# Bias Audit Report: COMPAS Recidivism Dataset Using AI Fairness 360

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset has garnered attention for potential racial bias in predicting recidivism risk. This audit applied the AI Fairness 360 (AIF360) toolkit to assess racial bias, focusing on disparities between African-American (unprivileged) and Caucasian (privileged) individuals.  
  
We began by preprocessing the dataset, filtering out incomplete or irrelevant records and encoding the protected attribute "race." The outcome variable used was whether an individual reoffended within two years (two\_year\_recid). A logistic regression model was trained to predict this outcome based on features such as age, sex, prior convictions, and race.  
  
To evaluate fairness, we used classification metrics from AIF360. The audit revealed a higher false positive rate (FPR) for African-Americans compared to Caucasians. Specifically, African-Americans were more likely to be incorrectly classified as high-risk, despite not reoffending—demonstrating a key indicator of bias. Additional metrics such as Equal Opportunity Difference and Average Odds Difference also confirmed this disparity.  
  
To address this, we applied the Reweighing algorithm, a pre-processing technique that adjusts sample weights based on protected attributes. After reweighing, the metrics showed reduced disparities, though not entirely eliminated. This highlights the difficulty of achieving both fairness and accuracy simultaneously.  
  
In conclusion, the COMPAS dataset exhibits measurable racial bias in risk predictions. While methods like reweighing help mitigate bias, broader efforts are needed in dataset design, feature selection, and policy-level oversight. Fairness-aware tools like AIF360 are critical in promoting transparency and ethical AI, particularly in high-stakes domains like criminal justice.