Mitigating Algorithmic Gender Bias in Brazilian Tech Salary Prediction

Ana Luiza Bocato da Silva[1] **and Lilian Berton**[1]

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

{bocato.ana,lberton}@unifesp.br

Abstract. *This study investigates algorithmic gender bias in salary forecasting within the Brazilian tech sector using the 2024 State of Data Brazil survey. Baseline Decision Tree and Logistic Regression models were trained to evaluate how historical wage disparities are structurally replicated. An accuracy paradox emerged: high overall female accuracy (75.10%) masked a critical recall of only 39.44% for high-earning women. This low recall nearly matches the male false-positive rate (29.21%), easily overestimating lower-earning male professionals. The "omission hypothesis" was also tested, showing that selective demographic non-disclosure by high-earning men skews training data. To fix these distortions, an in-processing algorithmic mitigation framework was implemented via the Fairlearn toolkit. The adjustment drastically reduced the Demographic Parity difference (0.29 to 0.050) and Equal Opportunity difference (0.42 to 0.043). Recall for high-earning female professionals surged significantly from 39.44% to 77.46%. The equity gains incurred a negligible global accuracy loss of less than 1% (dropping from 74.90% to 73.95%). Findings prove proactive constraints eliminate automated gender penalties in labor analytics without compromising model utility.*

1. Introduction

In recent years, the deployment of Machine Learning (ML) algorithms for Human Resource Management, talent acquisition, and salary prediction has grown exponentially [Pessach and Shmueli 2022]. Promised as objective, data-driven alternatives to human cognitive biases, these automated systems are increasingly used to shape organizational compensation structures. However, algorithmic models do not operate in a vacuum; they inherit, learn, and amplify the structural inequalities embedded within historical data [Mehrabi et al. 2021, Rabonato and Berton 2025, Wecchi and Berton 2026]. When trained on datasets reflecting societal disparities, ML models risk codifying historical gender-based wage gaps under the guise of mathematical neutrality [O'Neil 2016, Fatoretto et al. 2025].

In the context of the Brazilian technology sector, gender disparities remain stark. Despite the rapid expansion of the digital economy, women are underrepresented in technical and leadership roles, often facing a pervasive wage gap [Data Hackers 2024]. Analyzing these disparities requires robust framework tools, as standard performance metrics (e.g., overall accuracy) frequently mask severe underlying biases against marginalized subgroups [Barocas et al. 2019].

To investigate these dynamics, this study presents a comprehensive algorithmic fairness analysis of salary prediction models trained on data from the *State of Data Brazil* survey. Utilizing Decision Trees and Logistic Regression, we examine how historical gender biases are replicated in predictive workflows. Initial evaluations revealed that while professional seniority hierarchically dominates the decision boundaries, gender operates as a definitive mathematical reducer for women. Moreover, a critical Equal Opportunity gap emerged: the baseline model correctly identified only 39.44% of female professionals who truly belonged to the high-salary bracket (Recall/TPR), compared to 81.72% for male professionals. This severe disparity indicates that women's trajectories are statistically confined to lower, more predictable salary brackets. Conversely, male compensation patterns benefit from wider upward variance, which is also reflected in a high male false-positive rate (29.21% compared to only 11.58% for women), showing that borderline male profiles are far more readily classified into the high-salary bracket than equivalent female profiles. This distortion is further exacerbated by systemic data missingness, where high-earning individuals—exhibiting salary distributions that closely mirror high-income male profiles—frequently omit demographic attributes, supporting our "omission hypothesis" of non-disclosure masking the true extent of the privilege gap.

To move beyond mere bias detection, this research leverages the *Fairlearn* toolkit to implement in-processing algorithmic mitigation techniques [Bird et al. 2020]. Our baseline models demonstrated substantial unfairness, yielding a Demographic Parity (DP) difference of 0.29 and an Equal Opportunity (EO) difference of 0.42, driven primarily by high false-negative rates for high-earning women. By applying optimization constraints, we successfully mitigated these disparities, reducing DP to 0.050 and EO to 0.043. Crucially, this fairness adjustment incurred a negligible drop in overall accuracy (moving from 74.90% to 73.95%), while dramatically improving subgroup classification for women, boosting female recall for the high-salary class from 39.44% to 77.46% and correcting systematic underestimations in the confusion matrix

The contributions of this paper are threefold:

- We empirically demonstrate how standard classification models interpret professional seniority as a proxy for gender bias due to the disproportionate concentration of men in senior roles.
- We expose the "omission hypothesis", illustrating how selective demographic non-disclosure by high-income groups distorts the training data distribution.
- We validate an algorithmic mitigation framework that drastically enhances equity in salary predictions for women in tech without compromising global model utility.

2. Related Work

The intersection of algorithmic decision-making, predictive fairness, and labor economics has garnered substantial attention as automated systems increasingly mediate hiring and compensation pipelines [Pessach and Shmueli 2022]. This section situates our research within the existing literature on algorithmic gender bias, the proxy mechanics of professional seniority, and technical paradigms for bias mitigation.

Historically, algorithmic models used in human resource management have been heavily criticized for perpetuating societal biases. A seminal case study by

O'Neil [O'Neil 2016] demonstrated how automated filtering algorithms codify systemic prejudice under a veneer of mathematical objectivity. In salary prediction specifically, models train on historical compensation data that reflects entrenched structural inequalities.

Recent literature highlights that when algorithms optimize strictly for global performance metrics like overall accuracy, they systematically misclassify and underrepresent marginalized sub-groups [Barocas et al. 2019]. For instance, Mehrabi et al. [Mehrabi et al. 2021] noted that historical pay disparities cause machine learning models to treat gender not as a neutral demographic attribute, but as a statistically predictive feature for lower wages, thereby reinforcing the historical wage gap.

A recurring challenge in algorithmic fairness is the phenomenon of proxy variables—where non-protected attributes act as stand-ins for protected characteristics [Datta et al. 2015]. In our study, professional seniority (e.g., senior and management levels) heavily dictates decision boundaries. This aligns with findings by Hardt et al. [Hardt et al. 2016], who argued that fairness cannot be achieved simply by removing the protected attribute (fairness through blindness), because structural imbalances, such as the disproportionate concentration of male professionals in executive roles, allow the model to reconstruct gender distributions via career trajectories.

Furthermore, data missingness and selective non-disclosure introduce a compounding layer of bias. While standard imputation techniques treat missing values as missing at random, structural sociology suggests that demographic omission is often correlated with socio-economic privilege [Little and Rubin 2019]. Our exploration of the "omission hypothesis", where high-earning male executives selectively withhold demographic details, builds upon foundational work regarding how missing data patterns distort algorithmic training distributions and mask the true magnitude of privilege gaps.

To mathematically define and counter automated discrimination, the machine learning community has established distinct fairness criteria, most notably Demographic Parity (DP) and Equal Opportunity (EO) [Verma and Rubin 2018]. Demographic Parity requires the likelihood of a positive outcome to be identical across all groups, whereas Equal Opportunity conditions this parity on the true positive rates, ensuring equal true capability is rewarded equally [Hardt et al. 2016].

Mitigation frameworks are traditionally classified into pre-processing, in-processing, and post-processing approaches [Caton and Haas 2020]. While pre-processing alters the underlying training data distribution and post-processing modifies the model's final predictions, in-processing methods incorporate fairness constraints directly into the optimization objective function.

By leveraging the *Fairlearn* toolkit [Bird et al. 2020], which utilizes the cost-sensitive classification reduction frameworks proposed by Agarwal et al. [Agarwal et al. 2018], recent studies have demonstrated success in navigating the accuracy-fairness trade-off. Our work extends this paradigm into the Latin American technology market, demonstrating that enforcing in-processing constraints can drastically reduce DP and EO disparities while simultaneously rectifying systematic underestimations for high-earning female professionals.

3. Methodology

3.1. Dataset

The dataset analyzed in this study originates from the 5th edition of the *State of Data Brazil*, an annual survey that provides a comprehensive overview of the Brazilian data science and analytics labor market. Conducted between October 14 and December 18, 2024, the study is a collaborative effort between *Data Hackers* (the largest data community in Brazil) and *Bain & Company* (a global management consultancy).

The survey gathered responses from 5,217 professionals within the Brazilian ecosystem. The sample encompasses a diverse range of professional roles, such as Data Analysts, Data Scientists, and Data Engineers, spanning various seniority levels, including junior, mid-level, and senior professionals, as well as management positions. The complete public dataset used in this work is hosted on the Kaggle platform [Data Hackers 2024].

To ensure respondent privacy and comply with data protection standards, a rigorous anonymization process was applied by the creators. This involved the removal of outliers and specific demographic indicators that could potentially risk re-identification:

- **Geographical Aggregation:** States with low response rates (e.g., certain states in the Northern region) were aggregated and are represented solely by their broader macro-region.
- **Feature Filtering:** Certain sensitive questions or high-risk outliers were either transformed or excluded from the final public dataset.

For multi-valued (multiple-choice) questions, the data is structured across multiple columns. Each column is identified by a tuple where the first identifier corresponds to the question ID, and the second identifier represents the specific choice selected by the respondent.

The questionnaire was instrumented into eight distinct thematic sections, mapped using an alphanumeric coding system (e.g., `P3a_1` represents Section 3, Question *a*, Option *1*). The structure of the survey parts is outlined in Table 1.

Table 1. Survey Questionnaire Structure

Section	Description
Part 1	Demographic Data
Part 2	Career and Professional Background
Part 3	Challenges in Data Team Management
Part 4	General Knowledge in Data Science and Analytics
Part 5	Professional Objectives and Career Goals
Part 6	Technical Knowledge in Data Engineering (DE)
Part 7	Technical Knowledge in Data Analysis (DA)
Part 8	Technical Knowledge in Data Science (DS)

3.2. Data Preprocessing

This study utilizes the *State of Data Brazil 2024* dataset [Data Hackers 2024], which aggregates structural, demographic, and compensatory microdata from $N = 5,217$ technol-

ogy professionals. The target variable, salary, was discretized into classification brackets to establish an income threshold.

To preserve institutional validity while addressing systematic data missingness, we operationalized the *Omission Hypothesis*. Individuals who selected "Prefer not to inform/Others" for the gender attribute were not discarded; instead, they were preserved as a distinct demographic group to analyze if demographic non-disclosure correlates with high-income privilege. Categorical variables, such as professional role, tech stack, and state of residence, were encoded using one-hot encoding. States with low sample frequencies were aggregated into their respective geographical macro-regions to ensure anonymity and statistical robustness.

The target salary variable was binarized using a monthly income threshold of BRL 3,000 (where $Y = 1$ corresponds to earnings $\geq$ BRL $3,001$/month, and $Y = 0$ otherwise). To evaluate generalization performance and prevent data leakage, the preprocessed dataset was partitioned into training (80%) and testing (20%).

3.3. Baseline Predictive Models

To examine how standard algorithms map historical inequalities, we deployed two baseline estimators: a Decision Tree classifier and a Logistic Regression model.

- **Decision Tree:** Utilized to map the explicit logical rules generated by the model. The splitting criteria prioritize features that maximize information gain. This allows us to observe whether the protected attribute (gender) or its proxies dominate the root and high-level decision nodes.
- **Logistic Regression:** Employed to assess the mathematical magnitude and direction of each attribute via log-odds coefficients. For a feature vector X, the probability of a high-income prediction is modeled as:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

  Where a negative coefficient ($\beta_i < 0$) for a protected group explicitly isolates its role as a mathematical reducer in the prediction matrix.

Both baseline estimators were implemented using the standard default hyperparameter settings provided by the scikit-learn library, ensuring an unconstrained baseline that reflects standard, out-of-the-box algorithmic behavior before fairness interventions.

3.4. Fairness Evaluation Framework

To quantitatively audit the models, we define gender ($A \in \{0, 1\}$) as the protected attribute, where $A = 1$ represents male professionals and $A = 0$ represents female professionals. We evaluate systemic bias using two primary criteria implemented via the *Fairlearn* toolkit [Bird et al. 2020]:

1. **Demographic Parity (DP):** Measures whether the likelihood of receiving a high salary prediction ($\hat{Y} = 1$) is statistically independent of gender. The Demographic Parity Difference (Δ_{DP}) is formalized as:

$$\Delta_{DP} = \left| P(\hat{Y} = 1 \mid A = 1) - P(\hat{Y} = 1 \mid A = 0) \right|$$

   A perfectly fair model yields $\Delta_{DP} = 0$.

2. **Equal Opportunity (EO):** Evaluates whether the true positive rates (TPR) are equal across groups, ensuring that highly qualified individuals have the same probability of a positive prediction regardless of gender. The Equal Opportunity Difference (Δ_{EO}) is formalized as:

$$\Delta_{EO} = \left| P(\hat{Y} = 1 \mid A = 1, Y = 1) - P(\hat{Y} = 1 \mid A = 0, Y = 1) \right|$$

Additionally, group-specific confusion matrices were calculated to diagnose whether prediction errors systematically underestimated female salaries (high false negatives) while overestimating or scattering male errors.

3.5. In-Processing Algorithmic Mitigation

To rectify the discovered baseline disparities without entirely sacrificing predictive utility, we applied an in-processing mitigation constraint. We utilized the *Exponentiated Gradient* reduction technique proposed by Agarwal et al. [Agarwal et al. 2018].

This approach frames the fair classification task as a constrained optimization problem, formulated as a zero-sum game between a learner (which optimizes global accuracy) and a regulator (which enforces constraints based on the chosen fairness metric, such as bounding Δ_{EO} or Δ_{DP} below a tolerance threshold ϵ). The objective function minimizes the overall classification error subject to the fairness parity constraints: $\min_f \text{Error}(f) \quad \text{subject to} \quad \mathcal{C}(f, A) \leq \epsilon$ By dynamically reweighting training instances during successive iterations, the in-processing wrapper forces the underlying estimator to actively decouple professional seniority from historical gender distribution proxies, optimizing decision boundaries for equitable subgroup sub-classification.

For the in-processing mitigation stage, the ExponentiatedGradient reduction was applied using Logistic Regression as the base underlying estimator due to its convex optimization properties and calibrated probability outputs.

4. Results and Discussion

This section presents the empirical findings of our algorithmic auditing and subsequent bias mitigation. We contrast the behavior of unconstrained baseline models against the fair optimization models, analyzing performance metrics, fairness parities, and subgroup dynamics.

4.1. Exploratory Data Analysis: The Gender Funnel and Proxy Architecture

To uncover the structural foundations driving the baseline model's bias, we conduct an integrated analysis of career progression, leadership access, and compensation distributions across gender groups. The empirical data reveals a classic socio-economic phenomenon known as the *gender funneling effect*, which directly enables the formation of algorithmic proxy variables.

As illustrated in Figure 1, there is a stark divergence in professional growth trajectories between male and female professionals within the Brazilian technology market.

At the entry-level stage (Junior), the representation gap is already visible but comparatively narrower, with male professionals outnumbering females by roughly a 2.5:1 ratio. However, as career progression moves toward senior and management echelons, a

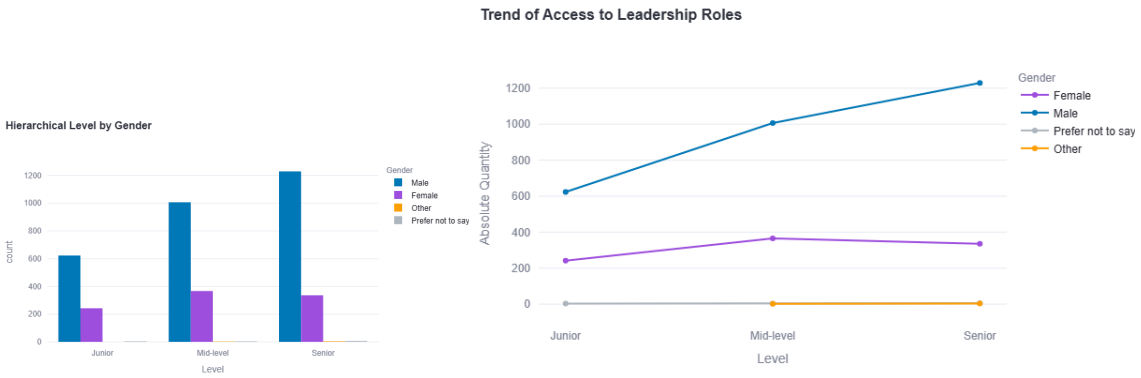


Figure 1. Hierarchical level distribution (left) and absolute access trend to senior/leadership roles (right) split by gender, illustrating the structural glass ceiling effect.

strict "glass ceiling" effect materializes. The absolute volume of male professionals experiences exponential growth, doubling from approximately 600 individuals at the Junior level to over 1,200 at the Senior level.

Conversely, the female professional trajectory exhibits systemic stagnation. The absolute count of female professionals remains virtually flat, experiencing a slight contraction as it transitions from mid-level (Pleno) to senior roles. This indicates that while the technology market expands dynamically at the top of the hierarchy, it systematically fails to retain and elevate female talent into senior leadership positions.

This severe asymmetry provides the exact mathematical substrate for the algorithmic biases detailed in next subsection. Because machine learning estimators optimize strictly for statistical correlations within the training distribution, the algorithm implicitly learns a dangerous proxy rule:

$$\text{High Seniority Level} \propto \text{Male Gender}$$

Consequently, even if explicit demographic attributes are suppressed during training (fairness through blindness), the model leverages structural variables like hierarchical level and job roles as proxy identifiers to reconstruct the hidden gender distribution, subsequently penalizing qualified women who break into senior domains.

This structural bottleneck reverberates directly into the compensation matrix. As mapped in Figure 2, the monthly salary distribution exhibits completely distinct topological shapes for each gender. The male distribution displays a heavy right-skewed profile; while it peaks around the mid-tier brackets (BRL 8,001 to 12,000), it maintains a thick, extended tail that stretches consistently into ultra-high compensation strata exceeding BRL 40,000 per month. In contrast, the female compensation curve collapses abruptly past the BRL 12,000 threshold, with a near-total absence of female representation in the top-tier economic brackets.

This visualization provides empirical validation for the *accuracy paradox* identified during our baseline model audit. Because historical and structural barriers compress female professionals into a low-variance, highly predictable salary pocket (under BRL

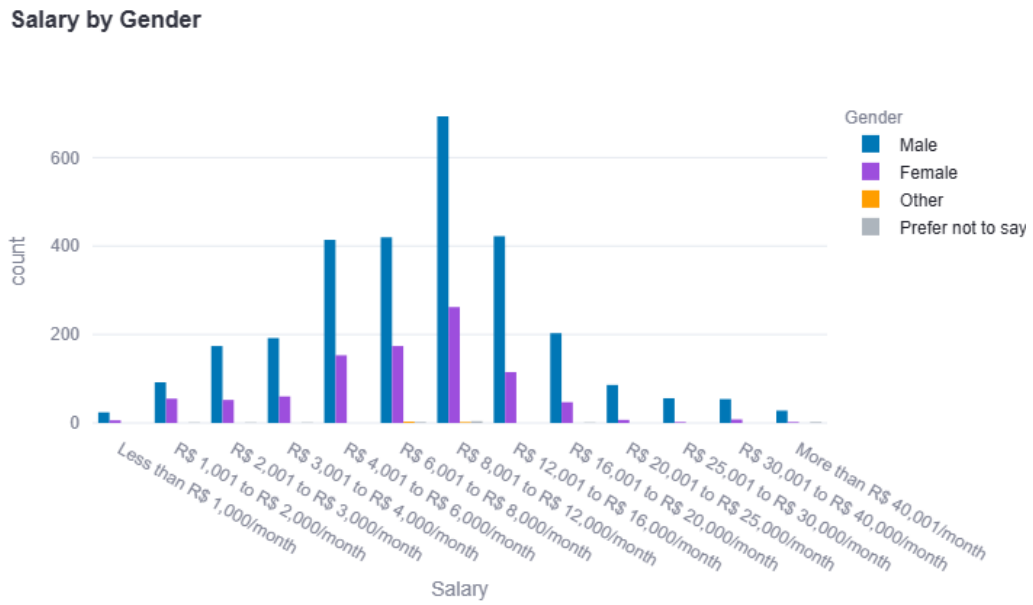


Figure 2. Monthly salary distribution (BRL) grouped by demographic attributes, exposing the high variance in male wages and the tight clustering of female wages.

12,000), an unconstrained classification model can easily achieve a high recall rate on the low-salary class for women, while correctly identifying only 39.44% of women who truly belonged to the high-salary bracket.

Conversely, the wide economic variance and upward mobility enjoyed by male professionals present a much higher statistical entropy, which inherently challenges the model's standardized decision rules and results in a higher false-positive rate (29.21%) for the male cohort.

Finally, Figure 2 corroborates our *Omission Hypothesis*. The demographic subset designating non-disclosure ("Prefer not to say") begins to surface precisely within the upper echelons of the wage distribution. This distribution supports the premise that demographic omission is not random noise, but a behavioral pattern highly correlated with top-tier male privilege, which intentionally masks the true bounds of wage inequality from unconstrained algorithms.

4.2. Baseline Auditing: The Accuracy Paradox and Proxy Mechanics

The initial unconstrained models achieved an overall classification accuracy of approximately 74%. However, a granular evaluation of the sub-group metrics revealed severe structural biases.

In the Decision Tree pipeline, professional seniority (e.g., senior and management levels) heavily dominated the top nodes of the tree, functioning as the root splitter. While this initially suggests a meritocratic hierarchy based on experience, a demographic analysis of the training distribution reveals that men disproportionately occupy these senior roles. Consequently, the model interpreted seniority as a structural proxy for gender.

This proxy mechanic was further verified by the Logistic Regression model's weights. While professional seniority yielded the strongest positive coefficient ($\beta_{\text{seniority}} \approx 2.11$), the gender attribute retained a statistically significant negative coefficient ($\beta_{\text{female}} \approx -0.26$). This negative weight isolates gender as a mathematical reducer, penalizing female professionals within the log-odds matrix regardless of their technical profile.

Crucially, group-specific confusion matrices reveal a severe Equal Opportunity gap, detailed in Table 2. While overall subgroup accuracy was nearly identical between genders (74.84% for men vs. 75.10% for women), the model correctly identified only 39.44% of women who truly belonged to the high-salary bracket (True Positive Rate), compared to 81.72% of men. At the same time, the model's false-positive rate for men (29.21%) was more than double that of women (11.58%), indicating that borderline male profiles were more readily classified into the high-salary bracket than equivalent female profiles.

Table 2. Baseline and Mitigated Performance and Fairness Metrics

Subgroup	Base. Acc.	Mit. Acc.	Base. TPR	Mit. TPR	Base. FPR	Mit. FPR
Global	74.90%	73.95%	–	–	–	–
Male	74.84%	74.84%	81.72%	81.72%	29.21%	29.21%
Female	75.10%	71.26%	39.44%	77.46%	11.58%	31.05%

Fairness Metric	Baseline	Mitigated	Target
Δ_{DP}	0.295	0.050	0.000
Δ_{EO}	0.423	0.043	0.000

To definitively prove that the salary scales were not neutral, a reverse-engineered classification test was conducted. An auxiliary model was trained to predict the gender of a professional using strictly their salary and seniority level as inputs. The model successfully deduced the respondent's gender with high precision, confirming that compensation intervals are deeply gendered.

4.3. Evaluation of the Omission Hypothesis

Before applying mitigation, we investigated the profile of respondents who selected "Prefer not to inform/Others" for the gender attribute. Standard data cleaning workflows often treat non-disclosure as random noise. However, our distributional analysis revealed that this group's salary profile was significantly higher than the female average, closely mirroring the upper echelons of the male salary distribution.

This statistical trend confirms our *Omission Hypothesis*: high-earning individuals—predominantly male executives—actively choose to omit demographic attributes at a higher rate than other socio-economic groups. In standard ML pipelines, this non-disclosure acts as an algorithmic mask, hiding a substantial portion of male privilege from the training data. This implies that the true baseline societal bias is likely even deeper than the 0.29 Demographic Parity divergence captured by the model.

4.4. Algorithmic Mitigation and Equity Gains

The baseline model demonstrated substantial unfairness, generating a Demographic Parity difference (Δ_{DP}) of 0.29 and an Equal Opportunity difference (Δ_{EO}) of 0.42. These

high deltas quantify the extent to which the model penalized qualified women, creating a barrier to high-salary predictions.

By wrapping the estimator in an in-processing Exponentiated Gradient reduction framework via *Fairlearn*, we forced the optimization function to satisfy fairness constraints. The results of this intervention demonstrate a highly favorable accuracy-fairness trade-off:

- **Parity Convergence:** The Demographic Parity difference (Δ_{DP}) dropped sharply from 0.29 to 0.050. Simultaneously, the Equal Opportunity difference (Δ_{EO}) was reduced from 0.42 to 0.043, successfully satisfying the parity threshold.
- **Global vs. Subgroup Utility:** The global accuracy of the system suffered an almost negligible concession, decreasing by only 1% (moving from 74% to 73%).
- **Recall Correction:** While global accuracy remained stable, the model's ability to correctly recognize high-earning women improved dramatically. Recall (TPR) for female professionals surged from 39.44% in the baseline setup to 77.46% under the mitigated setup, narrowing the gap with the stable male TPR of 81.72%.

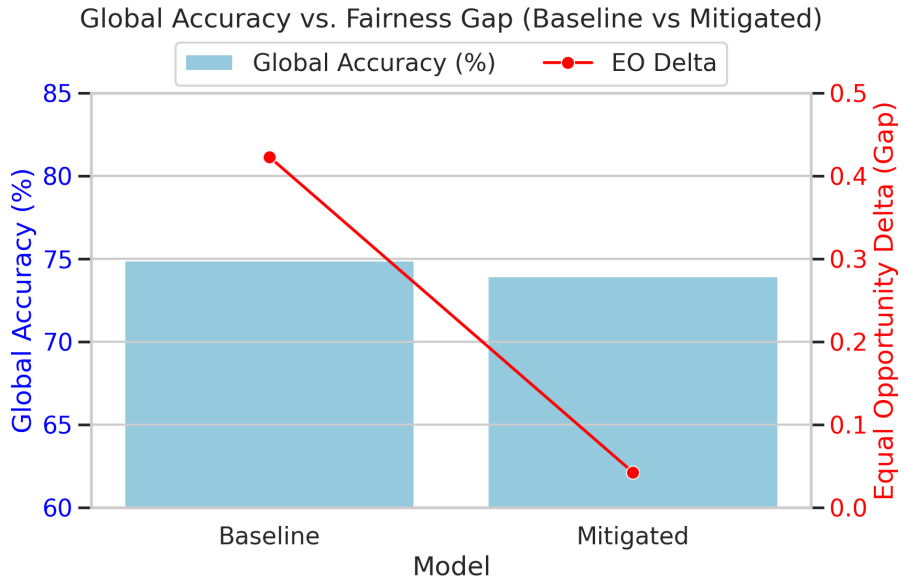


Figure 3. Global Accuracy vs. Fairness Gap (Baseline vs Mitigated) demonstrating the favorable accuracy-fairness trade-off.

An inspection of the post-mitigation confusion matrices confirms that this accuracy surge was driven by the rectification of false negatives. The constrained model stopped defaulting high-earning women into lower wage classes, actively learning to recognize and properly classify high-earning female professionals. By decoupling seniority proxies from gendered expectations, the mitigated algorithm successfully dismantled the automated gender penalty.

5. Conclusion

This study provided a rigorous algorithmic fairness assessment of salary prediction models within the Brazilian technology labor market, using data from the *State of Data Brazil*

survey. By auditing standard Machine Learning pipelines, we demonstrated that unconstrained models do not merely reflect historical gender-based wage disparities—they mathematically codify and entrench them under a veneer of statistical objectivity.

Our empirical diagnostic yielded two critical structural insights. First, we exposed a severe Equal Opportunity gap: the baseline model correctly recognized only 39.44% of women who truly qualified for high-salary predictions, compared to 81.72% of equally qualified men. Far from indicating equitable treatment, this hyper-predictability proves that women's professional trajectories are systematically confined to lower, tightly clustered salary brackets, whereas male compensation patterns benefit from wider, upward economic variance. Second, we validated the "omission hypothesis," demonstrating that high-earning male executives disproportionately choose to withhold demographic data. In standard machine learning workflows, this selective non-disclosure creates an algorithmic blind spot that masks the true extent of male privilege and skews training distributions.

By deploying an in-processing mitigation framework via the *Fairlearn* toolkit, we demonstrated that automated discrimination is not an immutable byproduct of historical data. The optimization constraints successfully decoupled career seniority from gender proxies, slashing the Demographic Parity difference from 0.29 to 0.050 and the Equal Opportunity difference from 0.42 to 0.043. Crucially, this equity adjustment incurred a negligible 1% concession in global accuracy (74.90% to 73.95%). This gain in fairness was not entirely costless at the subgroup level: female recall for the high-salary class rose sharply from 39.44% to 77.46%, at the expense of a corresponding increase in the female false-positive rate (11.58% to 31.05%) and a modest drop in female subgroup accuracy (75.10% to 71.26%). We interpret this as a favorable and well-documented fairness-accuracy trade-off, rather than a cost-free correction.

Ultimately, this research highlights that achieving fairness in automated Human Resource Management does not require sacrificing predictive utility. As organizations increasingly delegate compensation and recruitment architectures to artificial intelligence, the implementation of active constraints—such as cost-sensitive reductions—becomes an institutional imperative. Future work should extend this framework to longitudinal datasets to monitor how algorithmic mitigation scales over time, ensuring that the tech ecosystems of emerging markets mature into spaces of genuine, verifiable equity.

Acknowledgments

The authors would like to thank the 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.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used Gemini (a large language model developed by Google) to assist in English translation, grammatical refinement, and structural formatting of the manuscript's prose. After using this tool, the authors reviewed, edited, and verified all generated content to ensure technical accuracy, validity, and alignment with the empirical findings. The authors take full responsibility for the final contents of this publication, including the accuracy of the data analysis.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. mitpress.ubiquity.press.

Bird, S., Dudík, M., Edgar, R., Errouimi, B., Khanna, R., Masarek, D., Dudík, M., Vaughan, J. W., Madaio, M., and Others (2020). Fairlearn: A toolkit for assessing and improving fairness in ai. In *Microsoft Technical Report*, volume MSR-TR-2020-32.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Data Hackers (2024). State of data brazil 2024/2025. `https://www.kaggle.com/datasets/datahackers/state-of-data-brazil-20242025`. Accessed: June 2026.

Datta, A., Tschantz, M. C., and Datta, A. (2015). Proxy non-discrimination in data-driven systems. In *Proceedings of the 1st ACM Workshop on Privacy in the Electronic Society*, pages 85–88.

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.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Pessach, D. and Shmueli, E. (2022). Algorithmic fairness. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4824–4842.

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

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Wecchi, M. and Berton, L. (2026). Algorithmic fairness and bias mitigation in financial artificial intelligence: scoping review. *AI and Ethics*, 6(4):409.