Report: Fairness-Aware Classification on COMPAS Dataset Using Reweighing

Introduction

This project explores fairness in machine learning classification, specifically addressing bias related to race in criminal recidivism prediction using the COMPAS dataset. The goal is to build a logistic regression model that predicts whether a defendant will reoffend, while mitigating racial bias through a preprocessing fairness technique called **Reweighing**.

Dataset

The COMPAS dataset contains features about defendants and whether they reoffended. The sensitive attribute considered is **race**, simplified to a binary variable (privileged group: White = 1, unprivileged group: Black = 0). The target variable indicates if the defendant recidivated.

Methodology

1. **Data Loading and Preparation**
   The dataset was loaded using the fetch_compas() function from the AIF360 library, with binary race attribute. Features (X), target labels (y), and race were extracted.

2. **Train-Test Split**
   The data was split into training and testing sets (70% train, 30% test) with stratification to preserve distribution.

3. **Bias Mitigation using Reweighing**
   The Reweighing algorithm from AIF360's sklearn interface was applied to the training data to adjust sample weights. This technique assigns different weights to instances from privileged and unprivileged groups to compensate for bias during model training.

4. **Feature Scaling and Model Training**
   Features were scaled using StandardScaler. A logistic regression classifier was trained on the reweighed and scaled training data, utilizing the sample weights from the Reweighing step.

5. **Prediction and Evaluation**
   The trained model predicted outcomes on the test set. Fairness evaluation was performed using AIF360's ClassificationMetric, calculating key metrics:

   - **False Positive Rate Difference (FPRD):** Difference in false positive rates between privileged and unprivileged groups.

   - **Equal Opportunity Difference (EOD):** Difference in true positive rates between privileged and unprivileged groups.

6. **Visualization**
   The false positive rates for both racial groups were visualized via a bar chart to illustrate disparity.

While running the code, some errors were encountered that affected the full execution of the program. Specifically, the graphs visualizing the False Positive Rate by race group did not render as expected in the current environment. Despite this, the key fairness metrics (False Positive Rate Difference and Equal Opportunity Difference) were successfully calculated and printed.

These issues are likely due to compatibility or environment-specific display limitations. With more time or a different setup, the visualization could be fixed to better illustrate the fairness results.