Fairness on a Budget, Across the Board: A Cost-Effective Evaluation of Fairness-Aware Practices Across Contexts, Tasks, and Sensitive Attributes

Alessandra Parziale[1], Gianmario Voria[1], Giammaria Giordano[1], Gemma Catolino[1], Gregorio Robles[2], Fabio Palomba[1]

[1] *Software Engineering (SeSa) Lab - Department of Computer Science, University of Salerno, Italy*

[2] *Universidad Rey Juan Carlos, Madrid, Spain*

Abstract

Context. Machine Learning (ML) is widely used in critical domains like finance, healthcare, and criminal justice, where unfair predictions can lead to harmful outcomes. Although bias mitigation techniques have been developed by the Software Engineering (SE) community, their practical adoption is limited due to complexity and integration issues. As a simpler alternative, fairness-aware practices, namely conventional ML engineering techniques adapted to promote fairness, e.g., MinMax Scaling, which normalizes feature values to prevent attributes linked to sensitive groups from disproportionately influencing predictions, have recently been proposed, yet their actual impact is still unexplored. **Objective.** Building on our prior work that explored fairness-aware practices in different contexts, this paper extends the investigation through a large-scale empirical study assessing their effectiveness across diverse ML tasks, sensitive attributes, and datasets belonging to specific application domains. **Methods.** We conduct 5,940 experiments, evaluating fairness-aware practices from two perspectives: *contextual bias mitigation* and *cost-effectiveness*. Contextual evaluation examines fairness improvements across different ML models, sensitive attributes, and datasets. Cost-effectiveness analysis considers the trade-off between fairness gains and performance costs. **Results.** Findings reveal that the effectiveness of fairness-aware practices depends on specific contexts' datasets and configurations, while cost-effectiveness analysis highlights those that best balance ethical gains and efficiency. **Conclusion.** These insights guide practitioners in choosing fairness-enhancing practices with minimal performance impact, supporting ethical ML development.

Keywords: Software Engineering for Artificial Intelligence; Machine Learning Fairness Engineering; Cost-Effectiveness; Empirical Software Engineering.

Email addresses: `alparziale@unisa.it` (Alessandra Parziale[1]), `gvoria@unisa.it` (Gianmario Voria[1]), `giagiordano@unisa.it` (Giammaria Giordano[1]), `gcatolino@unisa.it` (Gemma Catolino[1]), `grex@gsyc.urjc.es` (Gregorio Robles[2]), `fpalomba@unisa.it` (Fabio Palomba[1])

1. Introduction

Artificial Intelligence (AI), with Machine Learning (ML) at its core, is rapidly integrating into daily life, automating decision-making processes [49, 73, 85]. However, its widespread adoption has raised ethical concerns regarding *fairness*, defined as an ML model's ability to make unbiased decisions without discriminating against specific groups [44]. Often, bias often arises from ML algorithms' reliance on historical data, leading to skewed representations [50, 54]. Typically, bias is linked to *sensitive attributes* such as gender, race, or age [1, 29]; indeed improper handling of these attributes can reinforce discrimination [21], as seen in documented ethical incidents like Facebook's discriminatory labeling of Black men and Amazon's biased ranking of LGBTQIA+ books [9, 47, 65, 75]. These cases highlight the urgent need for fair ML software.

To address these concerns, the Software Engineering (SE) and AI research communities have developed *bias mitigation techniques*, which operate at different ML development stages to reduce bias. These techniques fall into three categories: *pre-processing* (modifying data before training), *in-processing* (adjusting learning algorithms during training), and *post-processing* (modifying outputs after training) [32, 44]. Typically implemented in fairness toolkits [38], these solutions have demonstrated effectiveness in empirical experiments [17, 33, 82]. However, fairness is highly context dependent, i.e., the effectiveness of mitigation strategies often varies based on the specific dataset, task, model, and sensitive attribute involved [25]. This variability may notably impact practitioners, as it complicates the selection of interventions and limits the generalizability of findings. In addition, bias mitigation algorithms may affect the implementation costs [22], other than degrading model performance and reducing user trust [43, 44]. Because of the reasons above, fairness toolkits and bias mitigation algorithms remain underutilized [23, 38], with developers either applying fairness measures inconsistently or avoiding them altogether.

To overcome these challenges, recent SE research [70] has proposed fairness-aware practices: conventional ML engineering practices that are adapted to promote fairness without requiring specialized toolkits. Examples include data balancing, which addresses class imbalances, and mutation testing, which reveals fairness violations by evaluating prediction consistency under slight input variations. As these practices build on techniques familiar to practitioners, they might lower the barriers and make fairness enhancement more accessible. These practices are organized across the six ML development stages defined by Burkov [10]: they range from early stages like *"Requirements Elicitation"* and *"Data Preparation"* (e.g., *Multi-objective Optimization*, *Data Balancing*) to later stages like *"Model Maintenance & Evolution"* (e.g., *Model Outcomes Analysis*). While these practices have been deemed promising by practitioners [71], who acknowledged their fairness benefits and low implementation effort. On the one hand, *their effectiveness across diverse contexts, ML tasks, and sensitive attributes has not been investigated*. On the other hand, *their ability to improve fairness without compromising model performance remains unclear*. Addressing these questions is crucial to assess the viability of fairness-aware practices and to provide actionable guidance.

> ◎ **Research Objective.** *Our objective is to empirically evaluate the extent to which fairness-aware practices can increase ML fairness while not deteriorating performance for datasets belonging to specific contexts, on different ML tasks, and considering various sensitive attributes.*

In a preliminary investigation on the matter [52], we evaluated fairness-aware practices from two perspectives. First, we assessed their contextual impact, demonstrating that the effectiveness of individual practices varies depending on the datasets and application domains in which they are applied. Second, we conducted a cost-effectiveness analysis, providing trade-offs between fairness improvements and performance degradation. However, the scope of that study was limited to a single ML task, and only one sensitive attribute. In this paper, we extend our previous work by providing a more comprehensive investigation involving multiple ML tasks, models, and sensitive attributes. Moreover, rather than assessing fairness-aware practices as aggregated groups, we analyze the impact of each practice individually across diverse datasets belonging to critical contexts. To support this evaluation, we select widely used datasets from prior fairness research [17, 22, 25], each representative of different real-world application domains, i.e., Recidivism Prediction [25], Economics [4], Marketing [45], Finance [31], and Crime [59]. For each of these, we consider the ML tasks most associated with it in the literature, such as classification with Random Forests or clustering with K-means. We then select a set of fairness-aware practices informed by practitioner insights regarding their fairness impact and adoption frequency [71]. Finally, we conduct an extensive empirical evaluation involving 5,940 training runs across combinations of datasets, practices, ML tasks, and sensitive attributes, measuring both fairness and performance outcomes. Particularly, we adopt a group fairness [44] perspective, evaluating disparities between groups rather than focusing on individual-level fairness [44].

Our results indicate that Mutation Testing improves fairness across classification tasks, particularly for the datasets of the Recidivism, Finance, and Crime domains. MinMax Scaling is the most effective for clustering, especially in the Economics domain. Furthermore, Select Best and MinMax Scaling generally provide a balance between fairness and performance. Regularization and Mutation Testing shows promising results in balancing fairness improvements and predictive accuracy, while Simple and Iterative Imputers contribute to fairness in specific cases.

To summarize, our research provides the following major contributions:

1. A comprehensive empirical study with 5940 experiments of fairness-aware practices with different combinations of ML tasks, sensitive attributes, and datasets;
2. A dual-perspective analysis focusing on improvements of fairness and potential loss in performance;
3. An online appendix providing all data and scripts to replicate and verify our study [51];
4. Practical, evidence-based suggestions for practitioners aiming to enhance fairness in real-world ML systems through a tool that makes our findings actionable, available in our online appendix [51].

2. Background and Related Work

This chapter presents the fundamental concepts that guide our study. First, we formalize key notions such as individual and group fairness, clarify the role of sensitive attributes, and discuss how biases are present in the machine learning (ML) pipeline. Then, we review the state of the art in ML fairness, analyzing the frameworks, bias-mitigation algorithms, and the evaluations that motivate our experimental design. This provides the necessary context to understand the methodology and contributions presented in this study.

2.1. Terminology and Background

ML fairness seeks to ensure that predictions are unbiased with respect to individuals or groups [74].

Individual Fairness is the principle that any two similar individuals —according to specific characteristics— should receive similar outcomes with respect to a given task [24].

Group Fairness refers to the principle that distinct groups—e.g., groups defined by demographics or opportunities—should receive equal treatment, regardless of their characteristics [44]. In this study, we focus on **group fairness** because it is widely adopted in empirical research evaluating fairness-aware methods [50, 54], supported by widely-adopted fairness toolkits [44], and well-aligned with the types of metrics and datasets selected in our work [25].

Additionally, fairness definitions vary based on which and how sensitive attributes are treated. **Sensitive (or protected) attributes** are personal characteristics of groups or individuals that may lead to discriminatory treatment or influence decision outcomes for specific tasks [44]. Typical examples include particular genders, ethnicities, ages, religions, disabilities, or sexual orientations [44, 74]. For example, **Fairness through unawareness** excludes them from decisions [16, 74, 83], while **Fairness through awareness** explicitly incorporates them to ensure equitable outcomes [74, 81]. Fairness is now a critical concern in SE and AI, seen as a non-functional requirement for AI-integrated systems [17, 27, 32, 54, 62]. Bias, i.e., systematic distortion in data or models, can lead to unfair outcomes [44]. Persistent issues, like gender bias in hiring [48] or racial bias in facial recognition [64], highlight the need for fairness-aware practices. Unfairness can arise throughout the ML pipeline, from biased data collection to feature selection that embeds correlations with sensitive attributes [54, 62].

In previous research bias mitigation techniques were classified into pre-processing, in-processing, and post-processing approaches.

Pre-processing methods reduce bias by adjusting training data before model learning. Examples include Fair-SMOTE, which generates synthetic samples [13], and reweighting techniques that modify instance weights [35]. These approaches help address group underrepresentation [64, 77] by improving population representation in training data. **In-processing** techniques modify algorithms during training to mitigate bias. For instance, Zhang et al. [80] used adversarial learning, while Chakraborty et al. [14] applied multi-objective optimization. These methods help prevent reinforcing inequalities [48]. **Post-processing** methods adjust model outputs to improve fairness without retraining. Tools like Themis [28] and Aequitas [67] are useful when retraining is costly or impractical.

2.2. Related Works

Recent research has advanced quantitative evaluations for fairness improvement methods. Hort et al. [33] introduced Fairea, a tool to benchmark bias mitigation methods. Chen et al. [18] used Fairea in a large-scale study with seven algorithms, finding that mitigation methods can reduce accuracy, with effectiveness varying. Zhang and Sun [82] adapted ML fairness methods for multiple sensitive attributes. Chen et al. [17] benchmarked fairness improvements across eight techniques, while Hort et al. [32] proposed a new approach to enhance both fairness and accuracy. De Martino et al. [22] benchmarked bias mitigation algorithms and explored the trade-offs among social sustainability, i.e., fairness, economic sustainability, and environmental sustainability. Finally, Fabris et al. [25] performed an analysis of the algorithmic-fairness literature, screening papers and datasets, such as Adult, COMPAS, and German Credit. Their study introduces fairness tasks, sensitive attributes, and best-practice recommendations. On this basis, the authors propose practical guidelines for selecting datasets according to the domain and the fairness notion under study. Le Quy et al. [37] extend this perspective with an empirical analysis of the more commonly used tabular datasets. Mapping the dependencies between protected attributes, quantifying the trade-off between predictive utility and fairness, and exposing specific biases of the datasets. Their findings underscore that robust fairness evaluation must consider multiple application domains and sensitive attributes. With these results [37] and the guidelines of Fabris et al. [25], we designed the dataset-selection strategy adopted in this study.

Despite extensive research, fairness toolkits and bias mitigation techniques remain underused in practice [23, 38]. This gap stems from context-dependent effectiveness, potential performance trade-offs, implementation costs, and integration challenges. To address this, Voria et al.[70] compiled a catalog of fairness-aware practices—standard ML engineering techniques adapted to address bias—mapped to the six stages of the ML life-cycle[10], including *Data Balancing*, *Parameter Regularization*, and *Causal Validation*. These are familiar to practitioners and commonly used in everyday workflows. Voria et al. [71] also surveyed practitioners on each practice's perceived effectiveness, usage frequency, and implementation effort.

However, their evaluation remains primarily qualitative, lacking empirical validation of fairness impact across datasets of diverse application domains. Specifically, it does not assess effectiveness across application domains, ML tasks, or sensitive attributes, nor examine trade-offs between fairness and performance. Building on our earlier work [52], which provided preliminary insights into contextual effectiveness and cost-performance trade-offs, this paper offers a broader empirical evaluation across multiple tasks, models, and sensitive attributes in real-world application contexts. In this way, we explicitly integrate the methodological observations of the previous studies [10, 70, 71] using them as a foundation for our final study. Indeed, acknowledging the dataset- and task-dependence of ML fairness [26], our work goes beyond fixed dataset–model evaluations [17, 22]. The *scientific novelty* of this study lies in its comprehensive, fine-grained empirical assessment of individual fairness-aware practices, providing evidence-based insights into their fairness impact and cost-effectiveness across varied settings.

Our Contribution.

We extend prior research [52] by evaluating fairness-aware practices across tasks, contexts, and sensitive attributes. We assess their effectiveness in mitigating bias, conduct a cost-effective analysis to examine the performance-fairness trade-offs, and offer insights to select suitable fairness strategies based on contextual information.

3. Research Design

The *goal* of this empirical study is to evaluate the effectiveness of fairness-aware practices in mitigating bias across different datasets and with different tasks and sensitive attributes, following and expanding the design of preliminary research [52]. Its *purpose* is to assess their impact and associated performance trade-offs across different application scenarios. The study addresses the *perspective* of both researchers — interested in performance implications under specific settings — and practitioners — seeking guidance on integrating fairness practices into ML workflows. To this end, we define two research questions.

First, we aimed to quantitatively assess the impact of fairness-aware practices on mitigating bias on specific ML tasks. Building on prior qualitative work based on expert opinions [71], as well as empirical studies evaluating different specific techniques [17, 22], we sought to offer comprehensive with a systematic assessment to determine whether these practices improve fairness across different tasks, sensitive attributes, and application domain. This evaluation was performed in the context of our first research question:

> **RQ_1 - Fairness Evaluation**
>
> *To what extent can fairness-aware practices mitigate bias when applied to different tasks, contexts, and sensitive attributes?*

Our second objective was to investigate the performance trade-offs associated with fairness-aware practices, as it is a fundamental challenge in fairness research [22]. The results of the first **RQ** guided our investigation, revealing which fairness-aware practices effectively mitigate bias. However, improving fairness often comes at the cost of reduced model performance [17, 22], raising a critical challenge for both researchers and practitioners. Understanding the trade-off between fairness gains and performance loss is essential for making informed decisions about adopting fairness-aware practices in real-world applications. Without this knowledge, practitioners risk applying techniques that enhance fairness but render models impractical for deployment. Therefore, we needed to examine the extent to which fairness improvements come at the cost of performance, allowing us to assess the feasibility of these practices on datasets across different contexts, leading to the definition of our second research question:

> **RQ_2 - Cost-Effectiveness Evaluation**
>
> *What is the cost in terms of performance loss against fairness improvements given by the application of the practices?*

Figure 1 provides an overview of our research approach, illustrating the method used to address these research questions. The process begins with the selection of datasets and related ML tasks [25], and then the fairness-aware practices [71]. Afterward, we train models related to the tasks without any practice. Once trained, these models are evaluated based on fairness by using the sensitive attributes available in the datasets and performance metrics to get a comparison baseline. Finally, we repeat the same process for each fairness-aware practice selected, applying it before training the models. Our study follows the empirical research standards, adhering to the guidelines of Wohlin et al. [76] and the *ACM/SIGSOFT Empirical Standards* [56],[1] specifically aligning with the *"General Standard"* due to the nature of our investigation.

3.1. Objects of the Study

The fairness-aware practices evaluated in this study [70] were selected based on a recent expert survey [71], which assessed their fairness impact, usage frequency, and implementation effort. We selected practices that presented a balanced mix of positive fairness impact, good adoption in practice, and feasible integration into an automated evaluation pipeline. Many excluded practices, although potentially valuable, were not suited for scalable experimentation due to their reliance on substantial human intervention or lack of mature tool support. For example, practices in the 'Requirements Engineering' or 'Software Testing' categories require domain-specific manual setup or infrastructures that are either not publicly available or not generalizable across multiple learning tasks and datasets. The practices selected were well-suited for a detailed quantitative evaluation across diverse datasets, tasks, and sensitive attributes. Below, we outline each category, the selected practices, rationale, and implementation choices; Table 1 summarizes this information.

[1] Available at: `https://github.com/acmsigsoft/EmpiricalStandards`.

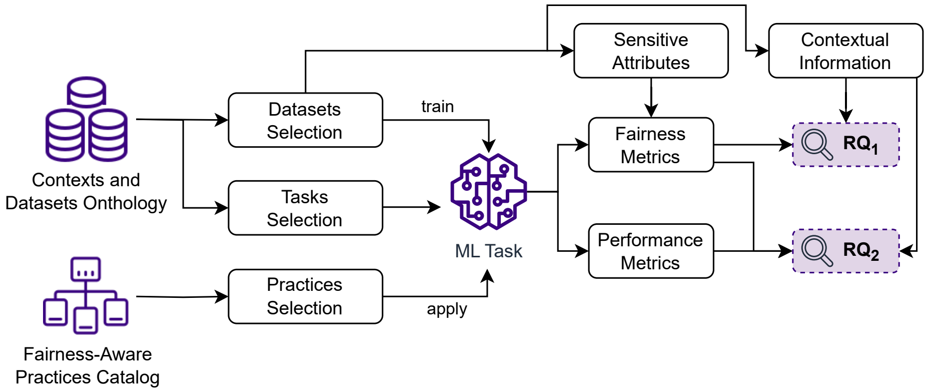

Figure 1: Overview of the Research Method Proposed for Our Study.

- *Data Balancing* mitigates bias in unbalanced datasets [68]. It is considered effective, with medium-high fairness impact and low implementation effort [71]. *Oversampling* increases the minority class frequency; we apply *Simple Oversampling*, which duplicates underrepresented samples but may risk overfitting. *Undersampling* reduces the dominant class size [25, 68]; we use *Simple Undersampling* to achieve class balance by removing majority class instances.

- *Data Transformation* aims to homogenize feature distributions [8]. Though it requires medium-to-high effort, its fairness impact is significant [71]. Techniques include: *Iterative Imputer*, which estimates missing values from other features; *Select Best*, which chooses features based on statistical relevance; and *Simple Imputer*, which fills missing values with the mean, median, or mode [8, 25, 46].

- *Feature Standardization* ensures all features contribute equally to the model [39]. It offers medium-to-high fairness impact with low implementation effort [71]. We use *MinMax Scaling*, which normalizes values to ensure uniform feature contributions [61].

- *Parameter Regularization* promotes fairness across subpopulations [57, 69]. Despite its high implementation effort, it has strong fairness potential [71]. This practice introduces constraints, such as penalties, to reduce prediction disparities and mitigate bias [25].

- *Metamorphic/Mutation Testing* assess prediction consistency under data variations [8]. Chosen for their fairness impact and low implementation effort [71], these techniques modify data to test model robustness [25]—e.g., adding random noise to verify if a classifier preserves labels.

Table 1: Fairness-Aware Practices Selected for Our Study.

Practice Category	Practice Implementation	Description
Data Balancing	Oversampling Undersampling	Increases the frequency of the minority class Reduces samples from the dominant class
Data Transformation	Iterative Imputer Select Best Simple Imputer	Replaces missing values based on estimates from other features Selects the most relevant features Replaces missing values with mean, median, or mode
Feature Standardization	MinMax Scaling	Normalizes values to a specific range
Parameter Regularization	Regularization	Adds penalties to reduce prediction disparity
Metamorphic/Mutation Testing	Input Variation	Modifies input data (e.g., adding noise)

3.2. Subjects of the Study

Datasets Selection. To evaluate fairness-aware practices across domains, we selected widely used datasets in fairness research and in the literature [25, 37]. Beyond popularity, our selection was also guided by the goal of ensuring diversity across key dimensions: application domain (e.g., healthcare, education, economics), learning tasks (e.g., classification, regression), and sensitive attributes. These datasets offer variability in structure, target variables, and fairness concerns, supporting a multifaceted evaluation. Moreover, each dataset reflects a distinct context and includes sensitive attributes explicitly defined in the official documentation [25]. The selected datasets are visible in Table 2.

- *COMPAS dataset* (*Recidivism prediction*): Contains 2013–2014 data used to estimate recidivism risk. This justice-related dataset influences decisions that may perpetuate social and racial inequalities. Sensitive attributes: *Sex*, *Race* [25, 37].

- *Adult dataset* (*Economics*): Based on U.S. census data, it predicts whether an individual earns over $50,000, highlighting economic disparities. Sensitive attributes: *Sex*, *Race* [25, 37].

- *Bank Marketing dataset* (*Marketing*): Includes data from a Portuguese bank's 2008–2013 campaigns to predict deposit subscription, where biased targeting may arise. Sensitive attributes: *Marital status*, *Age* [25, 37].

- *German Credit dataset* (*Finance*): Evaluates credit risk to determine loan eligibility, where fairness is crucial for equitable access to financial services. Sensitive attributes: *Gender status*, *Age* [25, 37].

- *Communities and Crime dataset* (*Crime*): Gathers data from 46 U.S. states to predict violent crime rates, enabling analysis of indirect discrimination at the community level. Sensitive attribute: *Race* [25, 37].

Tasks Selection. The selection of machine learning tasks for each dataset was guided by the task–context ontology introduced by Fabris et al. [25], which systematically maps commonly used datasets to fairness-related tasks. To maintain consistency with this ontology and ensure reproducibility, we selected tasks that (i) had been previously implemented in the referenced studies and (ii) could be instantiated with available public data and standard tooling. Moreover, we prioritized tasks that appeared across multiple datasets, to preserve comparability and avoid dataset-specific bias in the evaluation. Table 2 summarizes these tasks and their association with the datasets.

- *Classification* is an ML task aiming to treat similar individuals similarly [11, 24]. Fairness is typically addressed by equalizing measures across subpopulations [11, 24, 25]. This work considers: *Random Forest*, a tree-based method; *Logistic Regression*, which models class probabilities via the logistic function; *Extreme Gradient Boosting (XGBoost)*, an iterative tree-based algorithm; *Decision Tree*, which splits data by feature values; and *Naïve Bayes*, a probabilistic classifier using Bayes' theorem.

- *Regression* is essential in predictive modeling [5]. Individual fairness provides similar predictions to similar individuals and distributing losses uniformly [5, 25]. We consider: *Decision Tree*, which splits data to minimize error on a continuous target; and *Linear Regression*, which models the relationship between variables using a linear equation.

- *Clustering* partitions data into homogeneous groups based on feature similarity [20]. Fairness is defined by balanced subgroup distribution or average distance to cluster centers [20, 25]. We consider: *K-means*, minimizing intra-cluster variance; *K-center*, reducing maximum point-centroid distance; and *K-median*, minimizing absolute differences.

While datasets such as Adult and COMPAS may appear similar in terms of features and sensitive attributes, the number and type of tasks assigned to each were based on their documented usage in prior fairness studies [18, 22]. For example, the Adult dataset is widely used across a broad range of fairness tasks—particularly clustering and regression—making it a good candidate for multi-task evaluation. In contrast, although COMPAS appears in the ontology with multiple tasks, many of them are either highly specialized (e.g., fairness in transfer learning) or difficult to apply consistently across other datasets. Therefore, tasks were assigned to datasets not only based on technical feasibility (e.g., clustering applicability), but also on relevance and replicability according to Fabris et al.'s mapping [25].

Table 2: Datasets representing each context of our study. For each dataset, we report the sensitive attributes and tasks selected for our evaluation. Task assignment was guided by the ontology from Fabris et al. [25], considering prior use in fairness studies, reproducibility, and technical feasibility across datasets.

Dataset	Sensitive Attributes	Tasks
Compas	Sex, Race	Classification - Random Forest Classification - Logistic Regression Classification - XGBoost
Adult	Sex, Race	Classification - Random Forest Classification - Logistic Regression Classification - XGBoost Clustering - K-mean Clustering - K-center Clustering - K-median
Bank Marketing	Age, Marital	Classification - Random Forest Clustering - K-mean Clustering - K-center Clustering - K-median
German Credit	Age, Gender	Classification - Random Forest Classification - Logistic Regression Classification - XGBoost Classification - Decision Tree
Communities and Crime	Race	Classification - Decision Tree Classification - Naïve Bayesian Classification - Logistic Regression Regression - Linear Regression Regression - Decision Tree

Metrics Selection. For each task, we selected both fairness and performance metrics at the group level, following established literature [2, 42, 63]. In order to evaluate disparities between different demographic groups. As shown in Table 3, we measured performance and fairness for each of the three ML tasks selected, namely classification, clustering, and regression.

- To assess **performance**, we employed task-specific metrics [41, 53, 72]. For *classification* models, we measured *Accuracy*, which quantifies the percentage of correctly classified instances; *Precision*, which indicates the proportion of true positive predictions among all predicted positives; *Recall*, which evaluates ability to identify all positive instances correctly; and *F1-score*, which represents the harmonic mean of Precision and Recall [72]. In *clustering* tasks, we used the *Silhouette Coefficient*, which captures both the cohesion within clusters and their separation from one another [53]. For *regression*, we relied on *Mean Squared Error (MSE)*, which computes the average squared difference between predicted and actual values, and *Median Absolute Deviation (MAD)*, which measures the median of absolute deviations from the predicted values [41].

- To assess **fairness**, we applied different metrics depending on the task [2, 42, 63]. All the selected fairness metrics operate at the **group level**, as our evaluation specifically focuses on measuring disparities between demographic groups. In *classification*, we evaluated *Average Absolute Odds Difference (AAOD)*, which quantifies disparities in true and false positive rates between demographic groups; *False Discovery Rate Difference (FDRD)*, which assesses imbalances in false positive rates, revealing disparities in incorrect classifications; and *Disparate Impact (DI)*, which compares the proportion of positive outcomes between protected and non-protected groups [42]. Unlike the other metrics, DI is centered around one rather than zero. To ensure consistency across fairness measures, we adjusted it by subtracting one, aligning its balance point with the other metrics without altering its fundamental meaning. For *clustering* tasks, fairness was evaluated using *Average Euclidean (AE)* distance and *Maximum Euclidean (ME)* distance, which measure the average and maximum distances between cluster centroids, respectively, as well as *Average Wasserstein (AW)* distance and *Maximum Wasserstein (MW)* distance, which provide analogous measures based on the Wasserstein distance [2]. Finally, in *regression* tasks, we assessed fairness using *Independence*, which verifies whether predictions are uncorrelated with membership in a protected group; *Separation*, which considers both the protected group and the target variable when evaluating fairness; and *Sufficiency*, which ensures that the model's predictions contain all necessary information to estimate the target value [63].

Table 3: Fairness and Performance metrics selected to evaluate each task.

Tasks	Fairness Metrics	Performance Metrics
Classification	Average Abs Odds Difference (AAOD) False Discovery Rate Difference (FDRD) Disparate Impact (DI)	Accuracy Precision Recall F1-score
Clustering	Average Euclidean (AE) distance Maximum Euclidean (ME) distance Average Wasserstein (AW) distance Maximum Wasserstein (MW) distance	Silhouette Coefficient
Regression	Separation Sufficiency Independence	Mean Squared Error (MSE) Median Absolute Deviation (MAD)

3.3. Data Collection and Analysis

For both research questions, we conducted experiments using the selected tasks and datasets. Each model of the selected task was trained independently for its respective dataset without applying any fairness-aware practices. When multiple sensitive attributes were available in a dataset, we conducted separate training runs for each attribute. Additionally, we only applied fairness practices that were compatible with the specific task—for example, techniques that modify the target variable were not used in unsupervised tasks like clustering. Each training sessions was repeated *20 times*. This repeated training was based on methodological guidance for the statistical analysis of results of non-deterministic algorithms in software engineering by Arcuri and Briand [3]. Hence, we adopted a replication strategy that balances reliable estimation, sufficient paired observations for non-parametric significance tests, and a computational budget that is feasible across all configurations. This balance led us to perform 20 independent training runs for configuration, as also done by other studies in the software engineering and fairness literature [7]. In particular, for classification and regression tasks, we used 10-fold cross-validation [36], averaging the results across the 10 evaluations for each of the 20 training runs. In contrast, clustering training runs were conducted 20 times, each with varying numbers of clusters. We then assessed fairness and performance levels to establish a *baseline* for both research questions.

Next, we retrained the same ML models, this time applying the fairness-aware practices individually. Similar to the baseline experiments, each training run was repeated *20 times*, allowing us to conduct a second round of evaluations for fairness and performance. In total, including both the baseline and the additional experiments, we conducted *5940 experiments*, expanding from our initial set of 45 experiments [52].

RQ$_1$ — Fairness Evaluation. To verify the significance of the obtained results, we adopted an approach consistent with the preliminary study [52], applying the Shapiro-Wilk and Wilcoxon signed-rank tests to assess fairness outcomes. These tests allowed us to determine that the differences in metric distributions across repeated runs are statistically significant and observe that they are unlikely to be due to random variation. Nonetheless, they do not in themselves indicate improvements in fairness. Rather, fairness improvements are grounded in the observed reductions of group fairness disparities, as measured by metrics such as Demographic Parity Difference and Equal Opportunity Difference. In other terms, statistical significance is used to support the robustness of these improvements across multiple runs.

Specifically, the context of our first research question, the evaluation focused exclusively on fairness metrics. The objective was to determine whether the application of fairness-aware practices resulted in measurable improvements in fairness across each of the selected datasets. Unlike the previous study [52], we broadened the scope of this evaluation, shifting the focus towards individual practices. First, we increased the robustness of the experiments by repeating each run *20 times* for every task and sensitive attribute. This allowed us to apply statistical tests to verify whether each practice led to a statistically significant increase in fairness compared to the baseline across all the analyzed metrics. Hence, the application of the tests can confirm the consistency and robustness of the statistical significance improvements across multiple training runs, rather than their effectiveness in isolation.

We began by analyzing the normality of the data to select the most appropriate statistical methods. The *Shapiro-Wilk test* [30] conducted with a significance level of $\alpha = 0.05$, revealed that not all datasets followed a normal distribution. As a result, we adopted non-parametric methods. Specifically, we applied the *Wilcoxon signed-rank test* [78] to compare the baseline with the experiments incorporating fairness-aware practices, testing the null hypothesis of no significant difference. The use of the Wilcoxon test allowed us to compute p-values and directly assess statistical significance. In addition to assessing statistical significance, we computed the effect size to quantify the magnitude of the observed differences. We employed the *Cliff's Delta test* [40], which quantifies the degree of overlap between two distributions, offering an intuitive interpretation of the probability that a randomly selected observation from one group will be greater than a randomly selected observation from the other. This allowed us to evaluate not only whether fairness-aware practices resulted in statistically significant improvements but also the practical relevance of these improvements against the baseline.

RQ$_2$ — Cost-Effective Evaluation. For the second research question, we followed the same approach adopted in the first concerning the experiments, based on the methodologies of the preliminary study [52]. After calculating the performance and fairness metrics, we conducted a *cost-effectiveness* analysis [60]. This technique, used to quantify the relationship between the cost and effectiveness of an intervention, was elaborated in the preliminary study. We evaluated each fairness-aware practice applied during the training of the ML model with respect to a specific sensitive attribute, assessing its effectiveness in improving fairness versus its cost in terms of model performance loss. This approach allowed us to quantify and identify the most efficient technique for balancing fairness and performance. For each experiment in which a fairness-aware practice was applied, we calculated two fundamental measures: **Cost**, that is, the difference in performance between the baseline model (B) without practices and the model (I) incorporating fairness-aware practices. **Effectiveness** that is, the difference in fairness metrics between fairness-aware models (I) and baseline models (B).

With these two measures, we computed a *cost-effectiveness (CE)* ratio as follows:

$$\text{Cost-effectiveness} = \frac{\text{Performance}_B - \text{Performance}_I}{\text{Fairness}_I - \text{Fairness}_B}$$

This metric allowed us to compare fairness-aware practices and identify the one that improves fairness with the least negative impact on performance. The formula was designed with the understanding that smaller fairness metrics indicate better equity, while higher performance metrics reflect greater model efficiency. For the Regression task, where MSE and MAD are error-based metrics, we inverted the performance loss value to align with the other metrics.

A CE ratio close to zero indicates an ideal trade-off, where fairness improvements are achieved with minimal performance loss. CE values greater than 1 indicate that fairness gains come at a disproportionate performance cost, potentially undermining model utility. Conversely, CE values less than -1 suggest performance improvements at the expense of fairness, which conflicts with ethical objectives. Therefore, practices with CE ratios between -1 and 1 suggest a balanced relationship, where fairness gains are typically made without significantly sacrificing performance, or even with gains in both fairness and performance. In particular, when both fairness and performance improve (i.e., $CE < 0$ and the denominator is positive), the fairness-aware practice yields a win–win outcome and is especially desirable. On the other hand, if both fairness and performance worsen (i.e., $CE > 0$ and both differences are negative), the practice should be reconsidered, as it may harm both model utility and ethical objectives.

We also note that the CE ratio should be interpreted with caution, especially when considered in isolation. In practical applications, it is important to examine the individual fairness and performance differences alongside the CE value, as this provides a more nuanced understanding of how a given practice behaves in a specific context.

For each combination of dataset, task, sensitive attribute, and practice, we calculated the CE for every performance and fairness metric across 20 experiments, capturing a comprehensive view of trade-offs. We then aggregated these CE values to derive a single general CE ratio per practice, representing its overall balance between fairness and performance.

4. Analysis of the Results

In this section, we present the results of the empirical study. All the data and scripts used to collect results and answer our research questions are available in our online appendix [51]. The discussion of the results is organized around each dataset to improve readability and clarity. However, since model-specific trends could provide additional insights into the effectiveness of fairness-aware practices, we provide additional analyses for further reading in our online appendix [51].

4.1. ***RQ*$_1$** — *Fairness Evaluation*

To answer **RQ$_1$**, we conducted a comprehensive experimental analysis. Each task was trained on its corresponding dataset, and fairness metrics were computed with and without applying the practices, considering sensitive attributes. To assess significance, we used the *Wilcoxon signed-rank test* [78], and calculated *Cliff's Delta* [40] to estimate effect size. Descriptive statistics were also computed and are available, alongside full experimental data, including non-significant results and all metrics, in our online appendix [51] for transparency and reproducibility.

Table 4: RQ$_1$ — Results for the COMPAS dataset on classification tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness (*delta* ≤ -0.5) based on effect size.

COMPAS Dataset	AAOD		FDRD		DI	
Classification - Random Forest	Sex	Race	Sex	Race	Sex	Race
Iterative Imputer	0.121	0.152 ⬆	0.057 ⬆	0.099	0.311	0.514
Oversampling	0.120	0.162	0.062	0.090 ⬆	0.299	0.524
Mutation Testing	0.123	0.072 ⬆	0.048 ⬆	0.035 ⬆	0.160 ⬆	0.105 ⬆
Regularization	0.209	0.250	0.043 ⬆	0.045 ⬆	0.488	0.985
Simple Imputer	0.125	0.153	0.057 ⬆	0.097	0.322	0.522
MinMax Scaling	0.119	0.162	0.064	0.092	0.298	0.524
Select Best	0.113	0.157	0.085	0.097	0.278 ⬆	0.498
Classification - Logistic Regression	Sex	Race	Sex	Race	Sex	Race
Undersampling	0.219	0.249	0.029	0.046	0.496 ⬆	0.944 ⬆
Simple Imputer	0.221	0.245	0.032	0.046	0.519	0.996
Regularization	0.208 ⬆	0.251	0.043	0.045	0.486	0.991
Ovesampling	0.218	0.248	0.027	0.046	0.494 ⬆	0.942 ⬆
Mutation Testing	0.228	0.125 ⬆	0.018 ⬆	0.069	0.819	1.053
Select Best	0.212 ⬆	0.256	0.023 ⬆	0.037 ⬆	0.496 ⬆	1.012
MinMax Scaling	0.218	0.252	0.034	0.043	0.500	0.991
Iterative Imputer	0.222	0.246	0.030	0.046	0.520	0.999
Classification - XGBoost	Sex	Race	Sex	Race	Sex	Race
Oversampling	0.163	0.198	0.057	0.075	0.395	0.718
Simple Imputer	0.155	0.191 ⬆	0.059	0.076	0.397	0.731
Regularization	0.207	0.250	0.043 ⬆	0.044 ⬆	0.485	0.987
Mutation Testing	0.124 ⬆	0.095 ⬆	0.059	0.041 ⬆	0.168 ⬆	0.148 ⬆
Select Best	0.154	0.204	0.046	0.075	0.388	0.764
Iterative Imputer	0.153	0.194	0.061	0.074	0.392	0.744
MinMax Scaling	0.159	0.199	0.055	0.074	0.395	0.732

Dataset for the Recidivism Context. We assessed the practices in this context through the COMPAS dataset using three classification tasks, i.e., Random Forest, Logistic Regression, and XGBoost. In particular, Table 4 shows the results of our statistical analysis, only reporting practices for which the fairness score significantly changed for at least one metric and sensitive attribute.

For the classification task using Random Forest, *Mutation Testing* emerged as the most effective approach, demonstrating significant fairness improvements across multiple metrics. It showed positive shifts toward greater fairness for both Sex and Race attributes in FDRD metrics (0.048 and 0.035, respectively), and for Sex in DI (0.160). *Regularization* also performed notably well, particularly in FDRD metrics for both Sex (0.043) and Race (0.045).

In Logistic Regression models, *Select Best* demonstrated the most consistent improvements. It showed particular strength in FDRD for both Sex (0.023) and Race (0.037). *Mutation Testing* also performed strongly, especially for FDRD Sex (0.018), with the most substantial fairness improvement across all techniques. *Undersampling* and *Oversampling* both showed significant fairness improvements for DI metrics, with particularly strong results for Race (0.944 and 0.942, respectively). XGBoost classification results revealed that *Mutation Testing* provided the most consistent fairness improvements, with significant positive shifts. *Regularization* also performed well across FDRD metrics for both Sex (0.043) and Race (0.044).

> ☰ **Recidivism Context — COMPAS Dataset.**
>
> *For this dataset, Mutation Testing consistently improves fairness across classifiers, especially by reducing discrimination across sensitive attributes. Regularization and Select Best also show strong results, particularly with certain algorithms.*

Table 5: RQ_1 — Results for the Adult dataset on clustering tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness (*delta* ≤ -0.5) based on effect size.

Adult Dataset	AAOD		FDRD		DI	
Classification - Random Forest	Sex	Race	Sex	Race	Sex	Race
Iterative Imputer	0.078	0.070	0.003	0.012	0.6878	0.574
Oversampling	0.094	0.078	0.004	0.010	0.683	0.558 ⬆
Mutation Testing	0.020 ⬆	0.144	0.102	0.130	0.160 ⬆	0.6717
Regularization	0.070 ⬆	0.067	0.010	0.008	0.705	0.591
Undersampling	0.147	0.085	0.019	0.035	0.670 ⬆	0.502 ⬆
MinMax Scaling	0.187	0.141	0.004	0.004 ⬆	0.683	0.566
Select Best	0.077	0.065	0.048	0.041	0.660 ⬆	0.535 ⬆
Classification - Logistic Regression	Sex	Race	Sex	Race	Sex	Race
Iterative Imputer	0.110	0.045	0.252	0.035	0.685	0.481
MinMax Scaling	0.190	0.120	0.006 ⬆	0.010 ⬆	0.853	0.708
Mutation Testing	0.102 ⬆	0.350	0.095 ⬆	0.075	0.084 ⬆	1.351
Regularization	0.159	0.086	0.166 ⬆	0.058	0.879	0.732
Undersampling	0.242	0.100	0.287	0.088	0.563 ⬆	0.457 ⬆
Select Best	0.176	0.075	0.013 ⬆	0.034	0.842	0.595
Oversampling	0.236	0.101	0.288	0.091	0.558 ⬆	0.456 ⬆
Classification - XGBoost	Sex	Race	Sex	Race	Sex	Race
MinMax Scaling	0.188	0.145	0.003 ⬆	0.009	0.673	0.561
Mutation Testing	0.021 ⬆	0.124	0.116	0.156	0.181 ⬆	0.594
Oversampling	0.151	0.100	0.012	0.011	0.700	0.556 ⬆
Undersampling	0.146	0.094	0.020	0.030	0.673 ⬆	0.524
Select Best	0.067 ⬆	0.074	0.006 ⬆	0.009	0.675	0.588
Regularization	0.071	0.070	0.011	0.008	0.678	0.577

Dataset for the Economics Context. In this context, we evaluated the Adult dataset [4] using three classification tasks and three clustering tasks. Tables 5 and 6 present our comprehensive results. For classification tasks, fairness-aware practices showed varied effectiveness. With Random Forest, *Mutation Testing* yielded significant improvements for Sex in AAOD (0.020) and DI (0.160). In Logistic Regression, *MinMax Scaling* was particularly effective for FDRD Sex

Table 6: RQ$_1$ — Results for the Adult dataset on clustering tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness (*delta* ≤ -0.5) based on effect size.

Adult Dataset	AE		ME		AW		MW	
Clustering - k-mean	Sex	Race	Sex	Race	Sex	Race	Sex	Race
MinMax Scaling	1.0744 ⬆	0.4013	1.0847 ⬆	0.4792 ⬆	0 ⬆	0.0123 ⬆	0 ⬆	0.0123 ⬆
Undersampling	6023	6928	43584	17716	0.1889	0.176	0.1889	0.176
Oversampling	7112	7614	59328	17642	0.1928	0.1567	0.1928	0.1567
Clustering - K-center	Sex	Race	Sex	Race	Sex	Race	Sex	Race
MinMax Scaling	1.188 ⬆	0.462 ⬆	1.331 ⬆	0.659 ⬆	0.065 ⬆	0.081 ⬆	0.065 ⬆	0.081 ⬆
Undersampling	19288	9881 ⬆	64080	16552	0.180	0.172	0.180	0.172
Oversampling	17876	18609	71572	31747	0.154	0.190	0.154	0.190
Clustering - K-median	Sex	Race	Sex	Race	Sex	Race	Sex	Race
MinMax Scaling	1.123 ⬆	0.411 ⬆	1.169 ⬆	0.561 ⬆	0.056 ⬆	0.034 ⬆	0.056 ⬆	0.034 ⬆
Oversampling	2607	5921	17819	14732	0.192	0.148 ⬆	0.192	0.148 ⬆
Undersampling	2428	6518	15278	24428	0.192	0.181	0.192	0.181

(0.006), and *Mutation Testing* showed notable gains in AAOD Sex (0.102), FDRD Sex (0.095), and DI Sex (0.084). In XGBoost, both *Select Best* and *MinMax Scaling* performed well in FDRD Sex (0.006 and 0.003), while *Mutation Testing* improved AAOD (0.021) and DI (0.181) for Sex.

In clustering tasks, only *MinMax Scaling* and sampling-based approaches (*Undersampling*, *Oversampling*) showed effectiveness. *MinMax Scaling* displayed remarkable consistency across K-means, K-center, and K-median, improving all fairness metrics.

The outstanding performance of *MinMax Scaling* in clustering is rooted in its technical properties [34, 84]. By scaling features to a uniform range, it prevents dominance by high-magnitude features in distance calculations [79], which is crucial in clustering algorithms reliant on such metrics. This mitigates bias from features correlated with sensitive attributes [19, 55].

> ☰ **Economics Context — Adult Dataset.**
>
> *For this dataset, the classification tasks demonstrated effectiveness, particularly with Mutation Testing, Select Best, and Sampling strategies. Moreover, the clustering results suggest that MinMax Scaling should be prioritized when addressing fairness concerns in unsupervised learning contexts.*

Dataset for the Marketing Context. This evaluation was performed using the Bank Marketing dataset [45] dataset using one classification task and three clustering tasks, as illustrated in Tables 7 and 8.

For classification tasks using Random Forest, several fairness-aware practices proved effective. *MinMax Scaling* notably improved FDRD for Age (0.004) and DI for Marital Status (0.013). Sampling-based methods consistently benefited FDRD: *Oversampling* improved Age (0.007) and Marital Status (0.008); *Undersampling* enhanced AAOD for Marital Status (0.014) and FDRD for Marital Status (0.009). *Select Best* was particularly effective for FDRD Marital Status (0.005), while *Regularization* improved FDRD Age (0.005) and DI Marital Status (0.016).

Table 7: RQ$_1$ — Results for the Bank dataset on classification tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness (*delta* ≤ -0.5) based on effect size.

Bank Marketing Dataset	AAOD		FDRD		DI	
Classification - Random Forest	Age	Marital	Age	Marital	Age	Marital
Iterative Imputer	0.021	0.019	0.020	0.021	0.050	0.020
MinMax Scaling	0.019	0.020	0.004 ⬆	0.013	0.011	0.013 ⬆
Mutation Testing	0.035	0.016 ⬆	0.050	0.015	0.100	0.100
Oversampling	0.020	0.020	0.007 ⬆	0.008 ⬆	0.011	0.026
Undersampling	0.013 ⬆	0.014 ⬆	0.018	0.009 ⬆	0.013	0.046
Regularization	0.021	0.021	0.005 ⬆	0.013	0.032	0.016 ⬆
Select Best	0.021	0.022	0.009 ⬆	0.005 ⬆	0.036	0.031
Simple Imputer	0.023	0.021	0.031	0.019	0.032	0.024

Table 8: RQ$_1$ — Results for the Bank dataset on clustering tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness (*delta* ≤ -0.5) based on effect size.

Bank Marketing Dataset	AE		ME		AW		MW	
Clustering - k-mean	Age	Marital	Age	Marital	Age	Marital	Age	Marital
MinMax Scaling	0.904 ⬆	0.348 ⬆	0.989 ⬆	0.361 ⬆	0.089 ⬆	0.007 ⬆	0.089 ⬆	0.989 ⬆
Clustering - K-center	Age	Marital	Age	Marital	Age	Marital	Age	Marital
MinMax Scaling	0.985 ⬆	0.384 ⬆	1.395 ⬆	0.536 ⬆	0.151	0.036 ⬆	0.151	0.036 ⬆
Oversampling	173.726 ⬆	65.915	267.648	218.714	0.179	0.106	0.179	0.106
Undersampling	-	56.890	-	110.410	-	0.1475	-	0.147
Clustering - K-median	Age	Marital	Age	Marital	Age	Marital	Age	Marital
Oversampling	357.980	7.808	631.949	24.427	0.141	0.01	0.141	0.01
MinMax Scaling	0.962 ⬆	0.348 ⬆	1.112 ⬆	0.366 ⬆	0.137	0.006	0.137	0.006
Undersampling	-	64.069	-	295.403	-	0.108	-	0.108

In clustering tasks, *MinMax Scaling* showed strong effectiveness across all three algorithms (K-means, K-centers, K-median) and fairness metrics.The notable effectiveness of *MinMax Scaling* in clustering is due to its transformation of the feature space [34, 84]. Normalization is key to reducing the influence of features correlated with sensitive attributes [19, 55, 79].

> **Marketing Context — Bank Marketing Dataset.**
>
> *The analysis of this dataset confirms that while several methods enhance fairness in classification, MinMax Scaling stands out in clustering for its consistent and comprehensive fairness improvements across all algorithms and metrics.*

Dataset for the Finance Context. We used the German Credit dataset [31] on four classification tasks. As shown in Table 9, our experiments revealed several patterns in fairness improvements.

Table 9: RQ$_1$ — Results for the German Credit dataset on classification tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness ($delta \leq -0.5$) based on effect size.

German Credit Dataset	AAOD		FDRD		DI	
Classification - Random Forest	Age	Gender	Age	Gender	Age	Gender
Select Best	0.118	0.117	0.054	0.099	0.487	0.456
Mutation Testing	0.036 ⬆	0.034 ⬆	0.029	0.0209 ⬆	0.484	0.389 ⬆
Oversampling	0.111	0.124	0.031	0.102	0.447	0.316 ⬆
MinMax Scaling	-	-	0.04	0.017 ⬆	-	0.497
Classification - Logistic Regression	Age	Gender	Age	Gender	Age	Gender
Undersampling	0.137	0.148	0.02	0.082	0.415 ⬆	0.312 ⬆
Oversampling	0.145	0.139	0.018	0.081	0.372 ⬆	0.308 ⬆
Select Best	0.121	0.143	0.027	0.035 ⬆	-	0.790
Regularization	0.119	0.133	0.032	0.049	-	0.752
Mutation Testing	0.034 ⬆	0.086 ⬆	0.011	0.013 ⬆	0.029 ⬆	0.309 ⬆
Iterative Imputer	0.128	0.139	0.042	0.053	-	0.742
MinMax Scaling	-	-	0.012	0.015 ⬆	0.706	0.807
Classification - XGBoost	Age	Gender	Age	Gender	Age	Gender
Mutation Testing	0.040 ⬆	0.031 ⬆	0.029	0.015 ⬆	0.300 ⬆	0.138 ⬆
Regularization	0.114	0.134	0.023 ⬆	0.044	0.548	0.755
Undersampling	0.138	0.127	0.032	0.058	0.345 ⬆	0.218 ⬆
MinMax Scaling	-	-	0.028	0.012 ⬆	-	0.406
Classification - Decision Tree	Age	Gender	Age	Gender	Age	Gender
Mutation Testing	0.039 ⬆	0.043 ⬆	0.010 ⬆	0.010 ⬆	0.023 ⬆	0.048 ⬆
Regularization	0.111	0.136	0.029	0.045	0.520	0.768
Select Best	0.133	0.130	0.0411	0.068	0.311	0.269
Oversampling	0.141	0.125	0.043	0.093	0.436	0.217
Undersampling	0.141	0.139	0.035	0.062	0.201	0.138
MinMax Scaling	-	-	0.018 ⬆	0.017 ⬆	0.267	0.155

Mutation Testing was consistently effective across all classification algorithms. For Decision Tree models, it delivered strong fairness improvements for both Age and Gender (AAOD Age: 0.039, AAOD Gender: 0.043, FDRD Age: 0.010, FDRD Gender: 0.010, DI Age: 0.023, DI Gender: 0.048). A similar pattern was observed in XGBoost, with *Mutation Testing*.

In Random Forest models, *Mutation Testing* again performed well, improving AAOD for Age (0.036) and Gender (0.034), FDRD Gender (0.0209), and DI Gender (0.389). Additionally, a separate implementation of *MinMax Scaling* showed strong results in FDRD Gender (0.017). In Logistic Regression, effective techniques were more varied: *Mutation Testing* maintained good performance across most metrics (AAOD Age: 0.034, AAOD Gender: 0.086, FDRD Gender: 0.013, DI Age: 0.029, DI Gender: 0.309), while *MinMax Scaling* improved FDRD for both Age (0.012) and Gender (0.015). Across all classification models, FDRD improvements were achieved through multiple techniques. For the DI metric, sampling-based methods (*Undersampling* and *Oversampling*) were particularly effective in Logistic Regression and XGBoost.

☰ Finance Context — German Credit Dataset.

For this dataset, Mutation Testing offers the most consistent and comprehensive fairness gains across classifiers and sensitive attributes. However, techniques like MinMax Scaling (for FDRD) and sampling-based methods (for DI) demonstrate the value of tailoring fairness strategies to specific concerns and algorithms.

Table 10: RQ$_1$ — Results for the Crime dataset on classification tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness (*delta* ≤ -0.5) based on effect size.

Communities and Crime Dataset	AAOD	FDRD	DI
Classification - Decision Tree	Race	Race	Race
Iterative Imputer	0.301	0.320 ⬆	3.015
Select Best	0.349	0.319	3.157
Regularization	0.324	0.270 ⬆	3.701
Undersampling	0.322	0.361	2.027 ⬆
Mutation Testing	0.189 ⬆	0.029 ⬆	0.224 ⬆
Oversampling	0.300	0.319	3.110
MinMax Scaling	-	0.007 ⬆	3.019
Classification - Naïve Bayes	Race	Race	Race
Select Best	0.231	0.159 ⬆	6.076
Mutation Testing	0.231 ⬆	0.017 ⬆	0.223 ⬆
Regularization	0.372	0.232 ⬆	3.536
MinMax Scaling	-	0.006 ⬆	3.420
Classification - Logistic Regression	Race	Race	Race
Select Best	0.470 ⬆	0.10	8.758
Iterative Imputer	0.365	0.110	5.909
Regularization	0.401	0.083 ⬆	6.942
Undersampling	0.420	0.202	3.295 ⬆
Oversampling	0.401	0.200	3.376 ⬆
Mutation Testing	0.311 ⬆	0.021 ⬆	0.357 ⬆
MinMax Scaling	-	0.012 ⬆	5.91

Dataset for the Crime Context. We evaluated the Communities and Crime dataset [59] using three classification tasks and two regression tasks. As illustrated in Tables 10 and 11, our analysis revealed several notable patterns in fairness improvements across different algorithms and mitigation techniques.

For classification tasks, *Mutation Testing* consistently delivered superior fairness improvements across all three algorithms. In Decision Trees, it significantly improved all three metrics, with notable results in FDRD (0.029) and DI (0.224). Similarly, in Naïve Bayes, it achieved substantial gains in AAOD (0.231), FDRD (0.017), and DI (0.223). This trend continued in Logistic Regression (AAOD: 0.311, FDRD: 0.021, DI: 0.357).

In regression tasks, *Mutation Testing* improved all metrics in both Linear Regression and Decision Trees. In Linear Regression, it yielded optimal results for Separation (1.009), Sufficiency (1.009), and Independence (1.003). This pattern held in Decision Tree regression as well

Table 11: RQ$_1$ — Results for the Crime dataset on regression tasks. Values in the cells indicate the mean value for each metric across the 20 experiments. Only fairness-aware practices with a statistically significant difference against the baseline for at least one metric and sensitive attribute were reported. A Light Purple cell marks a significant difference. The arrow-up (⬆) marks a shift toward greater fairness (*delta* ≤ -0.5) based on effect size.

Communities and Crime Dataset	Separation	Sufficiency	Independence
Regression - Linear Regression	Race	Race	Race
Iterative Imputer	10.983 ⬆	4.376 ⬆	1.141
Undersampling	14.908	5.057	1.179
Simple Imputer	10.494 ⬆	4.225	1.115
Select Best	17.034	8.553	1.001 ⬆
MinMax Scaling	10.737 ⬆	4.343	1.104
Oversampling	15.947	5.485	1.120
Mutation Testing	1.009 ⬆	1.009 ⬆	1.003 ⬆
Regularization	12.781 ⬆	5.124	1.098
Regression - Decision Tree	Race	Race	Race
Select Best	1.385 ⬆	1.142 ⬆	1.147 ⬆
Regularization	3.197	1.762	1.518 ⬆
Undersampling	3.13	1.770	1.79
Mutation Testing	1.020 ⬆	1.019 ⬆	1.003 ⬆
Oversampling	2.393	1.456	1.756

(Separation: 1.020, Sufficiency: 1.019, Independence: 1.003). *Select Best* also performed well.

Crime Context — Communities and Crime Dataset.

For this dataset, Mutation Testing yields the most consistent fairness improvements across classification and regression. MinMax Scaling is particularly effective for classification FDRD metrics, while Select Best shows strength in regression tasks.

RQ$_1$ — Summary of the Results.

Overall, *Mutation Testing* consistently delivers strong fairness improvements across Recidivism, Finance, and Crime datasets in both classification and regression. In Economics, *MinMax Scaling* is key for unsupervised learning and performs reliably in Bank Marketing clustering. While these two methods excel across datasets, *Select Best* and sampling also show promise in specific scenarios, highlighting the need for context and dataset-specific fairness strategies across diverse ML tasks.

*4.2. **RQ$_2$** — Cost-Effective Evaluation*

For **RQ$_2$**, we evaluated the models' performance using a cost-effectiveness (CE) analysis to identify practices that improve fairness with the least negative impact on performance. For each task, we consider the average CE ratio for every combination of performance and fairness metrics, calculated across different practices, sensitive attributes, and datasets.

Undersampling. The analysis of the results obtained with the practice of *Undersampling*, visible in Table 12, highlights a significant variability in CE values, influenced by the dataset, the model used, and the sensitive attribute considered. In some cases, such as Logistic Regression on COMPAS with the Race attribute (CE = 0.042), it had minimal impact on performance, suggesting a good trade-off between fairness and accuracy. However, in other scenarios, the loss in performance outweighed the gains in fairness, as seen with XGBoost on Adult for the Sex attribute (CE = -45.841). On the other hand, led to simultaneous improvements in both fairness and performance, such as Logistic Regression on COMPAS for the Sex attribute (CE = 4.262). Compared to *Oversampling*, which also showed mixed results, *Undersampling*, in all practices, generally demonstrated fewer extreme performance drops, making it a more stable approach in certain application domains.

Table 12: RQ$_2$ — Results for the Undersampling practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

Undersampling			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Logistic Regression	Recidivism Prediction (COMPAS Dataset)	Sex	4.262
Classification - Logistic Regression	Recidivism Prediction (COMPAS Dataset)	Race	0.042
Classification - Random Forest	Economics (Adult Dataset)	Sex	-7.842
Classification - Random Forest	Economics (Adult Dataset)	Race	-2.784
Classification - XGBoost	Economics (Adult Dataset)	Sex	-45.841
Classification - Logistic Regression	Economics (Adult Dataset)	Sex	-0.204
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Age	3.549
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Marital	2.516
Classification - Decision Tree	Crime (Communities and Crime Dataset)	Race	0.163
Classification - Logistic Regression	Crime (Communities and Crime Dataset)	Race	0.248
Classification - XGBoost	Finance (German Credit Dataset)	Age	-2.311
Classification - XGBoost	Finance (German Credit Dataset)	Gender	-2.710
Classification - Logistic Regression	Finance (German Credit Dataset)	Age	0.311
Classification - Logistic Regression	Finance (German Credit Dataset)	Gender	-14.655
Clustering - K-center	Economics (Adult Dataset)	Race	-2.910

Oversampling. The analysis of *Oversampling* showed mixed results, with significant variability depending on the dataset, model, and sensitive attribute. In some cases, it balanced fairness and performance, such as Random Forest on COMPAS for Race (CE = 2.199) and German Credit for Gender (CE = 2.494). However, other cases saw severe performance drops, particularly XGBoost on Adult for Race (CE = -158.781) and K-median clustering on Adult for Race (CE = -50.889). While some models, like Logistic Regression on COMPAS for Race (CE = 0.574), showed minor improvements. Notably, all results highlight the need for careful evaluation before applying *Oversampling*, as exhibited by a wider range of CE values, indicating higher risk in terms of loss of performance but also greater potential fairness benefits in selected cases.

Iterative Imputer. The *Iterative Imputer* exhibited a high cost in terms of performance loss, with the obtained results reported in Table 14. For example, Random Forest on COMPAS for the Race attribute recorded a CE of -27.131, indicating a significant negative impact. However, not all results were negative: Decision Tree on Crime for the Race attribute (CE = 1.469) showed fairness benefits without excessively compromising performance. Moreover, Linear Regression on the Crime dataset for the Race attribute presented a CE of -0.268, suggesting it may be less detrimental to performance. This method can lead to significant fairness improvements, but it tends, in general, to introduce more drastic performance losses.

Table 13: RQ_2 — Results for the Oversampling practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

Oversampling			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Random Forest	Recidivism (COMPAS Dataset)	Race	2.199
Classification - Logistic Regression	Recidivism (COMPAS Dataset)	Sex	-4.074
Classification - Logistic Regression	Recidivism (COMPAS Dataset)	Race	-0.574
Classification - Random Forest	Economics (Adult Dataset)	Race	1.463
Classification - Logistic Regression	Economics (Adult Dataset)	Race	-0.930
Classification - Logistic Regression	Economics (Adult Dataset)	Sex	-0.215
Classification - XGBoost	Economics (Adult Dataset)	Race	-158.781
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Age	-0.431
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Mrital	0.482
Classification - Random Forest	Finance (German Credit Dataset)	Gender	2.494
Classification - Logistic Regression	Finance (German Credit Dataset)	Age	-0.008
Classification - Logistic Regression	Finance (German Credit Dataset)	Gender	-0.228
Classification - Logistic Regression	Crime (Communities and Crime Dataset)	Race	0.107
Clustering - K-median	Economics (Adult Dataset)	Race	-50.889
Clustering - K-center	Marketing (Bank Marketing Dataset)	Age	-0.760

Table 14: RQ_2 — Results for the Iterative Imputer practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

Iterative Imputer			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Random Forest	Recidivism (COMPAS Dataset)	Race	-27.131
Classification - Random Forest	Recidivism (COMPAS Dataset)	Sex	-3.338
Classification - Decision Tree	Crime (Communities and Crime Dataset)	Race	1.469
Regression - Linear Regression	Crime (Communities and Crime Dataset)	Race	-0.268

Simple Imputer. The *Simple Imputer* showed variable results, as represented in Table 15. In some cases, it significantly improved fairness, such as with Random Forest on COMPAS for the Sex attribute (CE = 10.474). This suggests that the imputation strategy effectively mitigated bias without overly compromising performance. On the other hand, its impact was more limited when applied to XGBoost on COMPAS for the Race attribute, which yielded a CE of 2.229. Generally led to more positive CE values, making it a preferable imputation strategy in application domains where minimizing performance loss is crucial.

Table 15: RQ_2 — Results for the Simple Imputer practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

Simple Imputer			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Random Forest	Recidivism Prediction (COMPAS Dataset)	Sex	10.474
Classification - XGBoost	Recidivism Prediction (COMPAS Dataset)	Race	2.229
Regression - Linear Regression	Crime (Communities and Crime Dataset)	Race	0.008

Select Best. The *Select Best* technique generally yielded better results than other practices, as can be observed in Table 16. Many experiments reported CE values that provide a good trade-off between fairness and performance. For instance, Linear Regression on Crime for Race (CE = -0.741) and Logistic Regression on German Credit for Gender (CE = 3.190) demonstrated fairness improvements with minimal performance loss. Similarly, Random Forest on Bank Marketing for Marital (CE = 2.558) showed positive fairness outcomes. While some models experienced performance drops, such as XGBoost on Adult for Sex and Logistic Regression on COMPAS for Race. Overall, *Select Best* stands out as one of the more effective techniques, yielding positive fairness outcomes with fewer performance losses.

Table 16: RQ$_2$ — Results for the Select Best practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

Select Best			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Random Forest	Recidivism (COMPAS Dataset)	Sex	-3.884
Classification - Logistic Regression	Recidivism (COMPAS Dataset)	Sex	-12.091
Classification - Logistic Regression	Recidivism (COMPAS Dataset)	Race	-5.009
Classification - Random Forest	Economics (Adult Dataset)	Race	-0.283
Classification - Random Forest	Economics (Adult Dataset)	Sex	-24.700
Classification - XGBoost	Economics (Adult Dataset)	Sex	-101.155
Classification - Logistic Regression	Economics (Adult Dataset)	Sex	-0.768
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Age	-1.145
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Marital	2.558
Classification - Logistic Regression	Finance (German Credit Dataset)	Gender	3.190
Classification - Naïve Bayes	Crime (Communities and Crime Dataset)	Race	-1.282
Classification - Logistic Regression	Crime (Communities and Crime Dataset)	Race	-3.282
Regression - Linear Regression	Crime (Communities and Crime Dataset)	Race	-0.741

MinMax Scaling. The application of *MinMax Scaling* showed more balanced CE values compared to other practices, as visible in Table 17. For Adult, most values were positive, such as XGBoost on Sex (CE = 2.916) and Random Forest on Race (CE = 0.730). However, some algorithms had negative impacts, such as Logistic Regression on Race (-0.946). In the financial dataset, the technique produced some highly positive values, such as Decision Tree on Age (CE = 25.943), indicating an improvement in fairness with some performance loss. For Crime, the results were more contained, with Logistic Regression (CE = 0.083) showing a balanced compromise. Compared to other techniques, *MinMax Scaling* exhibited a more consistent balance between fairness improvements and performance retention.

Regularization. The analysis of *Regularization* produced highly variable CE values, as observed in Table 18, indicating that its impact on fairness and performance strongly depends on the dataset and model. Some models showed significant fairness gains, such as Naïve Bayes on Crime for Race (CE = 31.642) and Logistic Regression on COMPAS for Sex (CE = 4.262). Similarly, Random Forest on Adult for Sex (CE = 5.980) improved fairness with minimal performance cost. However, other cases experienced sharp performance declines, as XGBoost on COMPAS for Sex and Random Forest on Marketing for Age. Overall, *Regularization* appears effective in improving fairness but requires careful analysis.

Mutation Testing. The analysis of *Mutation Testing* showed significant trade-offs between fairness and performance, as seen in Table 19, with several CE values close to zero, indicating

Table 17: RQ$_2$ — Results for the MinMax Scaling practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

MinMax Scaling			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Random Forest	Economics (Adult Dataset)	Race	0.730
Classification - XGBoost	Economics (Adult Dataset)	Sex	2.916
Classification - Logistic Regression	Economics (Adult Dataset)	Race	-0.946
Classification - Logistic Regression	Economics (Adult Dataset)	Sex	-0.719
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Age	0.0791
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Marital	1.673
Classification - Random Forest	Finance (German Credit Dataset)	Gender	0.132
Classification - Logistic Regression	Finance (German Credit Dataset)	Gender	-0.064
Classification - XGBoost	Finance (German Credit Dataset)	Gender	0.057
Classification - Decision Tree	Finance (German Credit Dataset)	Age	25.943
Classification - Decision Tree	Finance (German Credit Dataset)	Gender	0.428
Classification - Decision Tree	Crime (Communities and Crime Dataset)	Race	-0.712
Classification - Naïve Bayes	Crime (Communities and Crime Dataset)	Race	-0.918
Classification - Logistic Regression	Crime (Communities and Crime Dataset)	Race	0.083
Regression - Linear Regression	Crime (Communities and Crime Dataset)	Race	0.042
Clustering - k-mean	Economics (Adult Dataset)	Race	0.616
Clustering - k-mean	Economics (Adult Dataset)	Sex	0.770
Clustering - K-median	Economics (Adult Dataset)	Race	0.247
Clustering - K-median	Economics (Adult Dataset)	Sex	0.192
Clustering - K-center	Economics (Adult Dataset)	Race	0.764
Clustering - K-center	Economics (Adult Dataset)	Sex	0.913
Clustering - k-mean	Marketing (Bank Marketing Dataset)	Age	1.035
Clustering - k-mean	Marketing (Bank Marketing Dataset)	Marital	0.312
Clustering - K-median	Marketing (Bank Marketing Dataset)	Age	1.405
Clustering - K-median	Marketing (Bank Marketing Dataset)	Marital	0.067
Clustering - K-center	Marketing (Bank Marketing Dataset)	Age	0.574
Clustering - K-center	Marketing (Bank Marketing Dataset)	Marital	0.452

minimal performance loss while achieving fairness improvements. For example, Logistic Regression on COMPAS for Sex (CE = 0.412) and Random Forest on Finance for Gender (CE = 0.002) showed positive fairness effects without significantly harming performance. Similarly, XGBoost on Finance for Age (CE = 0.369) demonstrated a slight improvement in fairness. However, some cases exhibited performance deterioration, such as Random Forest on COMPAS for Sex (CE = -2.137) and Decision Tree on Economics for Race (CE = -0.709). Despite these outliers, *Mutation Testing* generally led to more balanced results in all cases.

RQ$_2$ — Summary of the Results.

The analysis highlights various techniques to improve fairness while minimizing performance loss, though their effectiveness varies across models and datasets. *Select Best* and *MinMax Scaling* emerged as generally reliable methods, often achieving a favorable trade-off. *Regularization* and *Mutation Testing* also showed promise, with many cases balancing fairness improvements and performance. Finally, *Simple Imputers* and *Iterative Imputers*

Table 18: RQ$_2$ — Results for the Regularization practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

Regularization			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Random Forest	Recidivism Prediction (COMPAS Dataset)	Sex	-18.00
Classification - Random Forest	Recidivism Prediction (COMPAS Dataset)	Race	1.404
Classification - Logistic Regression	Recidivism Prediction (COMPAS Dataset)	Sex	4.262
Classification - XGBoost	Recidivism Prediction (COMPAS Dataset)	Sex	-29.259
Classification - XGBoost	Recidivism Prediction (COMPAS Dataset)	Race	1.423
Classification - Random Forest	Economics (Adult Dataset)	Sex	5.980
Classification - Logistic Regression	Economics (Adult Dataset)	Sex	0.214
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Age	0.260
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Age	-3.562
Classification - Decision Tree	Crime (Communities and Crime Dataset)	Race	0.163
Classification - Naïve Bayes	Crime (Communities and Crime Dataset)	Race	31.642
Classification - Logistic Regression	Crime (Communities and Crime Dataset)	Race	0.248
Regression - Linear Regression	Economics (Adult Dataset)	Race	-0.818
Regression - Decision Tree	Economics (Adult Dataset)	Race	0.406

demonstrated fairness benefits in specific scenarios.

5. Discussion and Implications

Our findings provide multiple practical implications for practitioners (💼) and researchers (💡), which we discuss in this section.

5.1. On the Importance of Data Preparation

A key finding is the varied effectiveness of Sampling practices across scenarios. While *Undersampling* and *Oversampling* improved fairness in terms of Disparate Impact in datasets like Adult and models like Logistic Regression, they were less effective in Clustering tasks, where *Scaling* worked better. Dataset characteristics also mattered: sampling improved fairness in COMPAS and Adult but had limited impact on German Credit, suggesting class imbalance plays a larger role in some cases. These results align with prior work linking ML bias to dataset properties [13, 50, 54]. *Undersampling* showed more consistent results than *Oversampling*, which yielded a broader range of outcomes. While *Oversampling* led to fairness gains in some cases—e.g., Random Forest on COMPAS and German Credit—it also caused notable performance drops in others, such as XGBoost. These findings highlight the potential of *Oversampling*, but also the need for cautious application to avoid instability in performance [13].

💼 The original data distribution, especially class imbalance, strongly affects fairness. Practitioners should evaluate imbalances and apply suitable balancing techniques for each task.

Imputation practices showed a trade-off between fairness and accuracy. The *Iterative Imputer* improved fairness in some cases but often reduced performance—especially with Random Forest on COMPAS—due to increased data variability. In contrast, with Decision Trees, it improved

Table 19: RQ$_2$ — Results for the Mutation Testing practice. The table reports, for different combinations of tasks, datasets, and sensitive attributes, the average cost-effectiveness ratio across the 20 experiments.

Mutation Testing			
Task	Dataset (Context)	Sensitive Attribute	Cost-Effectiveness
Classification - Random Forest	Recidivism Prediction (COMPAS Dataset)	Sex	-2.137
Classification - Random Forest	Recidivism Prediction (COMPAS Dataset)	Race	0.085
Classification - Logistic Regression	Recidivism Prediction (COMPAS Dataset)	Sex	0.412
Classification - Logistic Regression	Recidivism Prediction (COMPAS Dataset)	Race	-1.699
Classification - XGBoost	Recidivism Prediction (COMPAS Dataset)	Sex	-0.733
Classification - XGBoost	Recidivism Prediction (COMPAS Dataset)	Race	-0.015
Classification - Random Forest	Economics (Adult Dataset)	Sex	0.001
Classification - Logistic Regression	Economics (Adult Dataset)	Sex	0.576
Classification - Logistic Regression	Economics (Adult Dataset)	Race	-0.173
Classification - XGBoost	Economics (Adult Dataset)	Sex	0.003
Classification - Random Forest	Marketing (Bank Marketing Dataset)	Marital	-0.120
Classification - Random Forest	Finance (German Credit Dataset)	Gender	0.002
Classification - Random Forest	Finance (German Credit Dataset)	Age	0.092
Classification - Logistic Regression	Finance (German Credit Dataset)	Age	-0.184
Classification - Logistic Regression	Finance (German Credit Dataset)	Gender	0.0009
Classification - Decision Tree	Finance (German Credit Dataset)	Age	0.021
Classification - Decision Tree	Finance (German Credit Dataset)	Gender	0.001
Classification - XGBoost	Finance (German Credit Dataset)	Age	0.369
Classification - XGBoost	Finance (German Credit Dataset)	Gender	-0.006
Classification - Decision Tree	Crime (Communities and Crime Dataset)	Race	-0.022
Classification - Naïve Bayes	Crime (Communities and Crime Dataset)	Race	-0.141
Classification - Logistic Regression	Crime (Communities and Crime Dataset)	Race	-0.004
Regression - Linear Regression	Economics (Adult Dataset)	Race	0.274
Regression - Decision Tree	Economics (Adult Dataset)	Race	0.709

fairness with minimal accuracy loss, highlighting the model's role. The *Simple Imputer* offered a more balanced outcome, improving fairness while maintaining performance, underscoring the value of simpler approaches and careful data integrity analysis [12, 58].

> 💡 The impact of imputation on fairness and stability is highly model-dependent. Researchers should study how imputation interacts with model architectures, as simpler methods can sometimes outperform more complex ones.

5.2. *Fairness-Aware Improvements are Algorithm-Specific*

In our analysis, the role of specific characteristics in the learning algorithms proved to be important. Indeed, the *Feature Selection* practices, such as *Select Best*, proved particularly effective for linear models like Logistic Regression, often improving fairness while maintaining acceptable accuracy levels. In contrast, this same technique led to the loss of performance in classification models such as XGBoost, demonstrating that its success is highly dependent on the dataset and model structure. These findings emphasize the need for alignment between fairness interventions and different algorithms because well-matched techniques can lead to improvements in both fairness and performance [6, 13, 50].

To deepen this analysis, we conducted a series of model-focused evaluations. First, we grouped all experimental results by model and visualized the average effect size (Cliff's Delta) of each fairness-aware practice. This analysis revealed that certain practices—such as *Select Best* and *Undersampling*—consistently achieved moderate-to-large fairness improvements in models like Decision Trees and Logistic Regression. In contrast, practices such as *Regularization* and *Iterative Imputation* exhibited more inconsistent results, sometimes offering gains and at other times showing negligible or negative impact. These results confirm that mitigation strategies interact differently with the inductive biases and optimization dynamics of each algorithm.

Next, we analyzed the consistency of each practice across datasets. For every model-practice combination, we measured how often a statistically significant fairness improvement was observed across all datasets. We found that practices like *Undersampling* and *Select Best* were not only effective in average effect size but also demonstrated high cross-dataset consistency—especially when applied to tree-based models. This reinforces the idea that some model–practice pairings generalize better than others and may be preferred in real-world applications where robustness is critical.

One of the most striking findings was the exceptional impact of *MinMax Scaling* in clustering tasks. Across all Clustering algorithms and in datasets, this method consistently improved fairness, whereas other practices failed to yield comparable results. This suggests that fairness concerns are often related to disparities in feature magnitudes, which *MinMax Scaling* effectively mitigates. While its effects in Classification were not as pronounced, they still demonstrated positive contributions, particularly in Financial datasets. *MinMax Scaling's* effectiveness in both Clustering and Classification further supports the notion that different learning models require distinct fairness interventions [6, 13].

> 💡 Researchers should focus on developing fairness interventions that dynamically adjust based on model architecture and data. The observed differences highlight the need for a deeper understanding of how fairness-aware practices interact with model learning dynamics.

Regularization practices exhibited highly variable effects, underscoring the need for context-aware interventions. While effective in improving fairness for models like Naïve Bayes on Crime, and Random Forest, Logistic Regression, and XGBoost on COMPAS, its impact was minimal across other datasets and algorithms, often accompanied by significant accuracy losses—highlighting model complexity as a key factor in fairness optimization [54].

Among all tested practices, *Mutation Testing* emerged as one of the most consistently effective interventions. By perturbing data, it addresses discriminatory patterns that influence predictions. It showed reliable fairness improvements across Random Forest, Logistic Regression, XGBoost, and Decision Tree models, enhancing metrics such as AAOD, FDRD, and DI simultaneously. Notably, it was especially impactful for datasets like COMPAS and German Credit, where historical biases are deeply embedded. Furthermore, *Mutation Testing* demonstrated greater stability than most techniques, often improving fairness without substantial performance loss—making it a promising approach for fairness-aware learning [66].

> 💼 Practitioners should carefully select fairness interventions based on the specific machine learning task and model characteristics. Techniques like *Mutation Testing* and *MinMax*

Scaling often improve fairness, but their effectiveness varies. Therefore, it should be integrated into the model selection and evaluation process rather than applied as a one-size-fits-all solution.

5.3. Differences in Sensitive Features

As an additional aspect, fairness practices vary by sensitive attribute. In COMPAS and Adult, gains were greater for Gender than Race, while in German Credit, improvements were more balanced across Age and Gender. Certain metric-attribute pairs, e.g., FDRD for Gender or DI for Race, consistently performed better. Strong fairness gains for Race in Communities and Crime highlight dataset-specific biases. These findings align with prior work [17, 22] and stress the need for attribute- and dataset-specific interventions.

Fairness intervention effectiveness depends on the sensitive attribute, requiring careful strategy selection. 💼 Practitioners should identify the most bias-prone attributes in their data, while 💡 Researchers should explore how metrics interact with these attributes to design targeted, context-aware solutions.

5.4. Toward a Context-Specific Fairness-Aware Recommender

Significant differences across datasets show that fairness in machine learning is not a one-size-fits-all issue but a complex challenge shaped by technical and social factors. Interventions depend on the model, dataset, sensitive attributes, and societal influences like data biases and historical inequalities. Future research should develop adaptive frameworks that tailor interventions to specific datasets within specific contexts, balancing performance with social expectations and addressing the complexity of fairness metrics.

To support this, we propose a preliminary framework to guide the selection of bias-mitigation strategies. It offers a structured view of the fairness-aware practices we evaluated, along with their fairness and performance metrics. Powered by a dataset covering various ML application domains, tasks, and sensitive attributes (see Section 3), and built on our experimental results (Section 4), the tool allows users to specify their domain and receive tailored suggestions.

Recommendations are presented as *Best Practices* and *Worst Practices*, alongside a graphical visualization to support data-driven decisions. By grounding suggestions in empirical evidence, the tool assists practitioners and lays the groundwork for expanding fairness-aware recommendation systems. An executable version is available in the online appendix [51].

The proposed tool bridges research and practice. 💼 For practitioners, it offers guidance on selecting fairness-aware methods suited to specific ML application domains. 💡 For researchers, it facilitates interpretation of results and supports exploration of interventions across settings. By combining empirical evidence with user-driven recommendations, this tool provides the basis for the development of fairness-aware recommender systems.

6. Threats to Validity

This section discusses potential threats to the validity of our empirical study and the strategies implemented to mitigate them.

Internal Validity. Internal validity concerns whether our results genuinely reflect the factors under study. One of the principal threats in this regard is the specific implementation choices made when applying fairness-aware practices. To counter this, we conducted a thorough examination of existing definitions of fairness-aware practices [70] and ensured that our implementation decisions were based on the original design of the cataloged practices. The selection of fairness metrics and performance measures can introduce biases in the evaluation process, as different metrics may lead to varying interpretations of fairness and trade-offs in performance. To mitigate this risk, we adopted a diverse set of metrics [2, 15, 17, 42, 53, 63, 72], aligned with previous research [52]. Furthermore, also reliance on a limited number of ML models could impact the results. To address this, we compared multiple models, including Random Forest, Logistic Regression, K-means, K-center, and Decision Tree [25]. Nonetheless, we acknowledge that alternative implementation choices could produce different results, influencing both the fairness and performance outcomes. An additional threat concerns the exclusion of certain practices due to their limited tool support or high implementation complexity. Although our selection prioritized scalability and reproducibility, we recognize that some excluded practices might yield different outcomes. Their evaluation remains an important direction for our future research agenda.

External Validity. External validity pertains to the generalizability of our findings beyond the study's specific setup. To enhance generalizability, we selected diverse datasets covering different application domains [25] and that are frequently utilized in fairness-related investigations [17, 22, 42], various ML tasks (classification, clustering, anomaly detection, and regression) [25], and different protected attributes. Nonetheless, our experimentation may not cover all possible contexts, and could studies are needed to validate the broader applicability of our findings. To support replication and further research, all data and scripts are publicly accessible through our online appendix [51].

Construct Validity. Construct validity reflects how well the study's measurements align with the constructs being evaluated. One potential threat is the selection of datasets to represent different contexts. To address this, we selected widely used datasets [25] that are pertinent to our focus on fairness-performance trade-offs [13, 17, 22, 42]. Another crucial consideration is the choice of fairness metrics and performance metrics. In particular, our selection is based on different metrics for each specific ML task, which are well-established within the literature and serve as robust measures of fairness [2, 15, 17, 42, 53, 63, 72]. Additionally, the choice of ML models could influence the results. To ensure reliability, we employed different models that are common in fairness research [17, 25].

Conclusion Validity. Conclusion validity refers to the reliability of the inferences we draw. One major threat is the use of statistical tests to determine the significance of fairness improvements. Specifically, our study uses the Wilcoxon signed-rank test [78] to assess statistical significance. This test assumes certain data distribution characteristics, and any violation of these assumptions could compromise the reliability of our results. To address this issue, we evaluated the data distribution using the Shapiro-Wilk test [30] to check for normality, ensuring we selected the most appropriate test for reliable conclusions.

7. Conclusion and Future Work

We extend prior work by empirically evaluating fairness-aware ML practices across high-stakes domains, examining their effectiveness across tasks, datasets, and sensitive attributes. Results show that impact varies: some practices significantly improve fairness in certain settings but not others, highlighting the need for context-specific approaches. Through cost-effectiveness analysis, we highlight trade-offs between fairness gains and performance loss, offering actionable recommendations to help practitioners balance both. These findings lay the foundation for future work, including broader experiments across more fairness-aware practices and datasets, and the design of tools to support the application of fairness-aware practices in diverse ML scenarios.

Credits

Alessandra Parziale: Formal analysis, Investigation, Data Curation, Validation, Writing - Original Draft, Visualization. **Gianmario Voria**: Formal analysis, Investigation, Data Curation, Validation, Writing - Original Draft, Visualization. **Giammaria Giordano**: Supervision, Validation, Writing - Review & Editing. **Gemma Catolino**: Supervision, Validation, Writing - Review & Editing. **Gregorio Robles**: Supervision, Validation, Writing - Review & Editing. **Fabio Palomba**: Supervision, Validation, Writing - Review & Editing.

Conflict of interest

The authors declare that they have no conflict of interest.

Data Availability

The data collected in the context of this research, along with the scripts and the results of the experiments, are publicly available in our online appendix [51].

Acknowledgement

We acknowledge the use of ChatGPT-4 to ensure linguistic accuracy and enhance the readability of this article. We acknowledge the support of the European Union - NextGenerationEU through the Italian Ministry of University and Research, Project PRIN 2022 PNRR “FRINGE: context-aware FaiRness engineerING in complex software systEms” (grant n. P2022553SL, CUP: D53D23017340001), Project PRIN 2022 “QualAI: Continuous Quality Improvement of AI-based Systems” (grant n. 2022B3BP5S, CUP: H53D23003510006), Project FAIR (PE0000013) under the NRRP MUR program funded by the EU - NGEU, the Spanish Ministry of Science, Innovation, and Universities under the Excellence Network AI4Software (Red2022-134647-T) and the Dependentium project (PID2022-139551NB-I00).

References

[1] [n. d.]. 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.

[2] Savitha Sam Abraham, Deepak P, and Sowmya S Sundaram. 2020. Fairness in Clustering with Multiple Sensitive Attributes. arXiv:1910.05113 [cs.LG] https://arxiv.org/abs/1910.05113

[3] Andrea Arcuri and Lionel Briand. 2011. A practical guide for using statistical tests to assess randomized algorithms in software engineering. In *Proceedings of the 33rd International Conference on Software Engineering* (Waikiki, Honolulu, HI, USA) *(ICSE '11)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/1985793.1985795

[4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. arXiv:1706.02409 [cs.LG] https://arxiv.org/abs/1706.02409

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. 2019. Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 453–459. https://doi.org/10.1145/3306618.3314234

[7] Arpita Biswas and Suvam Mukherjee. 2019. Fairness Through the Lens of Proportional Equality. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1832–1834.

[8] S. Biswas and H. Rajan. [n. d.]. Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline, Spinellis D. (Ed.). *ESEC/FSE 2021 - Proceedings of the 29th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering* ([n. d.]). https://doi.org/10.1145/3468264.3468536

[9] Yuriy Brun and Alexandra Meliou. [n. d.]. Software fairness. In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*.

[10] Andriy Burkov. 2020. *Machine learning engineering*. Vol. 1. True Positive Incorporated.

[11] Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* 21 (2010), 277–292.

[12] Simon Caton, Saiteja Malisetty, and Christian Haas. 2022. Impact of Imputation Strategies on Fairness in Machine Learning. *J. Artif. Int. Res.* 74 (Sept. 2022), 25 pages. https://doi.org/10.1613/jair.1.13197

[13] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. [n. d.]. Bias in machine learning software: why? how? what to do?. In *Proceedings of the 29th ACM ESEC/FSE*.

[14] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: a way to build fair ML software. In *Proceedings of the 28th ACM ESEC/FSE*. 654–665.

[15] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surv.* 41, 3, Article 15 (July 2009), 58 pages. https://doi.org/10.1145/1541880.1541882

[16] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 339–348. https://doi.org/10.1145/3287560.3287594

[17] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. [n. d.]. Fairness Improvement with Multiple Protected Attributes: How Far Are We?. In *Proceedings of the IEEE/ACM 46th ICSE*.

[18] Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 1–30.

[19] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. 2021. An overview of fairness in clustering. *IEEE Access* 9 (2021), 130698–130720.

[20] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 5036–5044.

[21] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (April 2020), 82–89. https://doi.org/10.1145/3376898

[22] Vincenzo De Martino, Gianmario Voria, Ciro Troiano, Gemma Catolino, and Fabio Palomba. [n. d.]. Examining

the Impact of Bias Mitigation Algorithms on the Sustainability of Ml-Enabled Systems: A Benchmark Study. *Available at SSRN 4966447* ([n. d.]).

[23] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM. https://doi.org/10.1145/3531146.3533113

[24] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[25] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. [n. d.]. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 ([n. d.]). https://doi.org/10.1007/s10618-022-00854-z

[26] Carmine Ferrara, Francesco Casillo, Carmine Gravino, Andrea De Lucia, and Fabio Palomba. [n. d.]. ReFAIR: Toward a Context-Aware Recommender for Fairness Requirements Engineering. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*.

[27] Carmine Ferrara, Giulia Sellitto, Filomena Ferrucci, Fabio Palomba, and Andrea Lucia. 2023. Fairness-aware machine learning engineering: how far are we? *Empirical Software Engineering* 29 (11 2023). https://doi.org/10.1007/s10664-023-10402-y

[28] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. [n. d.]. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th FSE*.

[29] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR, 22–34.

[30] Elizabeth González-Estrada and Waldenia Cosmes. 2019. Shapiro–Wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation* 89, 17 (2019), 3258–3272.

[31] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[32] Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. [n. d.]. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* 1, 2 ([n. d.]).

[33] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2021. 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* (Athens, Greece) *(ESEC/FSE 2021)*. Association for Computing Machinery, New York, NY, USA, 994–1006. https://doi.org/10.1145/3468264.3468565

[34] A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. *ACM Comput. Surv.* 31, 3 (Sept. 1999), 264–323. https://doi.org/10.1145/331499.331504

[35] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* (2012).

[36] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Montreal, Canada, 1137–1145.

[37] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), e1452.

[38] Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 699, 13 pages. https://doi.org/10.1145/3411764.3445261

[39] P. Ma, S. Wang, and J. Liu. 2020. Metamorphic testing and certified mitigation of fairness violations in NLP models, Bessiere C. (Ed.). *IJCAI International Joint Conference on Artificial Intelligence* 2021-January (2020), 458–465. cited By 31.

[40] Guillermo Macbeth, Eugenia Razumiejczyk, and Rubén Daniel Ledesma. 2011. Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations. *Universitas Psychologica* 10, 2 (2011), 545–555.

[41] Yasir Mahmood, Nazri Kama, Azri Azmi, Ahmad Salman Khan, and Mazlan Ali. 2021. Software effort estimation accuracy prediction of machine learning techniques: A systematic performance evaluation. *Software Prac. Experience* 52 (06 2021). https://doi.org/10.1002/spe.3009

[42] Suvodeep Majumder, Joymallya Chakraborty, Gina R Bai, Kathryn T Stolee, and Tim Menzies. [n. d.]. Fair enough: Searching for sufficient measures of fairness. *ACM Transactions on Software Engineering and Methodology* ([n. d.]).

[43] Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz,

Anna Maria Vollmer, and Stefan Wagner. 2022. Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 31, 2 (2022), 1–59.

[44] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[45] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31.

[46] Fatemeh Nargesian, Abolfazl Asudeh, and H. V. Jagadish. 2022. Responsible Data Integration: Next-generation Challenges. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA) *(SIGMOD '22)*. Association for Computing Machinery, New York, NY, USA, 2458–2464. https://doi.org/10.1145/3514221.3522567

[47] ABC News. 2009. *Amazon restores rankings for gay-themed books*. https://abcnews.go.com/Technology/story?id=7343222&page=1

[48] BBC News. 2018. *Google AI project taken down after ethics outcry*. https://www.bbc.com/news/technology-45809919 Accessed: 2025-03-10.

[49] Jianjun Ni, Yinan Chen, Yan Chen, Jinxiu Zhu, Deena Ali, and Weidong Cao. 2020. A survey on theories and applications for self-driving cars based on deep learning methods. *Applied Sciences* 10, 8 (2020), 2749.

[50] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. [n. d.]. 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 ([n. d.]).

[51] Alessandra Parziale, Gianmario Voria, Giammaria Giordano, Gemma Catolino, Gregorio Robles, and Fabio Palomba. [n. d.]. Online Appendix. https://figshare.com/s/20a84f94dff6fe7cd4b8

[52] Alessandra Parziale, Gianmario Voria, Giammaria Giordano, Gemma Catolino, Gregorio Robles, and Fabio Palomba. 2025. Contextual Fairness-Aware Practices in ML: A Cost-Effective Empirical Evaluation. *arXiv preprint arXiv:2503.15622* (2025).

[53] Iva Pauletic, Lucia Nacinovic Prskalo, and Marija Brkic Bakaric. 2019. An Overview of Clustering Models with an Application to Document Clustering. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 1659–1664. https://doi.org/10.23919/MIPRO.2019.8756868

[54] Dana Pessach and Erez Shmueli. [n. d.]. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* ([n. d.]).

[55] Bozidar Radunovic and Jean-Yves Le Boudec. 2007. A Unified Framework for Max-Min and Min-Max Fairness With Applications. *IEEE/ACM Transactions on Networking* 15, 5 (2007), 1073–1083. https://doi.org/10.1109/TNET.2007.896231

[56] Paul Ralph, Sebastian Baltes, Domenico Bianculli, Yvonne Dittrich, Michael Felderer, Robert Feldt, Antonio Filieri, Carlo Alberto Furia, Daniel Graziotin, Pinjia He, Rashina Hoda, Natalia Juristo, Barbara A. Kitchenham, Romain Robbes, Daniel Méndez, Jefferson Seide Molléri, Diomidis Spinellis, Miroslaw Staron, Klaas-Jan Stol, Damian A. Tamburri, Marco Torchiano, Christoph Treude, Burak Turhan, and Sira Vegas. 2020. ACM SIGSOFT Empirical Standards. *CoRR* abs/2010.03525 (2020). arXiv:2010.03525 https://arxiv.org/abs/2010.03525

[57] S. Raza, D.J. Reji, and C. Ding. 2022. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics* (2022). https://doi.org/10.1007/s41060-022-00359-4 cited By 0.

[58] Roozbeh Razavi-Far, Maryam Farajzadeh-Zanajni, Boyu Wang, Mehrdad Saif, and Shiladitya Chakrabarti. 2021. Imputation-Based Ensemble Techniques for Class Imbalance Learning. *IEEE Transactions on Knowledge and Data Engineering* 33, 5 (2021), 1988–2001. https://doi.org/10.1109/TKDE.2019.2951556

[59] Michael Redmond. 2002. Communities and Crime. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53W3X.

[60] Stephanie Riegg Cellini and James Edwin Kee. 2015. Cost-effectiveness and cost-benefit analysis. *Handbook of practical program evaluation* (2015), 636–672.

[61] Mohammed Shantal, Zalinda Othman, and Azuraliza Abu Bakar. 2023. A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization. *Symmetry* 15, 12 (2023). https://doi.org/10.3390/sym15122185

[62] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. [n. d.]. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. arXiv:2103.12016 [cs.HC]

[63] Daniel Steinberg, Alistair Reid, and Simon O'Callaghan. 2020. Fairness Measures for Regression via Probabilistic Classification. arXiv:2001.06089 [cs.LG] https://arxiv.org/abs/2001.06089

[64] The New York Times. 2015. *Google Photos Mistakenly Labels Black People 'Gorillas'*. https://archive.nytimes.com/bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/ Accessed: 2025-03-10.

[65] The New York Times. 2021. *Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men*. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

[66] Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. 2022. Fairness-aware configuration of machine learning libraries. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) *(ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 909–920. https://doi.org/10.1145/3510003.3510202

[67] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. [n. d.]. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*.

[68] I. Valentim, N. Lourenco, and N. Antunes. [n. d.]. The Impact of Data Preparation on the Fairness of Software Systems. *Proceedings - International Symposium on Software Reliability Engineering, ISSRE* ([n. d.]). https://doi.org/10.1109/ISSRE.2019.00046

[69] M. Vega-Gonzalo and P. Christidis. 2022. Fair Models for Impartial Policies: Controlling Algorithmic Bias in Transport Behavioural Modelling. *Sustainability (Switzerland)* 14, 14 (2022). https://doi.org/10.3390/su14148416 cited By 0.

[70] Gianmario Voria, Giulia Sellitto, Carmine Ferrara, Francesco Abate, Andrea De Lucia, Filomena Ferrucci, Gemma Catolino, and Fabio Palomba. 2024. A Catalog of Fairness-Aware Practices in Machine Learning Engineering. *arXiv preprint arXiv:2408.16683* (2024).

[71] Gianmario Voria, Giulia Sellitto, Carmine Ferrara, Francesco Abate, Andrea De Lucia, Filomena Ferrucci, Gemma Catolino, and Fabio Palomba. 2025. Fairness-aware practices from developers' perspective: A survey. *Information and Software Technology* (2025), 107710.

[72] Zeljko Vujovic. 2021. Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications* Volume 12 (07 2021), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670

[73] Pin Wang, En Fan, and Peng Wang. 2021. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern recognition letters* 141 (2021), 61–67.

[74] Xiaomeng Wang, Yishi Zhang, and Ruilin Zhu. 2022. A brief review on algorithmic fairness. *Management System Engineering* 1, 1 (2022), 7.

[75] Mengyi Wei and Zhixuan Zhou. 2022. AI Ethics Issues in Real World: Evidence from AI Incident Database. arXiv:2206.07635 [cs.AI]

[76] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, Anders Wesslén, et al. 2012. *Experimentation in software engineering*. Vol. 236. Springer.

[77] Jacky Wong. 2015. *Computers Are Showing Their Biases, and Tech Firms Are Concerned*. https://www.wsj.com/articles/computers-are-showing-their-biases-and-tech-firms-are-concerned-1440102894 Accessed: 2025-03-10.

[78] Robert F Woolson. 2005. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics* 8 (2005).

[79] Junjie Wu, Hui Xiong, and Jian Chen. 2009. Adapting the right measures for K-means clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Paris, France) *(KDD '09)*. Association for Computing Machinery, New York, NY, USA, 877–886. https://doi.org/10.1145/1557019.1557115

[80] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. [n. d.]. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

[81] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation testing-based discrimination discovery: a causal inference approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2718–2724.

[82] Mengdi Zhang and Jun Sun. 2022. 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* (Singapore, Singapore) *(ESEC/FSE 2022)*. Association for Computing Machinery, New York, NY, USA, 6–17. https://doi.org/10.1145/3540250.3549103

[83] Yan Zhang. 2018. Assessing fair lending risks using race/ethnicity proxies. *Management Science* 64, 1 (2018), 178–197.

[84] Yi Zhang, Weixuan Liang, Xinwang Liu, Sisi Dai, Siwei Wang, Liyang Xu, and En Zhu. 2022. Sample Weighted Multiple Kernel K-means via Min-Max optimization. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 1679–1687. https://doi.org/10.1145/3503161.3547917

[85] Jianlong Zhou and Fang Chen. 2018. *Human and Machine Learning*. Springer.