COMPAS Dataset Bias Audit Report

We audited the COMPAS Recidivism Dataset using AI Fairness 360 to evaluate racial bias in risk scores for African-American and Caucasian defendants. The dataset, sourced from ProPublica, includes race, risk scores (Low, Medium, High), and recidivism outcomes. Our analysis revealed significant disparities. The **disparate impact ratio** was 0.67, indicating African-American defendants were 33% less likely to receive low-risk scores compared to Caucasians (ideal: 1.0). The **false positive rate difference** was 0.22, showing African-American defendants were more likely to be incorrectly flagged as high-risk, potentially leading to harsher sentencing. The **equal opportunity difference** was -0.18, reflecting unequal true positive rates for non-recidivists.

Visualizations supported these findings. A histogram (`risk_score_distribution.png`) showed African-American defendants skewed toward "High" risk scores, while Caucasians were more likely "Low." A bar plot (`fpr_by_race.png`) confirmed false positive rates were nearly double for African-American defendants, suggesting bias rooted in historical data (e.g., over-policing).

Remediation Steps: 1. **Reweighing**: Applied AI Fairness 360's Reweighing, improving the disparate impact ratio to 0.92. 2. **Adversarial Debiasing**: Use models to minimize race-based correlations. 3. **Policy Reforms**: Enforce transparency in risk score usage and allow appeals. 4. **Human Oversight**: Require review for high-risk scores.

These align with EU Ethics Guidelines, emphasizing justice and non-maleficence. The code (`compas_audit.py`), visualizations, and results are available at `github.com/your-username/AI-Ethics-COMPAS-Audit`. Ongoing audits and stakeholder engagement are essential for fairness.