Audit Report: COMPAS Dataset Bias Analysis

Objective: This audit analyzed the COMPAS recidivism risk score dataset for racial bias, specifically focusing on disparities between Caucasian and non-Caucasian defendants.

Findings:
The initial analysis confirmed significant racial bias within the dataset. The key metric, **Disparate Impact**, was far from the ideal value of 1.0, indicating an imbalanced outcome distribution. More critically, the evaluation of a model trained on this data revealed a substantial disparity in error rates. The **False Positive Rate (FPR)** for non-Caucasian defendants was significantly higher than for Caucasian defendants. This means the model was incorrectly labeling a disproportionate number of non-Caucasian defendants as "high-risk" who did not subsequently re-offend. This finding aligns with the real-world controversy surrounding COMPAS, where it was shown to be unfairly punitive towards Black defendants.

Remediation Steps:

1. **Data Pre-processing:** Techniques like **Reweighing** were applied, which assigns weights to training instances to compensate for the bias in the dataset. This successfully moved the Disparate Impact metric closer to 1.0.
2. **Model In-processing:** Algorithms like **Prejudice Remover**, which incorporates a fairness constraint directly into the learning objective, can be used to train a less discriminatory model from the outset.
3. **Post-processing:** The model's output thresholds can be adjusted differently for different demographic groups to equalize error rates, a technique known as **equalized odds post-processing**.

Conclusion:
The COMPAS dataset, as a product of a biased criminal justice system, contains deeply embedded racial biases that are learned and amplified by AI models. Mitigation is possible but requires active and continuous intervention. It is recommended that any deployment of such a system be accompanied by rigorous, ongoing bias audits, transparency in error rates, and should not be used as the sole basis for sentencing or detention decisions.

Part 4: Ethical Reflection (5%)

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

For a future project involving an AI-powered resume screening tool for a university's admissions office, ensuring ethical adherence is paramount. I would integrate ethics from the initial design phase, guided by the following principles:

First, **Fairness and Non-maleficence** would be addressed by conducting a thorough bias audit on the training data, which would consist of historical application records. I would use tools like AIF360 to check for disparate impact against protected attributes like gender, ethnicity, and socioeconomic background (using proxies like school ZIP code). Mitigation strategies like reweighting would be applied. The model would be designed to screen for academic merit and potential, not to replicate past demographic imbalances.

Second, **Transparency and Explainability** would be built-in. The model would not be a "black box." I would use SHAP or LIME to provide feature-level explanations for why an application was ranked a certain way. This allows admissions officers to understand the AI's reasoning and overrule it with justification, maintaining a human-in-the-loop.

Finally, I would embed **Justice and Accountability** by defining clear success metrics that include both predictive accuracy and fairness metrics (like equalized odds). A feedback loop would be established for rejected applicants to appeal, ensuring continuous monitoring for unintended consequences and upholding the principle of accountability for the system's outcomes.

Bonus Task (Extra 10%)

Policy Proposal: Ethical AI Use in Healthcare

1. Guiding Principles:
Our healthcare AI policy is founded on four pillars: **Patient Welfare (Beneficence), Privacy and Autonomy, Justice and Equity, and Transparency and Accountability.** All AI systems must enhance patient care, respect human dignity, and reduce—not exacerbate—health disparities.

2. Patient Consent Protocols:

- **Informed Consent:** Patients must provide explicit, informed consent before their data is used for AI model training or clinical decision-making. Consent forms must be clear, concise, and explain the AI's role, potential risks, and benefits in layman's terms.
- **Dynamic Consent:** Patients should have the right to withdraw their consent and data at any time through a manageable digital portal.
- **Tiered Consent:** Offer options for consent, allowing patients to opt-in for specific use-cases (e.g., for their direct care but not for research).

3. Bias Mitigation Strategies:

- **Diverse Data Procurement:** Actively seek training data from diverse populations across races, ethnicities, genders, ages, and socioeconomic statuses to ensure model generalizability.
- **Mandatory Pre-Deployment Audits:** All AI diagnostic or treatment recommendation tools must undergo rigorous, independent bias audits using standardized metrics (e.g., Disparate Impact, Equalized Odds) before clinical use.
- **Continuous Monitoring:** Establish an MLOps pipeline for continuous monitoring of model performance and fairness across different patient subgroups post-deployment, with triggers for re-training if performance disparities are detected.

4. Transparency Requirements:

- **Explainability by Design:** For any AI used in diagnosis or treatment planning, a "right to an explanation" is mandatory. Clinicians must be able to access and understand the primary factors behind an AI's recommendation.
- **Algorithmic Transparency:** High-level information about the AI's intended use, performance characteristics, and known limitations must be publicly disclosed to build trust.
- **Clinical Accountability:** The final decision-making authority must always rest with a qualified healthcare professional. The AI is a tool to augment, not replace, clinical judgment. All AI-assisted decisions must be documented in the patient's record, including the AI's recommendation and the clinician's rationale for following or diverging from it.