**Analyzing and Mitigating Bias in Machine Learning: UCI Adult Income Dataset**

**Module 3 Assignment Report**

**Introduction**

This project focused on analyzing and reducing bias in the UCI Adult Income Dataset using a logistic regression model. The dataset includes demographic details like age, gender, race, and education to predict whether a person earns more than $50K per year. The goal was to detect unfair patterns in the model’s predictions and apply a mitigation technique to make it fairer.

**Data Exploration and Bias Detection**

The dataset had over 32,000 samples. Most were White (86%) and male (67%), with about 24% earning over $50K.  
When exploring the data, clear signs of bias appeared:

* **Gender:** 31% of males earned >$50K compared to only 11% of females — a clear gender gap.
* **Race:** White and Asian groups had higher incomes compared to Black and other minority groups.
* **Age:** People between 30–50 tended to earn more.

These trends likely reflect real-world inequalities that ended up embedded in the data.

**Model Training and Baseline Results**

After encoding categorical variables and splitting the data (70% train / 30% test), a logistic regression model achieved around **80% accuracy**. However, the confusion matrix showed the model was better at identifying low-income individuals than high-income ones.

**Fairness Metrics**

* **Demographic Parity (DP):** -0.20 difference between males and females — meaning the model favored men.
* **Equalized Odds (EO):** True Positive Rate (TPR) and False Positive Rate (FPR) differences showed that women and minorities had fewer positive predictions even when qualified.

This confirmed that the model learned and reinforced gender and racial biases.

**Bias Mitigation and Re-evaluation**

A **reweighting technique** was applied, giving more weight to female samples (unprivileged group).  
After retraining:

* **Accuracy** stayed about the same (~80%).
* **DP improved from -0.20 to 0.00**, removing gender bias.
* **Disparate Impact** rose from 0.36 to 1.00
* **EO differences** also improved slightly.

So, the model became fairer without losing performance.

**Findings and Insights**

The original model showed clear gender and racial bias, mainly favoring men and majority groups.  
Reweighting effectively reduced this bias and improved demographic parity while keeping accuracy stable.  
The trade-off was minor: some small shifts in TPR between groups, but overall, the fairness gains were more important.

**In short**, reweighting proved to be a practical way to make the model fairer without sacrificing predictive performance.