AI-Fairness360 Report

The audit analyzed racial bias in the COMPAS recidivism dataset using IBM's AIF360 toolkit. Baseline measurements revealed disparity: the **statistical parity difference** and **disparate impact** indicated that the unprivileged group (Black defendants) experienced systematically different predicted risk rates compared to the privileged group (White defendants). Classifier-level metrics showed a notable **false positive rate (FPR) gap**, meaning the model labeled more unprivileged individuals as high-risk despite not reoffending — a dangerous outcome for fairness in criminal justice contexts.

Two remediation strategies were tested. First, **Reweighing** — a preprocessing technique that assigns sample weights to correct historical imbalances — was applied to the training data. Retraining with those weights reduced FPR disparities and improved parity metrics without large reductions in overall accuracy. Second, **post-processing** methods such as Equalized Odds were recommended (and are implementable) to adjust decision thresholds to equalize error rates across groups; this is especially useful when changing the model or dataset is infeasible.

Key findings: (1) Bias in the COMPAS dataset arises from both historical/legal-system artifacts (label bias) and differing base rates; (2) Reweighing can reduce some disparity but may not fully resolve tradeoffs between fairness definitions; (3) No single metric suffices — we recommend reporting statistical parity, disparate impact, FPR/FNR differences, average odds difference, and calibration by group.

Recommended remediation steps: remove or carefully handle features that act as proxies for race, run repeated fairness-aware training (inprocessing or preprocessing) with cross-validated fairness metrics, and if deployed, require human review of high-risk outputs. Finally, mandate transparent reporting, regular audits, and provide affected individuals avenues to contest decisions. Together, these measures reduce harm and increase the legitimacy of risk-assessment systems.