Part 3: Practical Audit Report - COMPAS Recidivism Bias

Summary of Findings

This audit, simulated using principles from the **AI Fairness 360 (AIF360)** toolkit, analyzed a conceptual dataset mirroring the known racial disparities in the COMPAS recidivism risk assessment tool. The objective was to quantify bias against the unprivileged group (Black defendants, coded as 0) relative to the privileged group (White defendants, coded as 1). The 'favorable' outcome was defined as a **Low-Risk** score (Non-Recidivism).

The analysis yielded two critical metrics:

1. **Disparate Impact Ratio (DIR):** The calculated DIR was significantly below the typical fairness threshold of (e.g., in our simulation). This indicates a strong **statistical disparity**, meaning the unprivileged group (Black defendants) was assigned the favorable outcome (Low-Risk) at a rate much lower than the privileged group. This confirms the systemic practice of assigning high-risk scores disproportionately to Black individuals, even before controlling for true recidivism.
2. **Equal Opportunity Difference (EOD):** The calculated EOD was substantially negative (e.g., in our simulation). The EOD measures the difference in the True Positive Rate (correctly predicting low-risk when the outcome is truly low-risk). This negative value signifies that the model was significantly **less accurate** at correctly identifying low-risk Black defendants compared to low-risk White defendants, demonstrating a crucial failure of equal opportunity.

Remediation Strategy

To address the observed bias, a **Pre-processing Mitigation** strategy, specifically **Reweighing** (a technique available in AIF360), is recommended.

Reweighing works by adjusting the weights of individual data points in the training set. It assigns higher weights to data points where the predicted outcome is currently unfair to the unprivileged group, effectively forcing the model to pay more attention to those cases during training. This creates a more balanced training distribution, thereby encouraging the resulting model to satisfy statistical parity and equal opportunity metrics before it is deployed. Post-mitigation, the model would be re-audited to confirm the DIR and EOD are within acceptable bounds (and).