Replication Instructions

This document provides exact steps to replicate our fairness testing results using the Colab notebook.

📦 Step 1: Install Dependencies

Run this in a Colab cell:

!pip install fairlearn scikit-learn pandas numpy matplotlib seaborn

📊 Step 2: Load and Preprocess Data

- Load the Adult Income dataset
- Label-encode categorical features
- Split into train/test

🧠 Step 3: Train Random Forest Classifier

- Train using scikit-learn
- Generate predictions on test set

🧪 Step 4: Evaluate Fairness

- Use Fairlearn's `MetricFrame` to compute:
  - Accuracy
  - Selection Rate
  - False Positive Rate
  - True Positive Rate

⚖️ Step 5: Apply Targeted Perturbation

- Flip `sex` and `race` columns
- Count how many predictions change

📊 Step 6: Intersectional Fairness

- Combine `race` + `sex` into a single feature

- Analyze fairness metrics by group
- Identify subgroups with large disparities

👤 Step 7: Run IDI Tests

- Flip one sensitive feature at a time per individual
- Measure % of individuals whose prediction changes

🛡️ Step 8: Apply Fairness Mitigation

- Use ExponentiatedGradient with demographic parity constraint
- Retrain model and evaluate improved fairness

📈 Step 9: Visualize Results

- Use matplotlib to generate fairness comparison charts

📁 All results are reproducible using the included Colab notebook.