Summary Report

COMPAS Dataset Bias Audit Summary

We conducted a fairness audit on the COMPAS dataset using the AI Fairness 360 toolkit. The goal was to detect racial bias in risk predictions for recidivism. Our analysis focused on three key metrics: **False Positive Rate Difference (FPRD)**, **Equal Opportunity Difference (EOD)**, and **Disparate Impact (DI)**.

Findings: The model exhibited a **significant disparity in false positive rates** between African-American and Caucasian defendants, with African-Americans more likely to be incorrectly labeled as high-risk. This violates the ethical principle of **non-maleficence**, potentially leading to unjust sentencing.

The **Equal Opportunity Difference** further confirmed that the model performs unequally in terms of true positive rates — therefore unprivileged groups would be disadvantaged. Additionally, a **Disparate Impact value** far from 1.0 indicates overall unfair treatment in outcome distributions.

Remediation Strategy: To reduce this bias, we applied the **Reweighing** technique to the training dataset. This adjusts sample weights to counteract historical imbalances. We retrained a logistic regression classifier on this reweighed data and observed a **reduction in bias metrics**, though some disparity remained.

Recommendations: While reweighing is a good first step, we recommend combining it with other mitigation strategies such as:

- Fairness-aware classifiers (e.g., Adversarial Debiasing)
- Transparent model types with explainability (e.g., decision trees)
- Regular audits with evolving datasets to monitor drift

Ultimately, bias mitigation must be accompanied by **policy, oversight, and human review** to ensure justice in AI systems used in high-stakes areas like criminal justice.