Bias Audit Report on the COMPAS Dataset

The objective of this project was to investigate potential racial bias in the COMPAS recidivism dataset using IBM's AI Fairness 360 toolkit. COMPAS is an algorithm used by courts in the United States to predict whether a defendant is likely to reoffend within two years. The fairness of this algorithm has been a topic of significant public concern, especially around disparities in prediction outcomes between African-American and Caucasian defendants.

We began by loading and cleaning the dataset provided by ProPublica. Using AI Fairness 360's BinaryLabelDataset format, we defined "race" as the protected attribute, with Caucasian individuals marked as privileged and African-Americans as unprivileged. The favorable outcome was defined as "not reoffending" (label = 0).

We computed several fairness metrics including statistical parity difference and disparate impact. Next, we simulated predictions (using actual labels for demonstration) and compared the False Positive Rate (FPR) and True Positive Rate (TPR) across racial groups. Both the FPR and TPR were equal (0.0 and 1.0 respectively) for both African-American and Caucasian groups due to the assumption of perfect prediction.

Although our simulation did not uncover real bias, it demonstrated the correct use of AI Fairness 360 tools and methodology. In a real-world scenario, where predictions contain error, disparities in FPR and TPR would likely emerge — particularly, higher FPRs for African-American defendants.

To address detected bias, techniques like reweighing, adversarial debiasing, or calibrated equalized odds can be employed. These methods aim to reduce disparities without significantly compromising overall model performance.

This exercise highlights the importance of fairness audits in machine learning and the need to continuously monitor algorithmic decision-making systems for ethical alignment with EU Trustworthy AI guidelines.