Bias Audit Report on COMPAS Recidivism Dataset

The COMPAS dataset, widely used to predict the likelihood of recidivism, has been analyzed using the AI Fairness 360 (AIF360) toolkit by IBM to uncover potential racial bias.

Objective:

To assess and visualize the extent of racial disparities, particularly focusing on false positive rates between African-American and Caucasian individuals.

Findings:

The audit revealed a notable disparity in false positive rates (FPR). African-American individuals had a significantly higher FPR compared to Caucasian individuals. This suggests that Black individuals were more likely to be wrongly classified as high risk when they were not, exposing systemic racial bias in the model's predictions.

Statistical parity difference and disparate impact metrics further confirmed the presence of bias, indicating unequal treatment of individuals based solely on race.

Remediation:

To address these disparities, we applied the Reweighing algorithm, which adjusts the weights of training examples to reduce bias before training. This technique showed promising reductions in disparity metrics while maintaining acceptable levels of overall accuracy.

Conclusion:

The analysis emphasizes the importance of fairness audits in high-stakes decision-making systems. Using AIF360, we demonstrated the presence of racial bias in COMPAS predictions and

implemented a viable mitigation strategy. Future work should explore post-processing or in-processing algorithms to further improve fairness without sacrificing performance.