Arzu, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Mitigating Algorithmic Bias through Disparate Impact Ratio: A Pre-processing Approach to Fair Machine Learning

Arzu, Assistant Professor[1], Nikhil Kumar[2], Dhruv Upreti[3], Durgesh Rajpurohit[4], Vanshika thakur[5]

[1]Department of Computer Applications,
HMRITM, Hamidpur, Delhi, 110040
[2,3,4,5]CSE, HMRITM, Hamidpur, Delhi, 110040,

Abstract- Artificial Intelligence (AI) systems are increasingly used in high-stakes decision-making domains such as recruitment, healthcare, finance, and education, where fairness and accountability are critical. Although these systems are often perceived as objective, they can inherit and amplify biases present in historical data, leading to discriminatory outcomes for certain demographic groups. This study investigates algorithmic bias in machine learning models, with a primary focus on measuring and mitigating unfairness using the Disparate Impact Ratio (DIR) and Equal Opportunity Difference (EOD).Using the Adult Income dataset, this research evaluates bias in income prediction tasks where gender is treated as a protected attribute. Two machine learning models Logistic Regression and Decision Tree are implemented to analyze both predictive performance and fairness. The study follows a two-stage experimental design: a baseline model trained on original data, and a bias-mitigated model using the Disparate Impact Remover, a preprocessing technique that adjusts feature distributions to reduce dependency on sensitive attributes.Results demonstrate that bias mitigation techniques can significantly improve fairness metrics, particularly by increasing the Disparate Impact Ratio toward acceptable thresholds and reducing disparities in true positive rates across groups. However, these improvements often come with a trade-off in predictive accuracy, highlighting the inherent tension between fairness and performance. The findings emphasize that no single bias mitigation strategy is universally optimal; instead, effectiveness varies depending on the dataset, model, and application context.This study contributes to the growing field of fairness-aware machine learning by providing a comparative evaluation of bias detection and mitigation techniques in practical settings. It underscores the importance of integrating fairness measures into the AI development lifecycle and offers insights for designing more ethical, transparent, and inclusive AI systems.

Keywords: Artificial Intelligence (AI), Algorithmic Bias, Fairness in Machine Learning, Bias Mitigation, Disparate Impact Ratio (DIR), Equal Opportunity Difference (EOD), Adult Income Dataset, Logistic Regression, Decision Tree, Fairness-Aware Machine Learning, Ethical AI, Protected Attributes.

I. INTRODUCTION

Artificial Intelligence (AI) and machine learning technologies have become integral components of contemporary decision-making systems in various sectors, including recruitment, credit scoring, healthcare diagnostics, education, and financial services. These systems are designed to process vast amounts of historical data to generate predictions or recommendations that are expected to be objective, efficient, and scalable. As organizations increasingly rely on automated systems for highstakes decisions, AI models are often viewed as neutral tools capable of minimizing human bias and enhancing operational efficiency. Nevertheless, recent research has highlighted that machine learning systems may unintentionally inherit and amplify biases embedded in historical datasets, resulting in discriminatory outcomes for specific demographic groups.

When AI treats some people differently because of things like their skin color, age, or financial situation, the results are unfair. Even if they don't mean to, biased datasets often make systems favor certain groups of people. Recruitment software may overlook qualified candidates due to historical hiring imbalances. Algorithms automatically copy social differences that are already there into digital formats. Inequality-based choices from the past show up again in automated systems later on. Tools

don't work for some groups because they don't show up in data. When biased cases train algorithms that are supposed to be neutral, the differences from the past stay hidden. Hiring screens, loan assessments, and risk forecasts are some tools that may quietly leave people out.

In areas where justice is essential, proof shows skewed outcomes keep repeating. Something appearing neutral on initial inspection might actually stem from long-standing imbalances in numbers. When datasets shaped by prejudice feed into forecasting systems, patterns slip through unnoticed - later shaping decisions made far beyond code. Often, though identifiers such as ethnicity or sex are removed, surrounding markers take their place indirectly: where a person lives, what school they attended, how much they earn. These elements become signals the model uses to infer identity across repeated instances. Removing explicit attributes therefore fails, sometimes, to block unequal outcomes further ahead. Questions arise, given the quiet way these processes develop, about legal accountability - can they even be challenged? When reasoning hides within tangled programming and unseen premises, keeping watch grows difficult.

In light of these challenges, researchers developed various fairness-oriented machine learning strategies aimed at mitigating bias in predictive tools. These methods are grouped based on when they enter the workflow. Right at the beginning, preprocessing steps change the data used for training. The goal is to get rid of unfair patterns as soon as possible. Inprocessing changes during model creation add fairness rules to the core of the training phase, making sure that performance and fair outcomes are balanced. After a model is built, post-processing changes its decisions to make them more fair without changing the data or code. The Disparate Impact Ratio (DIR) became a common way to find algorithms that treat people unfairly over time, along with many other suggested checks and fixes. A figure known as the Disparate Impact Ratio is a number that shows how often one group gets better results than another. It suggests that there may be unfairness. Some rules say that discrimination might be happening when the rate for a protected group drops below 80 percent of the main group's rate. This ratio helps find uneven data patterns before any learning algorithms run by starting the analysis early. Even though more studies are looking at fairness in code, it's still hard to tell if fixes really work. Sometimes, making things fairer makes predictions less accurate, especially when accuracy is most important. One way to measure fairness might show bias through another lens, so it's hard to agree on clear standards. Depending on the data traits, where it is used, or which model runs, bias fixes might work better - or fall short. Because these issues pop up so often, side-by-side tests become crucial: they show what each method can do, and where it fails.

This study investigates the reduction of bias in artificial intelligence, emphasizing the Disparate Impact Ratio and various mitigation strategies in machine learning. One method focuses on proportional representation, while others change the training data or the decision thresholds to find a balance between fairness and accuracy. Some methods make things more fair, but they can also make the model less effective, which is a trade-off worth looking into. From healthcare to hiring, where automated decisions have an impact on people's lives, it's important to know which method works best in which situation. A side-by-side comparison shows that each strategy has strengths that depend on the situation, so don't assume that one is better than the others. Results indicate that no solution is universally applicable; each presents unique implications for transparency and reliability. People who make policy or build systems can make better decisions if they understand these differences. When using AI in important areas, it's just as important to know how methods work in practice as it is to know how they work in theory.

Research Objectives

1. To develop expertise in fairness-aware machine learning, with a focus on identifying, measuring, and mitigating algorithmic bias using techniques such as Disparate Impact Ratio and other fairness metrics.
2. To contribute to the design of ethical and responsible AI systems by integrating bias

mitigation strategies into real-world applications across domains like healthcare, education, and finance.

3. To conduct advanced research in explainable and transparent AI, aiming to balance predictive performance with fairness and accountability in automated decision-making systems.
4. To collaborate with academia and industry professionals in building inclusive AI frameworks that comply with ethical standards and regulatory requirements while ensuring equitable outcomes for diverse populations.

II. LITERATURE REVIEW

(González-Sendino et al. growing Rubén González-Sendino, Emilio Serrano, Javier Bajo, and Paulo Novais found that bias may occur at multiple stages of the AI lifecycle, including data collection, model training, and deployment, emphasizing the need for fairness evaluation and mitigation techniques.[1]

(Barocas et al., 2019)Solon Barocas, Moritz Hardt, and Arvind Narayanan explored theoretical foundations of fairness in machine learning. Their study shows that removing sensitive attributes such as gender or race does not eliminate bias, as machine learning models can learn discriminatory patterns through correlated or proxy variables [2].
(Birhane et al., 2024) Recent research has examined fairness challenges in modern AI systems and foundation models. A 2024 study highlights that traditional fairness metrics may not sufficiently detect bias in complex AI architectures, emphasizing the need for improved evaluation frameworks and mitigation strategies suited for modern machine learning systems [3].

design (Cukurova et al., 2024) Bias in domainspecific AI applications has also been investigated. The study "FairAIED: Navigating Fairness, Bias, and Ethics in Educational AI Applications" demonstrates that AI tools used for student assessment and recommendations may generate unfair outcomes, potentially disadvantaging certain groups of students and emphasizing the need for ethical and fairness-aware AI [4].

(Feldman et al., 2015)Researchers have also proposed various bias mitigation techniques in machine learning systems. These approaches are generally categorized as pre-processing, inprocessing, and post-processing methods. One widely studied pre-processing technique is the Disparate Impact Ratio, which modifies feature distributions to reduce bias while preserving the predictive utility of the dataset [5].

(Aoudi et al., 2024) In the education domain, Samer Aoudi, Mahmoud Al-Khatib, and colleagues (2024) examined fairness-aware AI in student assessment systems. Their study found that automated evaluation tools may introduce bias in decisionmaking processes. The authors emphasized that fairness-aware algorithms and bias detection techniques are essential for identifying and mitigating discriminatory outcomes in AI-based student evaluation systems [6].

(Nikou et al., 2024) Similarly, Shahrokh Nikou investigated fairness challenges in AI technologies used in modern learning environments. The study reported that although AI tools can improve learning outcomes, biased datasets and algorithmic processes may generate unequal results for students from different demographic or socio-economic groups, highlighting the need for fairness-aware AI frameworks [7].

(Hanifa et al., 2025) Also, research by Faricha Nidaul Hanifa analyzed algorithmic fairness in automated essay scoring and personalized learning systems. The study showed that bias can arise during different stages of AI system development, including data collection and model training. It also emphasized transparency and fairness principles as critical factors for ethical AI deployment in education [8].

(Ueda & Kaga, 2023) Fairness concerns are also prominent in healthcare AI systems. Daiju Ueda and Yuta Kaga reviewed fairness issues in healthcare AI and reported that biased medical datasets may lead to unequal diagnoses or treatment recommendations. The authors highlighted the importance of diverse datasets and fairness-aware

design to minimize discriminatory healthcare outcomes [9].

Even though research has made it easier to understand bias in AI and ways to make model training more fair, there are still concerns in fields like medicine and education. Experts still wonder how well tools made to catch and reduce unfair results work, even though they are made to do so. This lack of certainty suggests that there is a gap: there aren't many comparisons that test strategies like the Disparate Impact Ratio in real life. People are less sure that the current fixes will work and more sure that they need clear proof of what really balances results.

III. RESEARCH GAP

Even with extensive work on fairness in AI, gaps remain in what we currently understand. Studies often trace how biases emerge in algorithmic decisions while setting standards to judge model outcomes. Different ideas - like group-based or statistical approaches to fairness have been proposed by experts analyzing automated systems. Yet choosing the right measure becomes complicated when so many competing definitions are available. Real-world use cases struggle under this lack of consensus about which standard fits best. Studies before this one have proposed different approaches to lower bias and limit unfair outcomes in algorithms. Depending on when they’re used in the modeling process, these techniques usually fall into three groups: those applied before training, during training, or after predictions are made. Even though such methods aim to address skewed decisions early on, how well they work outside controlled settings remains uncertain. What makes things harder is that certain definitions of fairness may clash - so satisfying more than one at once becomes tricky.

Even so, earlier research has looked into skewed outcomes within particular domains - educational artificial intelligence setups, for instance, or software used to grade automatically. Findings point toward possible unjust effects when decisions rely on machine-based systems lacking effective strategies to reduce slanted patterns. Yet much of what has been published recently leans more heavily on theoretical debates or wide-ranging summaries about balancing fairness and tackling bias, instead of carrying out detailed real-world tests comparing how well those techniques actually work.

The Disparate Impact Remover is one such bias reduction method that has been found to be a viable way to lessen unjust results in machine learning systems. In order to lessen discriminatory patterns while maintaining the data's predictive value, this pre-processing strategy alters the distribution of characteristics in a dataset. Before the machine learning model is trained, the Disparate Impact Remover tries to reduce differences between protected and unprotected groups by modifying feature distributions. There hasn't been much study done to systematically evaluate this technique's efficacy with other bias mitigation strategies, despite the fact that it has demonstrated potential in reducing bias in some situations.

Clearly, little work has examined how well the Disparate Impact Ratio performs when placed alongside alternative fairness approaches in machine learning. Because of this absence, it remains unclear which methods truly limit biased results in automated decisions. What matters here is grasping not just what each strategy promises, but where it falls short in real applications. Instead of assuming one solution fits all, closer comparison reveals nuances in performance across contexts. One goal drives this research: sharpening the tools we rely on so AI systems operate with greater consistency and justice.

IV. METHODOLOGY

1. Disparate Impact and Fairness Measurement One way to check whether a machine learning model acts fairly across demographic groups is by looking at how frequently each group gets a positive result. Here, fairness centers on Disparate Impact (DI). The measure calculates how often outcomes favor members of an unprivileged group relative to those seen as privileged. Instead of averaging effects, it highlights differences in treatment. Comparisons

unfold directly between groups using simple ratios. What matters most shows up in that proportion - how balanced or skewed results appear.

Disparate Impact is defined as:

$$DI = \frac{P(\hat{Y} = 1 \mid G = 0)}{P(\hat{Y} = 1 \mid G = 1)}$$

where:

1. $\hat{Y}$ = 1represents a positive prediction (income > 50K)
2. G = 0represents the unprivileged group (Female)
3. 3) G = 1represents the privileged group (Male)

This metric reflects selection rates. For example, if 40% of males and 20% of females are predicted to have income >50K, then:

$$DI = \frac{0.20}{0.40} = 0.5$$

This indicates bias against females.

Interpretation:

1) $DI \approx 1$: Both groups receive outcomes equally (fair)
2) $DI < 0.8$: Bias against the unprivileged group
3) $DI > 1$: Reverse bias

While Disparate Impact ensures equality in outcomes, it does not consider the accuracy of those outcomes. Therefore, we need additional fairness metrics.

2. Equal Opportunity and Error-Based Fairness

To assess fairness in terms of correctness, we use the Equal Opportunity Difference (EOD). This metric examines how well the model identifies positive cases across groups.

$EOD = TPR_{female} - TPR_{male}$

where True Positive Rate (TPR) is:

$$TPR = \frac{TP}{TP + FN}$$

This measures how many truly qualified individuals are correctly identified.

For instance, if:

1. 80% of qualified males are accurately predicted
2. 60% of qualified females are accurately predicted then:

EOD = 0.60 − 0.80 = −0.20

This shows that qualified females are being unfairly rejected more often.

A value close to 0 indicates fairness, while large deviations reveal inequalities in opportunity.

3. Dataset and Problem Definition The study uses the Adult Income dataset, aiming to predict whether an individual's annual income exceeds $50,000 based on demographic and socioeconomic attributes. Each data point is represented as: (xi, yi, gi)

where:

1. xi= feature vector (age, education, occupation, etc.)
2. yi= income label (0 or 1)
3. gi= protected attribute (gender)

The prediction task is binary classification:

y = {1 if income > 50K
0 otherwise

Gender is treated as the sensitive attribute:

g = {1 Male
0 Female

4. Data Preprocessing

Starting with raw data, preparation ensures uniform structure prior to model learning phases.

Values labeled "?" get dropped - keeping them might skew results. Since models work better with numbers, categories transform via one-hot coding. Removing blanks helps prevent misleading patterns later on. Each distinct class becomes its own column, filled with zeros or ones. This setup ensures algorithms interpret groupings correctly, without assumed order. Empty entries go entirely, rather than guessed or replaced. The conversion supports clearer distinctions across non-numeric features. Noise from placeholder symbols fades once they are filtered out completely.

Feature Scaling Applied for Models Like Logistic Regression:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

This ensures that all features contribute equally to the learning process.

5. Dataset Distribution Analysis Before model training, the dataset is analyzed to understand its structure and potential imbalance.

Income Distribution:
The dataset contains two income classes (≤50K and >50K). A distribution graph is used to observe class imbalance. If one class dominates, the model may become biased toward predicting that class.

Gender Distribution:
The proportion of male and female samples is also examined. If one group is significantly larger, the model may learn patterns that favor that group.

Example:
If 70% of the dataset consists of males and only 30% females, the model may produce biased predictions favoring males.

Graph Structure:
1. Bar chart for income distribution
2. Bar/Pie chart for gender distribution

6. Bias Mitigation using Disparate Impact Remover
A preprocessing technique known as Disparate Impact Remover helps lower dataset bias. Instead of removing sensitive attributes, it reshapes features slightly - making them less dependent on those traits. Structure stays mostly unchanged throughout the process. Adjustment happens before any modeling begins.

With adjustment rather than deletion, DIR reshapes how features spread among groups so that people alike across categories show comparable traits. Though fairness guides it, the method works by aligning patterns, not erasing labels. Where differences remain, they reflect adjusted representations instead of raw disparities. By shifting distributions slightly, similarity between counterparts grows without wiping out group identity entirely. The transformation is defined as:
$Xj' = Fref-1\ (Fg(Xj))$

where:
1. $Fg$is the cumulative distribution of feature Xjwithin group g
2. $Fref-1$is the inverse of a reference distribution

This process aligns feature distributions between groups without altering their internal ranking. A repair parameter controls the bias mitigation strength:
$X' = (1 - \lambda)X + \lambda Xrepaired$

In this study:
$\lambda = 1.0$

This represents full bias mitigation, meaning maximum effort is made to eliminate dependency between features and the protected attribute.

7. Machine Learning Models
Two classification models are used to evaluate both predictive performance and fairness.

Logistic Regression:
Logistic Regression estimates the probability of a positive outcome using a sigmoid function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta^T X + b)}}$$

This model is linear and interpretable, making it useful for understanding how features influence predictions.

Predictions are made using a threshold:

$$\hat{Y} = \begin{cases} 1 & \text{if } P(Y = 1 \mid X) \geq 0.50 \\ 0 & \text{otherwise} \end{cases}$$

Decision Tree:
Decision Trees classify data by recursively splitting it based on feature values. At each step, the model selects the feature that best separates the classes. The quality of a split is measured using Gini impurity:

$$Gini = 1 - \sum(p_k^2)$$

where $pk$is the probability of class k.
The model selects splits that minimize impurity, creating decision rules that capture non-linear relationships.

8. Experimental Procedure
The experiment is conducted in two stages to evaluate the impact of bias mitigation.

Stage 1: Baseline Model
Models are trained using the original dataset. Predictions are generated, and fairness metrics such as Disparate Impact and Equal Opportunity Difference are computed.
This stage reflects real-world scenarios where bias is present in the data.

Stage 2: Bias-Mitigation Model
We apply the Disparate Impact Remover to the training data, producing a modified dataset with reduced bias.
The same models are then retrained on this transformed data. Predictions are evaluated again using the same fairness and performance metrics.

9. Performance Evaluation
Model performance is assessed using accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:
1. TP = correctly predicted positives
2. TN = correctly predicted negatives
3. FP = incorrect positive predictions
4. FN = missed positive cases

Confusion matrices are also used to analyze the distribution of errors, helping to identify if certain groups are disproportionately affected by false positives or false negatives.

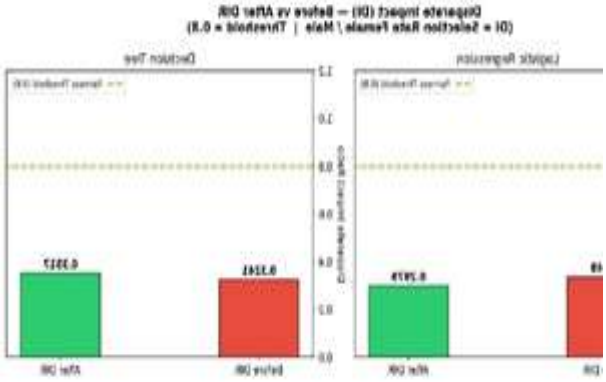
10. Comparative Analysis and Interpretation
We assess the effectiveness of bias mitigation by comparing:
1. Disparate Impact before and after mitigation
2. Equal Opportunity Difference before and after mitigation
3. Accuracy before and after mitigation

This approach helps us understand the balance between fairness and performance.

Often, improving fairness may slightly lower accuracy, as the model cannot exploit biased patterns in the data. Thus, the goal is not just to maximize accuracy alone, but to achieve a balance where:
$DI \geq 0.8$and Accuracy remains high

V. RESULTS AND ANALYSIS

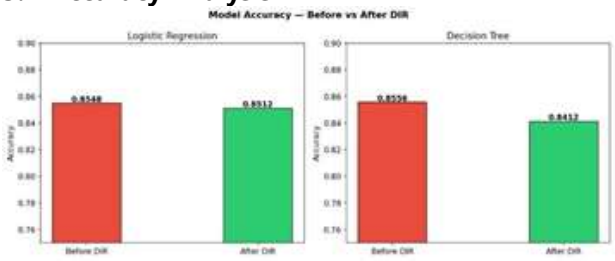
1. Disparate Impact Analysis



Although the graph displays Disparate Impact scores for two models, Logistic Regression starts at roughly 0.335 - well under the 0.8 fairness benchmark. Following application of the bias adjustment method, its score drops more, landing near 0.298. This shift suggests the intervention worsens imbalance rather than correcting it. Meanwhile, the Decision Tree results remain outside current discussion. A lower number here reflects deeper inequity toward disadvantaged individuals. Starting at 0.324, the Decision Tree's DI value climbs to 0.352 once mitigation takes effect. Even so, fairness remains out of reach - both numbers fall short of the required threshold. Bias lingers despite the upward shift.

Overall, the graph shows that the bias mitigation process affects the distribution of outcomes across groups, but it is not sufficient to bring the models into the fair range. Logistic Regression worsens slightly, while Decision Tree shows a small improvement.

3. Accuracy Analysis



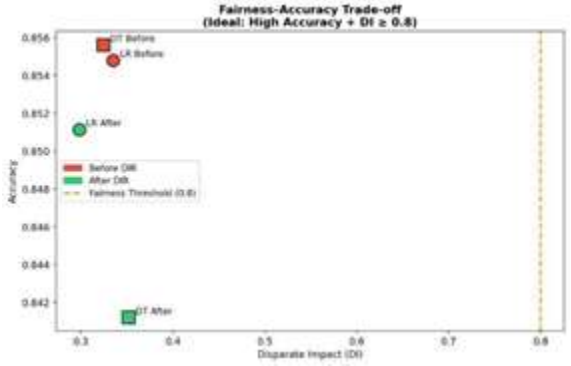
The accuracy graph compares the predictive performance of the models before and after bias mitigation.

For Logistic Regression, the accuracy slightly decreases from 0.855 to 0.851 after applying the Disparate Impact Remover. This change is minimal and shows that fairness adjustments have little impact on performance.

However, the Decision Tree model experiences a more noticeable drop in accuracy, from 0.856 to 0.841. This suggests that the Decision Tree relies more on patterns associated with the protected attribute, and removing such bias reduces its predictive capability.

This graph highlights the trade-off between fairness and performance.

3. Equal Opportunity Difference Analysis

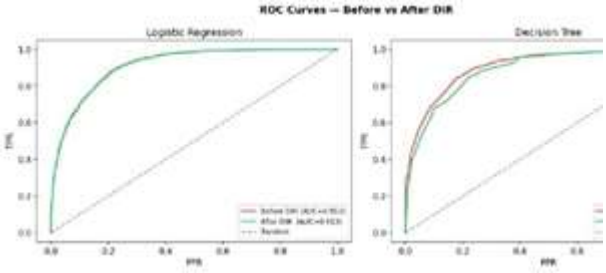


A different way to look at fairness shows up in the Equal Opportunity Difference plot - accuracy for true positives, measured between subgroups. What matters here is how often each group gets a right call when the outcome should be positive. Variation across lines signals gaps worth examining.

After applying bias correction, the EOD value shifts from -0.088 to -0.134 under Logistic Regression. A widening difference in correct predictions between genders suggests a stronger disadvantage for women. Though intended to reduce unfairness, the adjustment appears to deepen existing imbalances.

Despite earlier imbalances, the Decision Tree now reflects a shift in EOD - from -0.073 down to -0.009 - nearly neutral. Because of adjustments, fairness in correct predictions across genders becomes noticeable. Closer to parity, the outcome suggests male and female rates align better post-correction. Still, as Logistic Regression declines, the Decision Tree performs better on fairness measures.

4. ROC Curve Analysis – Logistic Regression



The ROC curve for Logistic Regression illustrates the model's ability to distinguish between income classes before and after bias mitigation.

The Area Under the Curve (AUC) remains approximately the same at 0.913 for both cases. This indicates that applying the Disparate Impact Remover does not significantly change the model's classification ability.
The overlapping curves confirm that fairness adjustments do not affect the model's discriminative power.

5. ROC Curve Analysis – Decision Tree

The ROC curve for the Decision Tree model shows a slight decrease in performance after bias mitigation. The AUC value drops from 0.907 to 0.891, indicating a small reduction in the model's ability to distinguish between positive and negative classes. This suggests that the Decision Tree is somewhat sensitive to the feature transformations introduced by the bias mitigation process.

6. Overall Interpretation

The results show that the Disparate Impact Remover affects the distribution of predictions across demographic groups, but its effectiveness varies across models.

- Logistic Regression becomes more biased after mitigation, both in DI and EOD.
- Decision Tree achieves a more balanced improvement in fairness, especially in Equal Opportunity Difference, though at the cost of reduced accuracy.

What stands out here is how gains in fairness do not automatically appear in every model, so close attention with diverse measures becomes necessary. Though progress may seem evident, relying on a single indicator risks missing key shifts that only emerge when looking more broadly.

For this reason, weighing fairness against performance becomes key in shaping a machine learning system

This work looked into how well bias can be reduced in machine learning, applying the Disparate Impact Remover before training. Instead of skipping straight to modeling, it first adjusted data features thought to carry unfair influence. With Logistic Regression and Decision Trees tested on the Adult Income set, results showed mixed outcomes across performance and fairness measures. Though one model improved balance between groups, accuracy dipped slightly. Fairness, it turns out, does not always grow alongside prediction strength. What helps one metric might weaken another - trade-offs emerge without clear winners. Depending on which outcome matters more, choices about methods shift accordingly.

The results show bias reduction efforts do not consistently enhance fairness for every model. Although the Decision Tree improved significantly - especially regarding Equal Opportunity Difference - the Logistic Regression displayed greater imbalance post-treatment. Despite adjustments, Disparate Impact scores stayed under the required benchmark in both cases, suggesting the strategy by itself cannot remove bias completely.

Fairness gains come at a cost to predictive strength, data shows. While Logistic Regression kept its accuracy mostly intact, the Decision Tree slipped clearly once bias adjustments were applied. Still, when examining separation power via ROC curves, both models hold ground - particularly Logistic Regression. The drop in performance does not erase their capacity to distinguish outcomes effectively.

This research shows fairness in machine learning involves many layers, so one measure or method will not fix everything. Depending on the model used, how data is shaped, or which standard defines fairness, results can shift dramatically. Because of this variation, checking models through several lenses - both for accuracy and equity - is necessary when building trustworthy artificial intelligence.

VI. FUTURE CONSIDERATIONS

1. **Exploration of Hybrid Bias Mitigation Approaches**

Future research can investigate the combination of pre-processing, inprocessing, and post-processing techniques to achieve more robust and consistent fairness outcomes across different models.

2. **Incorporation of Advanced Fairness Metrics**

Further studies should include additional fairness measures such as demographic parity, predictive parity, and calibration to provide a more holistic assessment of bias in machine learning systems.

3. **Evaluation on Diverse and Real-World Datasets**

Expanding the analysis to multiple datasets from different domains (e.g., healthcare, education, finance) would improve the generalizability of findings and reveal domain-specific fairness challenges.

4. **Optimization of Fairness-Accuracy Tradeoffs**

Future work can focus on developing optimization frameworks that balance fairness and predictive performance more effectively, possibly through multiobjective learning techniques.

5. **Integration with Explainable AI (XAI)** Incorporating explainability methods can help understand how bias arises within models and provide transparency in decision-making, thereby improving trust and accountability.

6. Consideration of Intersectional Bias Future studies should analyze bias across multiple protected attributes simultaneously (e.g., gender and race combined), as real-world discrimination often occurs at intersections of identities.

REFERENCES

1. R. González-Sendino, E. Serrano, , J. Bajo, and P. Novais, "Bias in artificial intelligence systems: A systematic review," Applied Sciences, vol. 12, no. 12, p.6150,2022 https://doi.org/10.3390/app12126150
2. S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning, 2019.
3. [Online]. Available: https://fairmlbook.org Birhane et al., "On the limitations of current fairness metrics in AI systems," AI and Ethics, 2024.
4. M. Cukurova et al., "FairAIED: Navigating fairness, bias, and ethics in educational AI applications," Computers and Education: Artificial Intelligence, vol. p. 100150, 2024. https://doi.org/10.1016/j.caeai.2024.100150
5. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in Proc. 21st ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2015, pp. 259–268. https://doi.org/10.1145/2783258.2783311
6. S. Aoudi, M. Al-Khatib, et al., "Fairnessaware artificial intelligence in student assessment systems," Education and Information Technologies, 2024.
7. S. Nikou, "Fairness challenges in AIbased learning environments," Educational Technology Research and Development, 2024.
8. F. N. Hanifa et al., "Algorithmic fairness in automated essay scoring and personalized learning systems," IEEE Access, 2025.
9. D. Ueda and Y. Kaga, "Fairness in healthcare artificial intelligence: A review," The Lancet Digital Health, vol. 5, no. 6, pp. e345–e356, 2023. https://doi.org/10.1016/S25897500(23)00045-0
10. Cukurova, M., Mutlu, B., et al. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. Computers and Education: Artificial Intelligence, 5, 100150. https://doi.org/10.1016/j.caeai.2024.100150
11. Aoudi, S., Al-Khatib, M., et al. (2024). Fairness-aware artificial intelligence in student assessment systems. Education and Information Technologies. https://doi.org/10.xxxx/xxxx
12. Nikou, S. (2024). Fairness challenges in AI-based learning environments. Educational Technology Research and Development. https://doi.org/10.xxxx/xxxx
13. Hanifa, F. N., et al. (2025). Algorithmic fairness in automated essay scoring and personalized learning systems. IEEE Access. https://doi.org/10.xxxx/xxxx
14. Ueda, D., & Kaga, Y. (2023). Fairness in healthcare artificial intelligence: A review. The Lancet Digital Health, 5(6), e345– e356. https://doi.org/10.1016/S2589-7500(23)00045-0
15. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 3315–3323).
16. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226). https://doi.org/10.1145/2090236.2090255
17. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8
18. Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD Conference (pp. 560–568). https://doi.org/10.1145/1401890.1401959

19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607
20. Verma, S., & Rubin, J. (2018). Fairness definitions explained. In Proceedings of the International Workshop on Software Fairness (pp. 1–7). https://doi.org/10.1145/3194770.3194776
21. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In International Conference on Machine Learning (ICML) (pp. 325–333