**AI Ethics & Bias Evaluation Report**

**Project:** Task 5 : AI Ethics and Bias Evaluation
**Intern:** Ghani Abdul Rehman Khan
**Organization:** Intern Intelligence

## 1. Introduction

The goal of this project was to evaluate an AI model for fairness and bias, and to apply mitigation techniques to improve fairness. For this purpose, I used the **Adult Income Dataset** from the UCI Machine Learning Repository, which predicts whether a person earns more than $50K per year.

This dataset is widely studied because it contains **bias against women and minority groups**.

## 2. Dataset

- **Dataset Size:** 45,222 rows × 98 features
- **Protected Attribute:** Sex (Male = privileged, Female = unprivileged)
- **Target:** Income (>50K or ≤50K)

## 3. Methodology

1. Loaded dataset using **AI Fairness 360**.
2. Trained a baseline **Logistic Regression** model.
3. Evaluated fairness using:
    - **Accuracy Difference** (Male vs Female)
    - **Disparate Impact (Female/Male)**
4. Applied **Reweighing** as a bias mitigation technique.
5. Compared before vs after results.

## 4. Results

**Before Reweighing:**

- Accuracy: **85%**
- Accuracy Difference (Male vs Female): **-0.115**
- Disparate Impact: **0.276** (unfair – <0.8 threshold)

**After Reweighing:**

- Accuracy: **79%**
- Accuracy Difference: **-0.149**
- Disparate Impact: **0.526** (improved fairness, but still unfair)

**5. Discussion**

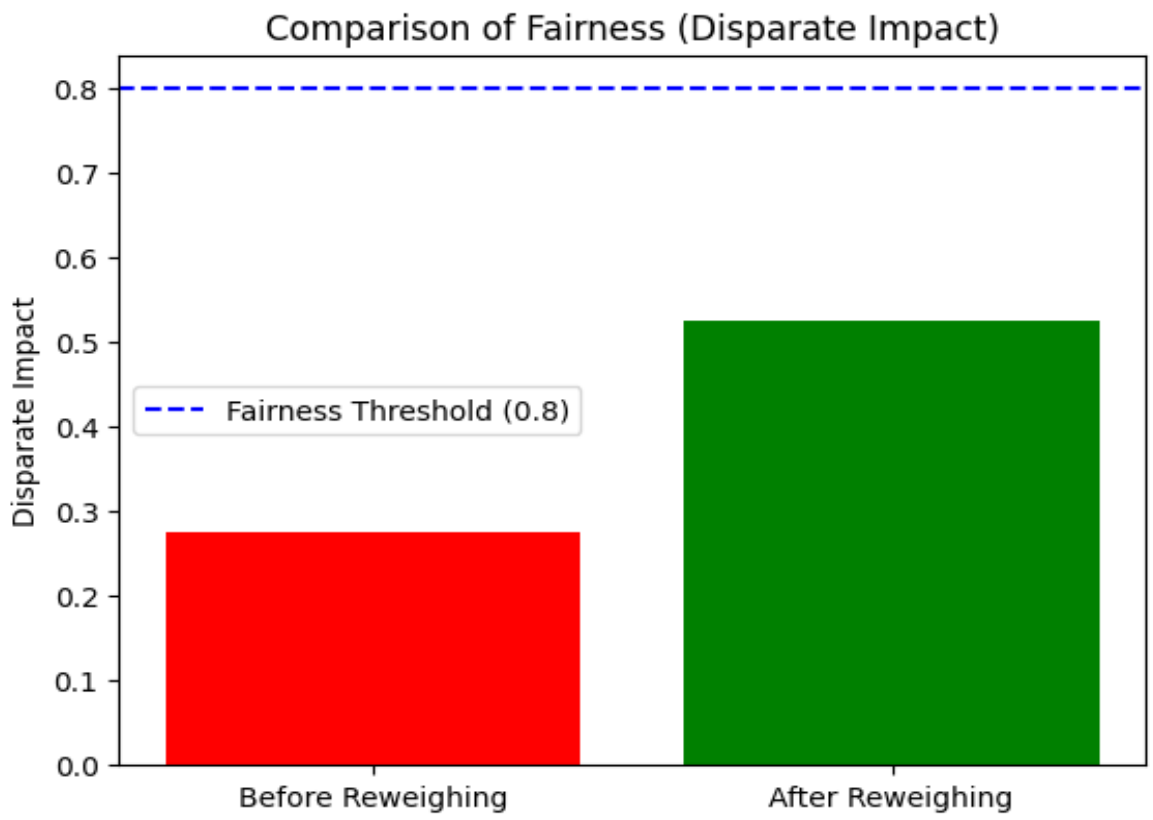
- The baseline model achieved high accuracy but showed **significant gender bias**.
- After applying **Reweighing**, fairness improved (Disparate Impact increased from 0.27 → 0.52).
- However, there was a **trade-off**: accuracy dropped slightly.
- This highlights a key challenge in **AI Ethics**: balancing performance with fairness.

**6. Recommendations**

- Always test AI models for fairness across sensitive groups.
- Use fairness metrics (Disparate Impact, Equal Opportunity, etc.) alongside accuracy.
- Apply bias mitigation techniques like **Reweighing, Adversarial Debiasing, or Prejudice Remover**.
- Document ethical considerations and limitations when deploying AI.

**7. Conclusion**

This project demonstrates that AI models can unintentionally discriminate against certain groups, even when achieving high accuracy. Bias mitigation techniques help reduce unfairness, but they may reduce accuracy. Ethical AI development requires careful evaluation of both accuracy and fairness before deploying models in real-world applications.


Comparison of Fairness (Disparate Impact)
Disparate Impact
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Fairness Threshold (0.8)
Before Reweighing
After Reweighing