Benchmarking Ranking Frameworks for Algorithmic Fairness

Gabriel Garcia Almeida[1] **and Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brasil

{gabriel.almeida12,lberton}@unifesp.br

Abstract. *Evaluating algorithmic fairness in machine learning pipelines presents a complex multi-criteria optimization challenge, often characterized by severe trade-offs between predictive utility and disparate impact mitigation. While numerous mathematical metrics and mitigation frameworks exist, systematically identifying the optimal configuration for a given operational context remains an open problem. This paper investigates the viability of ranking-based decision-support frameworks as a reliable paradigm for auditing and selecting fairness-aware machine learning models. We systematically evaluate three distinct multi-criteria aggregation paradigms, Multi-Criteria Performance Measure (MCPM), Multi-Objective Evaluation (MOE), and non-linear Borda-style ranking frameworks, across four benchmark datasets under varying sensitive attributes. Our empirical results reveal critical trade-offs between paradigm-specific stability and sensitivity, demonstrating that model selection is highly non-trivial and acutely sensitive to both the choice of evaluation framework and the baseline bias topology of the underlying data.*

1. Introduction

Machine learning (ML) models have fundamentally transformed high-stakes socio-economic domains, ranging from credit scoring and automated recruitment to predictive healthcare and recidivism risk assessment. Although these data-driven architectures routinely match or exceed human predictive performance, they risk codifying, amplifying, or institutionalizing historical and systemic biases present within empirical training distributions [Barocas and Selbst 2016, dos Santos Porphirio et al. 2025, Silva and Berton 2025]. When algorithms learn from historical data that reflects societal prejudices or structural inequalities, they systematically reproduce these disparities, leading to discriminatory outcomes for marginalized demographics defined by protected attributes such as sex, race, or age.

To formalize ethical constraints within statistical learning, the ML community has proposed a diverse taxonomy of fairness metrics alongside algorithmic interventions designed to mitigate disparate impact. These interventions span the entire machine learning pipeline, falling into pre-processing, in-processing, and post-processing methodologies [Rabonato and Berton 2025, Pérez 2025]. However, this proliferation of metrics and mitigation strategies introduces a secondary critical challenge: the model selection dilemma. Practitioners face a bewildering hyperparameter space where optimizing for one metric (e.g., demographic parity) frequently degrades another (e.g., equalized odds) or drastically reduces predictive utility (e.g., F1-score) [Fatoretto et al. 2025].

Consequently, establishing a principled framework to evaluate, rank, and select the optimal model configuration is paramount. This paper provides a comprehensive, comparative audit of multi-criteria ranking frameworks designed to solve this selection bottleneck. We evaluate how robust these decision-support frameworks are when tasked with balancing accuracy and parity across heterogeneous datasets, mapping out the boundaries where ranking paradigms remain consistent, and where they diverge under the pressure of conflicting optimization objectives.

2. Related Work

2.1. Fairness Mitigation Approaches

Algorithmic interventions for bias reduction are fundamentally categorized by their point of execution within the machine learning lifecycle [Pérez 2025].

2.1.1. Pre-Processing Fairness Mitigation

Pre-processing interventions modify the empirical data distribution prior to model training, seeking to minimize structural bias such that any downstream standard estimator inherits a fair feature space [Caton and Haas 2024].

Suppression (Fairness Through Unawareness): This baseline strategy removes explicit protected attributes A from the feature set X. The underlying design assumption is that an estimator $f(X)$ is fair if it lacks access to sensitive variables [Kusner et al. 2017]. Formally:

$$f(X \setminus \{A\}) \perp A$$

However, this approach fails significantly in practice due to the presence of redundant encodings and proxy variables within the remaining feature space (e.g., geographical zip codes correlating strongly with race), enabling downstream models to implicitly reconstruct the protected attribute.

Instance Reweighing: Proposed by Kamiran and Calders [Kamiran and Calders 2011], this technique assigns non-uniform importance weights to samples within the training set to counteract statistical imbalances. For a binary class label $Y \in \{0, 1\}$ and a binary sensitive attribute $S \in \{a, b\}$, the weight $W(s, c)$ for a sample with sensitive status s and class label c in a dataset D is formally defined as:

$$W(s,c) = \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(\text{Class}) = c\}|}{|D| \times |\{X \in D \mid X(S) = s \text{ and } X(\text{Class}) = c\}|}$$

Disparate Impact Removal: This geometric approach modifies the marginal distributions of continuous features to eliminate correlation with protected attributes while preserving the rank-order utility of the predictors [Feldman et al. 2014]. This is achieved by mapping group-specific cumulative distribution functions (CDFs) to a baseline median distribution via quantile matching. The intensity of this transformation is governed by a user-defined repair fraction $\lambda \in [0, 1]$ [MathWorks 2024]:

$$x' = \lambda F_m^{-1}(F_g(x)) + (1 - \lambda)x$$

where F_g represents the group-specific CDF and F_m^{-1} denotes the inverse quantile function of the median distribution.

Synthetic Data Generation: Oversampling frameworks generate synthetic instances to balance historical representation disparities within under-represented subgroups [Trinh and Zhang 2024].

- **SMOTE** (Synthetic Minority Over-sampling Technique) synthesizes continuous feature vectors by interpolating between a target minority sample and its k-nearest neighbors within the same class [Chawla et al. 2002].
- **SMOTE-NC** (Nominal and Continuous) generalizes this framework to mixed-type feature spaces by integrating an adjusted distance metric that penalizes categorical mismatches and assigning nominal attributes based on local modal frequencies [Chawla et al. 2002].

2.1.2. In-Processing Fairness Mitigation

In-processing methods inject mathematical fairness constraints directly into the optimization objective during the training phase [Barocas et al. 2023].

Exponentiated Gradient: This paradigm frames fair classification as a zero-sum game between a learner and a regulator, mapping the constrained optimization problem into a sequence of unconstrained, cost-sensitive learning steps [Agarwal et al. 2018]. By dynamically updating Lagrange multipliers associated with linear fairness constraints (e.g., demographic parity bounds), the algorithm yields a randomized ensemble of classifiers that optimally balances empirical risk against the target fairness threshold.

2.1.3. Post-Processing Fairness Mitigation

Post-processing strategies modify the output space of a frozen, pre-trained model $\hat{Y} = f(X)$, offering high utility for proprietary or computationally expensive black-box architectures.

Threshold Optimization: This technique replaces global decision boundaries with group-specific thresholds t_g. Predictions are mapped as $\hat{Y} = 1 \iff R \geq t_g$, where R is the continuous model score, and t_g is optimized to satisfy parity constraints such as equal opportunity [Hardt et al. 2016].

Calibration (Isotonic Regression): This alignment ensures that risk scores reflect identical empirical event rates across all demographic strata [Leeuw et al. 2009]. It applies non-parametric monotonic transformations to group-wise probabilities to prevent systematic risk misestimation.

Reject Option Classification (ROC): ROC selectively overrides classifications within a designated critical uncertainty region around the global decision boundary, $[\tau_{\text{low}}, \tau_{\text{high}}]$. Instances falling inside this margin are assigned favorable outcomes if they belong to unprivileged groups, and unfavorable outcomes if they belong to privileged groups [Kamiran et al. 2012].

2.2. Evaluation Space and Metrics

2.2.1. Predictive Performance Metrics

Statistical classification performance is computed using standard confusion matrix components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Metrics include Accuracy, Precision, Recall, F1-Score, Area Under the ROC Curve (AUC), and the Kolmogorov-Smirnov (KS) statistic.

2.2.2. Algorithmic Fairness Metrics

Let $G \in \{a, b\}$ denote demographic group membership, $\hat{Y} \in \{0, 1\}$ the model prediction, and $Y \in \{0, 1\}$ the ground truth. The primary parity formulations are defined as follows:

- **Demographic Parity Difference (Δ_{DP}):**

$$\Delta_{DP} = |P(\hat{Y} = 1 \mid G = a) - P(\hat{Y} = 1 \mid G = b)|$$

- **Equalized Odds Difference (Δ_{EO}):**

$$\Delta_{EO} = \max(|TPR_a - TPR_b|, |FPR_a - FPR_b|)$$

- **Equal Opportunity Difference (Δ_{EOpp}):**

$$\Delta_{EOpp} = |P(\hat{Y} = 1 \mid Y = 1, G = a) - P(\hat{Y} = 1 \mid Y = 1, G = b)|$$

- **False Positive Rate Difference (Δ_{FPR}):**

$$\Delta_{FPR} = |P(\hat{Y} = 1 \mid Y = 0, G = a) - P(\hat{Y} = 1 \mid Y = 0, G = b)|$$

- **False Negative Rate Difference (Δ_{FNR}):**

$$\Delta_{FNR} = |P(\hat{Y} = 0 \mid Y = 1, G = a) - P(\hat{Y} = 0 \mid Y = 1, G = b)|$$

- **Predictive Parity Difference (Δ_{PP}):**

$$\Delta_{PP} = |P(Y = 1 \mid \hat{Y} = 1, G = a) - P(Y = 1 \mid \hat{Y} = 1, G = b)|$$

2.3. Performance Criteria and Multi-Criteria Decision Frameworks

2.3.1. Multi-Criteria Performance Measure (MCPM)

MCPM geometrically aggregates three disparate metrics onto a normalized two-dimensional plane [Minatel 2024]. The metrics are represented as vectors originating from a shared origin, angularly separated by $120°$. The magnitude of each vector corresponds to the metric value. The scalar MCPM score is defined as the total geometric area of the triangle formed by connecting the vector termini:

$$\text{Area} = \frac{\sqrt{3}}{4}(M_1M_2 + M_2M_3 + M_3M_1)$$

2.3.2. Multi-Objective Evaluation (MOE)

MOE formulations explicitly model the Pareto efficiency frontier across conflicting objectives [Chia et al. 2022]. Models that are strictly dominated by competitors (i.e., performing worse across all objectives) are pruned. To resolve the non-comparability of the remaining non-dominated frontier, a linear regression model estimates the empirical trade-off trajectory of a base reference metric (m_b) relative to a secondary constraint metric (m_1). The final scalar model score is formalised as:

$$S_p(m_b, m_1) = (1 - w_1) \cdot m_b - w_1 \cdot EV_b^1(m_1)$$

where $EV_b^1(m_1)$ represents the expected value of the reference metric mapped from the secondary metric's distribution, and $w_1 \in [0, 1]$ controls the relative optimization weight.

3. Methodology

3.1. Datasets

Four standard, real-world credit risk and financial demographic datasets were utilized to audit the performance of the ranking algorithms. Table 1 outlines the structural dimensionality of each space before pipeline preprocessing.

Table 1. Dimensionality and structural summary of baseline experimental datasets.

Dataset	Instances	Features	Favorable Label	Reference
Default of Credit Card Clients	30,000	24	Not in default	[Kim and UCI Machine Learning 2005]
Credit Card Approval Prediction	438,557	18	< 60 days past due	[MoneyMan 2019]
Credit Card Customers	10,127	22	Account remains open	[Goyal 2020]
Statlog German Credit	1,000	20	Good Credit	[Hofmann 1994]

3.1.1. Default of Credit Card Clients

This dataset captures credit default histories in Taiwan [Kim and UCI Machine Learning 2005]. Preprocessing followed an institutional pipeline: continuous financial features (`BILL_AMT`, `PAY_AMT`, `LIMIT_BAL`) were scaled using standard z-score normalization. Marital variables were mapped to a binary flag `is_married`. Sub-populations older than 60 were pruned to reduce extreme variance, and the remaining individuals were binned into *young* (< 30) and *adult* (≥ 30). Sex was operationalized as a binary dummy variable `sex_male`.

3.1.2. Credit Card Approval Prediction

Constructed via an inner-join between demographic records and historical payment ledgers [MoneyMan 2019]. Target definition established a binary threshold: favorable outcomes (0) required payments to be active or less than 60 days past due, whereas unfavorable outcomes (1) marked profiles with structural default delays exceeding 60 days. Age vectors derived from continuous day counts were pruned and discretized using identical bounds to the Default dataset.

3.1.3. Credit Card Customers

This dataset targets bank customer churn dynamics [Goyal 2020]. Features containing structural leakage profiles (e.g., downstream Naive Bayes class estimations) were systematically dropped. The target target label mapped active profiles to 0 and churned states to 1.

3.1.4. Statlog German Credit

A classic low-sample financial benchmark dataset [Hofmann 1994]. The target risk label maps good credit to 0 and bad credit risks to 1. Complex categorical relationship matrices (e.g., Personal Status attributes) were disentangled to generate decoupled binary indicators for biological sex and age brackets.

3.2. Experimental Setup

3.2.1. Data Splitting and Pipelines

All computational models were developed using Python's scientific ecosystem (`scikit-learn`, `pandas`, `xgboost`). To guarantee exact statistical replicability across multi-stage pipelines, a global seed of $\psi = 19$ was locked. Feature allocation followed a stratified $70/30$ train-test allocation split. Continuous inputs passed through a standard scaling transformer, while categorical matrices were vectorized using one-hot encoding blocks with first-element drop flags enabled. Biological sex (`sex_male`) and standardized age brackets (*adult* vs. *young*) were optimized as individual sensitive coordinates independently; when evaluating sex-based parity, age indicators were completely omitted from the network, and vice versa.

3.2.2. Artificial Data Biasing

To evaluate the resiliency of the downstream mitigation mechanisms under adversarial or worst-case environments, data biasing was introduced on the training partitions of the *Default of Credit Card Clients* and *Credit Card Approval Prediction* domains. Subpopulation imbalances were synthetically amplified using `SMOTENC` to intentionally skew baseline target label distributions across unprivileged groups, testing the limits of the mitigation frameworks.

3.2.3. Base Learners and Space Optimization

Model hyperparameter tuning was conducted via a systematic 5-fold stratified cross-validation grid search optimizing for the Macro F1-Score. Estimators without native continuous probability calibration (e.g., standard Perceptrons) were mapped through a Platt scaling sigmoid calibrator via `CalibratedClassifierCV` to generate valid probability score distributions. Base estimators included Ridge Classification, Perceptron models, Random Forests, and Gradient Boosted Trees (`XGBoost`). Table 2 outlines the grid search boundaries.

Table 2. Model Hyperparameters and Structural Search Spaces

Model Class	Hyperparameter Group	Grid Search Values
Ridge (L2)	Regularization Strength (C)	0.1, 1, 10
Perceptron	Penalty; Alpha (α); Learning Rate (η_0)	{None, l1, l2}; {0.001, 0.01}; {0.1, 0.01}
Random Forest	Estimators (N); Maximum Depth (D)	{100, 200}; {3, 10}
XGBoost	Learning Rate (η); Max Depth (D)	{0.03, 0.05, 0.1}; {3, 5}

3.3. Ranking Paradigms and Selection Strategies

We audited model selection across three independent ranking topologies applied over 8 experimental variations (4 datasets $\times$ 2 distinct sensitive vectors). Fairness metrics were inverted via $1 - |\Delta_{\text{Metric}}|$ to yield a monotonic positive structure where a value of 1.0 indicates perfect statistical parity.

3.3.1. MCPM Configurations

- **Configuration A (M-1):** Synthesized macro F1-Score, Equalized Odds Difference (Δ_{EO}), and Demographic Parity Difference (Δ_{DP}).
- **Configuration B (M-2):** Synthesized classification Accuracy, Macro F1-Score, and Equal Opportunity Difference (Δ_{EOpp}).

3.3.2. MOE Pareto Formulations

- **Configuration A (E-1):** Paired primary performance (Macro F1) against structural Equalized Odds (Δ_{EO}).
- **Configuration B (E-2):** Paired crude performance (Accuracy) against continuous Demographic Parity (Δ_{DP}).

Feasibility bounding filters enforced operational utility: models were systematically pruned if they failed to maintain $\geq 90\%$ of the unmitigated baseline's predictive performance, or if they expanded baseline demographic disparity errors by $> 20\%$.

3.3.3. Non-Linear Borda Ranking Schemes

Models were ordered based on individual performance dimensions. Ordinal points incremented non-linearly based on relative performance placement within the model pool. Total scores were aggregated across performance and fairness coordinates to select the highest-scoring model.

- **Test 1 (R-1):** Ranked using F1-Score, Δ_{DP}, and Δ_{EO}.
- **Test 2 (R-2):** Ranked using Accuracy, F1-Score, and Δ_{EOpp}.

4. Empirical Results and Comparative Discussion

4.1. Multi-Criteria Performance Measure Evaluation

Table 3 summarizes the top-performing model pipelines selected via the MCPM Test 1 criteria.

Table 3. Top performing models per scenario: MCPM Test 1 (Metrics: F1, EOD, and DPD)

Dataset	SMT	Var.	Model	Technique	F1 ↑	EOD ↑	DPD ↑	MCPM ↑
Default Credit	Yes	Age	Random Forest	Suppression	0.522	0.986	0.990	**1.738**
Default Credit	Yes	Sex	Logistic Reg.	Thresh. Opt. EO	0.514	0.993	0.996	**1.742**
Credit Approval	Yes	Age	XGBoost	Thresh. Opt. EO	0.045	0.975	0.993	**0.914**
Credit Approval	Yes	Sex	Logistic Reg.	Inst. Reweighing	0.087	0.985	1.000	**1.002**
Credit Customers	No	Age	XGBoost	Exp. Gradient DP	0.900	0.928	0.982	**2.278**
Credit Customers	No	Sex	XGBoost	None (Baseline)	0.898	0.989	0.995	**2.394**
German Credit	No	Age	XGBoost	Thresh. Opt. EO	0.594	0.952	0.998	**1.827**
German Credit	No	Sex	Logistic Reg.	None (Baseline)	0.597	0.930	0.998	**1.802**

The empirical results show that non-linear ensemble tree architectures (`XGBoost`) and regularized linear classifiers (`LogisticRegression`) frequently dominated the selection space. Crucially, a target dependency is revealed: within identical structural environments, altering solely the target sensitive vector requires completely distinct algorithmic models and mitigation interventions to achieve optimality. Notably, across the unskewed Credit Customers and German Credit benchmarks, unmitigated baseline classifiers (None) outperformed active fairness interventions when biological sex was optimized, suggesting that active mitigation when baseline disparity is minimal can cause unnecessary degradation in predictive performance.

Table 4 illustrates model trajectories under the secondary MCPM formulation.

Table 4. Top performing models per scenario: MCPM Test 2 (Metrics: Acc., F1, and EOpD)

Dataset	SMT	Var.	Model	Technique	Acc. ↑	F1 ↑	Eq. Opp. ↑	MCPM ↑
Default Credit	Yes	Age	Random Forest	Suppression	0.790	0.522	0.995	**1.487**
Default Credit	Yes	Sex	Random Forest	Exp. Gradient EO	0.794	0.517	0.998	**1.488**
Credit Approval	Yes	Age	Perceptron	Inst. Reweighing	0.880	0.037	0.966	**0.795**
Credit Approval	Yes	Sex	Logistic Reg.	Inst. Reweighing	0.983	0.087	0.985	**0.987**
Credit Customers	No	Age	XGBoost	Calibration	0.964	0.884	0.954	**2.264**
Credit Customers	No	Sex	XGBoost	None (Baseline)	0.968	0.898	0.989	**2.350**
German Credit	No	Age	XGBoost	Inst. Reweighing	0.777	0.595	0.968	**1.551**
German Credit	No	Sex	XGBoost	None (Baseline)	0.784	0.608	0.995	**1.611**

A comparative breakdown between the two MCPM configurations highlights a core vulnerability of geometric ranking paradigms: selection instability. When shifting from Test 1 to Test 2, model recommendations remained consistent in only two instances. In all other scenarios, modifying the underlying target evaluation metrics caused

the framework to favor entirely different algorithmic combinations, showing that geometric area maximization is highly sensitive to metric configuration.

4.2. Multi-Objective Pareto Frontiers

Table 5 isolates optimal configurations selected via Pareto trade-off optimization under Test 1.

By enforcing a strict feasibility filter, the MOE paradigm reduces multi-dimensional spaces to explicit non-dominated trade-off lines. Random Forest architectures establish dominance over low-sample spaces, while `XGBoost` captures complex multi-variate interactions across large-scale datasets.

Table 6 catalogs performance profiles under the alternative MOE constraint space (Accuracy vs. $|\Delta_{DP}|$).

Table 5. Optimal models per scenario: Multi-Objective Evaluation Test 1 (Metrics: F1 and EOD)

Dataset	SMT	Var.	Model	Technique	F1 ↑	EOD ↓	S_p Score ↑
Default Credit	Yes	Age	Random Forest	Suppression	0.522	0.014	**0.006**
Default Credit	Yes	Sex	Random Forest	Synth. Data Gen.	0.520	0.007	**0.001**
Credit Approval	Yes	Age	XGBoost	Inst. Reweighing	0.078	0.210	**0.003**
Credit Approval	Yes	Sex	XGBoost	Inst. Reweighing	0.147	0.144	**0.006**
Credit Customers	No	Age	XGBoost	Exp. Grad. DP	0.900	0.072	**0.005**
Credit Customers	No	Sex	XGBoost	None (Baseline)	0.898	0.012	**0.002**
German Credit	No	Age	XGBoost	Inst. Reweighing	0.595	0.037	**0.019**
German Credit	No	Sex	XGBoost	None (Baseline)	0.608	0.052	**0.000**

Table 6. Optimal models per scenario: Multi-Objective Evaluation Test 2 (Metrics: Acc. and $|DPD|$)

Dataset	SMT	Var.	Model	Technique	Acc. ↑	$\lvert DPD\rvert$ ↓	S_p Score ↑
Default Credit	Yes	Age	Random Forest	Exp. Grad. EO	0.789	0.005	**0.002**
Default Credit	Yes	Sex	XGBoost	Inst. Reweighing	0.800	0.000	**0.008**
Credit Approval	Yes	Age	XGBoost	Inst. Reweighing	0.869	0.009	**0.010**
Credit Approval	Yes	Sex	Logistic Reg.	Inst. Reweighing	0.983	0.000	**0.000**
Credit Customers	No	Age	XGBoost	Exp. Grad. DP	0.969	0.018	**0.002**
Credit Customers	No	Sex	XGBoost	Inst. Reweighing	0.967	0.003	**0.000**
German Credit	No	Age	Logistic Reg.	Exp. Grad. DP	0.781	0.006	**0.016**
German Credit	No	Sex	Logistic Reg.	None (Baseline)	0.784	0.002	**0.000**

Under the demographic parity constraint formulation, Instance Reweighing maps as a robust generalizer across 50% of all evaluated scenarios. In-processing via Exponentiated Gradient optimization also performs competitively, balancing accuracy constraints against structural shifts in demographic parity.

Table 7. Best performing models per scenario: Ranking Test 1 (Metrics: F1, EOD, and DPD)

Dataset	SMT	Var.	Model	Technique	F1 ↑	EOD ↑	DPD ↑	Score ↑
Default Credit	Yes	Age	Random Forest	Suppression	0.522	0.986	0.990	**123**
Default Credit	Yes	Sex	Logistic Reg.	Thresh. Opt. EO	0.514	0.993	0.996	**118**
Credit Approval	Yes	Age	Random Forest	Exp. Grad. EO	0.057	0.878	0.999	**111**
Credit Approval	Yes	Sex	Logistic Reg.	Inst. Reweighing	0.087	0.985	1.000	**128**
Credit Customers	No	Age	Random Forest	Calibration	0.861	0.943	0.996	**120**
Credit Customers	No	Sex	XGBoost	Synth. Data Gen.	0.892	0.986	0.998	**126**
German Credit	No	Age	XGBoost	Thresh. Opt. EO	0.594	0.952	0.998	**119**
German Credit	No	Sex	Logistic Reg.	None (Baseline)	0.597	0.930	0.998	**118**

Table 8. Best performing models per scenario: Ranking Test 2 (Metrics: Acc., F1, and EOpD)

Dataset	SMT	Var.	Model	Technique	Acc. ↑	F1 ↑	Eq. Opp. ↑	Score ↑
Default Credit	Yes	Age	Random Forest	Suppression	0.790	0.522	0.995	**122**
Default Credit	Yes	Sex	Random Forest	Exp. Grad. EO	0.794	0.517	0.998	**113**
Credit Approval	Yes	Age	Random Forest	Inst. Reweighing	0.849	0.059	0.891	**108**
Credit Approval	Yes	Sex	Logistic Reg.	Inst. Reweighing	0.983	0.087	0.985	**128**
Credit Customers	No	Age	XGBoost	Synth. Data Gen.	0.969	0.902	0.917	**125**
Credit Customers	No	Sex	XGBoost	None (Baseline)	0.968	0.898	0.989	**122**
German Credit	No	Age	XGBoost	Inst. Reweighing	0.777	0.595	0.968	**113**
German Credit	No	Sex	XGBoost	None (Baseline)	0.784	0.608	0.995	**128**

4.3. Borda Ordinal Ranking Analysis

Tables 7 and 8 document performance outputs mapped via independent multi-criterion ordinal point scoring structures.

The Borda ordinal framework provides a more stabilized selection distribution, insulating the evaluation from extreme outlier values in a single metric. Across Test 2, Instance Reweighing and unmitigated baselines emerge as preferred choices for balancing accuracy alongside localized group parity.

5. Conclusion and Future Horizons

This paper audited multi-criteria selection frameworks within fairness-aware machine learning pipelines. Our empirical evaluations demonstrate that no single algorithmic model or bias mitigation framework holds universal dominance. Optimal choices depend heavily on the operational dataset, the specific sensitive vectors selected, and the target evaluation metrics prioritized.

Furthermore, we highlighted a critical challenge in algorithmic auditing: model selection is highly unstable under geometric or linear aggregation frameworks. Modifying the criteria configuration can drastically shift recommended pipelines, meaning practitioners must rigorously validate their choice of selection framework. Additionally, applying mitigation techniques to datasets with minimal baseline bias (such as Credit Cus-

tomers) can degrade predictive performance without yielding significant improvements in equity. Consequently, algorithmic selection remains highly context-specific, requiring customized cross-validation workflows that match the ethical and statistical objectives of the target domain.

Acknowledgments

The authors would like to thank Serasa Experian, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) grant 2021/14725-3, for supporting the development of this study.

Declarations

During the preparation of this work, the authors utilized generative AI tool Gemini to refine linguistic clarity and academic readability. The authors reviewed and edited all outputs, and take full responsibility for the technical accuracy and content of this publication.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Chia, P. J., Attanasio, G., Tagliabue, J., Bianchi, F., Greco, C., Moreira, G. d. S. P., Eynard, D., and Husain, F. (2022). E pluribus unum: Guidelines on multi-objective evaluation of recommender systems. *Journal of the ACM (JACM)*, 37(4):1–15.

dos Santos Porphirio, J., dos Santos, D. J., Azevedo, S., and Berton, L. (2025). Analyzing data bias through data augmentation: A case study in financial data. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 950–961. SBC.

Fatoretto, M. B., Özbulak, G., Anjos, A., and Berton, L. (2025). Optimizing fairness and utility in healthcare machine learning models. *Applied Soft Computing*, 181:113426.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2014). Certifying and removing disparate impact.

Goyal, S. (2020). Credit Card customers. Kaggle. URL: https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Kamiran, F. and Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision theory for freedom of choice in fair classification: Reject option-based approach. In *2012 IEEE 12th International Conference on Data Mining (ICDM)*, pages 378–387. IEEE.

Kim, E. and UCI Machine Learning (2005). Default of Credit Card Clients Dataset. Kaggle. URL: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4066–4076.

Leeuw, J. d., Hornik, K., and Mair, P. (2009). Isotone optimization in r: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5).

MathWorks (2024). Disparate impact remover. `https://www.mathworks.com/help/stats/disparateimpactremover.html`. Acess in october 22th 2025.

Minatel, D. (2024). *O Funcionamento Diferencial dos Itens na Indução de Classificadores Mais Justos*. Tese (doutorado em ciências — ciências de computação e matemática computacional), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

MoneyMan (2019). Credit Card Approval Prediction. Kaggle. URL: https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction.

Pérez, P. (2025). *Enhancing Fairness in Machine Learning: Identifying and Mitigating Bias with a Focus on Gender Bias in Finance*. PhD thesis, Open University.

Rabonato, R. T. and Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics*, 5(3):1943–1954.

Silva, B. P. M. and Berton, L. (2025). Analyzing the trade-off between fairness and model performance in supervised learning: A case study in the mimic dataset. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 212–223. SBC.

Trinh, T. K. and Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. *Journal of Advanced Computing Systems (JACS)*, 4(2):36–49.