Fairness in Machine Learning: Gender Bias Injection and Mitigation on Student Performance Prediction

Giovanni Bonadeo

May 9, 2025

Abstract

This project uses the Python package Fairlearn[1] to investigate fairness in machine learning algorithms, evaluating a classification model trained on the UCI Student Performance dataset. A baseline model was trained with a random forest classifier. An artificial bias was then introduced by selectively removing a portion of successful female students to simulate an imbalance in the dataset. To counteract this, a bias mitigation strategy was applied, in particular using the Exponential Gradient with a Demographic Parity constraint. The analysis compares the models before and after bias injection and mitigation, revealing the impact on fairness indicators such as demographic disparity and equalized odds. The results highlight the success of the functions in the Fairlearn library.

1 Introduction

The goal of this research is to show in a practical way the implementations of functions from the Fairlearn[1] package. In order to do so, a study of the most common functions has been made. The results that modern technologies can achieve will be shown, highlighting not only the accuracy of the trained models, but also innovative metrics related to bias for a specific feature: gender. While some models often achieve impressive accuracy, they can unintentionally reproduce or amplify social biases present in the training data. This can lead to unfair outcomes for underrepresented or disadvantaged groups. This research goal is also to show the trade off that often must be made between accuracy of the model and reduction of bias for sensitive features.

In this project, we explore fairness in supervised learning by analyzing a law school dataset[2]. We begin by training a baseline classification model to predict students' academic success. We then simulate real-world unfairness by injecting bias against female students. Specifically, we drop a portion of

successful female students from the dataset, introducing a gender imbalance. Finally, we apply a mitigation technique to correct for this bias and evaluate its impact.

To assess the fairness of each model, we utilize several group fairness metrics such as demographic disparity, true positive rate (TPR) per group, and equalized odds difference.

2 Fairness Metrics and Mitigation Method

Demographic Parity Difference[3] measures the absolute difference in the probability of receiving a positive outcome between groups defined by a sensitive attribute (in this case, gender). A value close to zero implies that both groups are selected at similar rates, regardless of true outcome. This metric captures fairness in terms of equal treatment but may overlook underlying qualification differences.

Equalized Odds Difference compares the true positive rates and false positive rates across groups. It aims to ensure that the model makes errors (or correct decisions) at similar rates for each group. This metric is stricter than demographic parity because it incorporates outcome-dependent fairness.

Exponentiated Gradient Reduction[1] is a fairness-aware in-processing technique that frames learning as a constrained optimization problem. It works by combining a set of base learners, adjusting their weights through exponentiated gradient updates to minimize classification error while satisfying a fairness constraint (such as demographic parity or equalized odds). This iterative algorithm is both flexible and powerful, adapting the model to achieve better group fairness without requiring changes to the dataset.

3 Dataset Description

The dataset used in this project is the *Student Performance Data Set*, publicly available from the UCI Machine Learning Repository. It contains demographic, social, and academic information about Portuguese students in secondary school.

For our analysis, we focused on the "Mathematics" subset, which includes 395 instances with 33 attributes. We performed feature selection and preprocessing steps to transform the data into a binary classification task: predicting whether a student passes or fails the course, based on their final grade. A new binary label `pass` was introduced, where students with a final grade of 10 or more (out of 20) are considered to have passed.

We selected `sex` (gender) as the sensitive attribute for fairness analysis, examining disparities between male and female students.

4 Baseline Model

The first step involved training a baseline model on the original, unbiased dataset. After preprocessing the features, we employed a supervised machine learning classifier to predict the binary outcome of whether a student would pass the course. For this, a Random Forest Classifier from the Sklearn[4] package has been chosen. In order to avoid overfitting, a max depth of 5 has been introduced.

Model evaluation was performed using accuracy methods from Sklearn[4] on both training and test sets. Furthermore, we analyzed the fairness of the model using advanced metrics of the Fairlearn[1] package, such as demographic disparity and the true positive rate between groups defined by the sensitive characteristic `sex`. These fairness metrics provided insight into whether the model exhibited any performance imbalances between male and female students before the injection of an artificial bias.

5 Bias Introduction

To simulate a real-world scenario where data may contain historical or societal bias, we intentionally introduced a gender bias into the dataset. Specifically, 50% of the female students who had passed the course have been removed. This created a dataset in which the representation of successful female students was artificially reduced, leading to an imbalanced distribution across the sensitive feature `sex`.

The purpose of this intervention was to evaluate how such a bias would affect the model's performance and fairness, particularly in terms of demographic disparity and equal opportunity across groups.

The same evaluations and metrics from the base model have been applied to this new model.

6 Bias Mitigation

After introducing bias into the dataset, we applied a fairness-aware machine learning technique to mitigate its impact. Specifically, we employed the *Exponentiated Gradient Reduction* (EGR) algorithm, a post-processing method provided by the Fairlearn library. This algorithm seeks to optimize both predictive accuracy and fairness constraints simultaneously.

The EGR method was applied using the sensitive attribute `sex` as the fairness constraint. The goal was to reduce the disparities introduced by the biased data while maintaining acceptable performance. The results of this mitigation were analyzed and compared against the baseline and biased models.

7 Comparative Analysis

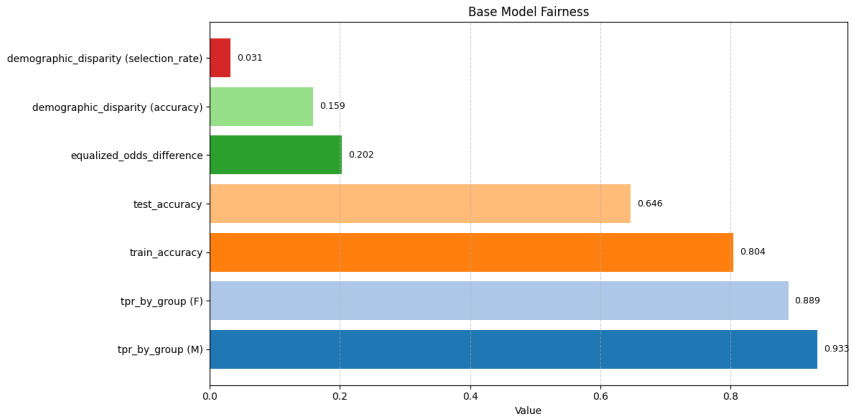

Figure 1: Fairness metrics visualization for the baseline model.

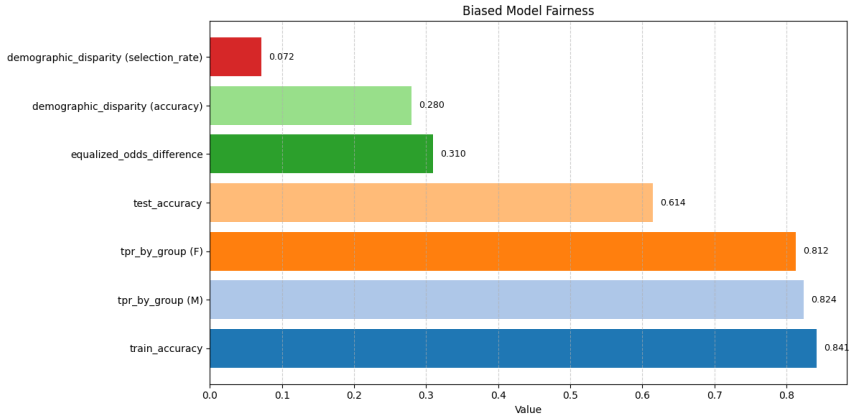

Figure 2: Fairness metrics visualization after bias injection.

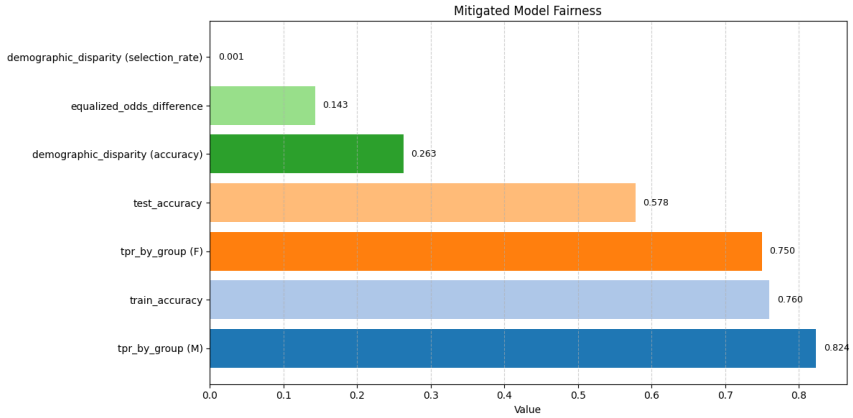

Figure 3: Fairness metrics visualization after bias mitigation.

Table 1: Performance and Fairness Metrics Across Models (TPR only for females)

Model	**Test Acc.**	**Train Acc.**	**DP Diff.**	**EO Diff.**	**TPR (F)**
Baseline	0.646	0.804	0.0313	0.2024	0.889
Biased	0.614	0.841	0.0717	0.3095	0.813
Mitigated	0.578	0.760	0.0012	0.1429	0.750

8 Results and Analysis

Table 1 summarizes the performance and fairness metrics for the baseline, biased, and mitigated models. The baseline model, trained on the original dataset, achieved a test accuracy of 64.6% and showed minimal demographic disparity (0.0313) and a moderate Equalized Odds (EO) difference (0.2024).

After introducing artificial bias by undersampling successful female students, the test precision dropped slightly to 61.4%, and the fairness metrics worsened: the demographic parity difference increased to 0.0717, and the EO difference to 0.3095, indicating a measurable decrease in the fairness of the model.

The application of the Exponentiated Gradient Reduction algorithm significantly improved fairness. The demographic parity difference was nearly eliminated (0.0012), and the EO difference reduced to 0.1429. However, this came at the cost of reduced test accuracy (57.8%).

9 Conclusions

In this project the impact of gender bias on machine learning models has been explored and the effectiveness of mitigation strategies using the Fairlearn[1] library has been evaluated. Starting from a balanced dataset, we introduced an artificial gender bias by reducing the representation of successful female students. As expected, the biased model showed worsened fairness metrics, with a noticeable drop in test accuracy for female students and an increase in the Equalized Odds Difference.

To address the bias, we applied the Exponentiated Gradient Reduction algorithm under a Demographic Parity constraint. The mitigation successfully reduced the Demographic Parity Difference from 0.0717 to 0.0012 and lowered the Equalized Odds Difference from 0.3095 to 0.1429, demonstrating a clear improvement in fairness. However, this came at the cost of a small decline in predictive accuracy, particularly for the majority class.

The experiment shows that bias mitigation methods can significantly improve fairness metrics, with a trade-off with the accuracy of the predictions. This reinforces the need to evaluate both performance and fairness when deploying machine learning models in sensitive applications.

References

[1] Fairlearn Contributors. Fairlearn documentation, 2020. Accessed: 2025-05-09.

[2] L. F. Wightman. Lsac national longitudinal bar passage study. LSAC Research Report Series, 1998. Dataset from https://github.com/tailequy/fairness$_d$$ataset/tree/main/Law_s$$chool$.

[3] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[4] Scikit learn Developers. Scikit-learn: Machine learning in python. `https://scikit-learn.org/stable/`, 2021. Accessed: 2021-05-09.