at the group level, organizational culture at the institutional level, and societal affordances at the broader societal level. Building on this stratified framework, we argue that fairness in ML systems serves as a key enabler of social sustainability because it directly influences individual and societal perceptions of equity and inclusivity. For example, fair ML systems that avoid discrimination promote greater trust, inclusiveness, and well-being for individuals, while at a societal level, they foster cohesion by mitigating systemic biases that harm marginalized or underrepresented groups [23]. This broader framing allows us to go beyond the conventional understanding of fairness as an isolated metric and instead position it as a measurable proxy for social sustainability.

While fairness metrics—such as demographic parity, equalized odds, and disparate impact—have been extensively studied in ML fairness research, their evaluation has traditionally focused on isolated trade-offs between fairness and accuracy. Existing studies have rarely contextualized fairness within a larger sustainability framework, and as such,

they may overlook the broader implications that fairness improvements have on other aspects. Our study bridges this gap by explicitly operationalizing social sustainability through fairness metrics. It is important to note that this is not merely a rephrasing of fairness evaluation but rather an extension that integrates fairness into a holistic sustainability construct. By aligning fairness metrics with the broader goals of social sustainability, we highlight that fair systems are not only ethically desirable but also socially sustainable, contributing to equity at multiple levels of impact.

In doing so, our work extends previous fairness studies in two important ways. First, we situate fairness within a broader theoretical construct, explicitly linking it to social sustainability, which has not been systematically addressed in fairness research. This allows us to connect fairness improvements to broader societal outcomes, such as psychosocial well-being and societal cohesion, reinforcing the role of fairness as a component of responsible software engineering. Second, by framing fairness as one dimension of sustainability, we establish a foundation for evaluating its interaction with environmental and economic dimensions. This multi-dimensional perspective provides practitioners and researchers with a more comprehensive understanding of the trade-offs introduced by bias mitigation methods, enabling more informed and balanced decision-making.

In summary, while fairness has traditionally been studied as a standalone objective, our work attempts to elevate it as an indicator of social sustainability. Through this framing, we demonstrate that fairness improvements extend beyond mitigating bias in ML models; they provide an important role in fostering socially sustainable software systems that promote equity, inclusiveness, and trust within society.

Environmental Sustainability: Environmental sustainability focuses on minimizing the ecological impact of software systems, particularly in terms of energy consumption and carbon emissions. In ML engineering, this dimension is becoming increasingly significant due to the exponential growth in the computational resources required to develop and deploy ML models [71]. Large-scale models, like those used in fairness-sensitive applications, consume substantial amounts of energy, contributing to global energy usage and carbon footprints [5]. These impacts raise concerns about the environmental costs associated with the adoption of fairness-enhancing methods, especially as organizations scale up ML solutions to address real-world ethical concerns. Bias mitigation algorithms, while addressing fairness and equity, may introduce additional computational complexity. For instance, in the training phase, pre-processing techniques involve transformations on the dataset, while in-processing methods may require modifications to the optimization process. These operations inherently increase training time and resource consumption, which in turn directly affects energy usage and carbon emissions. It becomes crucial, therefore, to examine whether the benefits of fairness achieved by bias mitigation methods are offset by their environmental footprint.

This evaluation is particularly relevant for practitioners and researchers because fairness in ML systems is often promoted as a means of advancing ethical and responsible AI practices. However, an unintended consequence of promoting fairness may be the increased environmental burden, which could compromise sustainability goals. If practitioners fail to account for these environmental trade-offs, fairness initiatives may conflict with global sustainability efforts, ultimately undermining the broader objectives of ethical AI and sustainable software engineering. In our study, we operationalize environmental sustainability by quantifying the energy consumption and carbon emissions associated with the execution of bias mitigation algorithms. These metrics serve as indicators of the environmental costs incurred when implementing fairness techniques. Indeed, evaluating bias mitigation methods through this lens enables us to uncover potential trade-offs between fairness improvements and environmental impacts, an aspect that has been underexplored in the literature. It is important to note that assessing environmental sustainability against bias mitigation methods is not merely a technical exercise; it holds practical implications for developers, organizations, and policymakers. For developers, understanding the environmental costs allows for more informed algorithm selection and optimization strategies. For organizations, balancing fairness with ecological efficiency helps align AI initiatives with corporate sustainability goals, enhancing their reputation and social responsibility. For policymakers, insights into these trade-offs can inform guidelines and regulations that encourage the development of fairness techniques that are both ethical and environmentally sustainable.

In conclusion, our study aims at addressing a current knowledge gap, providing a more holistic perspective on the impacts of fairness-enhancing methods. This broader view may enable practitioners to make informed decisions that consider equity, resource consumption, and long-term sustainability, ensuring that fairness initiatives do not come at the cost of environmental responsibility.


Economic Sustainability: Economic sustainability pertains to the financial viability and resource efficiency of software systems throughout their lifecycle. This dimension is particularly critical for ML-enabled systems, where

the computational and storage requirements of models can have significant cost implications. Economic sustainability emphasizes the need to maintain accessibility, cost-effectiveness, and resource efficiency, ensuring that systems remain operational and maintainable while aligning with broader economic objectives [61].

In the context of our work, we evaluate the impact of bias mitigation algorithms: while essential for improving fairness, these algorithms may introduce additional computational overhead that can directly impact the economic sustainability of ML systems. Traditionally, the economic aspects of ML systems have been indirectly evaluated using accuracy as a primary metric. Accuracy, however, provides only a partial view of economic impacts, as it does not account for the operational and development costs associated with implementing fairness techniques. For example, pre-processing techniques require additional steps in data preparation, while in-processing methods often necessitate changes to optimization processes, potentially increasing training times and computational costs. Without assessing these hidden costs, the practical feasibility of fairness-enhancing methods in real-world applications remains unclear.

In our study, we operationalize economic sustainability by incorporating metrics that capture the costs of implementing bias mitigation algorithms. Specifically, in addition to predictive performance metrics, we evaluate (1) training time, which reflects the additional computational overhead introduced during the model training phase when fairness techniques are applied; and (2) storage weight, which quantifies the size of the resulting ML models after applying bias mitigation techniques. Evaluating these metrics alongside accuracy provides a more comprehensive understanding of the economic trade-offs introduced by fairness techniques. For instance, while improving fairness is often ethically desirable, the financial costs associated with achieving such improvements may limit the feasibility of deploying bias mitigation methods in practice. This highlights the importance of balancing fairness with cost-efficiency to ensure that fairness-enhancing methods remain accessible and scalable for widespread use. This evaluation of economic sustainability is critical for several reasons. For practitioners, understanding the trade-offs associated with bias mitigation algorithms enables better decision-making regarding algorithm selection, optimization, and deployment. For organizations, economic sustainability aligns fairness-enhancing initiatives with financial and operational goals, ensuring that ethical AI practices remain practical and scalable. From a broader societal perspective, addressing economic sustainability democratizes access to fairness-enhancing methods, reducing barriers for organizations with limited resources and enabling their adoption at scale. While prior research has focused heavily on the fairness-accuracy trade-off, little attention has been paid to the broader economic impacts of implementing fairness techniques. By extending the evaluation of bias mitigation algorithms to include training time and storage weight, our study fills this gap, providing actionable insights into the economic sustainability of fairness-enhancing methods. This holistic perspective ensures that fairness initiatives are not only ethically sound but also financially viable, supporting their long-term adoption and scalability.

In conclusion, our work emphasizes the importance of evaluating economic sustainability as an integral dimension of fairness research. By quantifying the hidden costs of implementing bias mitigation algorithms, we provide a framework for assessing their broader resource and financial implications. This perspective enables practitioners to make informed decisions that balance fairness, performance, and resource efficiency, ensuring that fairness initiatives align with both ethical and economic sustainability goals.

 **Theoretical Framework:** The theoretical framework presented in this section goes beyond traditional fairness research by integrating fairness metrics into a broader sustainability construct. By aligning fairness with social sustainability, and evaluating its interactions with environmental and economic dimensions, our study addresses a knowledge gap in the current literature. This framework enables the identification of multi-dimensional trade-offs introduced by bias mitigation algorithms, offering a more comprehensive perspective on their implications. By linking these dimensions to actionable metrics, such as energy consumption, training time, and storage weight, we aim at providing a novel contribution that equips practitioners, organizations, and policymakers with the tools needed to make balanced and sustainable decisions.

2.2. Related Work

The rapid growth of the usage of AI has stimulated substantial changes across diverse industries, compelling both practitioners and the research community to confront the challenge of crafting products that mitigate rather than aggravate worldwide issues related to equality, inclusion, and the environment. This highlights the critical importance of

assessing the impact of AI on broader societal goals, notably the Sustainable Development Goals² (SDGs) established by the United Nations. While SDGs encompass a range of goals spanning poverty alleviation, health, inequalities, education, and environmental sustainability, the evaluation of AI’s contribution to these goals is paramount [82, 41, 33]. ML is a subfield of AI that focuses on developing algorithms that can learn from data [68]. However, the terms AI-enabled or ML-enabled systems are used to refer to a system or component that uses some type of AI or ML [59, 58]. Therefore, in this study, we consider these terms to be synonyms.

While researchers have made significant strides in addressing disparities and promoting equity in ML software, many have primarily focused on protected individual attributes [19, 25, 27, 66]. Conversely, intersectional fairness, i.e., ensuring fairness across multiple, overlapping social identities or demographic characteristics [40], is mostly underexplored [75]. In parallel, researchers have explored bias mitigation algorithms, which can be categorized into three main types: (i) pre-processing, which optimizes data before training; (ii) in-processing, which focuses on the learning process; and (iii) post-processing, which improves equity in decision-making outcomes [25, 7, 49]. For example, Biswas and Rajan [7] conducted a benchmark study using seven bias mitigation algorithms with shallow learning models from Kaggle, focusing on equity and accuracy metrics. Conversely, researchers have also delved into the environmental sustainability of ML-enabled systems, highlighting potential environmental risks associated with these systems [44]. Verdecchia et al. [81] elicited existing practices in the literature, which were further confirmed and synthesized by Järvenpää et al. [51].

Our study is mainly centered on bias mitigation algorithms and how these impact social, environmental, and economic sustainability. In terms of fairness and bias mitigation, the research community has been studying these matters from two main perspectives. A first line of research revolved around the improvement of existing algorithms or the definition of new fairness-aware techniques. For instance, Chakraborty et al. [19] introduced FAIR-SMOTE, an algorithm designed to correct biased labels and rebalance internal distributions, ensuring fair representation in both positive and negative classes based on protected attributes. Chen et al. [25] proposed MAAT, an ensemble technique to optimize the fairness-performance balance in ML software. FAIRWAY [20] combines pre-processing and in-processing methods to eliminate ethical bias from both training data and the trained model. Finally, Peng et al. [65] proposed FAIRMASK, a model-based extrapolation method for bias mitigation and explanation, comparing FAIRMASK against five algorithms across multiple protected attributes. These approaches focus on optimizing the trade-off accuracy and fairness, not considering other non-functional requirements of ML models. Perhaps more importantly, none of these approaches took into account the time required to improve a model, which directly affects the feasibility and scalability of implementing fair ML models in real-world scenarios.

A second line of research involved the design of benchmark studies for ML bias mitigation methods. In particular, Hort et al. [49] developed FAIREA, a tool for benchmarking ML bias mitigation methods through model behavior mutation. Their work included five pre-processing and in-processing algorithms, focusing on non-functional requirements. Later, Chen et al. [26] utilized FAIREA for doing a large study considering seven pre-processing and in-processing bias mitigation algorithms and found the bias mitigation methods significantly decreased the ML accuracy metrics of the studied scenarios. In addition, they described that the effectiveness of bias mitigation methods depends on the tasks, the models, the choice of protected attributes, and the set of metrics used to assess ML fairness and performance. Their evaluations were limited to scenarios involving only one protected attribute at a time. Zhang and Sun [87] adapted equity improvement methods previously proposed in the ML community to be able to handle more protected attributes. They compared six pre- and in-mitigation algorithms and the time needed to use their approach. However, they did not consider the time required to mitigate the bias of other algorithms. Recently, Chen et al. [24] conducted a benchmark study on improving fairness in relation to multiple protected attributes, looking at eight pre-processing and in-processing algorithms. Lastly, Hort et al. [50] have presented a new repair method of ML-based decision-making software that simultaneously improves fairness and accuracy.

More recently, researchers like Chen et al. [21] and Sarro [69] argued the definition of novel techniques that are able to balance fairness objectives with other sustainability principles, hence pioneering the research efforts in the area targeted by this work. Additionally, Candelieri et al. [16] proposed a hyper-parameter optimization approach that balances fairness metrics and environmental indicators. Our study builds on the current body of knowledge on ML fairness by proposing a broader, more comprehensive quantitative evaluation that integrates together the three dimensions of social, environmental, and economic sustainability.

²The Sustainable Development Goals: <https://sdgs.un.org/goals>

By framing fairness as a measurable proxy for social sustainability, we extend its evaluation beyond isolated metrics to include its broader societal implications, such as equity, inclusiveness, and trust. Moreover, our study operationalizes environmental sustainability through metrics like energy consumption and carbon emissions, providing insights into the ecological footprint of fairness-enhancing methods. Similarly, we assess economic sustainability by examining training time and storage weight, quantifying the resource and financial costs of implementing bias mitigation techniques. Our framework enables us to uncover complex trade-offs between fairness and other dimensions of sustainability, offering a multi-dimensional perspective that is absent in prior work. By merging insights from AI, software engineering, and sustainability, our work contributes a novel interdisciplinary perspective to fairness research. Our findings not only highlight the impacts of bias mitigation algorithms on equity but also provide actionable insights into their environmental and economic consequences. Our ultimate goal is to equip practitioners and researchers with the tools to make informed decisions about fairness-enhancing techniques, ensuring that they are not only ethically sound but also sustainable and scalable in real-world scenarios.

Novelty of the study: Our work introduces a new perspective by framing mitigation bias in the context of sustainability, moving beyond traditional equity metrics to assess their social, environmental, and economic dimensions. By systematically assessing these dimensions, we offer new insights into the complex interplay between fairness practices and sustainability goals, adding depth to the existing literature on responsible AI.

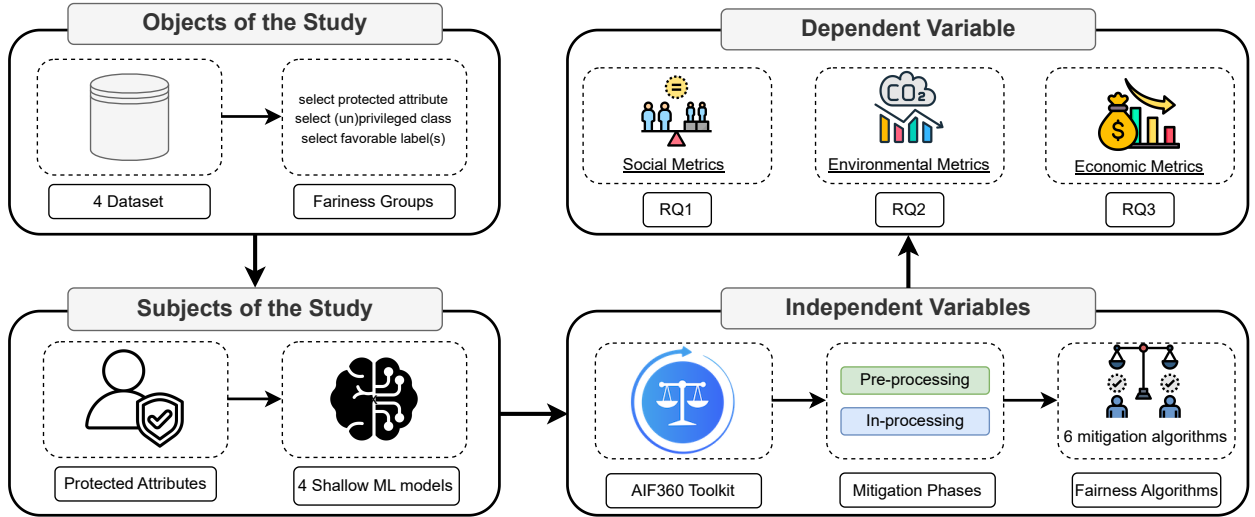


Figure 1: Overview of the research process.

3. Empirical Study Design

To define our research goal, we follow the Goal-Question-Metric (GQM) [14]. The *goal* of the empirical study is to assess the impact of bias mitigation algorithms on the sustainability of ML-enabled systems, with the *purpose* of providing insights into the trade-offs between addressing fairness-related concerns and their corresponding social, environmental, and economic impacts in the *context* of ML-enabled systems. The *perspective* is twofold: practitioners may leverage the outcomes of this study to better evaluate their activities, gaining knowledge on how mitigating risks due to unfairness might influence other critical system properties, potentially enhancing quality assurance processes. Meanwhile, researchers may benefit by exploring the interplay between multiple non-functional requirements of ML-enabled systems, gaining insights into current challenges and informing the design of next-generation quality assurance mechanisms. To achieve our objectives, we formulated three research questions. The first, **RQ₁**, sought to understand how bias mitigation algorithms impact *social sustainability* through fairness, particularly in scenarios

involving multiple protected attributes. Social sustainability, as defined in [61], encompasses properties that influence the individual’s physical and psychosocial well-being, making fairness a critical component of this dimension. Our analysis serves as both a *preliminary* and *confirmatory* step, designed to corroborate existing findings in fairness literature while framing fairness within the broader theoretical construct of social sustainability. This framing establishes a foundational understanding of fairness that enables the exploration of its interplay with environmental and economic dimensions in subsequent research questions. More particularly, bias mitigation algorithms are explicitly designed to deal with social sustainability properties and, therefore, we expected to observe a positive effect of all the algorithms [19]. Yet, a preliminary investigation may quantify the impact of existing bias mitigation algorithms on social sustainability properties, other than serve as an evidence-based assessment [56] of their performance and potential combinations thereof. At the same time, our preliminary analysis can support the findings observed by Chen et al. [24, 26] in recent investigations into the capabilities of bias mitigation algorithms, possibly discovering additional insights that may advance the current body of knowledge. While the fairness literature has focused on a single attribute at a time [19, 25, 66], the analysis of multiple protected attributes simultaneously needs more exploration [24]. Therefore, we focus all research questions on analyzing multiple protected attributes. Based on the argumentation above, we formulated our **RQ₁**:

RQ₁ – Social Sustainability

How do bias mitigation algorithms impact social sustainability regarding multiple protected attributes?

Building on the results of **RQ₁**, we turned our attention to environmental sustainability in **RQ₂** by investigating the impact of bias mitigation algorithms on metrics such as carbon emissions and energy consumption. These metrics are particularly relevant in the broad context of sustainability, as they quantify the trade-offs between addressing fairness and the ecological footprint of ML-enabled systems. By assessing the environmental impact of these algorithms, we aimed to enhance practitioners’ awareness of the energy and resource costs associated with bias mitigation and inspire the development of novel tools to balance fairness and environmental sustainability. Hence, we asked:

RQ₂ – Environmental Sustainability

How do bias mitigation algorithms impact environmental sustainability regarding multiple protected attributes?

Finally, **RQ₃** examined the economic sustainability of bias mitigation algorithms, extending the analysis beyond accuracy—a primary focus in traditional studies on fairness. While accuracy remains an important metric, our study incorporates additional dimensions, such as training time and model storage requirements, to provide a more comprehensive perspective. These metrics directly influence operational costs, resource efficiency, and the long-term feasibility of deploying fairness-enhancing algorithms in real-world scenarios. This more comprehensive view may offer insights into the trade-offs between fairness improvements and economic considerations, enabling researchers and practitioners to better evaluate the broader cost-effectiveness of bias mitigation techniques. By moving beyond accuracy, our study highlights the hidden costs associated with fairness-enhancing methods, thereby contributing to the understanding of how these techniques align with economic sustainability objectives. Therefore, we asked:

RQ₃ – Economic Sustainability

How do bias mitigation algorithms impact economic sustainability regarding multiple protected attributes?

Figure 1 shows an overview of our study design. As shown, we first determined the common ground for the quantitative benchmark study, i.e., the most appropriate datasets, the ML models to experiment with, and the metrics to estimate social, environmental, and economic sustainability. Afterward, we selected the bias mitigation algorithms to assess, picking those available within the well-known AIF360 toolkit [6]. Finally, we ran the selected bias mitigation algorithms against the benchmark and analyzed the corresponding results to address our research questions. In particular, a benchmark study is a standard tool for the evaluation and comparison of competing systems or components according to specific characteristics, such as performance, dependability, or security [55]. Benchmark studies can be tested repeatedly and quickly without requiring human subjects [77]. Consequently, the results section will present

findings that are based solely on quantitative analysis without any subjective interpretations or personal biases, as this type of research approach exploits numerical analysis and statistical techniques rather than descriptive or interpretive methods. However, we complement our findings with extensive discussions in a dedicated section (5). In terms of reporting, we employed the guidelines (i) by Wohlin et al. [83] and (ii) by the *ACM/SIGSOFT Empirical Standards*³.

Table 1: Datasets used in the experimentation.

Name	#Sample	#Feature	Protected Attributes	Favorable label	Description
Adult	32,560	104	sex,race	7508 (18,80%)	The goal is to classify individuals into specific annual income classes and determine whether their income is above or below \$50,000.
Mep15	15,830	42	sex,race	2718 (17.17%)	The goal is to predict individuals’ health care needs based on how Americans pay for medical care, health insurance, and out-of-pocket expenses.
Statlog (German Credit Card)	1,000	61	sex,age	700 (70%)	The purpose is to predict a bank customer’s ability to repay, or not, a loan.
COMPAS	7214	8	sex,race	3963 (54,93%)	the goal is to predict, through demographic information and criminal history, whether a defendant will reoffend within two years.

3.1. Objects of the study

The *objects* of our study were the datasets to be used for our experimentation. To address our research questions, we required the selection of fairness-relevant datasets—otherwise, we could not observe significant outcomes from the application of bias mitigation algorithms. Table 1 describes the main characteristics of the datasets selected for our study. As shown, we identified four popular datasets such as *Adult* [4], *Statlog* [45], *Mep15* [43], and *Compas* [1]. The reason behind this selection is manifold.

First, these datasets have been widely employed in previous literature on software engineering for artificial intelligence [59, 50] and ML fairness [7, 49, 24, 19]: as such, their adoption allowed us to have a common ground to compare our findings with previous achievements in the field, e.g., the four datasets are among those used by Chen et al. [24, 26]. In the second place, the selected datasets cover activities that pertain to individuals’ personal information in various fairness-sensitive domains, including finance, social, and medical sectors: as a consequence, we could assess the impact of bias mitigation algorithms in various contexts, possibly uncovering patterns that better describe the characteristics of these algorithms. Lastly, this fixed amount of datasets selected allowed us to relax the constraints on the number of investigations to be performed on different objects, hence increasing the number of experiments for each, improving our confidence in driving conclusions on these four datasets. Consistently with prior investigations [24, 26, 66], we chose ‘Sex’, ‘Age’ and ‘Race’ as protected attributes. As a consequence of this choice, our investigation could be performed on *multiple protected attributes*, which has been noted as a challenging and more realistic condition for the evaluation of any bias mitigation algorithm [24].

³Available at: <https://github.com/acmsigsoft/EmpiricalStandards>. Given the nature of our study and the currently available standards, we followed the “General Standard”, and “Benchmarking” guidelines.

3.2. Subjects of the Study

The *subjects* of the study were the machine learning models to experiment with to assess their sustainability properties. In this respect, we relied on previous literature in the field [8, 79, 19, 86, 78]. Particularly, the selection was informed by the work by Hort et al. [48], who provided a comprehensive survey of ML models frequently employed in fairness research. We selected and trained four different algorithms: *Logistic Regression* (LR), *Linear Support Vector Classification* (SVC), *Random Forest* (RF), and *XGBoostClassifier* (XGB). According to Hort et al. [48], these are among the most frequently used in the evaluation of bias mitigation methods. In particular, *LR* serves as a reference point for evaluating the relative effectiveness of bias mitigation techniques, being frequently employed as a baseline model due to its simplicity and interoperability; *SVC* is known for its strong classification capabilities and its ability to handle complex decision boundaries; *RF* is recognized for its robustness and resistance to overfitting, making it widely adopted for fairness evaluations across diverse datasets; lastly, *XGB* combines computational efficiency with strong predictive performance, making it a robust method for evaluating fairness and sustainability trade-offs. These models were trained using as features the attributes included in the four datasets described in Section 3.1. All of them pertained to binary classification tasks. For each ML algorithm, we defined seven variants, one for each of the independent variables of the study plus one baseline where we did not apply any treatment. In this way, we could experiment with one ML model for each bias mitigation algorithm considered in the study, i.e., each of them applied a different bias mitigation algorithm, against which we compared a baseline model that did not apply any bias mitigation algorithm. Following this design, we defined a total of 28 experimental models. In terms of data preprocessing, we removed missing or invalid values and converted continuous features to categorical ones. In addition, we adapted the preprocessing for equity analysis by identifying protected attributes, defining privileged and non-privileged groups, and clarifying the outcome or favorable label for predictions. We set sex, race, and age as privileged groups according to their availability in the selected datasets. As for hyper-parameters configuration, we used the default configurations coming from relevant studies [49, 24, 25, 26].

Table 2: Variables of the study

Name	Scale	Operationalization
<i>Independent variables:</i>		
Baseline Model	Nominal	Models without any bias mitigation algorithms.
Bias Mitigation Algorithms	Nominal	Reweighting, Meta Fair Classifier, Gerry Fair Classifier, Exponentiated Gradient Reduction, PrejudiceRemover, Grid Search Reduction.
<i>Dependent variables:</i>		
RQ₁ Mean Difference	Ratio	$P(Y = 1 D = 0) - P(Y = 1 D = 1)$
RQ₁ Equal Opportunity Difference	Ratio	$TPR_{D=\text{unprivileged}} - TPR_{D=\text{privileged}}$
RQ₁ Average Odds Difference	Ratio	$\frac{1}{2} \left((FPR_{D=\text{unprivileged}} - FPR_{D=\text{privileged}}) + (TPR_{D=\text{unprivileged}} - TPR_{D=\text{privileged}}) \right)$
RQ₂ Energy consumption	Ratio	Power usage \times Training time
RQ₂ Carbon Emission	Ratio	Carbon Intensity factor \times Energy consumption
RQ₃ Training Time	Ratio	$Time_{\text{end}} - Time_{\text{start}}$
RQ₃ Accuracy	Ratio	$(TP + TN) / (TP + TN + FP + FN)$
RQ₃ F1 score	Ratio	$(2 \times TP) / (2 \times TP + FP + FN)$
RQ₃ Storage Weight	Ratio	The size of the trained model

3.3. Empirical Study Variables

Once we had defined the context of our study, we proceeded with the definition of the independent and dependent variables of the empirical study. Table 2 summarizes the design choices taken in this respect.

Independent Variables. The independent variables of the study were represented by the bias mitigation algorithms that we aimed to assess with respect to their impact on sustainability. We selected techniques that have a potential impact on all evaluation measures relevant to our research questions. In particular, we focused on algorithms that operate *before* or *during* model training, as they directly influence the training phase and can affect key metrics such as energy consumption, storage weight, and other metrics considered. In contrast, post-processing algorithms, which operate exclusively during the *inference phase*, were excluded as they do not impact the metrics studied in this work. In fact, including post-processing methods would have introduced inconsistencies in our analysis and compromised its comprehensiveness. Therefore, our study is based on state-of-the-art pre-processing and in-processing algorithms [7, 26]. The former (referred to as ‘PRE’ in the remainder of the section) are methods that attempt to mitigate bias in training data to promote a more equitable model. The latter (referred to as ‘IN’) are methods to optimize training algorithms to improve fairness. In addition, our study also includes a baseline as an independent variable. More specifically, our study featured the following pre-processing and in-processing algorithms:

- *Baseline.* A baseline where no bias mitigation algorithms were applied.
- *PRE#1 - Reweighing* [52]. This allows the weight attributed to individual instances of the dataset to be changed, identifying instances belonging to protected and unprotected groups looking for ideal weights aimed at mitigating any type of discrimination present within the dataset.
- *IN#1 - Meta Fair Classifier.* [18]. It represents a classifier exploited in the in-processing phase to modify the standard training set into a training set which considers fairness metrics. Once the training set is modified, additional standard models are trained on the new modified training set to provide a training phase that primarily uses fairer training sets.
- *IN#2 - Gerry Fair Classifier* [54]. This is an algorithm for learning classifiers that are fair with respect to rich subgroups. The rich subgroups are defined by linear functions on the protected attributes, and the notions of fairness are statistical in nature.
- *IN#3 - Exponentiated Gradient Reduction* [3]. The algorithm is able to reduce fair classification to a sequence of cost-sensitive classification problems, returning a randomized classifier with the lowest empirical error subject to fair classification constraints.
- *IN#4 - PrejudiceRemover* [53]. The algorithm adds a term of discrimination-aware regularization to the learning objective.
- *IN#5 - Grid Search Reduction* [3]. It reduces fair classification to a sequence of cost-sensitive classification problems, returning the deterministic classifier with the lowest empirical error subject to fair classification constraints among the candidates sought.

The selection of this specific set of algorithms was motivated by two main factors. First, we prioritized algorithms with robust, standardized, and optimized implementations: we therefore selected the algorithms available in the AI FAIRNESS 360 (AIF360) toolkit [6]. This choice ensured methodological rigor: while we are aware of the existence of other bias mitigation algorithms, relying on unstandardized or non-optimized fairness libraries might have introduced inconsistencies and inaccuracies in the analysis, particularly for environmental metrics that are highly sensitive to implementation efficiency. For example, less efficient implementations might lead to inflated resource consumption, longer execution times, and higher energy usage, thereby skewing the results and compromising the validity of our conclusions. By leveraging AIF360, which has been validated in both academic research and industry practice [57, 32], we minimized such risks and ensured the reliability and reproducibility of our findings.

Second, the selected algorithms were required to handle *multiple protected attributes*. In this respect, the selected algorithms, as for their implementation, allowed for the specification of more than one protected attribute in the form of “a subset of features for which fairness is desired” as specified in the AIF360 documentation.⁴ More specifically, the chosen algorithms operate on AIF360 objects called `StandardDataset`, which include a parameter titled

⁴AIF360 documentation: <https://aif360.readthedocs.io/en/stable/>.

`protected_attribute_names(list)`, implying that more than an attribute can be indicated as protected. In cases where an algorithm requires a single attribute, this is explicitly stated—for example, the *DisparateImpactRemover* algorithm does not operate on a `StandardDataset` but instead includes the specific parameter `sensitive_attribute(str)`, which is the reason why we did not include it in our experiments. This selection ensures the flexibility and applicability of the chosen methods for scenarios involving multiple protected attributes, as argued in other studies on the matter, e.g., the recent study by Chen et al. [24].

Dependent Variables. The dependent variables of the study were defined based on the specific perspectives targeted, i.e., social, environmental, and economic sustainability, and based on the currently available metrics to measure each of them.

RQ₁. Social Sustainability Metrics. Table 2 overviews the social sustainability metrics employed when addressing RQ₁, which are known as ‘mean difference’ (alias of statistical parity difference), ‘average odds difference’, and ‘equal opportunity difference’. These metrics are recognized as standard instruments to measure the extent to which a ML model produces fair outcomes [37, 36, 64], providing insights into potential discriminatory patterns within the data that significantly influence the model’s knowledge and functionality. Let D be the protected attribute, where 1 indicates the privileged group and 0 indicates the unprivileged group. Let Y be the actual label, with 1 denoting the favorable class and 0 being the unfavorable class. Based on these definitions, the considered metrics operate as follows:

- ‘mean difference’ [13] calculates the disparity in favorable rates ($Y = 1$) between the privileged ($D = 1$) and unprivileged ($D = 0$) groups.
- ‘equal opportunity difference’ measures the maximum difference between privileged ($D = 1$) and unprivileged ($D = 0$) subgroups in true positive rates (TPR).
- ‘average odds difference’ is a measure that averages the differences between false positive rates (FPR) and TPR between privileged ($D = 1$) and non-privileged ($D = 0$) groups. It combines FPR and TPR to assess the overall performance of a model across different groups.

The metrics used in this context measure fairness between privileged and unprivileged groups. A metric value of zero signifies equal distribution and fairness, so the groups are fair. A value other than zero indicates a disparity in the benefit received by the non-privileged group or the privileged group. A value of 1 indicates a strong bias in favor of the privileged group. Otherwise, a value of -1 indicates a strong bias in favor of the non-privileged group. These metrics were computed through AI FAIRNESS 360 [6].

RQ₂. Environmental Sustainability Metrics. When addressing RQ₂, we estimated the environmental sustainability of ML models by quantifying the ‘energy consumption’ (measured in Joules) and ‘CO₂ emissions’ (measured in g/CO₂) of the model at training time. These metrics are inherently hardware-dependent and quantitatively represent the resources required to build a model. To quantify energy consumption and carbon emissions, we exploited CODECARBON,⁵ a library that leverages two widely used energy measurement tools: *RAPL* and *Nvidia pynvml* [38, 22, 67, 85]. *RAPL* is employed to calculate the resources utilized by the CPU and RAM, while *pynvml* is employed for GPU computations. CODECARBON provides implementations and patterns, with its central entity, i.e., *TrackerEmission*, capturing critical data on energy consumption and resources. These pieces of information are then employed by CODECARBON to generate a conclusive report on environmental sustainability in an `emissions.csv` file. While ‘CO₂ emissions’ is indeed positively correlated with ‘energy consumption’, we argue that both metrics are necessary because they provide different insights. Energy consumption captures the amount of energy required. At the same time, carbon emissions offer a more complete picture by considering the environmental impact, which depends on the carbon intensity of the energy source. By including both, we preserve critical information regarding emissions, which is vital for practitioners aiming to reduce not just energy use but also the associated carbon footprint. Moreover, since the tool we utilized computes both metrics, we opted to retain them for transparency and practical relevance to potential users, who may value the distinct insights provided by each. Table 2 describes how CODECARBON evaluates these metrics.

⁵The CODECARBON toolkit: <https://codecarbon.io/>.

RQ₃. Economic Sustainability Metrics. In **RQ₃**, we first estimated economic sustainability by assessing the *prediction quality* of the model, considering well-known metrics such as ‘Accuracy’ and ‘F1 score’. The rationale behind the selection of these metrics as a proxy for economic sustainability lies in the observation that higher values in these metrics indicate better efficiency and reliability, contributing to economic sustainability by reducing costs, saving time, enhancing customer satisfaction, providing a competitive edge, and aligning with business goals [11, 59]. In addition, we also computed ‘training time’ (measured in seconds) and ‘storage weight’ (measured in KB). These two metrics have a direct impact on resource consumption, operational costs, and overall efficiency, hence influencing the economic impact of ML models. Following the methodology used in prior studies, we computed the macro-averaged values for F1 score to facilitate a balanced performance comparison across favorable and unfavorable classes. This involves averaging both classes’ F1 score results. A higher score in this metric implies improved ML accuracy [24, 26].

3.4. Experimental Hypotheses, Execution, and Analysis

After defining the experimental subjects and objects, we designed our working hypotheses, which enabled the execution of the experiments and the subsequent data analysis. This section reports on these aspects, detailing the rationale and research methods employed to address the objectives of the study.

Experimental Hypotheses. We aimed at analyzing how the independent variables, namely the bias mitigation algorithms, affect the dependent variables, including social, environmental, and economic sustainability metrics. As such, we defined the following experimental elements.

Let μ_{B_i} and μ_{B_j} be the ML models built using a independent variables B_i and B_j , respectively, where $B_i, B_j \in \{\text{Baseline, Reweighting, Meta Fair Classifier, ...}\}$; let S be the set of sustainability metrics considered in the study. As for **RQ₁**, let S_{so} be a social sustainability metric in the set of sustainability metrics considered in the study. Our null hypothesis was the following:

$$H_0^{B_i, B_j, S_{so}} : \mu_{B_i}^{S_{so}} = \mu_{B_j}^{S_{so}} \quad \forall i \neq j$$

$$S_{so} \in \{\text{mean difference, average odds difference, equal opportunity difference}\}$$

The null hypothesis $H_0^{B_i, B_j, S_{so}}$ determines the effect of the chosen algorithms on the dependent variable S_{so} . Furthermore, μ_{B_i} and μ_{B_j} represents the average measurement result of variable S_{so} . This leads to the following alternative hypothesis, stating that for each dependent variable S_{so} , a statistically relevant difference can be observed between independent variables:

$$H_a^{B_i, B_j, S_{so}} : \mu_{B_i}^{S_{so}} \neq \mu_{B_j}^{S_{so}} \quad \forall i \neq j$$

$$S_{so} \in \{\text{mean difference, average odds difference, equal opportunity difference}\}$$

As for **RQ₂**, we defined a null and alternative hypothesis for each dependent variable related to environmental sustainability metrics S_{en} :

$$H_0^{B_i, B_j, S_{en}} : \mu_{B_i}^{S_{en}} = \mu_{B_j}^{S_{en}} \quad \forall i \neq j$$

$$S_{en} \in \{\text{energy consumption, carbon emission}\}$$

The corresponding alternative hypothesis was formulated as follows:

$$H_a^{B_i, B_j, S_{en}} : \mu_{B_i}^{S_{en}} \neq \mu_{B_j}^{S_{en}} \quad \forall i \neq j$$

$$S_{en} \in \{\text{energy consumption, carbon emission}\}$$

As for **RQ₃**, we described the null and alternative hypotheses for each dependent variable related to economic sustainability metrics S_{ec} :

$$H_0^{B_i, B_j, S_{ec}} : \mu_{B_i}^{S_{ec}} = \mu_{B_j}^{S_{ec}} \forall i \neq j$$

$$S_{ec} \in \{\text{accuracy, f1 score, training time, storage weight}\}$$

The corresponding alternative hypothesis was formulated as follows:

$$H_a^{B_i, B_j, S_{ec}} : \mu_{B_i}^{S_{ec}} \neq \mu_{B_j}^{S_{ec}} \forall i \neq j$$

$$S_{ec} \in \{\text{accuracy, f1 score, training time, storage weight}\}$$

It is important to note that the experiment was balanced with respect to its factor, as each treatment contains unique mitigation algorithms, each of which belonged to the same performance level represented by the treatment.

Experiment Execution. We designed the procedure depicted in Figure 2. Specifically, we conducted a benchmark study, which is defined as “*a standard tool for the competitive evaluation and comparison of systems or components based on specific characteristics such as performance, dependability, or security*” [55]. One of the primary advantages of benchmark studies is their capacity for repeated and rapid testing without relying on human subjects [74]. This characteristic aligns with the quantitative nature of our research, where results are derived entirely through objective, numerical analysis and statistical techniques, eliminating subjective interpretations or personal biases. By grounding our methodology in objectivity, the findings presented in the results section stem directly from the adopted approach, adhering to established standards for benchmark studies. We conducted experiments involving ML models: the four classifiers were trained using each of the independent variables of the study, i.e., a baseline model that did not include any bias mitigation algorithm plus the models trained using the six bias mitigation algorithms employed. A 70%-30% training-test ratio was applied to assess the sustainability implications of each model. The experiments were performed on a machine running Ubuntu Linux, equipped with an AMD Ryzen 7 5800H CPU, an RTX 3060 GPU, and 16 GB of RAM. We configured the environment to support Python 3.6, AIF 0.5.0, scikit-learn 1.2.2, scikit-posthocs 0.9.0, and xgboost 2.0.2 libraries.

Further clarifications on the execution are worth discussing to understand all the aspects included in this experimentation. First, we included 4 ML models and 7 bias mitigation techniques in the study. Evaluating each bias mitigation technique for the 4 ML models, we collected a total of 28 experimental models. Therefore, these experimental models are then evaluated considering 4 datasets, resulting in 112 experimentation objects. Finally, every single experimentation object is repeated 30 times to account for problems related to potential non-determinism caused by the hardware configuration. Therefore, we obtained a total of 3,360 experiments. This amount of re-executions was required to account for the potential non-determinism of the measurements caused by the hardware/software configuration of the experimental machine.

In addition, as energy consumption is strongly influenced by hardware temperature, we followed existing guidelines [29] and executed a five-minute Fibonacci sequence before each measurement with the aim of warming the CPU up. Before running our experiments, we stopped all the unnecessary background processes to let our machine reach a stable condition [38]. Then, we ran the models to get their energy consumption and other metrics written in a .csv file with CodeCarbon. We introduced a one-minute pause after each benchmark execution to minimize the impact of CPU/GPU warm-up and overall system acceleration.

We first trained ML models without bias algorithms to create a first benchmark baseline and then applied each treatment described in Section 3.3. After training the models, we collected measures on the test set to address our research questions. Finally, we stored ML models in .pickle files for further analysis.

Data Analysis. As a final step of our research method, we statistically analyzed the data coming from the 3,360 experiments. In this respect, we first independently considered the experiments executed on each of the datasets in our study. As such, we had 840 distributions to consider at the time: these provided information on the sustainability metrics computed during the 30 executions of the 28 ML models on a given dataset. The raw data produced are in our online appendix (see “*Measurements*” folder) [30].

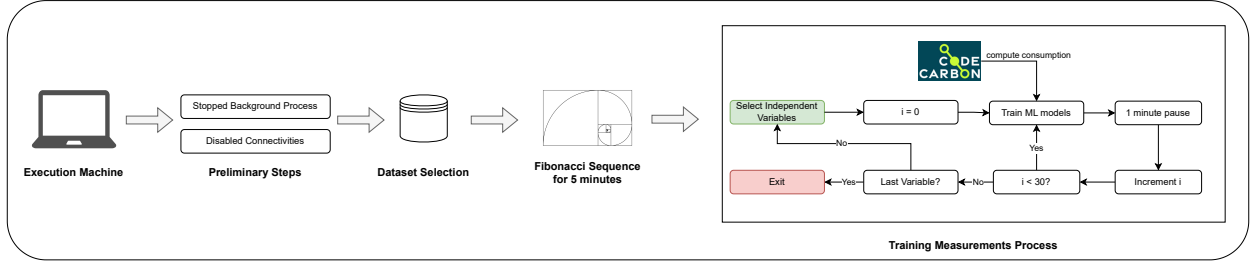


Figure 2: Execution process.

We statistically compared each pair of models against all the sustainability metrics in the study by employing (1) quantile-quantile (Q-Q) plots; and (2) the Shapiro-Wilk normality test, with a significance level of $\alpha = 0.05$ [72]. These methods allowed us to determine whether our metric data adhered to a normal distribution. Upon inspecting the Q-Q plots and the Shapiro-Wilk test results, it became evident that not all models exhibited normal distributions.

After identifying the non-normal distribution, we applied the Friedman test [76], a non-parametric statistical test used to detect treatment differences across multiple algorithms for each metric and dataset. However, Friedman’s test does not indicate which groups are significantly different from each other; for this reason, if the Friedman test result was significant, with α below $= 0.05$, indicating that there are differences among the solutions, we proceeded to conduct the Nemenyi post-hoc test [63, 46]. Nemenyi’s post-hoc test is a multiple comparison test that is used to identify which pairs of groups are significantly different from each other. From the Nemenyi results, the Holm-Bonferroni correction [47], a general method for controlling the family-wise error rate (FWER) in multiple comparisons, was applied. This correction can be applied regardless of whether the statistical tests are parametric or non-parametric. It adjusts the significance level α to account for the number of comparisons made, reducing the risk of Type I errors (false positives) in multiple comparisons, even in non-parametric settings like those involving the Nemenyi test.

To evaluate which specific solutions differed and interpret the post-hoc results, we used a representation based on Critical Difference (CD) diagrams [31]. These diagrams rank multiple groups along the x-axis, connecting statistically indistinguishable groups with horizontal crossbars. The designed plots retain the statistical rigor of the original CD diagrams while improving clarity. They highlight groups of methods that do not show statistically significant differences for each metric and emphasize the average performance of each algorithm. To make the results more accessible, we introduced visual cues, such as distinct colors, to indicate whether methods performed better, worse, or were statistically equivalent to the baseline. Additionally, algorithms that did not exhibit significant differences are presented in separate rows, further enhancing clarity.

4. Analysis of the Results

This section reports the quantitative insights from the data extraction and analysis phase. In particular, we address the study-specific research questions by visualizing and discussing the results of our experiments and statistical analyses. For each **RQ** and metric analyzed, we display a set of diagrams reporting a comparison of the application of each bias mitigation algorithm on each dataset through critical distance-like plots in the cases where the significance of the Friedman test is less than 0.05. The plots that depict our results organize algorithms into rows, with each row representing a group of methods that do not show statistically significant differences in performance for a given metric. To highlight the performance of each bias mitigation algorithm relative to the baseline, the algorithms are depicted in distinct colors. These colors indicate whether the algorithms performed better (●), worse (●), or were statistically equivalent (●) to the baseline. The dotted line represents the best possible value for each metric. Additionally, the plot includes a caption summarizing the average metric values for each algorithm, ranking them from best to worst.

4.1. **RQ₁** - How do bias mitigation algorithms impact social sustainability regarding multiple protected attributes?

The goal of our research question was to understand the impact of bias mitigation algorithms considering multiple protected attributes. To evaluate this aspect, we use mean difference, equal opportunity, and average odds difference. These metrics, used to assess fairness between privileged and unprivileged groups, are detailed in Section 3.3,

4.1.1. Mean Difference Evaluation

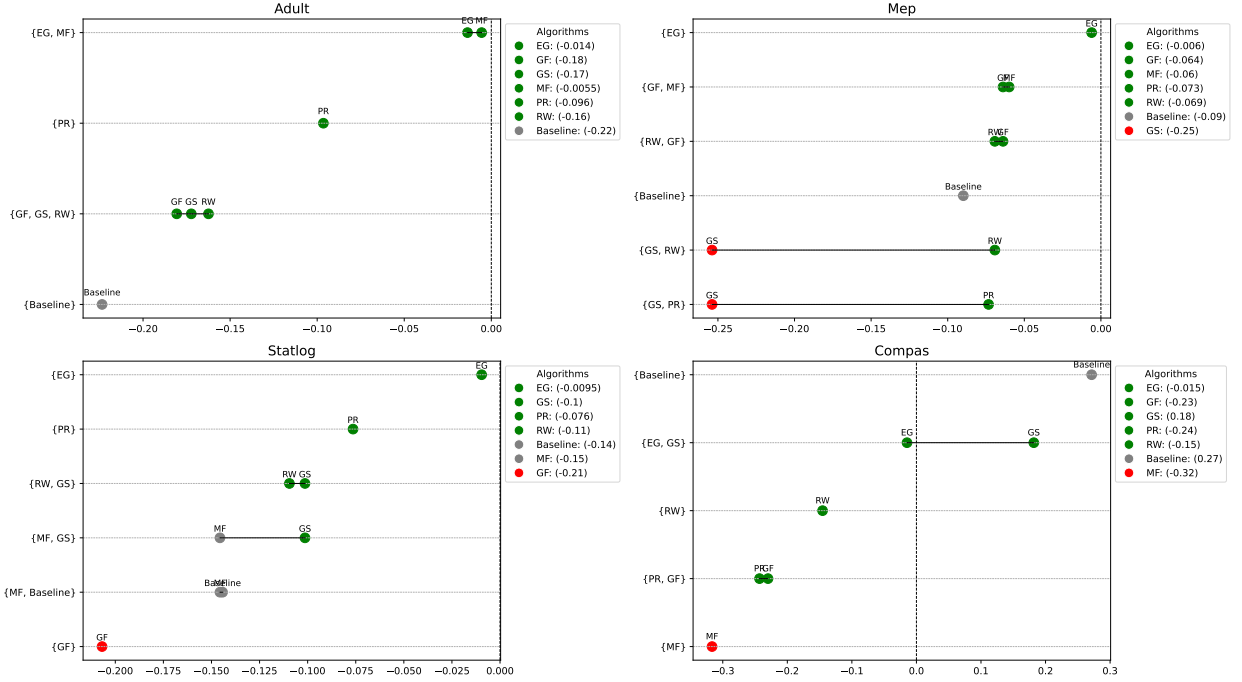


Figure 3: Impact of Fairness Algorithms on *Mean Difference* with different datasets. The *algorithms* are Reweighting (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot indicates a statistically significant improvement against the baseline, a red dot indicates a deterioration instead. Finally, a gray dot indicates that there is no statistically significant difference with the baseline.

Figure 3 shows a series of plots of the mean difference for the datasets described in section 3.1. Adult, Mep, and Statlog datasets exhibit negative values, indicating a predisposition favoring the privileged group across all independent variables. Conversely, the Compas dataset's independent variables span negative and positive values, delineating a spectrum from privileged to unprivileged groups for all the algorithms.

Within the Adult dataset, the analysis reveals a pronounced bias with the Baseline (-0.22) compared to other algorithms, with the most favorable algorithm identified as MF (-0.0055). The groups {GF, GS, RW} and {EG, MF} demonstrate no significant variance. However, a notable deviation exists between all algorithms in this dataset and both the Baseline and the PR algorithm, underscoring a substantial differential impact.

Regarding the Mep dataset, the GS algorithm is the least effective (-0.25), underperforming relative to the Baseline, while EG stands out as the most efficacious algorithm (-0.0006). Statistical analysis batches the algorithms into groups {GS, PR}, {GS, RW}, {RW, GF}, and {GF, MF}, within which no significant differences are observed. Exceptionally, EG distinguishes itself by showing statistically significant variation from its counterparts. The baseline shows significant variation from other algorithms. In the Statlog dataset, algorithm values within the groups {MF, Baseline}, {MF, GS}, and {RW, GS} appear homogenous, with no significant disparities. In contrast, GF, PR, and EG algorithms are statistically different, with GF presenting as the least preferable option (-0.21) and EG approximating the most equity (-0.0095).

Finally, the Compas dataset presents a broad range of values (-0.32 to 0.27), with the EG algorithm nearing 0 (-0.015). The algorithm groups {EG, GS} and {PR, GF} exhibit internal consistency, showing no significant variance. Yet, the Baseline, RW, and MF algorithms stand apart, demonstrating statistically significant differences from their counterparts, highlighting the nuanced performance across different evaluation metrics.

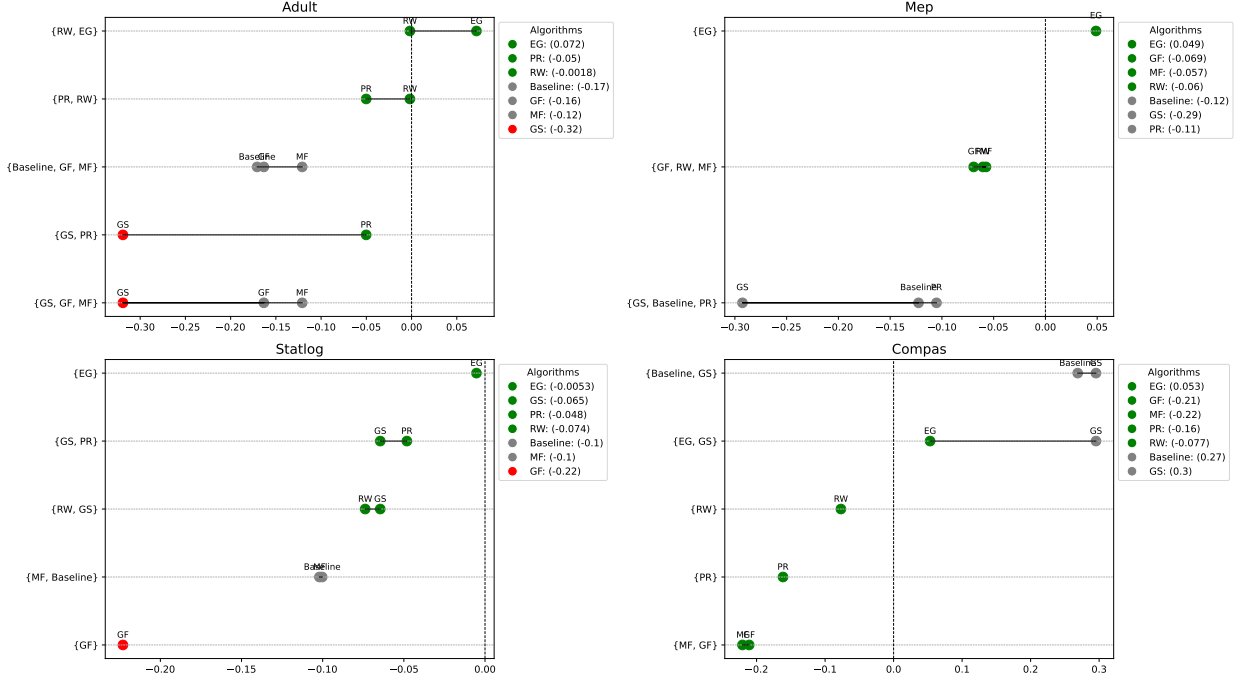


Figure 4: Impact of Fairness Algorithms on *Equal Opportunity Difference* with different datasets. The *algorithms* are Reweighting (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot \bullet indicates a statistically significant improvement against the baseline, a red dot \bullet indicates a deterioration instead. Finally, a gray dot \bullet indicates that there is no statistically significant difference with the baseline.

4.1.2. Equal Opportunity Difference Evaluation

For the equal opportunity difference metric, Figure 4 shows a series of critical distance diagrams highlighting how the algorithms differ. The datasets Adult, Mep, and Compas display a range of values indicative of biases toward both privileged and unprivileged groups across independent variables, while the Statlog dataset exhibits a bias favoring only the privileged group.

In the Adult dataset, the GS algorithm is the worst, decreasing the Baseline value from -0.17 to -0.32. The RW algorithm shows a slight advantage (0.0018), though this advantage is not significant when compared to EG (-0.072). The groups {Baseline, GF, MF}, {GS, PR}, {GS, GF, MF}, {PR, RW}, and {RW, EG}, do not show significant statistical variations. Each algorithm shares a statistical equivalence with at least one other in this scenario.

Regarding the Mep dataset, the GS algorithm is the least effective (-0.29), underperforming relative to the Baseline (-0.12), while EG stands out as the most efficacious algorithm (0.049). The algorithm groups {GS, Baseline, PR} and {GF, RW, MF} exhibit no discernible statistical differentiation. However, EG deviates, showing a statistically significant difference from other algorithms.

In Statlog, the algorithm values within the groups {MF, Baseline}, {RW, GS}, and {GS, PR} are statistically analogous, presenting no significant variance. In stark contrast, the GF and EG algorithms demonstrate significant statistical departures from the rest, with GF presenting as the least advantageous option (-0.22) and EG approximating the most equity (-0.0053).

Finally, for the Compas dataset, the algorithms GS (-0.22) and MF (-0.21) change the Baseline (0.27) from unprivileged to privileged values. For this dataset, EG is the best algorithm (0.053). The groups {MF, GF}, {EG, GS}, and {Baseline, GS} show no significant variance. However, PR and RW showed statistically significant differences from the others.

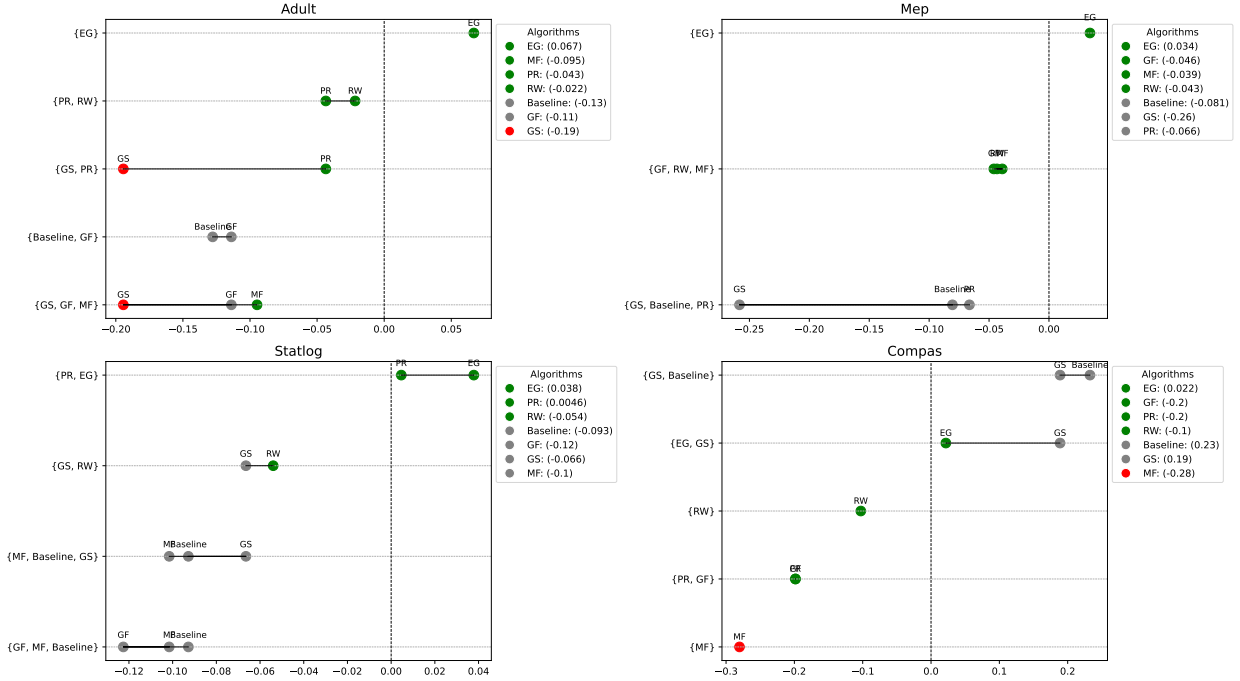


Figure 5: Impact of Fairness Algorithms on Average Odds Difference with different datasets. The algorithms are Reweighing (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot indicates a statistically significant improvement against the baseline, a red dot indicates a deterioration instead. Finally, a gray dot indicates that there is no statistically significant difference with the baseline.

4.1.3. Average Odds Difference Evaluation

Figure 5 shows the last social metric analyzed, the average odds difference. The independent variables in all the datasets exhibit different values of biases toward both privileged and unprivileged groups. Concerning the Adult dataset, the groups {GS, GF, MF}, {Baseline, GF}, {GS, PR}, and {PR, RW} show no significant variance. However, EG is significantly different from other algorithms. The worst algorithm is GS (-0.19), and the best is RW (-0.022), even if not significantly different from PR (-0.043). Even in the Mep dataset, GS (-0.26) is the worst algorithm, deteriorating the Baseline (-0.081). The best algorithm in this case is EG (0.034), which is also significantly different from the other algorithms. The algorithm groups {GS, Baseline, PR} and {GF, RW, MF} are not significantly different.

In Statlog, the GF (-0.12) is the worst algorithm, deteriorating the Baseline value (-0.093). Additionally, the groups {GF, MF, Baseline}, {MF, Baseline, GS}, and {GS, RW} are not significantly different. PR (-0.0046), who contributes to the group {PR, EG}, is the best algorithm in this case, and no algorithms are statistically significant compared to all others. For the Compas dataset, MF and RW have a statistically significant difference with respect to all algorithms. The baseline (0.23) has values trending toward the unprivileged group, and all algorithms reduce this value, even worsening as MF (-0.28) turns out to be the worst. The best algorithm is EG (0.022), suggesting that there is no statistically significant difference with GS (0.19). The groups {EG, GS} and {GS, Baseline} are not significantly different. Also, {PR, GF} show no statistically significant differences.

The analysis of the fairness metrics *mean difference*, *equal opportunity*, and *average odds difference* showed that the algorithms used have an impact and are not statistically significant to each other. Therefore, the null hypothesis is rejected $H_0^{B_i, B_j, S_{so}}$.

🔊 **Answer to RQ₁.** The findings from our RQ₁ indicate that models not utilizing bias mitigation algorithms are more susceptible to bias. Conversely, applying algorithms designed to mitigate bias across various protected attributes generally enhances social metrics, particularly with the *Exponentiated Gradient Reduction* algorithm showing notable improvements. However, the *Gerry Fair Classifier* and *Grid Search Reduction* algorithms exhibit a deterioration in the metrics across datasets when we consider multiple protected attributes. Therefore, the results show that there is a significant difference with at least one of the bias mitigation algorithms for each dataset.

4.2. RQ₂ - How do bias mitigation algorithms impact environmental sustainability regarding multiple protected attributes?

To answer our second research question (RQ₂), we analyzed the impact of bias mitigation algorithms on two environmental sustainability metrics, that are energy consumption and carbon emission, both described in Section 3.3. Through this detailed analysis, we provide an incisive perspective on the often overlooked environmental aspect of bias mitigation algorithms. To enhance readability, we have incorporated diagrams illustrating all the algorithms for energy consumption and carbon emission metrics. For additional diagrams and raw materials, please refer to the online appendix [30].

4.2.1. Energy Consumption Evaluation

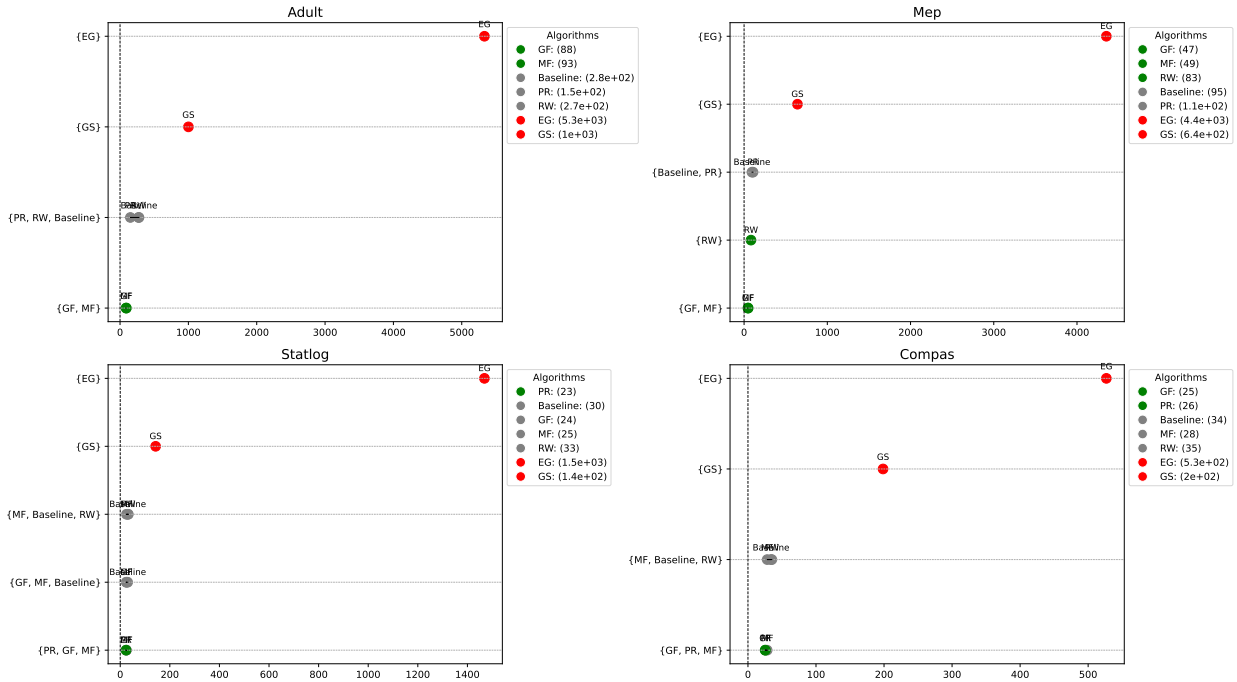


Figure 6: Impact of Fairness Algorithms on *Energy Consumption (Joule)* with different datasets. The *algorithms* are Reweighing (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot ● indicates a statistically significant improvement against the baseline, a red dot ● indicates a deterioration instead. Finally, a gray dot ● indicates that there is no statistically significant difference with the baseline.

We first analyze the energy consumption of the independent variables in each dataset. Figure 6 shows the critical difference plot for our results, which compares the energy consumption of seven independent variables with four datasets. The EG and GS algorithms are the worst for all datasets, consuming much more energy to remove bias from the models than the Baseline. Additionally, the EG and GS algorithms are statistically different.

In the Adult dataset, the algorithm group {GF, MF} and {PR, RW, Baseline} suggest no statistically significant difference. On the one hand, the algorithms GF (88 Joule) and MF (93 Joule) required less energy to train the models

compared with the Baseline (280 Joule). On the other hand, GS and EG required, respectively, over 1000 and 5000 Joule to mitigate the bias. In the same way, for the Mep dataset, the same algorithm groups {GF, MF} and {PR, Baseline} suggest that there is no statistically significant difference with GF (47 Joule) and MF (49 Joule), requiring less energy than the Baseline (95 Joule). Even in this case, GS and EG required more energy than the Baseline, respectively, over 640 and 4400 Joule. For the Statlog dataset, the algorithm groups {PR, GF, MF}, {MF, Baseline, RW} and {GF, MF, Baseline} are not significantly different, with PR (23 Joule) and GF (24 Joule) having the lowest values and EG and GS the highest.

In the same way, for the Compas dataset, the algorithm groups {GF, PR, MF} and {MF, Baseline, RW} are not significantly different; the algorithms GF (25 Joule) and PR (26 Joule) required less energy than the Baseline (34 Joule). Even in this case, the GS and EG algorithms use more energy than the Baseline, respectively, over 200 and 530 Joule.

4.2.2. Carbon Emission Evaluation

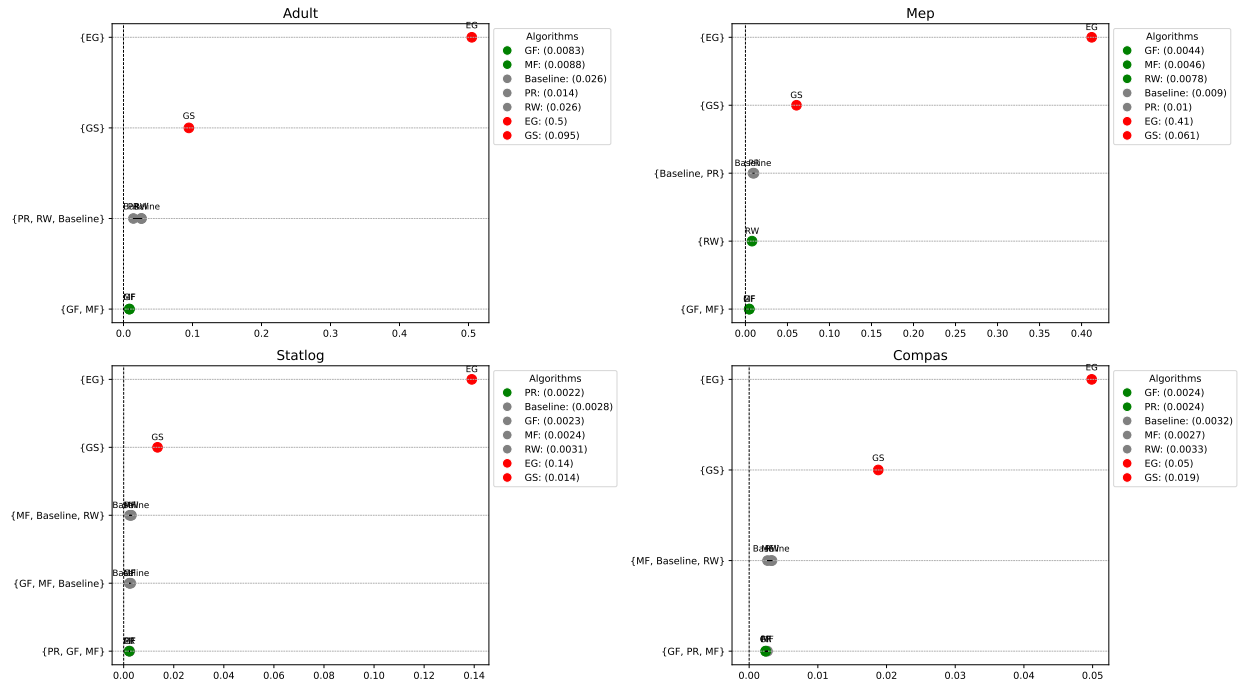


Figure 7: Impact of Fairness Algorithms on *Carbon Emission* (g/CO2) with different datasets. The *algorithms* are Reweighting (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot indicates a statistically significant improvement against the baseline, a red dot indicates a deterioration instead. Finally, a gray dot indicates that there is no statistically significant difference with the baseline.

Figure 7 shows a visual comparison of the environmental impact of several fairness algorithms, as measured by carbon emission in grams of CO2, across four datasets. The EG algorithm appears to produce the most carbon emissions for all datasets, reaching 0.5 grams of CO2 for the Adult dataset and 0.41 grams for the MEP dataset, highlighting its substantial environmental footprint.

Among the algorithms, EG and GS demonstrate statistically significant differences in carbon emissions across the datasets. For the Adult dataset, algorithms in the group {GF, MF} showed no statistical difference from each other, and the GF (0.0083 grams) and MF (0.0088 grams) algorithms emit significantly lower carbon dioxide than the Baseline (0.026 grams). Conversely, the PR algorithm (0.014 grams) cuts emissions compared to both the Baseline (0.026 grams) and RW (0.026 grams), hence leading to algorithms in the group {PR, RW, Baseline} having no statistical distinction between them. Similarly, two algorithm groups in the MEP dataset showed no statistical difference between their components: {GF, MF} and {Baseline PR}. Particularly, GF (0.0044 grams) and MF (0.0046 grams) reduce

emissions compared to the baseline. RW's emissions (0.0078 grams) are also lower than those of the Baseline (0.009 grams) and PR (0.01 grams), and these last two are statistically indistinguishable.

In the Statlog dataset, both PR (0.0022 grams), GF (0.0023 grams), and MF (0.0024 grams) show significantly lower emissions, resulting in no statistical difference within the group {PR, GF, MF}. The MF algorithm also undercuts the Baseline (0.0028 grams) and RW (0.0031 grams) in emissions without marked statistical separation. In addition, the algorithms GF, MF, and Baseline resulted as non-statistically different, hence resulting in two other groups: {GF, MF, Baseline} and {MF, Baseline, RW}. Finally, in the Compas dataset, we only have two groups of algorithms within which there is no statistical difference: {GF, PR, MF} and {MF, RW, Baseline}. GF and PR each emit (0.0024 grams) of CO₂, lower than the Baseline (0.0032 grams), alongside MF with (0.0027 grams) of emissions. Also, the second group, comprising MF (0.0027 grams) and RW (0.0033 grams), exhibits no significant differences in carbon emissions from the baseline. The analysis of the environmental metrics *carbon emission and energy consumption* showed that the algorithms used have an impact and are not statistically significant to each other. Therefore, the null hypothesis is rejected $H_0^{B_i, B_j, S_{en}}$.

👉 **Answer to RQ₂.** Our analysis of RQ₂ reveals the profound impact of the chosen algorithms on both CO₂ emissions and energy consumption during the model production phase. Interestingly, the GF algorithm is the most energy-efficient, boasting the lowest emission rates. On the other hand, the EG and GS algorithms present a sharp contrast, increasing energy consumption and, as a result, significantly exceeding the emission levels of the baseline model. In comparison, the EG and GS algorithms increase energy use, leading to higher emissions than the Baseline models. In contrast, the MF and PR algorithms were particularly efficient, requiring less energy for model training. In this case, the results show a significant difference with at least two bias mitigation algorithms for each dataset.

4.3. RQ₃ - How do bias mitigation algorithms impact economic sustainability regarding multiple protected attributes?

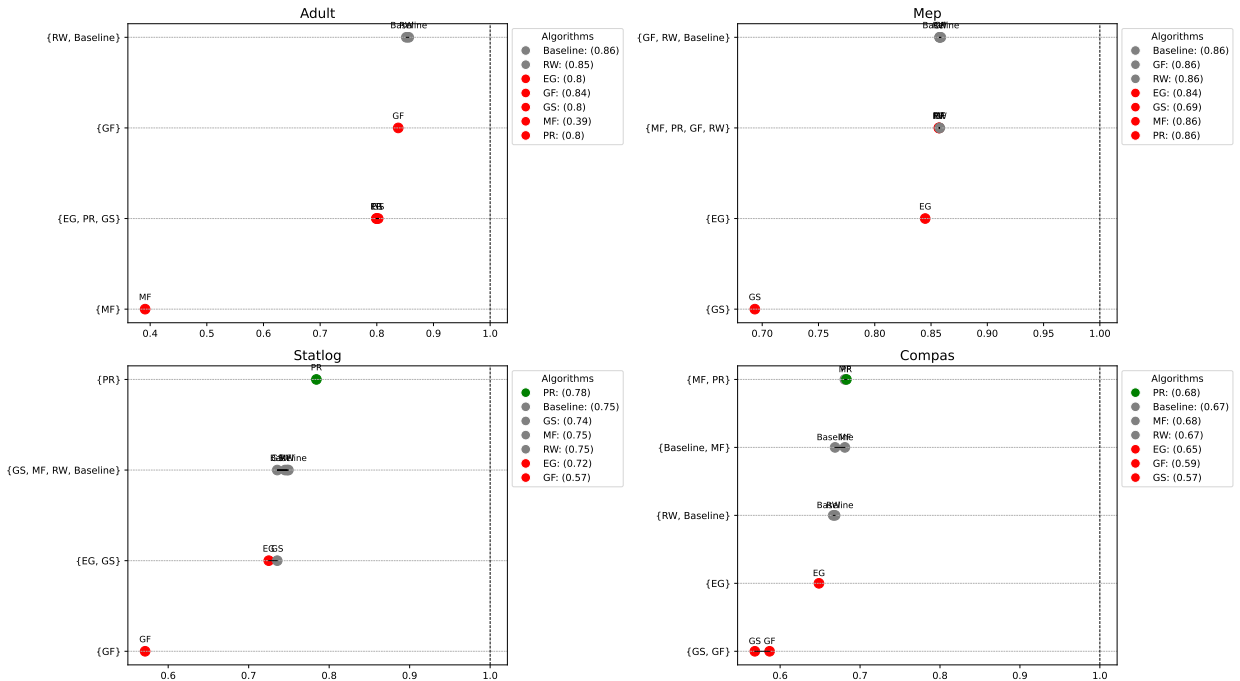


Figure 8: Impact of Fairness Algorithms on Accuracy with different datasets. The algorithms are Reweighting (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot indicates a statistically significant improvement against the baseline, a red dot indicates a deterioration instead. Finally, a gray dot indicates that there is no statistically significant difference with the baseline.

In our third research question (**RQ3**), we aim to explore the effects of bias mitigation algorithms on economic sustainability. We consider the models' accuracy, F1 score, training time, and storage weight to assess such aspects. These metrics are described in Section 3.3. Due to readability issues, we have incorporated diagrams illustrating all the algorithms for training time and storage weight metrics; additional diagrams and raw materials are in the online appendix [30]

4.3.1. Accuracy Evaluation

Our analysis begins with examining Accuracy as depicted in Figure 8. Starting with the Adult dataset, we observe a decline in accuracy metrics across all algorithms when compared to the Baseline. Notably, RW's accuracy dips slightly to (85%) from the Baseline's (86%), a difference that is not statistically significant. However, MF's accuracy significantly drops to (39%), indicating a substantial deviation. The algorithms in the group {EG, PR, GS} maintain an accuracy of 80% without significant statistical divergence. In Mep, the Baseline (86%) has the same accuracy as the others in the group {GF, RW, Baseline} and is not statistically different between them. Additionally, the group {MF, PR, GF, RW} is not statistically different. In this case, GS is the worst algorithm (69%) followed by EG (84%), with GS and EG statistically different. For the Statlog dataset, the PR algorithm shows an accuracy of (78%) from the Baseline's (75%), standing out as statistically significant. The groups {GS, MF, RW, Baseline} and {EG, GS} reveal no significant accuracy differences since they all hover around (75%) accuracy. Conversely, GF shows (57%) accuracy, underperforming significantly with respect to the others. In the Compas dataset, both PR and MF algorithms match in accuracy at (68%), outperforming the Baseline's (67%). The groups {RW, Baseline}, {Baseline, MF} and {GS, GF} are not statistically different. Here, GS and GF are the least effective, with accuracy of (57%) and (59%), respectively. EG is statistically different to the others.

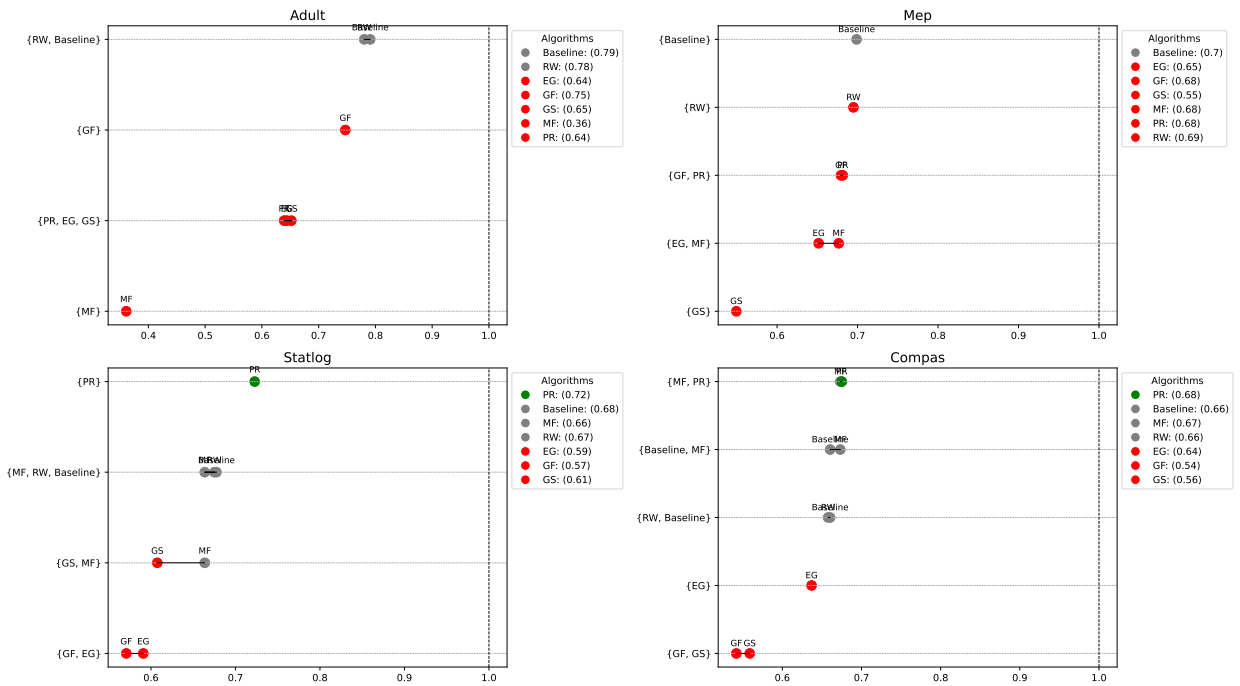


Figure 9: Impact of Fairness Algorithms on *F1 score* with different datasets. The *algorithms* are Reweighting (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot ● indicates a statistically significant improvement against the baseline, a red dot ● indicates a deterioration instead. Finally, a gray dot ● indicates that there is no statistically significant difference with the baseline.

4.3.2. F1 score Evaluation

Examining the F1 score, we found similar results with accuracy evaluation. Figure 9 shows the result of the F1 score across the datasets. In Adult, the Baseline (79%) shows a higher F1 score than the algorithms but is not

statistically different from RW (78%). Also, the group {PR, EG, GS} is not statistically different. Conversely, MF (36%) is the worst algorithm, decreasing the percentage considerably. The MF and GF are statistically different from the others.

For the Mep data set, the baseline (70%) also shows higher values, but in this case, it is statistically different. No statistical differences are found between the pairs {EG, MF} and {GF, PR}, indicating similar values within these groups. GS (55%) and RW (69%) are statistically different, with the GS algorithm being the worst in this case.

In Statlog, the PR (72%) algorithm surpasses the F1 score of the Baseline (68%), and PR is statistically different from other algorithms. The {MF, RW, Baseline}, {GS, MF}, and {GF, EG} groups show no statistical differences. GF (57%) is the worst in this case.

In Compas, PR (68%) and MF(78%) algorithms, besides not being statistically different, surpass the F1 score of the Baseline (66%), and there is no statistical difference in the group {Baseline, MF}. Also, the groups {RW, Baseline}, and {GF, GS} are not statistically different, with GF (54%) and GS (56%) having the lowest values. Only the EG algorithm, in this case, is statistically different.

4.3.3. Training Time Evaluation

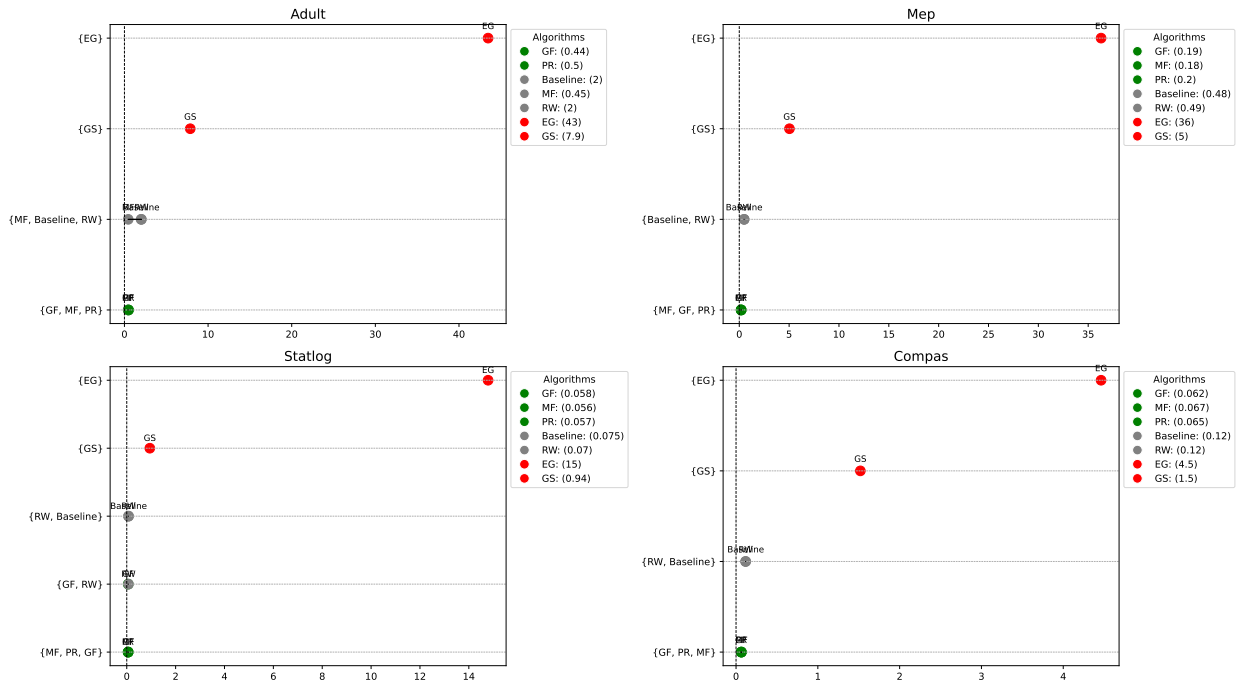


Figure 10: Impact of Fairness Algorithms on *Training Time (seconds)* with different datasets. The *algorithms* are Reweighting (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot indicates a statistically significant improvement against the baseline, a red dot indicates a deterioration instead. Finally, a gray dot indicates that there is no statistically significant difference with the baseline.

Figure 10 details the training time required for each dataset, indicating that EG and GS consistently require the longest training times across all datasets, marking them as statistically distinct from other algorithms. This observation aligns with the trends identified in RQ2. In the Adult dataset, the EG algorithm takes significantly longer to train (43 seconds) than the Baseline's (2 seconds). The groups {GF, MF, PR} and {MF, Baseline, RW} do not show statistical differences in training time, despite GF (0.44 sec), MF (0.45 sec), and PR (0.5 sec) being quicker than the others.

In Mep, the groups {MF, GF, PR} and {Baseline, RW} exhibit no statistical difference in training time, with MF (0.18 sec), GF (0.19 sec), and PR (0.2sec) requiring less time than Baseline (0.48 sec). Even here, EG (36 sec) and GS (5 sec) are the worst algorithms.

For the Statlog dataset, the groups {GF, RW}, {RW, Baseline} and {MF, PR, GF} are not statistically different, with MF (0.056 sec) and PR (0.047 sec) required less time than Baseline (0.075 sec). EG (15 sec) and GS (0.94 sec) are outliers, again needing longer than the rest.

In Compas, the groups {GF, PR, MF} and {RW, Baseline} are not statistically different with GF (0.062 sec), PR (0.065 sec), and MF (0.067 sec) that required less training time than Baseline (0.12 sec). The algorithms EG (4.5 sec) and GS (1.5 sec), even in this case, required more training time than others.

4.3.4. Storage Weight Evaluation

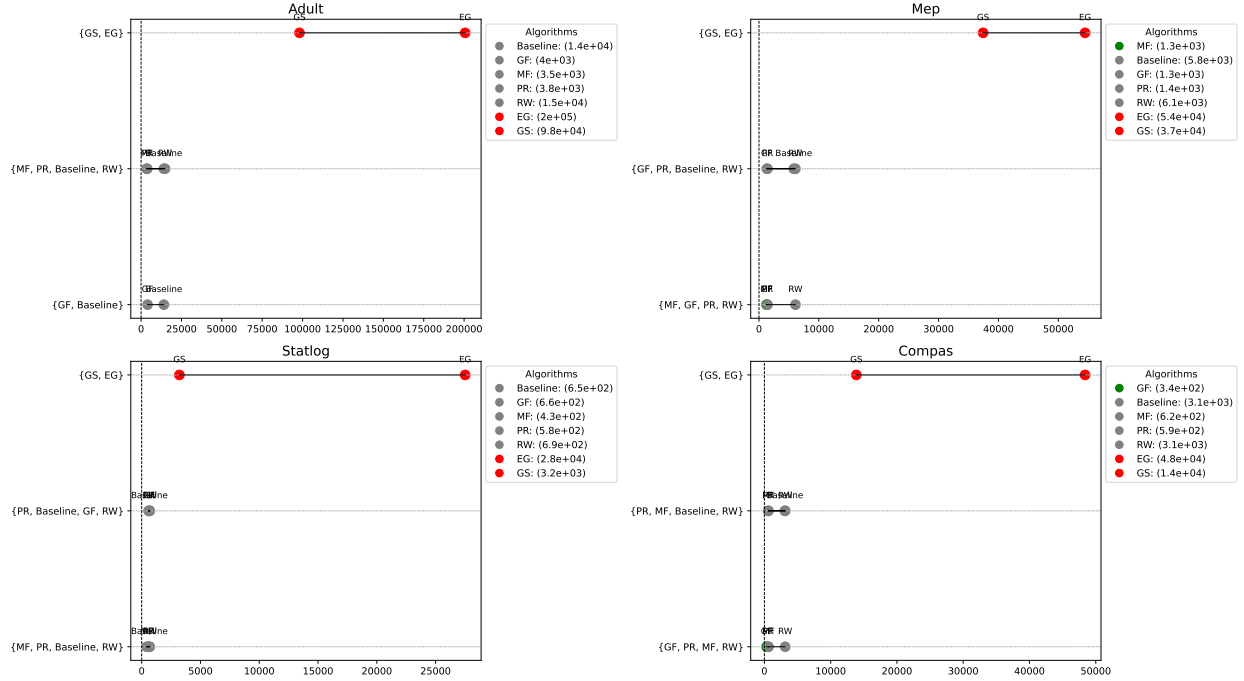


Figure 11: Impact of Fairness Algorithms on *Storage Weight* (KB) with different datasets. The *algorithms* are Reweighing (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). A green dot ● indicates a statistically significant improvement against the baseline, a red dot ● indicates a deterioration instead. Finally, a gray dot ● indicates that there is no statistically significant difference with the baseline.

The last metric considered for the RQ_3 is storage weight, depicted in Figure 11. In the Adult dataset, no statistical differences are shown in the groups {MF, PR, Baseline, RW}, {GF, Baseline} and {GS, EG}. The GS and EG algorithms increase the model storage weight, respectively, over 98.000 and 200.000 KB, comparing them with the Baseline, which is only 14.000 KB. On the other hand, MF (3.500 KB), GF (3.800 KB), and PR (4.000 KB), even if there are no statistical differences with the Baseline, reduced the storage weight. In Mep, the groups {GF, PR, Baseline, RW}, {MF, GF, PR, RW} and {GS, EG} showed no statistical differences. However, GS (37.000 KB) and EG (54.000 KB) algorithms are the worst compared with others. Even in these case the algorithms MF (1.300 KB), GF(1.300 KB), and PR (1.400 KB) reduced the storage weight, even though there were no statistical differences with the Baseline. In the Statlog dataset, the group {GS, EG}, {PR, Baseline, GF, RW} and {MF, PR, Baseline, RW} are not statistical difference. Even in this case, GS (3.200 KB) and EG (28.000 KB) have the worst values.

For the Compas dataset the GS (48.000 KB) and EG (14.000 KB) algorithms are the worst compared to others, and they also are not statistically different. On the other hand, the groups {GF, PR, MF, RW} and {PR, MF, Baseline, RW} are not statistically different. GF (3.400 KB) is the best algorithm, even if not statistically different from others. The analysis of the economic metrics *accuracy*, *f1 score*, *training time*, *storage weight* showed that the algorithms used have an impact and are not statistically significant to each other. Therefore, the null hypothesis is rejected $H_0^{B_i, B_j, S_{ec}}$.

☛ **Answer to RQ₃.** Our investigation of RQ₃ reveals the influence of the chosen algorithms on the economic metrics evaluated. This finding underscores the dual potential of these techniques to improve or decrease model quality. The Accuracy and F1 score of Prejudice Remover in Statlog and Compas is higher than the Baseline. On the other hand, the other algorithms significantly reduced these metrics. Exponentiated Gradient Reduction and Grid Search Reduction algorithms demonstrate higher training times across datasets, indicating increased computational costs, and are statistically different. Additionally, these algorithms notably increase storage weight. Conversely, some algorithms manage to slightly reduce storage weight, though not significantly compared to the Baseline.

5. Discussion and Implications

In the following section, we first discuss the main results of the study, attempting to provide insights into the added value brought by our research to the current state of the art. Secondly, we provide the implications that our findings have for researchers and practitioners.

5.1. Discussion of the Findings

At first, we discuss our findings in relation to the state of the art and the qualitative insights that may explain them.

On the relation with the state of the art. Our study builds on and extends the existing body of knowledge by addressing critical gaps and offering a broader perspective on the interplay between the three sustainability dimensions in ML-enabled systems. First, regarding RQ₁, our results corroborate previous findings that position fairness as a cornerstone of social sustainability. While our results align with earlier work, two factors distinguish their significance. By replicating and validating previous findings, our study enhances the *ecological validity* [70] of these insights, i.e., the extent to which experimental results can be generalized to real-world applications. Moreover, our analysis positions fairness as a foundational element for examining its interplay with other dimensions of sustainability, such as environmental and economic. We believe that such an integrated perspective enables the exploration of trade-offs among fairness, energy consumption, and resource efficiency, and our contribution leads to further investigation of these relationships. Second, as for RQ₂, our work addresses a knowledge gap by systematically examining the environmental implications of bias mitigation algorithms. By operationalizing environmental sustainability through metrics such as energy consumption and carbon emissions, we provide new insights into the ecological trade-offs introduced by these algorithms. Our findings reveal that while certain algorithms, e.g., *Exponentiated Gradient Reduction*, significantly improve fairness, they also exhibit substantial energy demands and higher carbon emissions. Conversely, computationally efficient algorithms, e.g., *Meta Fair Classifier* and *Gerry Fair Classifier*, demonstrate lower environmental impacts, providing practitioners with actionable strategies to balance fairness with ecological responsibility. These results contribute to a more holistic understanding of the sustainability of ML-enabled systems by emphasizing the need to select fairness techniques carefully during the design process, aiming at finding a suitable solution considering the trade-off between the sustainability aspects.

As for RQ₃, our research expands the scope of fairness studies by framing the evaluation of bias mitigation algorithms within the economic sustainability dimension. As reported earlier, previous work has predominantly assessed these techniques through the lens of accuracy, not considering broader operational implications such as training time and storage requirements. By incorporating these additional metrics, we offer a multi-faceted view of the economic trade-offs associated with fairness-enhancing techniques. Our findings reveal that while certain bias mitigation algorithms introduce computational overhead, such as *Exponentiated Gradient Reduction* and *Prejudice Remover*, they also open avenues for balancing fairness with economic feasibility. Such a perspective enriches the current body of knowledge, providing a framework to assess the practical and financial implications of implementing fairness techniques in real-world settings.

In summary, our empirical study advances the state of the art by integrating fairness into a comprehensive sustainability framework that spans its social, environmental, and economic dimensions. This integrated approach provides a holistic understanding of the trade-offs introduced by bias mitigation techniques, equipping practitioners and researchers with the tools to make informed decisions that balance ethical, operational, and environmental considerations in the design of sustainable and responsible ML-enabled systems.

Bias Mitigation Algorithms Versus Sustainability: A Qualitative Perspective. While our work primarily presented a benchmark study with a quantitative focus, it might also provide qualitative insights into the potential reasons underlying the observed results. Our findings suggest multiple considerations that highlight the complex interplay between bias mitigation techniques, dataset characteristics, and the stochastic nature of ML processes, which open avenues for further investigation. A first observation from our results is the variability in the performance of bias mitigation algorithms across different datasets, particularly in terms of energy consumption and resource efficiency. This suggests that the characteristics of the datasets themselves—such as feature dimensionality, group imbalances, and overall complexity—can significantly impact determining the results of these algorithms. Specifically, datasets with high dimensionality or imbalanced distributions between privileged and unprivileged groups may require additional computational resources for bias mitigation techniques to converge effectively. This observation is particularly relevant to **RQ₁**, as our results indicate that fairness outcomes are not solely determined by the algorithms but also by the data properties that interact with algorithmic mechanisms. For example, in the Compas dataset, all three social metrics used have higher values than the other datasets; this could be due to the number of features being lower than the other datasets. In this respect, future research should explore these interactions to provide more tailored guidance on selecting appropriate algorithms based on dataset characteristics.

Another perspective relates to the intrinsic computational characteristics of bias mitigation algorithms themselves. Pre-processing methods such as *Reweighting*, which involve data manipulation that reduces the dataset size, tend to exhibit lower computational overhead. Conversely, in-processing methods like *Exponentiated Gradient Reduction*, which rely on iterative optimization steps to improve discrimination, result in higher energy consumption and longer execution times. While these theoretical expectations provide a framework for anticipating the impact of bias mitigation algorithms on computational performance, our findings also suggest that empirical results may deviate from these expectations due to variations in implementation efficiency, interactions with specific datasets, and differences in model configurations. These insights emphasize the importance of our benchmark study, which uses quantitative measurements under controlled conditions to capture the practical effects of bias mitigation algorithms. This perspective is directly connected to **RQ₂**, as it illustrates how algorithm-specific factors interact with sustainability metrics.

In addition, the stochastic nature of ML processes introduces further variability into sustainability metrics. Variations in random initialization, data splits, and probabilistic elements within the algorithms themselves can lead to fluctuations in energy consumption and training time. These stochastic effects may also explain unexpected findings, such as certain algorithms achieving lower resource consumption despite their inherent computational complexity. Understanding how randomness influences sustainability outcomes is critical for improving the robustness and predictability of bias mitigation techniques in real-world applications. Future research could extend these insights by examining how individual ML algorithms are affected by different bias mitigation strategies, enabling more responsible algorithm selection and promoting better sustainability practices.

Finally, our findings indicate that some bias mitigation algorithms may inadvertently regularize model behavior, simplifying decision boundaries or reducing overfitting. For instance, algorithms such as *Meta Fair Classifier* or *Prejudice Remover* might promote less complex models that require fewer computational resources, thereby reducing energy consumption or training time. While this phenomenon is promising from a sustainability perspective, it raises important questions about the trade-offs between simplicity and fairness. This aspect raises the critical issue of whether such regularization aligns with the intended fairness objectives or results in unintended biases in other areas. These considerations are particularly relevant in the context of **RQ₃**, as they highlight the need to balance fairness improvements with resource efficiency and operational feasibility.

In conclusion, we acknowledge that future research should complement our findings with qualitative methods, like case studies, interpretability analyses, or interviews with practitioners, to uncover the underlying causes of the observed phenomena. The insights of these qualitative analyses would provide a more comprehensive understanding of the nature of bias mitigation algorithms and the trade-offs between fairness and sustainability in bias mitigation algorithms, possibly guiding the development of more effective and efficient approaches.

5.2. Implications of the Study

The results of our work have a number of practical implications for researchers, practitioners, and project managers, which we elaborate in the following.

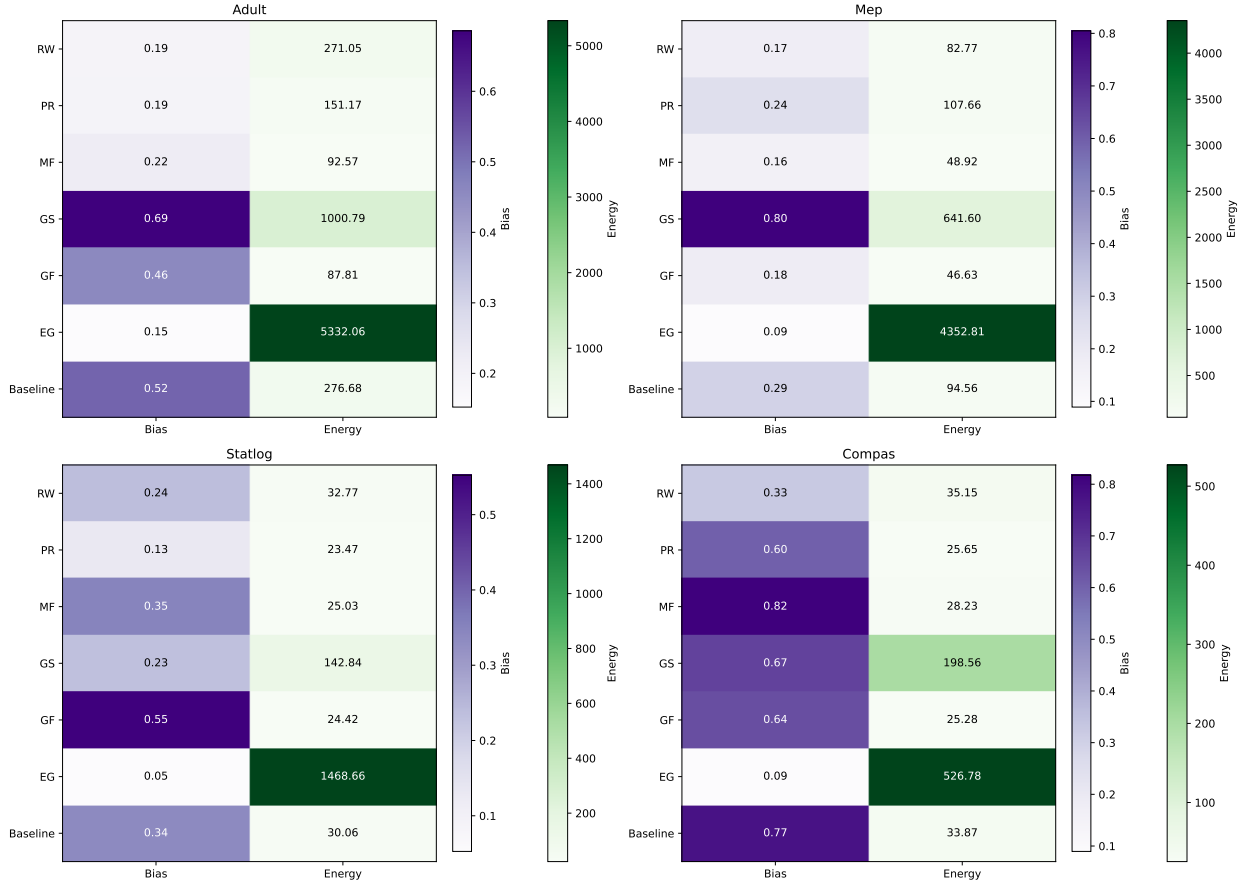


Figure 12: Energy Consumption Vs. Cumulative Bias for each dataset. The *algorithms* are Reweighing (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). The numerical values represent cumulative bias (purple), where lower values are better, and energy consumption (green), where lower values are better.

On the need for trade-off analysis to deploy models in real scenarios. The deployment of machine learning systems with bias mitigation algorithms in real-world scenarios often requires practitioners to navigate complex trade-offs among multiple non-functional requirements, such as fairness, energy efficiency, storage costs, and accuracy. Our findings highlight that these trade-offs are not merely theoretical but manifest concretely in different algorithms' performance across various sustainability dimensions. These observations highlight the importance of systematic trade-off analyses during the development and deployment phases of ML models. Current research has studied the balance between accuracy and fairness [24, 49], accuracy and energy efficiency [80] and, more recently, fairness and efficiency [16]. However, our study expands this understanding by jointly studying the three dimensions of sustainability and exploring interactions among additional non-functional requirements, such as storage weight, training time, and carbon emissions.

To address these challenges, practitioners must prioritize trade-offs based on application-specific requirements. For instance, in critical domains where fairness cannot be compromised, high-energy-consuming algorithms such as *Exponentiated Gradient Reduction* might be justified. Conversely, for edge devices or IoT applications where energy and storage constraints dominate, algorithms like *Reweighing* or *Gerry Fair Classifier* may provide a better balance of fairness and resource efficiency. These considerations highlight the importance of designing adaptable systems that can accommodate varying trade-off priorities depending on the deployment context.

Beyond the findings of this study, there is an urgent need for tools and frameworks that can assist practitioners in understanding and optimizing these trade-offs—for example, visualizations that compare cumulative fairness metrics with energy and storage requirements, as demonstrated in Figures 12 and 13, could help stakeholders make informed

decisions. Moreover, future work should explore adaptive, context-aware algorithmic approaches that dynamically adjust their behavior to balance multiple non-functional requirements under specific constraints.

✚ Implication 1. While researchers are starting investigating novel approaches to handle the trade-offs among multiple non-functional requirements, our findings represent a call for further research actions aiming to quantify, optimize, and visualize these trade-offs.

✚ Implication 2. Future research should focus on adaptive and context-aware bias mitigation algorithms capable of dynamically balancing fairness and other sustainability metrics based on application-specific requirements. These advancements could ensure that ML-enabled systems remain effective and efficient in diverse real-world scenarios.

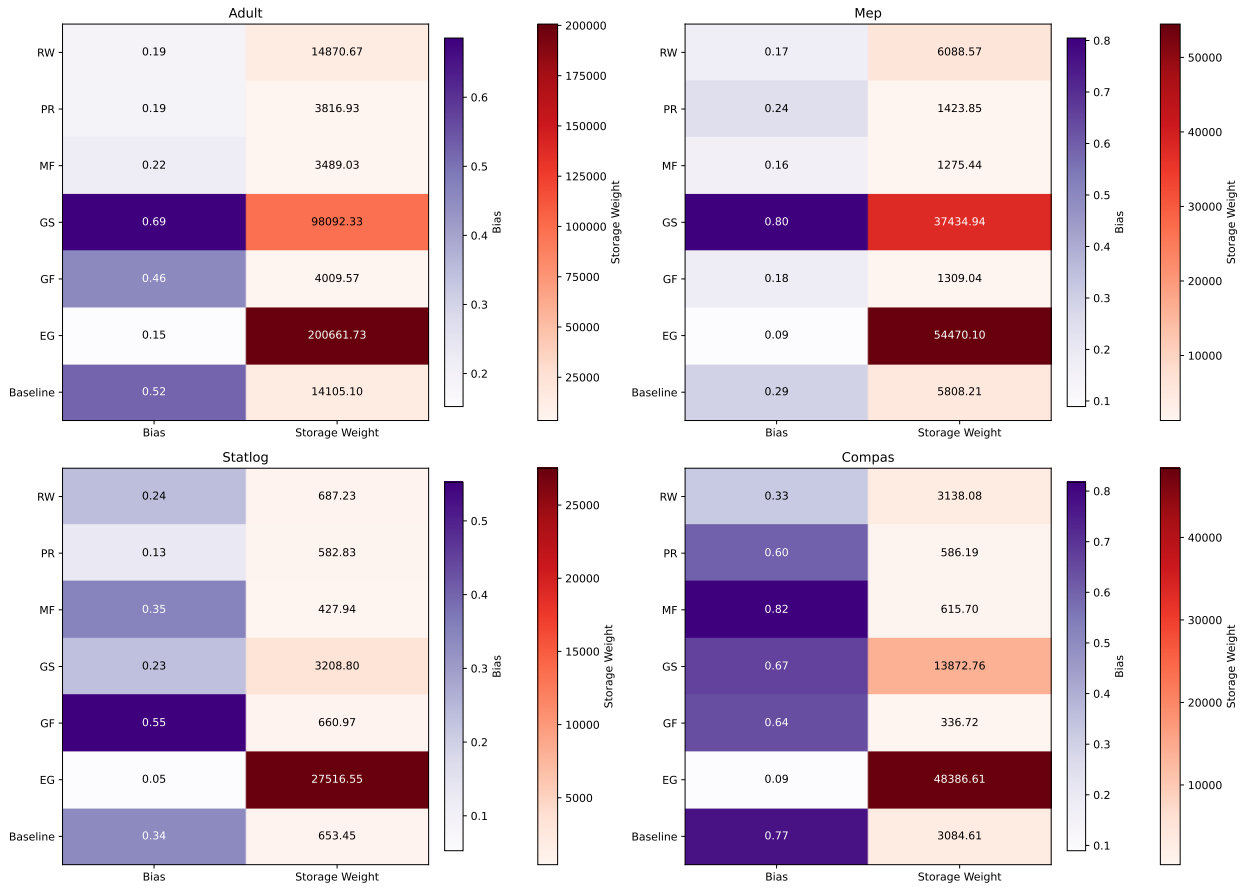


Figure 13: Cumulative Bias Vs. Storage Weight for each dataset. The *algorithms* are Reweighing (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). The numerical values represent cumulative bias (purple), where lower values are better, and storage weight (red), where lower values are better.

For the sake of completeness, let us further elaborate on these matters by discussing the figures in detail, highlighting the trade-offs between the various metrics analyzed in this study. Figure 12 provides a visualization of the trade-off between cumulative bias [7], calculated as the sum of the absolute values of the three social fairness metrics, and energy consumption for all algorithms across the datasets. As shown in the figure, a high energy demand is associated with fairness improvements, particularly for *Exponentiated Gradient Reduction*, which achieves the highest fairness improvements but consistently creates substantial energy costs. At the same time, *Reweighing* and *Prejudice Remover* demonstrate a more balanced trade-off by delivering moderate fairness gains while consuming significantly less energy. These insights are potentially interesting for practitioners operating in energy-constrained environments,

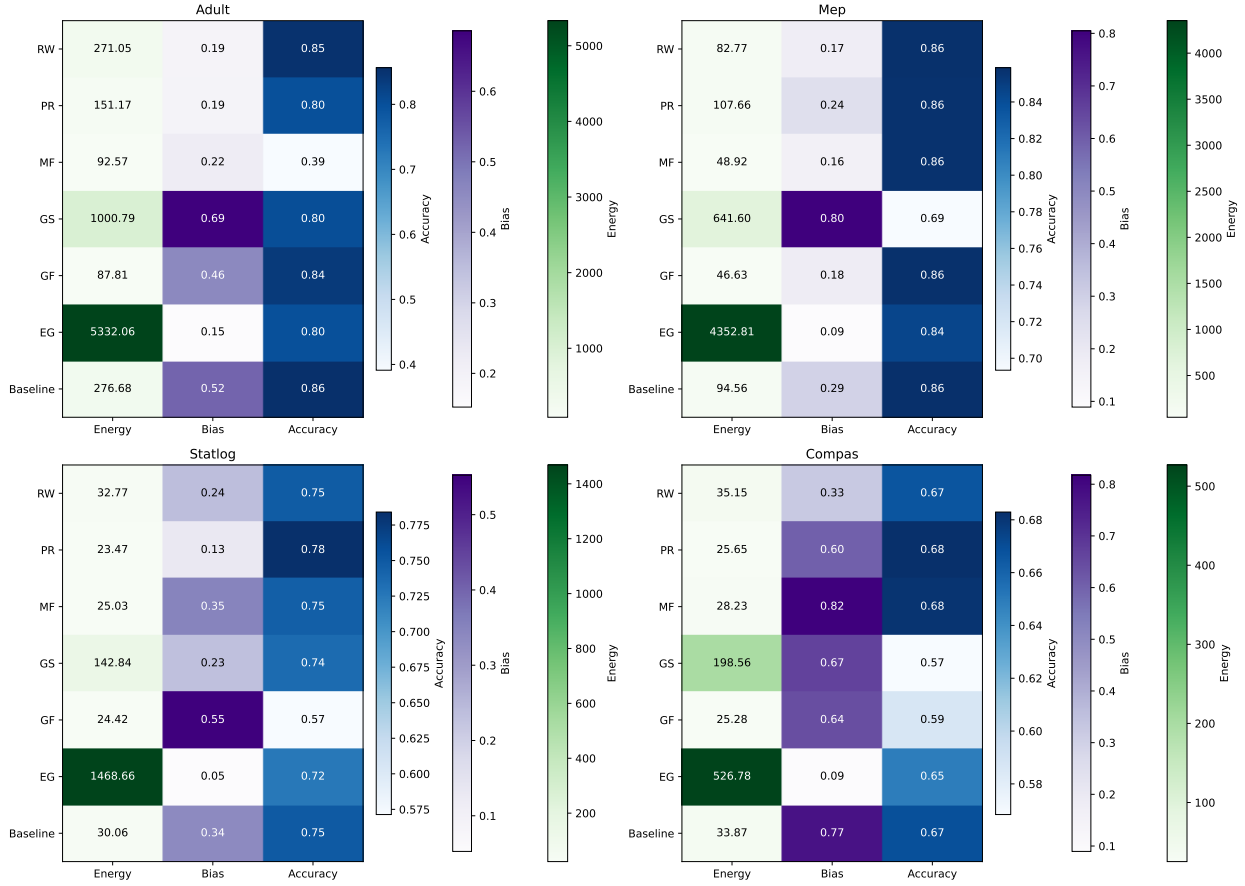


Figure 14: Cumulative Bias Vs. Energy Consumption Vs. Accuracy for each dataset. The *algorithms* are Reweighting (RW), Meta Fair Classifier (MF), Gerry Fair Classifier (GF), Exponentiated Gradient Reduction (EG), PrejudiceRemover (PR), Grid Search Reduction (GS). The numerical values represent cumulative bias (purple), where lower values are better; energy consumption (green), where lower values are better; and accuracy (blue), where higher values are better.

where energy efficiency must be weighed carefully against fairness objectives.

Figure 13 shows the relationship between cumulative bias and storage weight. The results indicate that algorithms like *Reweighting* and *Prejudice Remover* maintain lower storage footprints while achieving significant fairness improvements. These characteristics make them highly suitable for resource-constrained deployments, such as IoT or mobile applications. At the same time, algorithms like *Exponentiated Gradient Reduction* and *Grid Search Reduction* demonstrate a trend of higher storage costs relative to their bias reduction capabilities, emphasizing the importance of assessing average algorithmic behaviors when making deployment decisions in environments with strict storage constraints. This analysis highlights the importance of aligning algorithmic choices with resource-related non-functional requirements in real-world scenarios.

Figure 14 provides a multi-dimensional perspective by illustrating the interplay among cumulative bias, energy consumption, and accuracy. This analysis highlights the intricate trade-offs that practitioners must navigate. The results highlight that no algorithm achieves optimal performance across all three dimensions simultaneously. *Exponentiated Gradient Reduction*, for instance, delivers the highest fairness improvements but at the expense of accuracy and energy efficiency. In contrast, *Reweighting* and *Meta Fair Classifier* provide a more balanced performance, making them more appropriate for use cases that require a compromise between fairness, accuracy, and energy consumption. This figure reinforces the importance of a holistic evaluation framework that considers multiple non-functional requirements simultaneously, tailored to specific deployment contexts. As a consequence of this analysis, we may argue that there is a need for decision-making tools that help practitioners navigate these trade-offs based on their specific

operational constraints.

✚ **Implication 3.** Practitioners would need decision-support systems that prioritize deployment context when selecting bias mitigation algorithms. These systems should integrate multi-objective optimization techniques to evaluate trade-offs between fairness, energy consumption, storage efficiency, and accuracy dynamically. Such tools could guide algorithm selection based on the operational constraints and goals of specific applications, such as IoT deployments, energy-intensive systems, or high-stakes domains.

Implications for AI Engineering research. Recent research in Software Engineering for Artificial Intelligence has introduced approaches designed to manage multiple protected attributes concurrently [19, 25, 65]. Despite this, popular bias mitigation toolkits like AIF360 and Fairlearn have yet to integrate these advancements, and this results in a lack of experimentation of these solutions in studies like ours. Implementing these approaches in such toolkits would enable developers to reduce model bias in their systems while providing researchers with opportunities to explore additional non-functional requirements. Moreover, multi-objective optimization in bias mitigation should go beyond the traditional fairness-accuracy trade-off. By leveraging our findings and adopting our sustainability framework, researchers should create solutions that consider other non-functional requirements to produce higher-quality systems that align with business needs. Our study specifically addressed sustainability requirements, but extending multi-objective optimization to include a broader range of requirements could help overcome existing limitations and facilitate the creation of responsible and robust ML tools.

✚ **Implication 4.** Our work emphasizes the need to integrate existing bias mitigation approaches from the literature—particularly those capable of addressing multiple protected attributes simultaneously—into widely used tools. Furthermore, there is a pressing need for new approaches that can effectively handle multiple non-functional requirements, striking a balance between them without compromising any single requirement.

From a broader perspective, the results of our research may improve standards and guidelines designed to manage the non-functional requirements of ML-enabled systems. In particular, our findings have the potential to serve as valuable resources that support and augment existing or developing standards such as *ISO/IEC 25059* and *ISO/IEC 20226*.⁶ These standards, which are currently emerging in the field, aim to establish a clear and comprehensive set of criteria for assessing the quality of software systems. By integrating our findings, it is conceivable that these standards could be further enriched, providing a more robust and practical guide for operators and developers who need to handle non-functional requirements specific to ML-enabled systems. For example, adding guidelines for systematically evaluating trade-offs between energy consumption and accuracy could guide developers in creating more sustainable ML-enabled systems. This integration could facilitate a more systematic and standardized approach to ensuring that these systems are functional but also lightweight, unbiased, and accurate. Moreover, our results have implications at multiple levels. Beyond technical optimizations, they inform policy and strategic decisions within organizations, suggesting a framework for integrating non-functional requirements into the regular standards development workflow. This holistic approach can help institutions create more robust and practical guidelines for developing ML-enabled systems.

✚ **Implication 5.** Our RQs provide insights that may be useful in extending and supplementing emerging regulatory standards by describing relevant non-functional requirements and offering better criteria for evaluating the trade-offs of ML-enabled systems. Moreover, our results highlight current research gaps for further study of non-functional requirements of ML-enabled systems, suggesting that these requirements should be integrated into the development process. This work encourages deeper investigation into how non-functional factors influence system design and performance, potentially defining new directions for future research.

⁶The *ISO/IEC 25059* standard: <https://iso25000.com/index.php/en/iso-25000-standards/iso-25059> and *ISO/IEC 20226* standard: <https://www.iso.org/standard/86177.html>

6. Threats to Validity

Some design decisions might have introduced threats to the validity of our study. The following sections discuss these potential limitations and how we addressed them.

6.1. Construct Validity

The first threat in this category concerns the reliability of the subjects selected for the study. We relied on well-known datasets employed by previous researchers in the field [59, 7, 49, 24, 19]. Their use allowed us to perform an accurate comparison with the results achieved in literature (**RQ₁**), other than advancing the current body of knowledge with additional insights into the impact of bias mitigation algorithms on sustainability properties of ML-enabled systems (**RQ₂** and **RQ₃**). In addition, the datasets selected spanned across different application domains, allowing us to provide a larger sustainability analysis of the models experimented.

Similarly, our study features a set of machine learning models exploited by previous researchers [8, 79, 19, 86, 78]: this allowed us to contrast our findings with previous ones. As for the dependent and independent variables, we relied on the implementation provided by state-of-the-art toolkits and libraries such as AI FAIRNESS 360 toolkit [6], CODECARBON, and SCIKIT-LEARN. More in general, we followed the guidelines by Wohlin et al. [83] and the *ACM/SIGSOFT Empirical Standards* to reduce the risk that our explanation of operational constructs would be inadequate.

6.2. Internal Validity

When addressing internal validity concerns, we considered factors that may have potentially influenced the outcomes of the study. The real-time energy consumption and performance metrics computed in the context of **RQ₂** were subject to variables like background processes, which could significantly affect measurements. We sought to minimize these effects by enabling airplane mode and terminating non-essential processes, though fully controlling operating system workloads and background operations is complex.

This variability introduces an element of uncertainty in our results, which is a common issue in similar studies (e.g., [38, 85]). To enhance stability, we adhered to established protocols by implementing a warm-up execution and one-minute pause between each task execution, aligning with strategies from previous research [38, 67, 29]. Still, in terms of **RQ₂**, we are aware that the CODECARBON's energy consumption is slightly lower than the actual consumption. This limitation may have led to a slight underestimation of our findings. While we required to use CODECARBON because of the lack of dedicated hardware to make measurements with physical devices, future studies might assess the impact of software and hardware measurement on our findings [85]. Further aspect to discuss is the sustainability metrics selected in the study. Each dimension considered in the study, i.e., social, environmental, and economic, encompasses a large number of metrics: incorporating all of them would have been challenging because of (i) the unavailability of automated measurement instruments or (ii) the decrease of explainability that our study would have obtained, as analyzing all of them would have made data harder to be analyzed and explained.

To overcome this limitation, we defined nine metrics covering each dimension. As for social sustainability, we used three equity metrics that have been significantly adopted in the literature [24, 25]. As for environmental sustainability, we chose two metrics used in green software engineering research [85, 38]. Finally, as for economic sustainability, we employed the most widely used metrics [24, 25], complementing them with training time and storage weight, which are essential features when having to release a model in physical environments [17].

6.3. Conclusion Validity

Concerning threats regarding the relationship between treatment and study outcomes, we defined a number of working hypotheses that could have been tested by statistical tests. We applied the Shapiro-Wilk test for each subject {model, independent variable} on each dependent variable, and most of these tests produced values below the significance level of $\alpha = 0.05$ [72], which increased our confidence in the validity of the reported results. Additionally, we performed a visual assessment of normality using quantile-quantile (Q-Q) graphs. To deal with potential threats due to the experiment execution environment, for each study subject we performed 30 experiments for each combination of subject and treatment, for a total of 840 experiments to answer our RQs. Additionally, we used the Friedman test to verify the assumptions necessary for the Nemenyi post-hoc test [63, 46]. If the Friedman test yielded a significance level below $\alpha = 0.05$, we then applied the Nemenyi post-hoc test to assess the statistical significance between

independent variables. In this respect, it is important to remark that we accounted for the risk of introducing Type I errors (false positives) arising from multiple comparisons. We specifically applied the Holm-Bonferroni correction [47], which is a statistical approach that ensures that the family-wise error rate (FWER) is controlled, reducing the likelihood of spurious results. We have released a publicly available replication package [30] that can be exploited by researchers to reproduce our experiments and build on our results.

6.4. External Validity

When considering generalizability and transferability of our findings, we exploited five datasets frequently employed in related research [59, 7, 49, 24, 19]. While these datasets are well-established, they may have inherent limitations, such as potential biases or unrepresentative samples, which could skew our results [34, 24]. Our primary focus on ‘Sex’, ‘Race’, and ‘Age’ as protected attributes was dictated by their prevalence in fairness studies. However, this choice might limit the applicability of our findings to other protected attributes or the number of protected attributes. Our study featured a range of bias algorithms, yet this selection was not exhaustive. The inclusion of a more diverse array of algorithms might provide a broader perspective and potentially alter the findings - as such, our future research agenda includes a larger analysis of bias mitigation algorithms.

Regarding machine learning models, we opted for well-established models in fairness literature [85, 24]. As for machine learners, we kept default hyperparameters for traditional ML algorithms like SVM, LR, RF, and XGB to maintain consistency with previous works [24, 26]. For the bias mitigation algorithms proposed by AIF360 we also used the same parameters for all datasets, while for the MetaFairClassifier and GerryFairClassifier algorithms we had problems with the default parameters and entered the same values for all datasets. More details are within the replication package [30]

7. Conclusion

In this paper, we assessed the impact of six bias mitigation algorithms on social, environmental, and economic sustainability. Our findings reveal the potential trade-off between the fairness of different protected attributes and other non-functional requirements and the need to consider these attributes when developing ML pipelines in real-world environments. The conclusions of our work represent the input of our future research agenda, which first aims at further analyzing the impact of bias mitigation algorithms on sustainability. In addition, we aim at assessing additional trade-offs that practitioners should pay attention to when developing ML-enabled systems. Perhaps more importantly, our future work will aim at extending this article by including a qualitative investigation into the likely causes behind the quantitative results identified in this work.

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

This paper includes data as supplementary material. Datasets generated and analyzed in the context of this study, raw data, and detailed plots, as well as additional resources useful for reproducing our research, are available in the online appendix of this paper: [30].

Credits

Vincenzo De Martino: Conceptualization, Formal analysis, Investigation, Data Curation, Validation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Gianmario Voria:** Conceptualization, Formal analysis, Investigation, Data Curation, Writing - Review & Editing. **Ciro Troiano:** Formal analysis, Investigation, Data Curation, Validation. **Gemma Catolino:** Conceptualization, Supervision, Writing - Review & Editing. **Fabio Palomba:** Conceptualization, Supervision, Writing - Review & Editing.

Acknowledgement

This work has been partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, Projects PRIN 2022 PNRR "FRINGE: context-aware Fairness engineering in complex software systems" (grant n. P2022553SL, CUP: D53D23017340001). The opinions presented in this article solely belong to the author(s) and do not necessarily reflect those of the European Union or The European Research Executive Agency. The European Union and the granting authority cannot be held accountable for these views. We thank Gilberto Recupito and Giammaria Giordano for having reviewed our work and for being supportive and useful in enhancing it.

References

- [1] GitHub - propublica/compas-analysis: Data and analysis for 'Machine Bias' — github.com. <https://github.com/propublica/compas-analysis>. [Accessed 10-02-2024]
- [2] Significant EEOC Race/Color Cases(Covering Private and Federal Sectors) — eeoc.gov. <https://www.eeoc.gov/initiatives/e-race/significant-eeoc-racecolor-casescovering-private-and-federal-sectors#intersectional>
- [3] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International conference on machine learning, pp. 60–69. PMLR (2018)
- [4] Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996). DOI: <https://doi.org/10.24432/C5XW20>
- [5] Becker, C., Chitchyan, R., Duboc, L., Easterbrook, S., Penzenstadler, B., Seyff, N., Venters, C.C.: Sustainability design and software: The karlskrona manifesto. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, vol. 2, pp. 467–476. IEEE (2015)
- [6] Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., 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)
- [7] Biswas, S., Rajan, H.: Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2020, p. 642–653. Association for Computing Machinery, New York, NY, USA (2020). DOI 10.1145/3368089.3409704. URL <https://doi.org/10.1145/3368089.3409704>
- [8] Biswas, S., Rajan, H.: Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021, p. 981–993. Association for Computing Machinery, New York, NY, USA (2021). DOI 10.1145/3468264.3468536. URL <https://doi.org/10.1145/3468264.3468536>
- [9] Bruegge, B., Dutoit, A.H.: Object-oriented software engineering. using uml, patterns, and java. *Learning* **5**(6), 7 (2009)
- [10] Brundtland, G.H.: World commission on environment and development. *Environmental policy and law* **14**(1), 26–30 (1985)
- [11] Brynjolfsson, E., Mitchell, T.: What can machine learning do? workforce implications. *Science* **358**(6370), 1530–1534 (2017)
- [12] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: S.A. Friedler, C. Wilson (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, *Proceedings of Machine Learning Research*, vol. 81, pp. 77–91. PMLR (2018). URL <https://proceedings.mlr.press/v81/buolamwini18a.html>
- [13] Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* **21**, 277–292 (2010)
- [14] Caldiera, V.R.B.G., Rombach, H.D.: The goal question metric approach. *Encyclopedia of software engineering* pp. 528–532 (1994)
- [15] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. *Science* **356**(6334), 183–186 (2017)
- [16] Candelieri, A., Ponti, A., Archetti, F.: Fair and green hyperparameter optimization via multi-objective and multiple information source bayesian optimization. *Machine Learning* **113**(5), 2701–2731 (2024)
- [17] Castanyer, R.C., Martínez-Fernández, S., Franch, X.: Which design decisions in ai-enabled mobile applications contribute to greener ai? *Empirical Software Engineering* **29**(1), 1–34 (2024)
- [18] Celis, L.E., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: A meta-algorithm with provable guarantees. In: Proceedings of the conference on fairness, accountability, and transparency, pp. 319–328 (2019)
- [19] Chakraborty, J., Majumder, S., Menzies, T.: 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, p. 429–440. Association for Computing Machinery, New York, NY, USA (2021). DOI 10.1145/3468264.3468537. URL <https://doi.org/10.1145/3468264.3468537>
- [20] Chakraborty, J., Majumder, S., Yu, Z., Menzies, T.: Fairway: A way to build fair ml software. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2020, p. 654–665. Association for Computing Machinery, New York, NY, USA (2020). DOI 10.1145/3368089.3409697. URL <https://doi.org/10.1145/3368089.3409697>
- [21] Chen, C.f., Napolitano, R., Hu, Y., Kar, B., Yao, B.: Addressing machine learning bias to foster energy justice. *Energy Research & Social Science* **116**, 103653 (2024)
- [22] Chen, S., Liu, C., Haque, M., Song, Z., Yang, W.: Nmtslth: Understanding and testing efficiency degradation of neural machine translation systems. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022, p. 1148–1160. Association for Computing Machinery, New York, NY, USA (2022). DOI 10.1145/3540250.3549102. URL <https://doi.org/10.1145/3540250.3549102>

- [23] Chen, Z., Li, X., Zhang, J.M., Sarro, F., Liu, Y.: Diversity drives fairness: Ensemble of higher order mutants for intersectional fairness of machine learning software. arXiv preprint arXiv:2412.08167 (2024)
- [24] Chen, Z., Zhang, J., Sarro, F., Harman, M.: Fairness improvement with multiple protected attributes: How far are we? In: 46th International Conference on Software Engineering (ICSE 2024). ACM (2023)
- [25] Chen, Z., Zhang, J.M., Sarro, F., Harman, M.: Maat: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 1122–1134 (2022)
- [26] Chen, Z., Zhang, J.M., Sarro, F., Harman, M.: A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology* **32**(4), 1–30 (2023)
- [27] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* **5**(2), 153–163 (2017)
- [28] Chu, C.H., Donato-Woodger, S., Khan, S.S., Nyrup, R., Leslie, K., Lyn, A., Shi, T., Bianchi, A., Rahimi, S.A., Grenier, A.: Age-related bias and artificial intelligence: a scoping review. *Humanities and Social Sciences Communications* **10**(1), 1–17 (2023)
- [29] Cruz, L.: Green software engineering done right: a scientific guide to set up energy efficiency experiments. <http://luiscruz.github.io/2021/10/10/scientific-guide.html> (2021). DOI 10.6084/m9.figshare.22067846.v1. Blog post.
- [30] De Martino, V., Voria, G., Troiano, C., Catolino, G., Palomba, F.: Examining the impact of bias mitigation algorithms on the sustainability of ml-enabled systems (2024). DOI 10.6084/m9.figshare.25673658. URL https://figshare.com/articles/dataset/Examining_the_Impact_of_Bias_Mitigation_Algorithms_on_the_Sustainability_of_ML-enabled_Systems/25673658
- [31] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research* **7**, 1–30 (2006)
- [32] Deng, W.H., Nagireddy, M., Lee, M.S.A., Singh, J., Wu, Z.S., Holstein, K., Zhu, H.: Exploring how machine learning practitioners (try to) use fairness toolkits. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, p. 473–484. Association for Computing Machinery, New York, NY, USA (2022). DOI 10.1145/3531146.3533113. URL <https://doi.org/10.1145/3531146.3533113>
- [33] Di Vaio, A., Palladino, R., Hassan, R., Escobar, O.: Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. *Journal of Business Research* **121**, 283–314 (2020)
- [34] Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems* **34**, 6478–6490 (2021)
- [35] Fabris, A., Messina, S., Silvello, G., Susto, G.A.: Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* **36**(6), 2074–2152 (2022)
- [36] Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., De Lucia, A.: Fairness-aware machine learning engineering: how far are we? *Empirical Software Engineering* **29**(1), 9 (2024)
- [37] Franklin, J.S., Bhanot, K., Ghalwash, M., Bennett, K.P., McCusker, J., McGuinness, D.L.: An ontology for fairness metrics. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 265–275 (2022)
- [38] Georgiou, S., Kechagia, M., Sharma, T., Sarro, F., Zou, Y.: Green ai: Do deep learning frameworks have different costs? In: Proceedings of the 44th International Conference on Software Engineering, ICSE '22, p. 1082–1094. Association for Computing Machinery, New York, NY, USA (2022). DOI 10.1145/3510003.3510221. URL <https://doi.org/10.1145/3510003.3510221>
- [39] Ghosh, A., Genuit, L., Reagan, M.: Characterizing intersectional group fairness with worst-case comparisons. In: *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pp. 22–34. PMLR (2021)
- [40] Gohar, U., Cheng, L.: A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-2023, p. 6619–6627. International Joint Conferences on Artificial Intelligence Organization (2023). DOI 10.24963/ijcai.2023/742. URL <http://dx.doi.org/10.24963/ijcai.2023/742>
- [41] Goralski, M.A., Tan, T.K.: Artificial intelligence and sustainable development. *The International Journal of Management Education* **18**(1), 100330 (2020)
- [42] Habibullah, K.M., Horkoff, J.: Non-functional requirements for machine learning: understanding current use and challenges in industry. In: 2021 IEEE 29th International Requirements Engineering Conference (RE), pp. 13–23. IEEE (2021)
- [43] for Healthcare Research, A., Quality: Medical Expenditure Panel Survey Public Use File Details — meps.ahrq.gov. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181. [Accessed 08-01-2024]
- [44] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J.: Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.* **21**(1) (2020)
- [45] Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994). DOI: <https://doi.org/10.24432/C5NC77>
- [46] Hollander, M., Wolfe, D.A., Chicken, E.: Nonparametric statistical methods. John Wiley & Sons (2013)
- [47] Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* pp. 65–70 (1979)
- [48] Hort, M., Chen, Z., Zhang, J.M., Harman, M., Sarro, F.: Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.* **1**(2) (2024). DOI 10.1145/3631326. URL <https://doi.org/10.1145/3631326>
- [49] Hort, M., Zhang, J.M., Sarro, F., Harman, M.: Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 994–1006 (2021)
- [50] Hort, M., Zhang, J.M., Sarro, F., Harman, M.: Search-based automatic repair for fairness and accuracy in decision-making software. *Empirical Software Engineering* **29**(1), 36 (2024)
- [51] Järvenpää, H., Lago, P., Bogner, J., Lewis, G., Muccini, H., Ozkaya, I.: A synthesis of green architectural tactics for ml-enabled systems (2023)
- [52] Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* **33**(1), 1–33 (2012)
- [53] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II* **23**, pp. 35–50. Springer (2012)

- [54] Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International conference on machine learning, pp. 2564–2572. PMLR (2018)
- [55] v. Kistowski, J., Arnold, J.A., Huppler, K., Lange, K.D., Henning, J.L., Cao, P.: How to build a benchmark. In: Proceedings of the 6th ACM/SPEC international conference on performance engineering, pp. 333–336 (2015)
- [56] Kitchenham, B.A., Dyba, T., Jorgensen, M.: Evidence-based software engineering. In: Proceedings. 26th International Conference on Software Engineering, pp. 273–281. IEEE (2004)
- [57] Lee, M.S.A., Singh, J.: The landscape and gaps in open source fairness toolkits. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21. Association for Computing Machinery, New York, NY, USA (2021). DOI 10.1145/3411764.3445261. URL <https://doi.org/10.1145/3411764.3445261>
- [58] Lewis, G.A., Ozkaya, I., Xu, X.: Software architecture challenges for ml systems. In: 2021 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 634–638. IEEE (2021)
- [59] Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, A.M., Wagner, S.: Software engineering for ai-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)* **31**(2), 1–59 (2022)
- [60] Martino, V.D., Palomba, F.: Classification, challenges, and automated approaches to handle non-functional requirements in ml-enabled systems: A systematic literature review (2023)
- [61] McGuire, S., Schultz, E., Ayoola, B., Ralph, P.: Sustainability is stratified: Toward a better theory of sustainable software engineering. In: 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), pp. 1996–2008. IEEE (2023)
- [62] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* **54**(6), 1–35 (2021)
- [63] Nemenyi, P.B.: Distribution-free multiple comparisons. Princeton University (1963)
- [64] Pagano, T.P., Loureiro, R.B., Lisboa, F.V., Peixoto, R.M., Guimarães, G.A., Cruz, G.O., Araujo, M.M., Santos, L.L., Cruz, M.A., Oliveira, E.L., et al.: Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing* **7**(1), 15 (2023)
- [65] Peng, K., Chakraborty, J., Menzies, T.: Fairmask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering* **49**(4), 2426–2439 (2022)
- [66] Peng, K., Chakraborty, J., Menzies, T.: Fairmask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering* **49**(4), 2426–2439 (2023). DOI 10.1109/TSE.2022.3220713
- [67] del Rey, S., Martínez-Fernández, S., Cruz, L., Franch, X.: Do dl models and training environments have an impact on energy consumption? *arXiv preprint arXiv:2307.05520* (2023)
- [68] Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. *SN computer science* **2**(3), 160 (2021)
- [69] Sarro, F.: Search-based software engineering in the era of modern software systems. In: 2023 IEEE 31st International Requirements Engineering Conference (RE), pp. 3–5 (2023). DOI 10.1109/RE57278.2023.00010
- [70] Schmuckler, M.A.: What is ecological validity? a dimensional analysis. *Infancy* **2**(4), 419–436 (2001)
- [71] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green ai. *Communications of the ACM* **63**(12), 54–63 (2020)
- [72] Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). *Biometrika* **52**(3/4), 591–611 (1965)
- [73] Sim, S., Easterbrook, S., Holt, R.: Using benchmarking to advance research: a challenge to software engineering. In: 25th International Conference on Software Engineering, 2003. Proceedings., pp. 74–83 (2003). DOI 10.1109/ICSE.2003.1201189
- [74] Sim, S.E., Easterbrook, S., Holt, R.C.: Using benchmarking to advance research: A challenge to software engineering. In: 25th International Conference on Software Engineering, 2003. Proceedings., pp. 74–83. IEEE (2003)
- [75] Soremekun, E., Papadakis, M., Cordy, M., Traon, Y.L.: Software fairness: An analysis and survey (2022). URL <https://arxiv.org/abs/2205.08809>
- [76] Sprent, P., Smeeton, N.C.: Applied nonparametric statistical methods. CRC press (2007)
- [77] Tichy, W.F.: Where’s the science in software engineering? ubiquity symposium: The science in computer science. *Ubiquity* **2014**(March), 1–6 (2014)
- [78] Tizpaz-Niari, S., Kumar, A., Tan, G., Trivedi, A.: Fairness-aware configuration of machine learning libraries. In: Proceedings of the 44th International Conference on Software Engineering, ICSE ’22, p. 909–920. Association for Computing Machinery, New York, NY, USA (2022). DOI 10.1145/3510003.3510202. URL <https://doi.org/10.1145/3510003.3510202>
- [79] Udeshi, S., Arora, P., Chattopadhyay, S.: Automated directed fairness testing. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE ’18, p. 98–108. Association for Computing Machinery, New York, NY, USA (2018). DOI 10.1145/3238147.3238165. URL <https://doi.org/10.1145/3238147.3238165>
- [80] Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., Hotellier, E.: Data-centric green ai an exploratory empirical study. In: 2022 international conference on ICT for sustainability (ICT4S), pp. 35–45. IEEE (2022)
- [81] Verdecchia, R., Sallou, J., Cruz, L.: A systematic review of green ai. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* p. e1507 (2023)
- [82] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M., Fuso Nerini, F.: The role of artificial intelligence in achieving the sustainable development goals. *Nature communications* **11**(1), 1–10 (2020)
- [83] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering. Springer Science & Business Media (2012)
- [84] Wu, C.J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., et al.: Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* **4**, 795–813 (2022)
- [85] Xu, Y., Martínez-Fernández, S., Martinez, M., Franch, X.: Energy efficiency of training neural network architectures: An empirical study. *arXiv preprint arXiv:2302.00967* (2023)
- [86] Zhang, J.M., Harman, M.: “ignorance and prejudice” in software fairness. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), pp. 1436–1447 (2021). DOI 10.1109/ICSE43902.2021.00129
- [87] Zhang, M., Sun, J.: Adaptive fairness improvement based on causality analysis. In: Proceedings of the 30th ACM Joint European Software

Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022, p. 6–17. Association for Computing Machinery, New York, NY, USA (2022). DOI 10.1145/3540250.3549103. URL <https://doi.org/10.1145/3540250.3549103>