COMPAS Recidivism Risk Assessment – Bias Audit Report

This audit evaluates the presence of racial bias in the COMPAS recidivism risk scoring algorithm using IBM's AI Fairness 360 toolkit. The dataset used in this analysis consists of defendants labeled as either African-American or Caucasian, with risk scores categorized as "High" or "Low".

Our initial fairness assessment revealed significant disparities in outcomes between racial groups. The **Statistical Parity Difference** and **Disparate Impact** metrics indicated that favorable outcomes (i.e., being labeled "Low risk") were more likely to be assigned to Caucasian defendants. The **False Positive Rate** was notably higher for African-American individuals, meaning they were more often misclassified as high-risk despite not reoffending.

A simple logistic regression model trained on the dataset performed adequately in terms of accuracy, but fairness metrics painted a more troubling picture. African-American defendants experienced disproportionately higher false positives and lower true positives than their Caucasian counterparts. This reinforces concerns that the algorithm may reinforce systemic racial disparities present in the criminal justice system.

To address these issues, we recommend a multi-pronged remediation strategy:

1. **Preprocessing**: Apply techniques like reweighing to reduce bias before training.

2. **In-processing**: Use fairness-aware algorithms, such as adversarial debiasing, during model training.

3. **Post-processing**: Implement threshold optimization to equalize error rates across groups.

4. **Monitoring**: Continuously track fairness metrics in production systems.

5. **Inclusive Oversight**: Engage legal experts, ethicists, and community stakeholders in reviewing algorithmic decisions.

In conclusion, while the COMPAS tool is widely used to inform critical decisions in the criminal justice system, this analysis highlights the urgent need to implement fairness-aware practices. Bias mitigation strategies are essential to ensure that such tools promote justice, rather than unintentionally reinforcing inequality.