Title: *Bias Audit of COMPAS Recidivism Dataset Using AI Fairness 360*

Submitted by: James Njoroge

Summary:

In this audit, we evaluated the fairness of the COMPAS Recidivism dataset using IBM's AI Fairness 360 (AIF360) toolkit. The dataset is known to exhibit racial bias, particularly in predicting the likelihood of recidivism for Black versus White individuals. Our goal was to identify and mitigate potential disparities in false positive rates and other fairness metrics.

The initial analysis involved loading the dataset and computing the *disparate impact*, which measures the ratio of favorable outcomes between unprivileged (African-American) and privileged (Caucasian) groups. The original dataset exhibited a disparate impact below the generally accepted threshold of 0.8, indicating significant bias.

To address this, we applied the **Reweighing** bias mitigation technique, which assigns different weights to the training examples to reduce the influence of biased data distributions. We then trained a logistic regression classifier on the reweighed dataset and evaluated its predictions.

Post-mitigation results showed improved fairness, with the *disparate impact* moving closer to the ideal value of 1.0 and a reduction in the *false positive rate difference* between racial groups. This demonstrates that reweighing is an effective preprocessing step for enhancing fairness in predictive models.

Visualizations were generated to compare metrics before and after mitigation, helping stakeholders understand the improvements made. However, while reweighing helps, it does not eliminate all bias, and further interventions such as in-processing or post-processing methods might be necessary depending on deployment context.

In conclusion, this audit highlights the critical importance of dataset-level fairness assessments in the AI development pipeline. Without such scrutiny, models risk perpetuating historical injustices. Fair AI systems require both technical solutions and ethical oversight to ensure they serve all users equitably.