**Bias Audit Report: COMPAS Dataset**

*Author: [Karabo J Masipa]*

**Objective:**

Using IBM's AI Fairness 360 toolkit, we performed a fairness audit of the COMPAS Recidivism dataset to evaluate racial bias in recidivism risk prediction.

**Findings:**

Disparities in treatment between racial groups were found by preliminary analysis. The False Positive Rate (FPR) was substantially higher for African-Americans (an underprivileged group) than for Caucasian people (a privileged group). This indicates that even though Black people did not commit crimes again, they were more likely to be wrongly classified as high risk. Bias against the underprivileged group was also evident in metrics like Disparate Impact and Statistical Parity Difference.

A bar chart showed that the FPR for those who were less fortunate was significantly higher than that of those who were privileged. This is a reflection of real-world outcomes where biased AI tools could lead to systemic inequality and incorrect parole decisions.

**Mitigation Strategy:**

In order to balance fairness during training, we used the Reweighing algorithm, a pre-processing method that modifies instance weights. We noticed a notable decrease in the FPR disparity following the implementation of reweighing and retraining the classifier. The gap between privileged and underprivileged groups shrank, indicating how mitigation strategies can increase equity.

**Conclusion & Recommendations:**

There is detectable racial bias in the COMPAS dataset and its application to predictive risk scoring. Deploying such models in high-stakes situations without fairness audits is crucial. Future research should guarantee diverse representation in training data and

incorporate post-processing methods like calibrated equalized odds. Responsible AI implementation in the criminal justice system requires openness and frequent audits.