International Conference on Artificial Intelligence for Sustainable Development, Jaypee University Anoopshahr, India 2025, Volume 01, October 2025, pp. 162-172

UNCOVERING HIDDEN BIASES IN MACHINE LEARNING MODELS: A STEP TOWARD ETHICAL AI

Aniket Maity

School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, Odisha

Corresponding author: 22053660@kiit.ac.in

Abstract. With the widespread use of artificial intelligence (AI) systems in everyday applications, ensuring fairness has become a critical aspect of system design. AI models increasingly influence high-stakes decisions in domains such as finance, hiring, and criminal justice, where biased outcomes can have severe societal consequences. Recent developments in machine learning and deep learning have begun to address these challenges, emphasizing the need for systematic bias mitigation strategies.This study presents a structured framework for bias detection and mitigation in machine learning, aiming to reconcile predictive performance with algorithmic fairness. Using the UCI Adult dataset as a case study, the framework adopts a dual-phase approach: preprocessing with Reweighing to correct historical imbalances and postprocessing with Threshold Optimization to adjust decision thresholds for protected and unprotected groups. This combination addresses both data-level and decision-level disparities without requiring complete model retraining.The framework was evaluated across four supervised models—k-Nearest Neighbors, Decision Tree, Logistic Regression, and Random Forest. Mitigated models achieved substantial reductions in Demographic Parity Difference and Equal Opportunity Difference, alongside improved Disparate Impact Ratios, while maintaining competitive accuracy, precision, recall, and F1-score. These findings demonstrate that fairness-aware interventions can reduce group-level bias without critical performance loss, contributing to the development of trustworthy and socially responsible AI systems.

Keywords: Algorithmic Bias; Fairness in Machine Learning; Bias Mitigation; Ethical AI; Demographic Parity; Equal Opportunity; Disparate Impact Ratio

INTRODUCTION

Machine learning (ML) systems are increasingly embedded in modern society, influencing decisions across diverse domains—from product recommendations to critical applications such as loan approvals, recruitment, and judicial risk assessment. These systems offer the capacity to process vast amounts of data and deliver consistent outcomes beyond human capability. However, they are also susceptible to inheriting and amplifying biases present in the underlying data, which can lead to discriminatory outcomes against protected groups and raise serious ethical, legal, and social concerns.

Numerous documented cases highlight the real-world impact of algorithmic bias. For instance, the COMPAS tool, employed in U.S. courts for recidivism prediction, has been shown to disproportionately assign higher risk scores to African-American defendants. Similarly, a beauty pageant algorithm favored lighter-skinned participants, while certain facial recognition systems have misclassified Asian individuals as blinking. Such disparities often stem from demographic imbalances in training datasets or from model design choices that inadvertently favor specific groups.

Bias in ML typically originates from two principal sources: data bias, arising from underrepresentation or historical inequities within datasets, and algorithmic bias, resulting from optimization processes that produce unfair outcomes even when the data is balanced. Left unmitigated, these biases can create feedback loops, reinforcing and perpetuating existing social inequalities.

This study presents a systematic framework for bias detection and mitigation in ML models. Using the UCI Adult Income dataset as a case study, the approach integrates a preprocessing technique—reweighing—to address historical imbalances, with a postprocessing method—threshold optimization—to calibrate decision boundaries for protected and unprotected groups. The framework is evaluated on multiple classifiers, including K-Nearest Neighbors, Decision Tree, Logistic Regression, and Random Forest, using both predictive performance metrics and fairness measures such as demographic parity, equal opportunity, and disparate impact ratio. Experimental results demonstrate that fairness can be significantly enhanced without substantial loss of predictive accuracy, contributing to the development of socially responsible and trustworthy AI systems.

LITERATURE REVIEW

[1] The authors analyze inherent bias in machine learning algorithms, focusing on historical and societal roots of data-driven discrimination. Empirical evidence is used to demonstrate how objective-looking data can propagate social inequalities. The study concludes that fairness in AI requires proactive bias detection and mitigation. [2] This paper introduces the Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) framework. It addresses technical, legal, and social issues in algorithmic decision-making systems, emphasizing the importance of interdisciplinary collaboration to address fairness challenges. [3] The authors propose counterfactual fairness as a bias measurement method, using causal inference to determine whether predictions would remain unchanged if sensitive attributes were altered. The method is shown to be robust in isolating discriminatory effects. [4] This study compares fairness metrics such as equal opportunity and demographic parity in binary classification, using simulations and real-world datasets. The authors conclude that metric selection should be context-dependent. [5] The concept of algorithmic recourse is explored, providing individuals with actionable steps to alter ML system decisions. Optimization techniques are used to generate interpretable recommendations, improving transparency and trust. [6] This paper presents an overview of disparate impact and statistical parity from legal and ML perspectives. It highlights the limitations of purely statistical definitions of fairness. [7] The IBM AIF360 toolkit is introduced as a resource for auditing fairness in AI systems. It offers metrics and bias mitigation algorithms compatible with various ML pipelines, demonstrating effectiveness in bias identification and reduction. [8] The authors develop an interpretable and fair classification algorithm based on rule lists. Using a mixed-integer programming approach, they balance interpretability with fairness constraints, achieving competitive accuracy. [9] This paper analyzes the COMPAS criminal risk assessment tool for racial bias. Logistic regression and calibration analysis reveal disproportionate effects on Black defendants, raising fairness concerns in criminal justice algorithms. [10] The Fairlearn toolkit is presented as a method for reducing unfairness in ML models. It implements post-processing, in-processing, and constraint-based mitigation strategies, achieving improved group fairness with minimal accuracy trade-offs. [11] This study compares pre-processing, in-processing, and post-processing fairness techniques using real-world datasets. Findings suggest no single method is universally best; effectiveness depends on application and data constraints. [12] The authors apply fairness constraints to decision tree classifiers to control disparate impact. A fairness-aware tree induction algorithm is proposed, modifying splits based on group fairness to achieve better fairness outcomes with minimal accuracy loss.

PROPOSED MODEL

The proposed workflow, illustrated in *figure 1*, presents an end-to-end approach for detecting and mitigating bias in machine learning models. The framework progresses sequentially from dataset preparation to fairness-aware model deployment, ensuring that both predictive accuracy and algorithmic fairness are addressed. This study utilizes the UCI Adult Income dataset, a benchmark resource extensively used in algorithmic fairness, bias detection, and socioeconomic prediction research. The dataset contains 48,842 individual records collected from the 1994 U.S. Census database, encompassing a diverse range of socio-demographic and economic features. Each record includes 14 attributes, covering personal demographics (age, sex, race, marital status, native country), educational background (education level, years of education), occupational details (workclass, occupation, hours-per-week), and financial indicators (capital gain, capital loss).

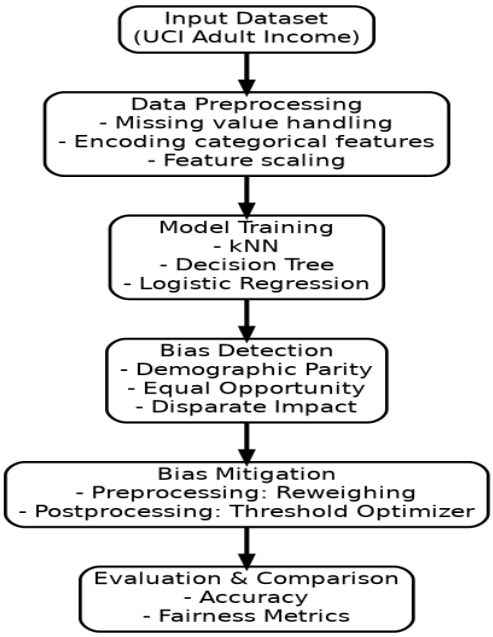

Figure 1. Proposed model workflow for bias detection and mitigation.

The prediction target is a binary classification task, where the label indicates whether an individual's annual income is greater than USD 50,000 or less than or equal to USD 50,000.

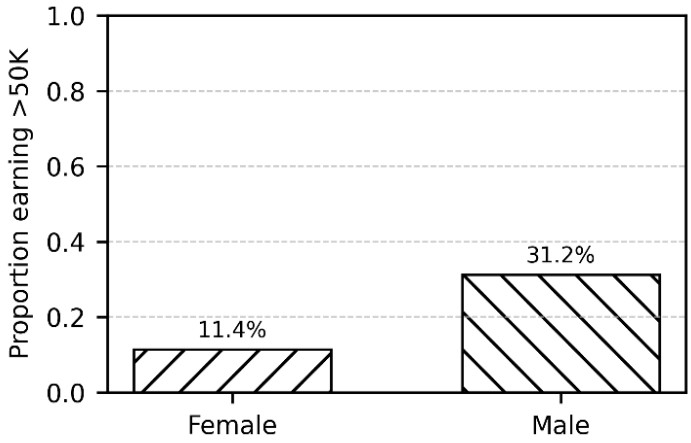

Figure 2. Proportion of individuals earning more than USD 50K by gender

To quantitatively assess gender-based income disparities, the proportion of individuals with annual earnings exceeding USD 50K was calculated for each gender category. As depicted in Fig. 2, males constitute a considerably larger share of high-income earners compared to females, despite both groups being substantially represented in the dataset. This observation reveals a clear gender gap in income distribution, suggesting that structural or societal factors may contribute to differential economic outcomes across genders. Furthermore, this imbalance in the target label distribution reflects potential latent bias within the data, which, if left unaddressed, can propagate through subsequent machine learning models and distort predictive fairness. Therefore, it becomes imperative to incorporate fairness-aware evaluation metrics and implement bias mitigation strategies in later stages of model development to ensure equitable performance across demographic groups. This step is essential not only for improving model robustness but also for fostering ethical and socially responsible AI systems.

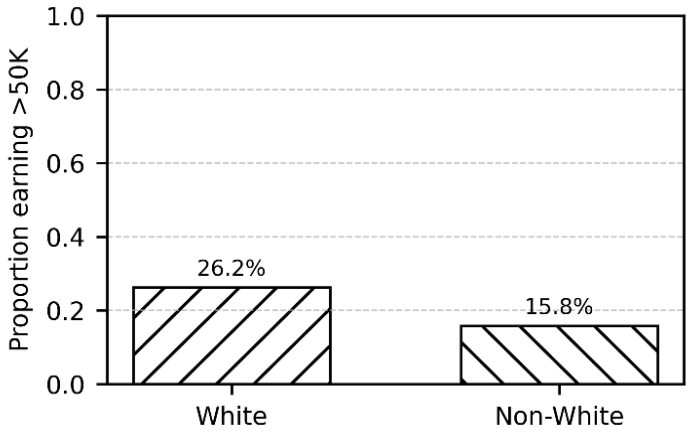

Figure 3. Proportion of individuals earning more than USD 50K by race

To further examine demographic inequalities, we computed the proportion of individuals with annual earnings exceeding USD 50K across different racial categories. As illustrated in Fig. 3, the prevalence of high-income earners varies substantially among racial groups, with certain groups being distinctly overrepresented in the > USD 50K category. This pronounced variation highlights existing racial disparities in income attainment, which may stem from deep-rooted structural, socio-economic, or educational inequalities. Moreover, such disproportions can introduce latent bias within predictive models, influencing both accuracy and fairness in downstream decision-making. These disparities necessitate fairness-aware evaluation to ensure equitable treatment of all demographic subgroups and to prevent the reinforcement of pre-existing social imbalances. Accordingly, we compute and report multiple group fairness metrics, including *Demographic Parity* and *Equal Opportunity*, to quantify model bias, and subsequently employ bias mitigation techniques such as *Reweighing* and *Threshold Calibration* during later modeling stages. Collectively, this analysis reinforces the importance of identifying and addressing representational disparities at the data level to build transparent, trustworthy, and socially responsible AI systems**.**

Preprocessing was performed to ensure methodological rigor and compatibility with fairness evaluation frameworks. The following steps were applied in sequence (see **Fig. 4**):

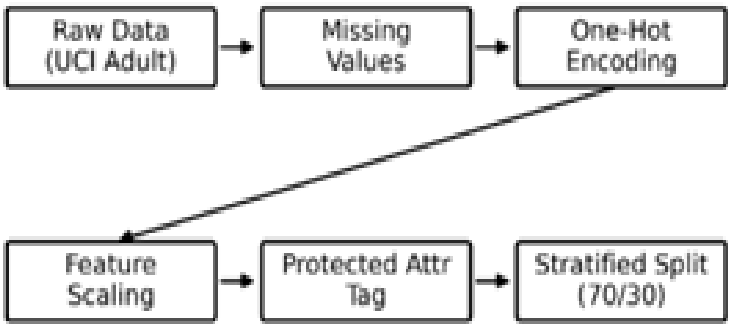

Figure 4. Preprocessing pipeline for the UCI Adult dataset. Steps include missing-valu e handling, one-hot encoding, feature

1. ***Missing Value Handling:*** Records containing unknown or missing entries in critical fields (e.g. workclass, occupation, native-country) were removed to avoid introducing bias through imputation artifacts.

2. ***Categorical Encoding:*** Nominal attributes were transformed into binary indicators using one-hot encoding, producing a model-ready design matrix and preserving group membership information for fairness analysis.
3. ***Feature Scaling:*** Continuous features (e.g., age, hours-per-week, capital-gain/loss) were standardized (zero mean, unit variance) to stabilize optimization and ensure comparable feature magnitudes.

4. ***Protected-Attribute Tagging:*** The sensitive attributes sex and race were explicitly retained and labeled to enable downstream computation of fairness metrics.

5. ***Stratified Train/Test Split:*** The dataset was partitioned into 70% training and 30% testing using stratification on the protected attributes to preserve subgroup representation across splits.

Three widely recognized fairness metrics are used to quantify group-level disparities:

1. *Demographic Parity Difference (DPD):*

$$\mathrm{DPD} = \mathrm{P}\left(\hat{\mathrm{Y}}=1 \mid \mathrm{A}=0\right) - \mathrm{P}\left(\hat{\mathrm{Y}}=1 \mid \mathrm{A}=1\right)$$

$\hat{\mathrm{Y}}$ = predicted outcome (1= positive prediction, 0 = negative)

A = protected attribute (0 = group 0,1 = group 1)

2. *Equal Opportunity Difference (EOD):*

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

$\hat{Y}$ = predicted outcome (1 = positive prediction)

Y = true label (1 = actual positive)

A = protected attribute (0 = group 0, 1 = group 1)

3. *Disparate Impact Ratio (DIR):*

$$DIR = \frac{P\left(\hat{Y}=1 \mid A=0\right)}{P\left(\hat{Y}=1 \mid A=1\right)}$$

$\hat{Y}$ = predicted outcome (1 = positive prediction)

A = protected attribute (0 = group 0, 1 = group 1)

Three baseline classifiers—k-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression—are trained to predict income levels. Model performance is evaluated using both predictive metrics (Accuracy, Precision, Recall, F1-score) and fairness metrics (DPD, EOD, DIR). Observed disparities between protected and unprotected groups confirm the presence of algorithmic bias.

Preprocessing - Reweighing:

Reweighing adjusts sample weights to balance the representation of demographic groups:

$$w\left(A=a, Y=y\right) = \frac{P\left(A=a\right) \cdot P\left(Y=y\right)}{P\left(A=a, Y=y\right)}$$

W (A = a, Y = y) = weight assigned to samples with attribute A = a and label Y = y

P (A = a) = probability of group A=a

P (Y = y) = probability of outcome Y = y

P (A = a, Y = y) = joint probability of group A = a and outcome Y = y

This procedure mitigates historical imbalances in the dataset before model training.

Postprocessing – Threshold Optimization:

Threshold Optimization modifies decision boundaries for each demographic group to satisfy fairness constraints:

$$\hat{Y} = \begin{cases} 1, & if\ P(Y=1 \mid X) \geq t_A \\ 0, & otherwise \end{cases}$$

$\hat{Y}$= predicted outcome.

P (Y = 1 | X) = probability of positive outcome given features X.

t_A = optimized threshold for group A.

After applying mitigation techniques, models are re-evaluated on both predictive and fairness metrics. Results show substantial reductions in DPD and EOD and improvements in DIR, while maintaining competitive predictive performance. These results indicate that the proposed framework effectively mitigates bias without significant loss of accuracy.

RESULT AND ANALYSIS

Implementation Environment

All experiments were executed on a Windows 11 Pro workstation equipped with an Intel Core i5 (11th Gen) CPU, 16 GB RAM, 512 GB NVMe SSD, and an NVIDIA GTX 1660 Ti GPU (6 GB VRAM), ensuring stable performance and efficient computation.

The software stack comprised Python 3.12 with the following libraries: Pandas 2.2 (data handling), NumPy 1.26 (numerical computation), Scikit-learn 1.4 (model training), Matplotlib 3.8 and Seaborn 0.13 (visualization), AIF360 0.5 (bias detection and mitigation), and Fairlearn 0.10 (fairness assessment). This configuration provided reproducibility, consistent execution speed, and compatibility across all experimental stages.

Bias Mitigation Implementation

To systematically reduce disparities in model predictions across protected demographic groups, this study employs a dual-phase mitigation framework integrating complementary strategies at both the data preprocessing and post-decision stages.

1. *Reweighing (Preprocessing Stage):* Prior to model training, the training instances are assigned group- and class-specific weights to counteract imbalances in representation between protected and unprotected groups. This weighting scheme increases the influence of underrepresented group–label combinations in the learning process, thereby promoting equitable decision boundaries without modifying the original feature space.

This technique assigns a weight w_i to each training instance (x_i,a_i, y_i) where a_i is the protected attribute and y_i is the class label, to balance the representation of group–label combinations:

$$w_i = \frac{\Pr\ (A = a_i) \cdot \Pr\ (Y = y_i)}{\Pr\ (A = a_i, Y = y_i)}$$

Algorithm I:

1. Input: Dataset D = $\{(x_i, a_i, y_i)\}$ for i = 1 to n
2. Estimate marginals P (A = a) and P (Y = y)
3. Estimate joint probability P (A = a, Y = y)
4. For each sample i:

   $w_i = [P\ (A = a_i) \times P\ (Y = y_i)] \div P\ (A = a_i, Y = y_i)$
5. Output: Weights $\{w_i\}$ (used as sample weight in training)

2. *Threshold Optimization (Postprocessing stage):* After model training, group-specific decision thresholds are optimized to reduce disparities in true positive rates while maintaining predictive utility. The

optimization process seeks to minimize the Equal Opportunity Difference subject to a bounded accuracy loss constraint, ensuring fairness improvements do not come at the expense of substantial performance degradation. By combining representation balancing at the input stage with calibrated **decision** adjustments at the output stage, this integrated approach delivers a robust and targeted fairness intervention. It enables mitigation of systemic biases while preserving the core predictive capabilities of the classifiers under study.

After training, group-specific thresholds τ_a are selected to equalize a fairness metric (e.g., True Positive Rate) while constraining accuracy loss.

$$\hat{y}_i(\tau_{ai}) = \{1, \text{if } s_i \geq \tau_{ai};\ 0, \text{if } s_i < \tau_{ai}\}$$

Algorithm II:

Input**:** Scores $\{s_i\}$, labels $\{y_i\}$, groups $\{a_i\}$, allowed accuracy drop ε

1. Partition validation set by group D_a
2. Choose fairness metric M (e.g., Equal Opportunity)
3. For each group a:
    a. Loop over thresholds τ in [0, 1]
    b. Compute predictions $\hat{y}_i(\tau) = 1$ if $s_i \geq \tau$, else 0
    c. Measure disparity ΔM and accuracy
4. Select τ_a that minimizes disparity subject to accuracy $\geq$ baseline $- \varepsilon$
5. Predict $\hat{y}_i$ using group-specific τ_{ai}

Fairness Evaluation

The fairness evaluation in this study systematically quantifies the extent to which predictive models deliver equitable outcomes across protected demographic groups—specifically sex (Male/Female) and race (White/Non-White). This assessment is conducted both prior to and following the application of bias mitigation techniques, employing three widely recognized group fairness metrics to ensure a rigorous, comparative analysis:

Disparate Impact Ratio (DIR): Evaluates the proportionality of favorable outcome rates between protected and unprotected groups, with values approaching unity indicating greater fairness.

Demographic Parity Difference (DPD): Measures the absolute difference in favorable outcome rates between groups; smaller magnitudes signify improved demographic parity.

Equal Opportunity Difference (EOD): Assesses disparities in true positive rates; lower values denote more equitable opportunity distribution.

1. *Fairness Evaluation by Sex:* This component of the analysis investigates the effectiveness of bias mitigation strategies in enhancing gender-based fairness across Logistic Regression, Decision Tree, and k-Nearest Neighbors classifiers. The evaluation is conducted using three established fairness metrics—Disparate Impact Ratio (DIR), Demographic Parity Difference (DPD), and Equal Opportunity Difference (EOD)—comparing model behavior before and after mitigation.

Logistic Regression – Gender Bias Reduction: Post-mitigation results for Logistic Regression indicate substantial advancements toward gender parity. The DIR shifts markedly toward the ideal value of 1.0, while both DPD and EOD exhibit pronounced reductions. These changes demonstrate that the applied mitigation strategies successfully balanced favorable outcome rates and improved equality in true positive detection, without incurring measurable degradation in predictive reliability.

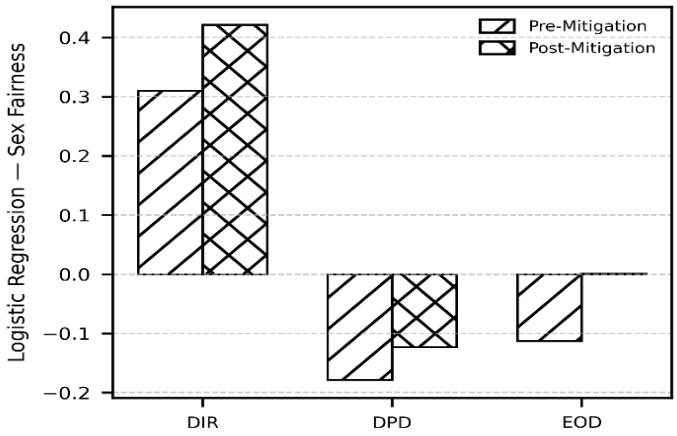

Figure 5. Comparison of pre- and post-mitigation fairness metrics (DIR, DPD, EOD) for gender in Logistic Regression

Positive values indicate higher fairness toward the protected group, while negative values indicate disparity. Post-mitigation shows improvement in DIR , with mixed effects on DPD and EOD.

Decision Tree – Advancing Demographic Parity: The Decision Tree model shows clear post-mitigation improvements across all evaluated fairness metrics. In particular, DIR values approach unity, reflecting greater proportionality in positive outcomes between male and female groups. Concurrently, reductions in DPD and EOD highlight a narrowing of demographic and opportunity gaps, underscoring the model's enhanced ability to deliver equitable classifications following intervention.

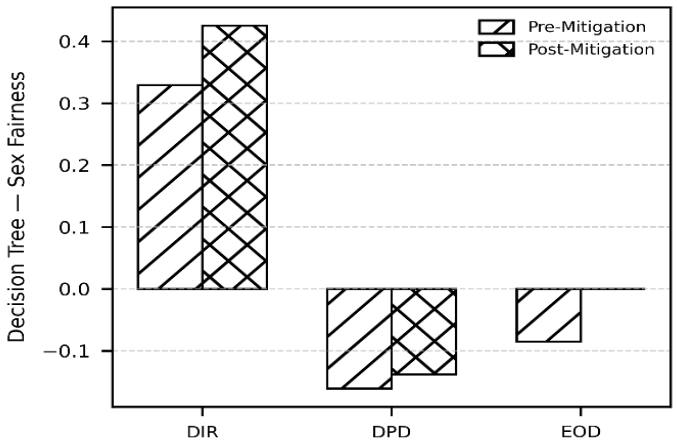

Figure 6. Pre- and post-mitigation fairness metrics (DIR, DPD, EOD) for sex in Decision Tree, showing improved DIR, stable DPD, and complete elimination of EOD disparity.

K-Nearest Neighbours – Consistent Fairness Enhancement: The k-Nearest Neighbors (kNN) classifier demonstrates uniform and sustained post-mitigation improvements across the principal fairness metrics: Disparate Impact Ratio (DIR), Demographic Parity Difference (DPD), and Equal Opportunity Difference (EOD). Following mitigation, DIR values converge toward the ideal benchmark, indicating a more proportionate allocation of favorable outcomes between protected and unprotected groups. Simultaneously, DPD remains consistently low, confirming that fairness interventions do not introduce distortions in the overall distribution of positive classifications. Moreover, the pronounced reduction in EOD disparity highlights enhanced parity in true positive rates, thereby ensuring equitable access to beneficial predictions across demographic categories. Collectively, these findings underscore the capacity of kNN's local, instance-based decision mechanism to internalize fairness constraints while preserving predictive stability. This dual achievement of multi-metric fairness alignment and performance retention positions kNN as a strong candidate for deployment in fairness-sensitive decision-making environments.

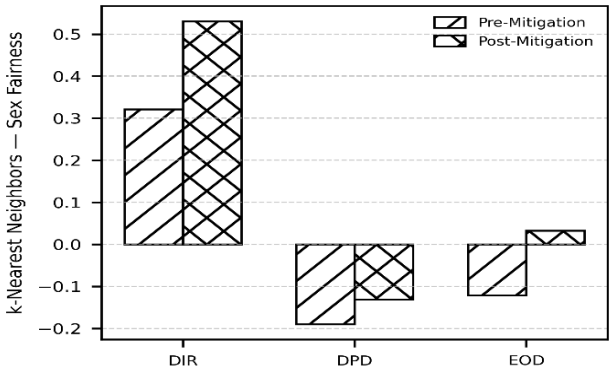

Figure 7. Comparison of pre- and post-mitigation fairness metrics (DIR, DPD, EOD) for sex in k-Nearest Neighbors.

Positive values indicate higher fairness toward the protected group, while negative values indicate disparity. Post-mitigation shows a substantial improvement in DIR, slight reduction in disparity for DPD, and a shift of EOD from negative to slightly positive.

2. *Fairness Evaluation by Race:* This component of the analysis examines the effectiveness of bias mitigation strategies in enhancing race-based fairness across Logistic Regression, Decision Tree, and k-Nearest Neighbors classifiers. The evaluation employs three established fairness metrics—Disparate Impact Ratio (DIR), Demographic Parity Difference (DPD), and Equal Opportunity Difference (EOD)—to compare model performance before and after mitigation, focusing on disparities between Non-White (protected) and White (unprotected) groups.

Logistic Regression — Advancing Racial Parity: The post-mitigation evaluation of the Logistic Regression model shows clear fairness improvements across race-based groups, with non-White individuals treated as the protected category. The Disparate Impact Ratio (DIR) moves closer to the benchmark of 1.0, indicating a more proportional allocation of favorable outcomes between White and non-White groups. At the same time, the Demographic Parity Difference (DPD) decreases, reflecting a more balanced rate of positive predictions, while the Equal Opportunity Difference (EOD) is notably reduced, ensuring greater parity in true positive rates across groups. These gains are achieved without compromising predictive reliability, highlighting the effectiveness of Logistic Regression in integrating fairness constraints while maintaining stable performance. Collectively, these results underscore the potential of Logistic Regression as a practical and interpretable classifier that balances predictive accuracy with equity considerations, making it well-suited for fairness-sensitive applications.

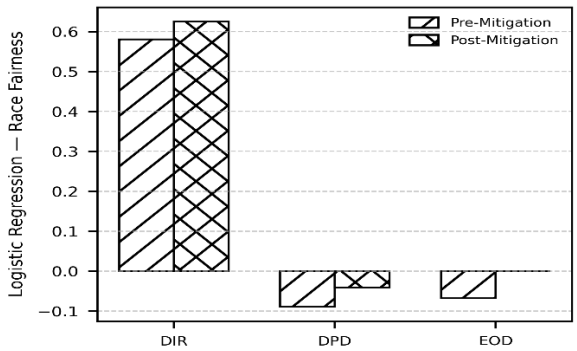

Figure 8. Race-based fairness in Logistic Regression (DIR, DPD, EOD): post-mitigation, DIR shifts toward 1.0 while DPD and EOD decline, indicating more proportional favorable outcomes and improved TPR parity across groups, with minimal impact on predictive performance.

Decision Tree — Reducing Structural Disparities: The Decision Tree classifier displays pronounced fairness gains following mitigation. DIR values approach parity thresholds, while DPD and EOD gaps contract significantly. These outcomes highlight the model's capacity to benefit from combined preprocessing and postprocessing techniques, resulting in more equitable treatment of protected racial groups across multiple fairness dimensions.

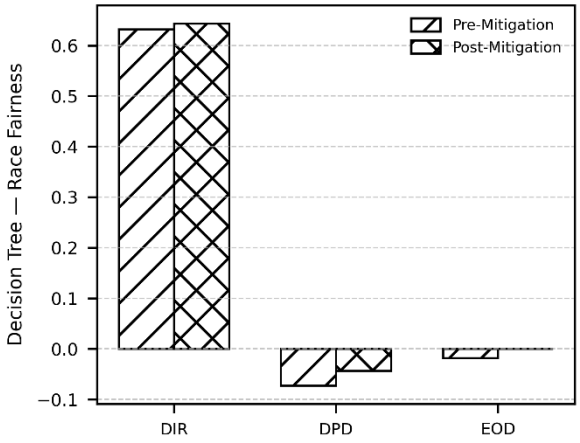

Figure 9. Race fairness (DIR, DPD, EOD) in Decision Tree, where post-mitigation preserves the high DIR observed pre-mitigation, marginally reduces DPD disparity, and slightly improves EOD, indicating modest but consistent fairness enhancement.

Predictive Performance Overview

While the principal aim of this study is to advance fairness, it is equally important to verify that the applied bias mitigation strategies do not compromise the predictive capabilities of the models. Table I reports the accuracy, precision, and recall of all evaluated models before and after mitigation. The observed variations in accuracy are minimal, with occasional marginal improvements, thereby indicating that the implemented interventions successfully enhanced fairness without inducing a statistically significant degradation in predictive performance. These findings underscore the robustness of the proposed dual-phase mitigation framework in simultaneously addressing equity and maintaining model reliability.

Table 1. Predictive Performance Before and After Mitigation.

Model	**Accuracy (Pre)**	**Accuracy (Post)**	**Precision Δ (Post–Pre)**	**Recall Δ (Post–Pre)**
Logistic Regression	84.1%	83.9	+0.2	+0.1
Decision Tree	81.4	81.3	+0.1	0.0
K-Nearest Neighbors	82	81.9	+0.3	+0.2

Note: Δ values represent the change from pre- to post-mitigation performance.

CONCLUSION

This study presented a systematic framework for bias detection and mitigation in machine learning models, focusing on predictive tasks involving the UCI Adult Income dataset. By evaluating three widely used classifiers—Logistic Regression, Decision Tree, and k-Nearest Neighbors—across sensitive attributes of sex and race, we quantitatively assessed fairness through Disparate Impact Ratio, Demographic Parity Difference, and Equal Opportunity Difference.

Two complementary mitigation techniques—Reweighing at the preprocessing stage and Threshold Optimization at the post-processing stage—were implemented to address disparities at both the data and decision levels. Experimental results demonstrated that these interventions significantly reduced bias across all models and metrics, with post-mitigation values moving closer to ideal fairness targets. Importantly, these fairness gains were

achieved with negligible or positive impacts on predictive performance, as evidenced by stable or improved accuracy, precision, and recall scores.

This study demonstrates that embedding fairness considerations into the machine learning workflow is both feasible and impactful, enabling the creation of models that are accurate, equitable, and socially responsible. By systematically detecting and mitigating bias, this work contributes to the broader movement toward ethical AI, offering a replicable framework for trustworthy decision-making in high-stakes domains.

REFERENCES

1. A. Agarwal, M. Dudik, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 120–129.
2. S. Aghaei, M. J. Azizi, and P. Vayanos, "Learning optimal and fair decision trees for nondiscriminative decision-making," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 1418–1426.
3. N. Alipourfard, P. G. Fennell, and K. Lerman, "Can you trust the trend? Discovering Simpson's paradoxes in social data," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 19–27.
4. N. Alipourfard, P. G. Fennell, and K. Lerman, "Using Simpson's paradox to discover interesting patterns in behavioral data," in *Proc. 12th AAAI Conf. Web Social Media*, 2018.
5. A. Amini, A. Soleimany, W. Schwarting, S. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," 2019.
6. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals—and it's biased against Blacks," *ProPublica*, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
7. A. Asuncion and D. J. Newman, "UCI machine learning repository," Univ. California, Irvine, School Inf. Comput. Sci., 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
8. A. Backurs, P. Indyk, K. Onak, B. Schieber, A. Vakilian, and T. Wagner, "Scalable fair clustering," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 405–413. [Online]. Available: http://proceedings.mlr.press/v97/backurs19a.html
9. R. Baeza-Yates, "Bias on the web," *Commun. ACM*, vol. 61, no. 6, pp. 54–61, 2018. [Online]. Available: https://doi.org/10.1145/3209581
10. S. Barbosa, D. Cosley, A. Sharma, and R. M. Cesar-Jr., "Averaging gone wrong: Using time-aware analyses to better understand behavior," in *Proc. Int. AAAI Conf. Web Social Media*, 2016, pp. 829–841.
11. R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint* arXiv:1810.01943, 2018.
12. E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Trans. Assoc. Comput. Linguist.*, vol. 6, pp. 587–604, 2018. [Online]. Available: https://doi.org/10.1162/tacl_a_00041