Bias Audit Report – COMPAS Recidivism Dataset

The COMPAS dataset, developed for predicting criminal recidivism risk, has faced criticism for racial bias in its predictions. Using IBM's **AI Fairness 360 (AIF360)** toolkit, this audit examined differences in predicted risk scores between African-American and Caucasian defendants.

The analysis began by defining **Caucasians as the privileged group** and **African-Americans as the unprivileged group**. Metrics such as **Disparate Impact (DI)** and **Mean Difference** were used to measure fairness. The initial results showed a **Disparate Impact below 0.8**, indicating substantial bias against African-American defendants. Visualization of positive outcome rates confirmed that African-American individuals were more likely to be classified as "high risk," even when they did not reoffend — implying a **higher false positive rate**.

To address these disparities, the **Reweighing algorithm** from AIF360 was applied to rebalance the dataset by adjusting sample weights based on group membership and labels. After applying this method, the **Disparate Impact** value improved, showing a more equitable distribution of predictions between racial groups.

This audit highlights the ethical risks of using historical or biased data in criminal justice AI systems. Such bias can lead to **unfair treatment, loss of trust, and potential violation of human rights**. Therefore, any AI system in this domain must include regular bias testing, transparent reporting of fairness metrics, and human oversight during deployment. Additionally, agencies should adopt clear **governance frameworks** ensuring that predictive tools are used responsibly and do not reinforce systemic discrimination.

Conclusion:
The COMPAS dataset contains measurable racial bias, but bias mitigation techniques like **Reweighing** can significantly improve fairness. Continuous monitoring and transparency are essential for ethical AI use in policing and justice.