2. Report — Example Summary of Findings and Remediation

Bias Audit Report on COMPAS Recidivism Dataset

The COMPAS dataset was audited using IBM's AI Fairness 360 toolkit to analyze racial bias in the risk scores predicting recidivism. The dataset distinguishes between privileged (White) and unprivileged (African-American) groups.

Initial analysis revealed a significant disparity in false positive rates (FPR) between racial groups. The FPR for African-American defendants was substantially higher than for White defendants. This implies the model is more likely to incorrectly flag African-Americans as high risk, potentially leading to unfair judicial outcomes.

To address this, we applied a bias mitigation technique called *Reweighing*. This method adjusts instance weights during training to balance the representation and reduce bias without altering the underlying data. After retraining a logistic regression model on reweighted data, the disparity in false positive rates notably decreased, indicating improved fairness.

Visualizations before and after reweighing clearly demonstrate the reduction in bias. Although perfect parity was not achieved, reweighing mitigated some of the racial unfairness.

Remediation steps recommended:

1. **Data preprocessing with reweighing or similar fairness algorithms** to balance biased distributions.

2. **Regular bias audits** on deployed models, especially those impacting critical decisions like recidivism.

3. **Incorporate fairness constraints into model training** to actively reduce disparate impacts.

4. **Human oversight** in decision-making to catch potential algorithmic errors or biases.

In conclusion, while AI models like COMPAS can assist judicial decisions, continuous evaluation and bias mitigation are essential to prevent reinforcing systemic racial disparities.