Report Summary

The COMPAS dataset contains risk scores predicting recidivism, often used in judicial decision-making. Using IBM's AI Fairness 360 toolkit, I analyzed racial bias in risk scores between white and non-white defendants.

Initial analysis showed significant disparities: non-white defendants were more likely to be falsely labeled as high risk (higher false positive rate). Metrics like disparate impact (~0.73) and statistical parity difference indicated bias favoring white defendants.

To mitigate this, I applied reweighing , a preprocessing technique that assigns weights to instances to reduce bias. Post-transformation, disparities were reduced: disparate impact improved to ~0.98, and statistical parity difference dropped significantly.

Visualizations confirmed that the reweighted model produced more balanced risk predictions across races. However, while reweighing helps, it does not eliminate bias entirely. Future steps include applying in-processing and post-processing techniques like adversarial debiasing or calibration to further enhance fairness.

This audit underscores the importance of proactive fairness evaluation in high-stakes domains like criminal justice. Ethical AI requires continuous monitoring and adjustment to ensure equitable outcomes.

Part 4: Ethical Reflection (5%)

Prompt: Reflect on a personal project (past or future). How will you ensure it adheres to ethical AI principles?

In a future project involving a recommendation system for job seekers, I will ensure ethical AI use by:

- Bias Auditing : Regularly evaluating the model for disparities across gender, race, and age.
- Transparency : Providing clear explanations for recommendations and allowing users to opt out.
- User Consent : Ensuring explicit consent for data usage and giving users control over their data.
- Accountability : Maintaining logs of decisions and implementing mechanisms for user feedback and appeal.

Bonus Task: Policy Proposal – Ethical AI Use in Healthcare (1 page)

Guideline for Ethical AI Use in Healthcare

1. Patient Consent Protocols

- Obtain informed, explicit consent before collecting or using patient data.
- Provide clear explanations of how AI will be used, including risks and benefits.
- Allow patients to opt out or request data deletion where feasible.

2. Bias Mitigation Strategies

- Use diverse, representative datasets to train AI models.
- Conduct fairness audits across demographic groups (race, age, gender, etc.).
- Apply bias correction techniques during training and testing phases.
- Involve clinical experts and ethicists in model development.

3. Transparency Requirements

- Ensure explainability of AI decisions, especially in diagnosis and treatment.
- Maintain audit trails of model decisions and data sources.
- Disclose limitations and uncertainties of AI systems to healthcare providers and patients.
- Publish model cards with performance metrics, training data, and use cases.

Conclusion

These guidelines ensure AI systems in healthcare respect patient rights, reduce disparities, and operate transparently. Ethical AI in healthcare is not just a technical challenge but a moral imperative to protect vulnerable populations and build trust in digital health solutions.