Bias Audit Report – COMPAS Dataset

In this audit, we used the COMPAS Recidivism Dataset and IBM's AI Fairness 360 toolkit to evaluate potential racial bias in predicting reoffending risk scores. The primary focus was on false positive rates—instances where the model predicted a high risk of recidivism, but the individual did not reoffend.

Our analysis revealed a significant disparity: the false positive rate for African-American individuals was substantially higher than that for Caucasian individuals. This indicates that African-Americans were more likely to be incorrectly labeled as high-risk, which could lead to unfair outcomes such as harsher sentencing or parole denial.

To address this, we applied the Reweighing algorithm, which adjusts instance weights in the training data to reduce bias. While reweighing improved fairness metrics, some disparities persisted, suggesting the need for deeper mitigation strategies, such as adversarial debiasing or post-processing corrections.

Visualizations clearly showed the gap in false positive rates across racial groups, reinforcing the importance of auditing AI models—especially in sensitive domains like criminal justice.

Remediation Recommendations:
Adopt pre-processing techniques (e.g., Reweighing, Disparate Impact Remover).

Include fairness-aware models during training (e.g., prejudice remover, adversarial debiasing).

Mandate fairness evaluation before deployment using tools like AIF360.

Establish policy oversight to review algorithmic decisions and ensure transparency.

In conclusion, ethical AI must go beyond accuracy—it must be accountable, fair, and trustworthy, especially when dealing with human lives and liberties.