Part 3: 300-word report

Bias Audit Report: COMPAS Recidivism Dataset

The COMPAS recidivism dataset is widely used in criminal justice to predict the likelihood of offenders reoffending. Given the high-stakes nature of these predictions, ethical concerns around racial bias are critical. Using IBM's AI Fairness 360 toolkit, we conducted a bias audit to analyze disparities in risk scores between racial groups, particularly focusing on African-American (unprivileged) versus White/Other (privileged) individuals.

Initial analysis revealed that the **disparate impact ratio** was below 1, indicating that Black defendants were disproportionately classified as high risk compared to their White counterparts. Further investigation of classification errors showed a **higher false positive rate** for Black defendants, meaning they were more likely to be incorrectly predicted as high risk, which could result in unfair sentencing. The false negative rate difference indicated that White defendants were under-classified, highlighting further inequity in risk assessment.

To address these disparities, we applied **reweighing**, a pre-processing bias mitigation technique that adjusts the weights of instances in the training dataset to balance the influence of privileged and unprivileged groups. Post-reweighing metrics showed improved fairness, with reduced differences in false positive and false negative rates while maintaining reasonable predictive accuracy.

Remediation steps include:

1. **Bias Mitigation:** Implement pre-processing techniques like reweighing or in-processing methods such as adversarial debiasing to minimize disparities.

2. **Regular Auditing:** Continuously monitor models for emerging biases, especially when datasets are updated or deployed in real-world contexts.
3. **Transparency:** Document decision-making processes and provide explanations for predictions to stakeholders to ensure accountability.
4. **Policy Integration:** Align AI systems with legal and ethical frameworks to prevent harm, particularly in sensitive domains like criminal justice.

Conclusion: This audit highlights the importance of evaluating AI systems not only for accuracy but also for fairness. Proactive mitigation strategies, coupled with ongoing monitoring and transparency, are essential to prevent AI from perpetuating systemic inequalities in society.