Question 6

Bias & Fairness Tools

False Negative Rate (FNR) Parity

What the Metric Measures:

- **False Negative Rate (FNR):** The proportion of positive instances (e.g., qualified candidates) incorrectly classified as negative (e.g., rejected) by the model. FNR = FN / (TP + FN), where FN is false negatives and TP is true positives.
- **FNR Parity:** Ensures that the FNR is equal across different demographic groups (e.g., gender, race). It measures whether the model is equally likely to miss positive cases for all groups.

Why It's Important:

- FNR parity is critical in applications where missing a positive outcome has significant consequences, such as hiring, loan approvals, or medical diagnoses. Unequal FNRs across groups can lead to unfair treatment (e.g., one group's qualified candidates are disproportionately rejected).
- It helps identify allocational harms, ensuring equitable access to opportunities or resources.
- Example: In a hiring model, if the FNR is higher for female candidates than male candidates, qualified women are unfairly rejected more often, perpetuating gender disparities.

How a Model Might Fail This Metric:

- A model might fail FNR parity if it is trained on biased data reflecting historical inequalities. For instance:
    - **Scenario:** A loan approval model is trained on data where minority applicants were historically denied loans due to systemic bias.
    - **Failure:** The model learns to reject minority applicants more often, resulting in a higher FNR for minorities (e.g., FNR = 0.4 for minorities vs. 0.2 for non-minorities).

    - **Impact:** Qualified minority applicants are disproportionately denied loans, exacerbating economic disparities.
- Causes of failure include biased training data, feature selection that correlates with protected attributes, or lack of regularization to enforce fairness.