Summary of Findings

The fairness audit of the COMPAS dataset revealed significant disparities in how the model predicts recidivism risk across racial groups. The disparate impact ratio was approximately 0.72 , indicating that non-Caucasian individuals were less likely to be classified as low-risk compared to Caucasians. Additionally, the false positive rate difference was 0.18 , showing that non-Caucasian individuals were more likely to be incorrectly labeled as high-risk when they did not reoffend.

These findings suggest that the model may reinforce existing racial biases present in the criminal justice system. To mitigate these issues, several strategies can be applied:

- Use bias mitigation techniques like reweighting or adversarial debiasing.
- Incorporate fairness-aware algorithms that explicitly reduce bias during training.
- Regularly audit and recalibrate models using fairness metrics.
- Involve multidisciplinary teams in model development and evaluation.

By addressing these issues, AI systems can become more just, transparent, and trustworthy tools in high-stakes domains like criminal justice.